Université de Montréal

Lexical Knowledge Patterns for Semi-automatic Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Study of English and French

Volume 1 of 2

par

Elizabeth Marshman

Département de linguistique et de traduction

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)

en traduction

option terminologie

novembre 2006

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Lexical Knowledge Patterns for Semi-automatic Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Study of English and French

présentée par :
Elizabeth Marshman

a été évaluée par un jury compose des personnes suivantes :

[Taper le nom] , Président-rapporteur
Marie-Claude L'Homme, Directrice de recherche
Sylvie Vandaele, Co-directrice
[Taper le nom] , Membre du jury
[Taper le nom] , Examinateur externe
[Taper le nom] , Représentant du doyen de la Faculté des études supérieures

Abstract

This research focuses on lexical knowledge patterns indicating conceptual relations of CAUSE–EFFECT and ASSOCIATION in specialized medical texts in English and French, for semi-automatic extraction of knowledge-rich contexts (KRCs) that can assist users such as terminologists in conceptual analysis and terminological description. Lists of patterns — prototypically composed of a lexical marker (i.e., a lexical unit or series of lexical units) that indicates the presence of a relation between two concepts (represented in a text by terms or other linguistic expressions) — may be compiled and used by computer tools to identify information-rich segments in corpora automatically or semi-automatically.

The objectives of this study were twofold: to evaluate possibilities and challenges for pattern-based tools in English and French, in order to study the feasibility of developing tools that can process corpora in the two languages; and to identify potential sources of interlinguistic variation that may affect these possibilities and challenges and require adjustments in pattern-based KRC extraction strategies to achieve comparable, high-quality results (i.e., acceptable precision and recall).

After a review of several research projects that evaluated the usefulness of lexical knowledge patterns in various domains, languages and applications, a number of issues affecting pattern-based tools were identified. These included the number of relation occurrences observed and of different markers indicating these relations, as well as their relative frequencies in the corpora, the types of markers observed, the precision with which they identify relations, the variability of the structures in which they may occur, the type and form of the elements they link, and the prevalence of a number of challenges for identifying useful and reliable information, including interruptions of one or more constituents of a pattern or the non-contiguity of the pattern elements, and the presence of expressions of uncertainty of information that may be extracted from the context.

These issues were then evaluated in a set of contexts extracted from corpora in each language using a candidate-term-based approach. The data gathered in the two languages in this analysis were then compared in order to evaluate the impact each issue may have on specific types of pattern-based tools and on the productivity of a pattern-based approach to semi-automatic KRC extraction in general.

Strong similarities in the data in the two languages were identified in respect to many of the criteria evaluated, confirming that pattern-based approaches are promising in both languages and that the development of bilingual tools seems to be both a worthwhile and an achievable goal.

Nevertheless, some differences noted indicate a need to carefully consider the strategies implemented when developing pattern-based tools, in order to satisfy the requirements of situations in which these tools may be used and to maintain acceptable performance in both languages. Awareness of specific characteristics of patterns and their components and some other elements of contexts that may affect the form of patterns and/or the usefulness of information they convey can help developers to make more informed choices when creating tools, whether for use in mono-, bi- or multilingual environments.

Keywords: conceptual relations, knowledge patterns, corpus-based terminology, medical language, interlinguistic comparison

Résumé

La présente recherche évalue les possibilités d'exploiter des patrons de connaissances lexicaux indiquant des relations conceptuelles de CAUSE–EFFET et d'ASSOCIATION dans des corpus médicaux en anglais et en français, et ce à des fins d'extraction semi-automatique de contextes riches en connaissances qui peuvent aider des utilisateurs (par exemple, des terminologues) dans l'analyse conceptuelle et la description terminologique. Ces patrons sont prototypiquement constitués d'un marqueur lexical (unité lexicale ou séquence d'unités lexicales) qui indique la présence d'une relation entre deux concepts, à leur tour réalisés sous forme de termes ou d'autres expressions linguistiques; ils sont exploités par des outils informatiques pour identifier des segments riches en information au sein de corpus.

Cette étude vise deux objectifs. Le premier est d'évaluer certaines possibilités d'élaboration et d'exploitation d'outils à base de patrons ainsi que certains défis rencontrés en anglais et en français, afin d'étudier la faisabilité de développer des outils qui peuvent traiter des corpus dans les deux langues. Le deuxième est de repérer et de caractériser les éventuelles variations interlinguistiques qui influenceraient ces possibilités et défis, et qui nécessiteraient une modification des stratégies d'extraction pour produire des résultats comparables et de haute qualité (c'est-à-dire, une précision et un rappel acceptables).

Grâce à une analyse de divers projets de recherche qui ont évalué l'utilité de ces patrons de connaissances lexicaux dans différents domaines, langues et applications, certaines questions reliées aux applications faisant appel aux patrons ont été identifiées. Parmi celles-ci on trouve : le nombre d'occurrences des relations observées et de marqueurs distincts associés, les fréquences relatives de ceux-ci dans les corpus, les types de marqueurs trouvés, leur précision, la variabilité des structures dans lesquelles ils participent, le type et la forme des éléments qu'ils relient, et la prévalence des défis

en ce qui concerne l'identification et l'utilité de l'information extraite (par exemple, la non contiguïté des composantes du patron ou la présence d'indications d'incertitude).

Ces facteurs sont évalués dans des contextes extraits de corpus dans les deux langues à l'aide de candidats-termes. Les observations sont par la suite comparées dans les deux langues, permettant d'évaluer leur influence éventuelle sur des outils faisant appel à des patrons, ainsi que la productivité générale de cette approche à l'extraction semi-automatique de contextes riches en connaissances.

Des similarités frappantes entre les données anglaises et françaises sont observées par rapport à plusieurs des critères évalués, confirmant le caractère prometteur d'une telle approche dans les deux langues. D'après ces résultats, le développement d'outils bilingues paraît un but à la fois intéressant et atteignable.

Néanmoins, certaines différences notées indiquent un besoin d'évaluer des stratégies de développement d'outils faisant appel à des patrons, afin d'adapter l'approche aux exigences des situations spécifiques d'utilisation et ainsi maintenir une efficacité satisfaisante dans les deux langues. Une connaissance des caractéristiques des patrons et de leurs composantes ainsi que d'autres éléments contextuels influençant la forme des patrons ou l'utilité de l'information véhiculée peut aider des créateurs d'outils à prendre des choix éclairés dans la conception d'outils uni-, bi- ou multilingues.

Mots-clés : relations conceptuelles, patrons de connaissances, terminologie basée sur corpus, langue médicale, comparaison interlinguistique

Table of Contents

List of Tables

List of Figures

To the memories of

Ingrid Meyer (1957–2004)

and

Winifred May Smith (1913–2003)

I spent four years prostrate to the higher mind [predisposing factor],
Got my paper [cause] *and* [lexical marker] *I was free* [effect].

- Indigo Girls

Acknowledgements

Avant tout, je tiens à remercier mes directrices, Marie-Claude L'Homme et Sylvie Vandaele, pour leur aide et encouragement constants. Les discussions — et parfois débats ! — que nous avons eus ont énormément enrichi le travail, et sans leur appui je n'aurais jamais pu arriver au bout. Je remercie Sylvie particulièrement pour ses conseils précieux en matière de l'interprétation des textes médicaux aux niveaux terminologique et conceptuel.

Je remercie aussi Patrick Drouin pour son généreuse et précieuse aide technique et de m'avoir permis d'utiliser TermoStat, Didier Bourigault de m'avoir généreusement permis d'utiliser son outil SYNTEX pour analyser mes corpus, Nathan Ménard pour des conseils en matière de statistiques, Jean-Yves Morin et Richard Kittredge pour leur rétroaction sur le projet de recherche et leurs suggestions pour la continuation, Pierre Zweigenbaum pour ses excellentes suggestions pour la construction et l'analyse des corpus, Alain Polguère et Igor Mel'čuk, pour leur rétroaction et leurs conseils, et les bibliothécaires à la Bibliothèque de lettres et sciences humaines à l'Université de Montréal, particulièrement Mme Brisebois, pour leur aide en matière de corpus.

I also thank Dr. Joan Marshman of the University of Toronto's Faculty of Pharmacy for her time and her counsel on the interpretation of the English corpus texts and for her advice on questions epidemiological, statistical and stylistic. Dr. Lynne Bowker of the University of Ottawa has also been extremely helpful throughout my M.A. and PhD studies, and I wish to thank her for all of her insight and encouragement.

I thank the Social Sciences and Humanities Research Council of Canada and the Université de Montréal for their generous financial support.

And finally, I thank my family and my friends for all of their help and support — some also financial, but the most precious emotional!

Typographical conventions

Throughout this document, the following typographical conventions will be used.

I. Conventions used in the text:

- **English translations of terms in other languages**: Eng. plus square brackets, e.g., *marqueur de relation* [Eng. relation marker]
- **Relation names**: small capitals, e.g., CAUSE–EFFECT
- **The class of markers**: [MARKER]
- **Elements related by pattern markers**: X, Y, Z, W, U…
- **Specific markers and marker forms referred to in the text**: italics, e.g., *lead to*
- **Terms and other lexical items referred to in the text**: italics, e.g., *atherosclerosis*
- **Concepts**: quotation marks, e.g., "knowledge-rich context"
- **Vocables**: all capitals, e.g., CAUSER
- **Parts of speech**: small capitals, e.g., NOUN
- **Paraphrases of semantic components and senses**: single quotation marks, e.g., 'cause'
- **Lexical functions**: Courier New 11 pt, e.g., Caus

II. Conventions used in the examples:

- **Markers**: bold, e.g., **lead to**
- **Elements of the examples being discussed**: underlining, e.g., <u>possible</u>

List of symbols and abbreviations

χ^2	Chi-square
ADJ.	adjective
CT	candidate term
CTKB	corpus-based terminological knowledge base
df	degrees of freedom
EN	English
EV	expected value
FR	French
IULA	Institut universitari de lingüística aplicada (Universitat Pompeu Fabra, Barcelona, Spain)
KRC	knowledge-rich context
LF	Lexical function
MeSH	Medical Subject Headings
POS	part of speech
PPL.A.	participial adjective
PREP.	preposition
TKB	terminological knowledge base
V	value
UMLS	Unified Medical Language System

Key to interpreting Chi-square tables in this thesis

The Chi-square tables in this thesis present the figures that were used to calculate the proportions of cases evaluated in which a given criterion was observed in the English and French data, so that these proportions could then be compared using the Chi-square test. The proportions are calculated by comparing the numbers of cases in which a given criterion was present with the total number of cases evaluated. A model of the standard Chi-square table used in this thesis is shown below, with a key explaining its contents.

Table 1: Comparison of the proportions of relation occurrences containing item A (A) in English and French

	EN	FR	Total
A+	*x*	*y*	*a*
A-	*z*	*w*	*b*
Total	*u*	*v*	*c*

Symbol	Description
A	Represents the criterion being evaluated.
A+	Indicates the row presenting numbers of cases in which item A was observed.
A-	Indicates the row presenting numbers of cases in which item A was not observed.
EN	Indicates the column presenting the English data.
FR	Indicates the column presenting the French data.
a	Total number of cases in which item A was observed in the two data sets. Unless otherwise indicated, this is a number of relation occurrences.
b	Total number of cases in which item A was not observed in the two data sets. Unless otherwise indicated, this is a number of relation occurrences.
c	Total number of cases analyzed. Unless otherwise indicated, this is the total number of relation occurrences in the two data sets together.
u	Total number of cases analyzed in the English data. Unless otherwise indicated, this is the total number of relation occurrences in the English data.
v	Total number of cases analyzed in the French data. Unless otherwise indicated, this is the total number of relation occurrences in the French data.
x	Number of cases in which item A was observed in the English data. Unless otherwise indicated, this is a number of relation occurrences.
y	Number of cases in which item A was observed in the French data. Unless otherwise indicated, this is a number of relation occurrences.
z	Number of cases in which item A was not observed in the English data. Unless otherwise indicated, this is a number of relation occurrences.
w	Number of cases in which item A was not observed in the French data. Unless otherwise indicated, this is a number of relation occurrences.

Introduction

This research was carried out with a view to acquiring knowledge that will contribute to the development of bilingual computer tools for analyzing corpora that can assist users such as terminologists, terminographers and others carrying out conceptual analysis and related tasks in specialized domains (specifically the field of medicine). As such, it draws on aspects of traditional terminology, as well as the somewhat newer fields of computer-assisted and corpus-based terminology.

Computer tools and knowledge patterns

A vast amount of information is now available in text form, and this resource is constantly growing. However, as the volume of texts increases, it is becoming more and more challenging to find specific kinds of information in the mass of data quickly and easily. This is very evident in domains such as medicine, in which a high volume of constantly evolving information is available.

One of the strategies for more efficiently exploiting data available to users in text form is the development of computer tools to aid in identifying specific types of information in texts. This research focuses on one technique that has been studied for developing such tools: the use of lexical knowledge patterns for the semi-automatic extraction of knowledge-rich contexts.

Knowledge-rich contexts (Meyer 2001) are contexts that provide information that is useful for conceptual analysis (e.g., information about a conceptual relation or a concept's attributes). Knowledge patterns may be used to identify such contexts in texts, so that these contexts may be presented to users seeking a specific type of information, for interpretation and evaluation.

Knowledge patterns (Meyer 1994, 2001) are linguistic structures that commonly indicate information that is pertinent for conceptual analysis. These generally involve two elements that are linked by some kind of relationship (e.g., a concept and one or more of its attributes, two or more concepts); these are realized in the text by terms or

other linguistic expressions and a marker of the relationship that exists between them. These markers may take various forms; this research will be concerned with *lexical knowledge patterns*, in which the marker takes the form of a lexical unit or sequence of lexical units. Thus, English knowledge patterns for the CAUSE–EFFECT relation include X *causes* Y, X *results from* Y, and *stimulation of* X *by* Y, in which the two elements linked by a relation are represented by the variables X and Y and the markers *causes*, *results from* and *stimulation of... by* indicate the presence of a CAUSE–EFFECT relation. Similarly, in the knowledge patterns *association between* X *and* Y, X *correlates with* Y and X *characterized by* Y, the markers *association between... and*, *correlates with* and *characterized by* indicate the presence of an ASSOCIATION relation.

Knowledge patterns in semi-automatic knowledge extraction for terminology work thus can help users locate potentially useful contexts, and moreover provide them with more information about of the type of relation present and the elements it links.

This kind of approach saves users time and effort, since they are not obliged to read and analyze a corpus in its entirety. Moreover, by giving users a general overview of the kinds of contexts that are most likely to indicate a given type of information in the corpus, the tool may bring out regularities and recurrences that might go unnoticed if each occurrence of a potentially interesting term or relation were analyzed individually. This can allow for more comprehensive and consistent conceptual analysis and description. Even information that is potentially pertinent but not certain enough to allow users to draw firm conclusions may be valuable, since once this information has been brought to their attention, users can pursue further research as needed.

Knowledge patterns cannot, however, indicate categorically that a given type of information is present or that this relation is necessarily pertinent for a given application. These decisions, along with many linked to the finer nuances of the information contained in a given context, are considered best left to terminologists (or domain experts), who can take into account the specific context in which research is being

carried out and thus better evaluate whether a given piece of information is useful — and reliable enough for use — in a given context.

The involvement of users in the evaluation of the results of extraction thus ensures that the highest quality product may be obtained; human interpreters of language will always be able to provide a more informed interpretation of that language than even the most highly developed automatic application, and can thus identify and correct many problems in automatic analysis. Moreover, human users are the best judges of the complex extralinguistic factors surrounding the results produced by an automatic tool, which in large part determine the pertinence of any piece of information in a use situation. An automatic tool is rarely equipped to judge whether a given piece of information may be useful for a specific user group or goal, while human users should be able to do so.

At a methodological level, applications that are intended for use in a semi-automatic context with the participation of users in the evaluation of the results of extraction can often retain a wider range of potentially useful information than more automated applications, which must strictly limit the information that is retained in order to minimize noise in the results.

For all of these reasons, a semi-automatic approach knowledge extraction was considered for the purposes of this research to be the most realistic and productive starting point for evaluating the usefulness of pattern-based tools in concept analysis and description in terminology work. It capitalizes on the strengths of both human and machine, taking advantage of the machine's ability to process large amounts of information quickly and uniformly, and the human's to carry out a detailed analysis of linguistic information that takes into account both intra- and extralinguistic factors.

On a concrete level, computer tools may identify knowledge-rich contexts using sets of knowledge patterns. They compare these with texts, in order to identify segments

of the text that correspond to these patterns and are thus likely to contain expressions of a specific type of information, such as a conceptual relation. The performance of these kinds of tools is generally evaluated on the basis of two main criteria: their recall (i.e., the proportion of useful contexts present in corpora that are identified by the tool), and their precision (i.e., the proportion of the contexts identified by the tool that are useful). These measures generally vary inversely, since restrictions imposed on applications to ensure that they provide precise results generally entail the exclusion of some pertinent contexts. Additionally, another factor that is important to evaluate is the investment of time and effort in the development and use of patterns.

In their simplest form, patterns used in computer tools may take the form of character strings representing the marker (e.g., *caus** to represent markers such as *cause* (VERB) or *cause* (NOUN)). However, in order to improve the performance (and particularly the precision) of these tools, patterns may be further developed with additional information and represented by more complex structures such as regular expressions. These may specify a term or terms of interest for a given research task (e.g., by searching for character strings representing terms in (relative) proximity to those representing pattern markers), or — in the case of what are often referred to as *lexico-syntactic knowledge patterns* — the part of speech classes of the pattern marker and potentially of the lexical items surrounding it (e.g., in structures such as NOUN PHRASE + *to cause* (CONJUGATED VERB) + NOUN PHRASE to represent cases in which the verbal marker *to cause* is preceded by a noun phrase, generally indicating the cause in question in a given context, and followed by another noun phrase, generally representing the effect).

In some types of applications, a tool may take on more responsibility for sorting contexts and attempting to identify those that contain pertinent information, and what that information might be. This may involve tasks such as sorting contexts according to the relation or sub-relation present or the potential usefulness of the information the

contexts contain (as indicated, for example, by expressions of uncertainty such as negation or modal verbs occurring in and around the patterns), or attempting to identify the elements that participate in a relation automatically by analyzing contexts' structures. Of course, with each additional task carried out by the tool, the complexity of representing markers and the structures in which they occur increases exponentially, and the precision and recall that can be expected of a tool may vary substantially.

Extracted knowledge-rich contexts may be used in many different applications. This research focuses on information that would be useful for conceptual analysis in the context of terminology work. Terminologists may use the information identified in knowledge-rich contexts to assist them in a variety of tasks, including the acquisition of domain knowledge, conceptual analysis of concepts covered in terminological resources, the construction of concept systems, the linking of related terms and term records, the formulation of definitions, and the selection of contexts for inclusion in term records. In addition, in bi- or multilingual work, these contexts may also be useful for the comparison of term and conceptual systems constructed in two or more languages and the establishment of equivalence between terms in multiple languages.

Need for bilingual research

The current situation in the field of terminology in Canada and around the world, with a strong focus on bi- and multilingual work, creates an obvious need for tools to assist terminologists with analysis in two or more languages. Tools that enable users to process languages in parallel may be invaluable for tasks such as those mentioned above. However, little is known about how different languages compare in terms of the number, types and characteristics of knowledge patterns used to express conceptual relations (for example, the relative frequencies of pattern markers, the part of speech classes of these markers, the nature and forms of elements they link), and how these factors will affect the development and use of semi-automatic knowledge extraction tools and the ultimate usefulness of the contexts extracted using them.

Conventional wisdom about interlinguistic differences (often reflected in works such as Vinay and Darbelnet (1958)) nevertheless raises questions that need to be considered. For example, if French is generally recognized to have a lower tolerance for repetition of lexical units than English, will more variety be found in the markers used to denote relations, therefore reducing the productivity of individual markers and patterns and requiring more of these in French to obtain comparable results in the two languages? Will the ways in which elements linked by relations are expressed in texts also show more variation (e.g., increased use of anaphoric expressions), therefore creating challenges for identifying and interpreting contexts? Is a general tendency to use nouns in French and verbs in English in certain contexts reflected in the markers that indicate conceptual relations, and if so, how should this be taken into account when designing pattern-based tools? If English sentence structures are more variable than those in French, will pattern structures be more variable in this language as well, requiring larger sets of pattern forms? Will the available means for expressing uncertainty about a statement affect possibilities for evaluating these levels of certainty automatically in extracted contexts?

Moreover, while all researchers in the field agree that pattern-based tools may confront difficulties, there is a lack of data on both the frequency with which these difficulties may occur, and the forms that they may take in the two languages. This lack of knowledge means that expectations for the development of bi- and multilingual tools cannot be set realistically, and it is — at best — very difficult to develop strategies for dealing with the adjustments that will need to be made and the problems that may occur.

Some previous work carried out on markers of conceptual relations in English and French (Marshman 2002, 2002a, 2004; Marshman et al. 2002) did provide some information that may be used as the basis for a preliminary evaluation of the potential for observing interlinguistic variation in knowledge patterns, through the comparison of markers observed in the two languages and the identification of a number of

characteristics of patterns and their markers that may differ from one language to another — including, for example, the types of markers observed (including the part of speech classes to which they belonged) and the frequency of these markers. These studies also provided data for the observation of additional challenges that may be encountered by pattern-based tools in one or both of the languages.

However, given the limited nature of these previous discussions, there is not yet sufficient data available to answer questions such as the following: How many and what kinds of markers are useful indicators of relations such as ASSOCIATION and CAUSE-EFFECT in the two languages? How often, and in what kinds of structures do they occur? What kinds of elements are linked by these markers? What kinds of external elements can occur within the pattern structures and contexts, how often, and how will they affect the identification, processing and usefulness of these contexts? Given these factors, can pattern-based tools be expected to perform differently in the two languages? If so, in what ways? What aspects of the application development and use will be affected? What factors could be further investigated — and ultimately what strategies could be developed — to adapt pattern-based tools to these realities in order to obtain comparable, high-quality results in the two languages?

This research will aim to gather information that sheds light on these issues in the context of bilingual work in English and French. We hypothesize that the analysis of data related to these questions will reveal differences between the types of markers used in English and French, the structures and contexts in which they appear, and the difficulties likely to be encountered in the identification, processing and use of these contexts, which will be pertinent for the development and use of pattern-based tools.

Objectives

The main objectives set for this research were thus twofold: to observe the types and characteristics of candidate knowledge patterns and their markers in English and French, as well as the challenges encountered in their identification and likely to occur in their

use and the use of the results they produce; and to compare these in the two languages to observe similarities and differences in respect to pertinent criteria in order to determine whether these similarities and differences may be expected to have an impact on the development and use of pattern-based tools for terminology work in a bilingual context.

Originality and contribution of the research

This research constitutes a rare systematic and comparative look not only at the types and characteristics of knowledge patterns, their markers and their occurrences in two languages but also at some challenges in their use in pattern-based approaches, which will begin to provide information that can help application developers to adjust expectations of pattern-based tool performance in a bilingual context, and to start to research and develop strategies that may be used in different languages in order to improve efficiency and obtain comparable results in the two languages.

In accordance with its descriptive and comparative orientation and the primary application envisaged (i.e., semi-automatic extraction of knowledge-rich contexts for terminology work), the research begins with a broad definition of what constitutes potentially useful information and the forms that this information may take, allowing for a comprehensive analysis of the issues that may be observed in pattern-based applications.

The semi-automatic extraction of knowledge-rich contexts nevertheless also constitutes the starting point for further automated processing of potentially useful contexts as determined by the application for which a tool is intended, and thus provides opportunities to evaluate the conditions that would surround these additional processing tasks and the possibilities and difficulties that would be encountered. Moreover, this kind of analysis also provides an opportunity to evaluate not only the contexts that would be retained in applications involving more highly developed processing of identified contexts, but also those that would be excluded, in order to estimate the proportion of potentially useful data that might be lost in such approaches in the two

languages. As such, while the methodology used in the research and the patterns retained were chosen as a function of an approach geared to semi-automatic extraction, the analysis of the observations may also form the basis for a discussion of the possibilities and difficulties likely to be observed in tasks such as the sorting of contexts or the automatic identification of the elements linked by a relation.

Details of the work and its methodology

The subject fields

Among the medical sub-domains of great interest today — and the objects of vast amounts of research that is then reported in text form — are heart disease and cancer. These two sub-domains (and more specifically atherosclerosis and breast cancer) were chosen as the subjects of the texts analyzed in this research. These texts focused on the development, effects, progression, diagnosis, prevention and treatment of the diseases.

The relations

Medical texts are rich in many types of information; in this research the focus will be placed on conceptual relations, and specifically the relation of ASSOCIATION and the CAUSE–EFFECT relation. ASSOCIATION is defined in this work as the significant co-occurrence of two variables, and is often a precursor to conclusions of CAUSE–EFFECT relations between these variables. As such, it is intrinsically linked — but not identical — to the CAUSE–EFFECT relation. The CAUSE–EFFECT relation is taken in this research to denote a relationship between two concepts in which one, the cause, exerts an influence that determines the existence or occurrence of the other, or changes this existence or occurrence. It includes several more specific types of influences, including CREATION, DESTRUCTION, MAINTENANCE or PERMISSION, PREVENTION, MODIFICATION, INCREASE, DECREASE and PRESERVATION (according to a subdivision established by Barrière (2002), which in turn calls upon an analysis by Talmy (1985)).

The research will focus on the identification of these conceptual relations as they are manifested in texts. This process may thus be informed not only by the description

and classifications of the relations from a conceptual perspective — that most coherently associated with the use of knowledge patterns in (relatively traditional) terminology work, as reflected in the work of researchers such as Meyer (Meyer et al. 1999; Meyer 2001), whose approach is closest to that used in this study — but also by analyses of the relations as reflected in the semantics of the two languages studied in this research.

Methodology

The first step in the study involved constructing corpora of English and French texts, and identifying in these corpora a set of contexts in which CAUSE–EFFECT and ASSOCIATION relations were present and were indicated by lexical knowledge patterns. These contexts were then analyzed and annotated according to a set of criteria — established by calling upon previous observations in other research projects and supplemented and refined in the light of the observations in this work — to identify knowledge patterns and pattern markers and their pertinent characteristics, as well as potential difficulties in pattern identification and use. The data thus obtained were analyzed quantitatively and qualitatively in each language, and finally the results in the two languages were compared, in order to identify similarities and differences. The potential pertinence of these for pattern identification and use, as well as for the subsequent use of the extracted contexts, was then evaluated.

Pattern based applications: Factors affecting design and performance

By considering the most basic of applications and knowledge pattern forms (i.e., the use of character strings to identify KRCs) as the starting point for this research, it is possible to observe not only the basic elements necessary for pattern-based tool development, but also characteristics of the markers observed and the contexts in which they occur that may affect possibilities for further refinements to pattern forms and further processing of contexts. The effects these may have on the development and performance of applications, as well as on the ultimate use of the information extracted, will be briefly presented below. They will be addressed in two main groups: first the characteristics of

the patterns themselves — i.e., that involve the form and nature of the markers and of the elements that they link, and their placement relative to one another — and second the additional challenges related to external elements that may be found within or around the structures of these patterns.

Before these groups are described, however, a more general factor is important to introduce: the number of relation occurrences observed in the initial process of identifying relation occurrences indicated by lexical markers in the two languages. This may indicate the relative densities of pattern occurrences that met the criteria for this research, which in turn reflects the potential productivity of pattern-based tools.

The pattern characteristics analyzed included the variety of the markers themselves and the number of occurrences of these markers, the types of markers observed, the variability of pattern forms, and the number and form of the elements that these markers link in a given context.

The range of markers used to express a given relation in a language affects the number of markers required for a tool to achieve a given level of recall: the more different markers are used, the more patterns will be necessary to locate the contexts containing a relation. An additional criterion for evaluating marker variety is the distribution of relation occurrences among the various markers, which can also indicate the number of markers required to attain a certain level of productivity in pattern sets. The frequency of markers is likely to be closely connected to marker variety: if relatively few different markers are used to express a relation, these markers are likely to be used relatively frequently, and thus to allow a tool to find a relatively large number of relation occurrences with relatively few patterns. Markers that are infrequent in general are of course not as useful for identifying KRCs as their more frequent counterparts. These two factors together thus are indicators of the expected productivity of pattern sets, and can help to guide pattern set development.

Characteristics of markers that can affect the development and performance of pattern-based applications include the part of speech classes to which they belong, their form (either simple or complex), and the variation that may be observed in marker form. All of these affect the design of pattern forms, and may also be linked to challenges for application performance, as variations from forms accounted for in pattern sets (e.g., changes in the form of markers or the order of their elements) can interfere with KRC identification. The process of pattern design and application performance are also affected by the variability in pattern structures (i.e., in the placement of pattern elements relative to one another), for similar reasons.

Two additional factors that affect tool design and performance are the precision of relation markers for identifying relations, and the closely related issue of marker polysemy. Clearly, not all occurrences of every potential marker will retrieve contexts in which complete information about the desired relation is present, and the reliability of individual markers for locating these useful contexts will affect the productivity of these markers — and thus of pattern-based tools — for extracting KRCs. The number of relations that are associated with a given marker also affects both the choice of patterns and the performance of pattern-based tools. Some markers may not denote only a single, specific relation or sub-relation, but may be used to express other (sub-)relations or other meanings entirely; this of course can lead to noise in the results of KRC extraction and/or to problems in classifying contexts according to the relation present.

The number and form of related elements linked by pattern markers may also affect the development and performance of pattern-based tools, as well as the ultimate usefulness of the contexts they extract. For example, in some cases, two or more elements share a role in a relationship (e.g., two or more causes or two or more effects are described in a single context, as in *X causes Y and Z*). Such contexts are generally more informative than basic structures because they indicate additional participants in the relation, in addition to the relationship that exists between these participants. This

phenomenon must be reflected in the design of patterns that include representations of related elements, and particularly by those that attempt to identify related elements automatically, or pertinent contexts may be missed or incorrectly analyzed. The ways these structures may differ in different languages (including the number and variety of the lexical units that indicate the relationships between the related elements that co-occur in a given role) may affect the complexity of developing such pattern forms.

The form of related elements themselves may also pose some challenges, particularly for applications that impose restrictions on these forms. For example, while in most cases related elements occur in noun form, in some contexts concepts may be expressed by other types of units (e.g., adjectives, verbs, clauses); moreover, in some contexts they may be represented by anaphoric expressions (e.g., pronouns). Pattern-based approaches must either use forms that allow for such variations or accept silences in the results of extraction.

Additional challenges for pattern-based tools include the interruption of pattern forms by external elements and the presence of expressions of uncertainty in the contexts in which the patterns occur.

Pattern forms — particularly lexico-syntactic patterns — that specify the structures in which markers occur may not recognize potentially useful contexts if these structures are interrupted by external elements (e.g., modifiers, relative clauses, references) that occur between the related elements and the marker. The situation becomes even more complicated if the interruptions of these pattern structures take the form of other patterns or pattern markers (e.g., as in X *leads to the suppression of* Y); this phenomenon may not only affect the form of the context, but also the type of relation that is expressed, and thus may pose challenges for the sorting and/or ultimate use of extracted contexts. (For example, the marker *lead to* generally indicates the CAUSE–EFFECT sub-relation of CREATION, but contexts that contain the structure illustrated above indicate PREVENTION.)

The presence of expressions of uncertainty within KRCs affects the usefulness of a given context for various applications. These expressions may take various forms, including quantification of related elements (e.g., *some* X*s play a role in* Y), hedging (e.g., *X plays a minor role in* Y), modal verbs (e.g., X *may play a role in* Y), or negation (e.g., X *does not play a role in* Y). For example, for some uses, only contexts in which no doubt about the relation present is expressed may be pertinent, while in other cases relations that are expressed as doubtful — or even denied — may be useful. However, the variability in the form and semantic impact of these expressions may pose significant challenges for automatic sorting of contexts according to the reliability of the information they express.

It is thus clear that a large number of factors may be pertinent in various kinds of applications for relation identification and context extraction, and may have varying impacts depending on the types of patterns used and the extent to which tools attempt to process the information located for users. An evaluation of the possibilities and challenges of pattern-based extraction of KRCs in multiple languages should thus consider as wide a range as possible of these factors in order to provide a comprehensive portrait of how pattern-based tool design and performance may be affected by similarities and differences in the two languages, and what impact these factors may have on the usefulness of the information extracted.

Structure of the thesis

This thesis is divided into five chapters. The basic concepts pertinent in this research (including descriptions of knowledge patterns and conceptual relations, and specifically those studied in this work) will be presented in Chapter 1. In Chapter 2, previous work on knowledge patterns and pattern-based applications will be presented and some pertinent aspects of these research projects compared with one another; more details of the objectives of this work and how it differs from previous research will also be presented in light of this comparison. Chapter 3 will present the methodology used in

this work, and Chapter 4 the results obtained, focusing on the interlinguistic comparison of the data gathered. Chapter 5 will present a discussion of the similarities and differences observed in the English and French data and the impact they may have on various aspects of the development and use of pattern-based tools, complemented by some additional observations. This will be followed by some conclusions and ideas for future research.

1 Basic concepts

This Chapter will present an overview of some basic concepts pertinent to this research. In Section 1.1, some basic concepts in terminology will be described briefly. Section 1.2 will present the definition of the term *knowledge pattern*, and Section 1.3 will introduce the subject of conceptual relations. Section 1.4 will present a selection of relation typologies used in terminology, and Section 1.5 will describe the relations studied in this research, ASSOCIATION (Section 1.5.1) and CAUSE–EFFECT (Section 1.5.2).

1.1 Basic concepts in terminology

The field of terminology is concerned with the study and description of communication in specialized fields, and the terms used in it. These fields, areas of human interest and study, can be identified using several criteria. Two of these involve the setting of what have been characterized as horizontal and vertical limits (e.g., Hoffmann 1976; Sager et al. 1980; Kocourek 1991). The former involve the delineation of these areas of study as opposed to others; this task is becoming more and more complex as interdisciplinary fields of study develop and the borders between different fields become more flexible and porous. The latter involve the distinction of levels of specialization, which may be characterized by criteria such as the background knowledge and training possessed by the participants (both senders and receivers) in communication and the goals of this communication (e.g., Pearson 1998). Delineating specialized domains involves choosing to set borders at a given point along the continua established according to these criteria, and this choice generally depends on the situation and goals of a given task.

Terms constitute the primary focus of terminology, and may be viewed from a number of different perspectives that target particular aspects of their natures and their roles in specialized discourse; perspectives include the communicative (e.g., Cabré 1992), sociocognitive (e.g., Temmerman 2000), and lexico-semantic (e.g., L'Homme 2004). All of these may reveal important aspects of terms and their functioning. For the purposes of this research, however, the focus will be placed on the role of terms as

linguistic units that represent concepts, and that are thus a means of referring to these concepts in communication. This is a perspective that is very close to that of traditional terminology (typical, for example, of the Vienna School and particularly of Wüster (e.g., 1981), and observable to a lesser extent in Sager (1990)), although it is somewhat softened from the more dogmatic view in which "[t]he primary objects of terminology, the terms, are perceived as symbols which represent concepts" (Sager 1990: 22). The point of view in this research does not deny that terms are more complex and multifaceted than simple symbols for concepts; rather, it focuses on the role of terms as representing concepts in specialized discourse. Moreover, the view taken in this research is also somewhat removed from the more restrictive views of traditional terminology, in that the prescriptive principles requiring bi-univocal relations between terms and concepts are considered here to be unduly restrictive. The reality of terminological variation in specialized discourse (e.g., Gaudin 1993; Daille 2005) is undeniable; concepts may clearly be denoted by various terms or term variants, and even by non-terminological linguistic units. As Daille observed, the scope of the variation that is acceptable and pertinent for a given project depends in large part on the goals of that project; for the purposes of information retrieval, it is important to take into account a wide range of such phenomena.

Terms then are seen here as (one of the) access points to knowledge in the form of concepts, i.e., units of thought that are created by a process of abstraction and generalization from observations of reality. Sager (1990: 22) describes the process that leads to the creation of a concept — and thus a starting point for a definition of what a concept is — as follows:

> Concept formation is a process of variously grouping and ordering the material and immaterial objects which we sense, perceive or imagine into abstract categories. In a first stage of observation of our environment we identify a number of individual objects as having certain properties or characteristics in common. From the individual objects we have identified as having certain common features, we abstract some of these properties in order to arrive at types of objects.... In a further stage of

> ordering, we may then group the already abstract types of objects into broader classes.... An important distinction is thus created between the individual objects of our sensation, perception and imagination and the abstract categories, i.e., the concepts which represent them. We therefore define concepts provisionally as 'constructs of human cognition processes which assist in the classification of objects by way of systematic or arbitrary abstraction'.[1]

Concepts are linked by various types of relationships, which determine the structures of knowledge in a given domain. In traditional terminological description, these systems have generally been represented using hierarchical, GENERIC–SPECIFIC relations between concepts, creating tree structures. However, concepts may also be related by a number of other pertinent relations, including those evaluated in the context of this research, CAUSE–EFFECT and ASSOCIATION.

The analysis of the structures in which concepts participate is the prototypical starting point for conceptual description in traditional terminology. Terminological resources such as term banks have traditionally been concept-centred, including entries (e.g., term records) that represent a single concept. They generally establish definitions of concepts in large part according to their place in a concept system (e.g., the generic concept to which they are linked, the characteristics that differentiate them from other specifics of this generic). Terms are then associated with these concepts, and equivalence between terms (within a language or between languages) evaluated in light of the term–concept relationship identified. Thus, the identification and analysis of the links between concepts constitutes a critical step in terminological research and development of terminological resources: these relations help to delineate, define and differentiate between concepts.

Moreover, conventional term bases and their representation of knowledge in structures focusing on GENERIC–SPECIFIC relations may be considerably enriched by

---

[1] The use of *object* here should be noted particularly: the term does not refer exclusively to concrete entities, but includes those that are both *material* and *immaterial*, realities that are sensed, perceived or imagined. (cf. also Wüster (2003) on this point).

information about additional relations. This kind of development can be carried out within a more classical terminological resource structure (e.g., by including this information in definitions or contexts in term records), or in one that is specifically developed with this kind of approach in mind. One proposal for this kind of development involves the creation of what Meyer et al. (1992) called the *terminological knowledge base* (TKB), a resource that could integrate a much larger part of the knowledge (e.g., about relations between concepts) that terminologists acquire in the process of domain research than is usually the case in conventional resources. By representing various kinds of relations between concepts, resources can reflect the kinds of links that have a particularly important role in defining the knowledge structure in a specific domain, and provide a more complete portrait of concepts and the roles they play in knowledge structures. Information of this kind can be particularly valuable for the description of concepts that lend themselves less easily or less well to definition by a classical model consisting of a generic and specific characteristics; these are likely to include in particular those that represent events such as processes and activities rather than concrete entities. An approach such as that used in a TKB can thus improve users' understanding of these concepts and help them to better express knowledge in the domain.

Given terminology's focus on communication in specialized domains, subject-field specialists have always been precious sources of information for terminologists and terminographers. However, given the limitations imposed by reliance on these specialists (e.g., their availability for consultation and ability to effectively convey information about a wide variety of aspects of terms, concepts and their usage), text-based resources are generally the primary resource for terminology work in the field today. Documents provide concrete, readily available and usable examples of specialized communication. Moreover, by collecting various types of documents, terminologists are able to develop a comprehensive view of a field and its discourse.

As a result, there is ever-growing interest in corpus-based terminology work (e.g., Meyer and Mackintosh 1996; Pearson 1998; Bowker and Pearson 2002). The creation of corpora of representative texts that can be used to evaluate the characteristics of the discourse in a field as a whole is a challenging and complex task (described in the works mentioned above, among others), but one that provides a wealth of information that can be used in many ways in terminology work.

The advantages of this kind of approach — as of any approach — are nevertheless accompanied by certain difficulties. Ensuring that a sample of texts is representative of the greater whole requires careful evaluation and selection according to a number of criteria, and it may not always be possible to include as wide a range of texts as desired or to eliminate all potential sources of bias in results. (Moreover, in projects with a bilingual or multilingual orientation, ensuring comparability between corpora in different languages increases this complexity substantially.) In addition, corpus texts may contain errors or other elements that can be difficult to interpret or even misleading to users. As with the use of any resource, a certain amount of both trust and critical evaluation are required in corpus-based approaches.

Redundancy in corpora is often helpful in confirming the validity of the information extracted; when information (be it factual or related, for example, to term form or usage) is identified repeatedly in a variety of texts and contexts, its validity is more certain. Because large corpora can provide more opportunities to observe a wider range of phenomena more frequently, they are particularly useful. However, as the size of corpora increases so does the need for tools that can provide quick and easy access to the information contained in them. Computer tools such as concordancers offer terminologists and terminographers one strategy for accessing specific kinds of information. More specialized tools may use techniques such as knowledge patterns to target specific types of information.

1.2 Knowledge patterns

Referred to under various names by different researchers (including Cruse's (1986) *diagnostic frames*, Ahmad and Fulford's (1992) *knowledge probes*, Bowden et al.'s (1996) *triggers*, Condamines' (2002) *conceptual relation patterns*, and the term that will be used here, Meyer's (Meyer et al. 1999; Meyer 2001) *knowledge patterns*), linguistic structures that indicate the presence of semantic and conceptual relations have been widely recognized as extremely useful tools.

In 1992, Ahmad and Fulford described *knowledge probes* as forms (in their case, character strings) that could be used in developing tools to aid in searching for information about relations in text corpora, as they reliably identify contexts in which relations are discussed. They identified sets of probes that could be used to identify a set of relations, including HYPERONYMY and HYPONYMY, PART–WHOLE and CAUSE–EFFECT.

In a 1994 article in *Terminology Update*, Meyer stressed both the importance and the challenges of concept analysis in terminography. She introduced the idea of "knowledge-rich context" and the possibilities that lie in "exploiting the many regularities in the way that 'linguistic patterns' found in specialized texts encode conceptual information" (1994: 8). In 1994, Meyer defined knowledge-rich contexts as "free (i.e., non-collocational) language combinations that frequently identify a particular conceptual relation or attribute" (8), giving examples such as *X is a kind of Y* and *As include Bs, Cs and Ds* as indicators of generic-specific relations, and *X is characterized by Y* and *the features of an X include Y and Z* for the association of a concept and its attributes. She also cited Ahmad and Fulford (1992) and their suggestion that such linguistic items could be used as search patterns for discovering conceptual information in corpora.

In her later work (e.g., Meyer et al. 1999; Meyer 2001), Meyer further developed this terminology, dividing what was first described in the 1994 definition of the knowledge-rich context into two separate concepts: "knowledge-rich contexts," text

segments that provide at least one piece of information about a concept (e.g., one of its attributes or a relation in which it participates), and "knowledge patterns," linguistic structures that frequently indicate a relation (or, more rarely, an attribute) (e.g., *X is a kind of Y*).

Knowledge patterns have been classified by Meyer (e.g., Meyer et al. 1999) into three categories: lexical, grammatical and paralinguistic. *Lexical knowledge patterns* prototypically take the form X + [MARKER] + Y (e.g., *X is a kind of Y* for the relation of HYPERONYMY), i.e., including two elements linked by a relation (here represented by the variables *X* and *Y*), and a lexical unit or sequence of lexical units that indicate the relation between them.[2] *Grammatical* (syntactic) patterns, which involve parts of speech or combinations thereof (e.g., NOUN + VERB for the FUNCTION relation), and *paralinguistic* patterns, involving for example formatting and punctuation (e.g., parentheses used to introduce a synonym, in a structure such as *X (Y)*), may also indicate relations.[3]

In many research projects, lexical markers and/or knowledge patterns are subject to syntactic restrictions, e.g., indicating the part of speech class to which the markers and potentially other elements of the structures in which they participate may belong; in this case a fourth category of *lexico-syntactic knowledge patterns* may be identified.

---

[2] In some cases, items such as derivational affixes (e.g., *pro-*, *anti-*) may also be included in this category, although they are not strictly speaking lexical units.

[3] It has been observed (e.g., L'Homme, personal communication) that grammatical knowledge patterns are somewhat different from the other two categories in that the link between the two elements participating in the relation is present at the level of the senses of the units that denote them, rather than external to them as in the case of lexical and paralinguistic knowledge patterns. (Lexical patterns may be seen then as equivalent to paraphrases of the relation that is inherent in the senses of the two items in grammatical knowledge patterns, e.g., *the computer processes data* / *the function of a computer is to process data*.) Grammatical knowledge patterns are nevertheless indicators of the presence of a given relation in a text segment, which is of course the application envisaged in the establishment of the categories of knowledge patterns.

As described in the Introduction, these patterns may then be used by computer tools to aid in the gathering of information for applications such as terminology work; most projects have focused on the identification of conceptual (or semantic) relations.

1.3 Conceptual relations

In the traditional view of terminology, the centre and starting point of any research is a conceptual system constructed with clearly delineated and described concepts. That is to say, research begins with mental images of classes of entities, processes, qualities, etc. in the real world and move on from there to study their attributes, the relations between them in the conceptual system of the domain, and finally the terms used to denote these concepts.

Sager's (1990) description of concept analysis is based on attributes and relations. Attributes are qualities or properties associated with the concept in isolation. These may include such things as colour, measurements, and other properties of the real-world element the concept represents. Relations are the ways in which the concept relates to other concepts, which determine its place in a concept system (knowledge structure). The most commonly studied relations are those of specific to generic, sometimes called *HYPERONYMY*, and of whole to part, or *MERONYMY*. However, other important conceptual relations also exist, including among many others those of FUNCTION and CAUSE–EFFECT. In medicine, as in any domain, important information can be provided by both concept attributes and conceptual relations.

In the less traditional theories of terminology, the conception of terminology as onomasiological and concept-based has been challenged, and the gaps between traditional theory and current practice have been recognized. It has been accepted that terminology work in the real world often takes a semasiological approach, which starts with the term itself. Accompanied by a drastically reduced insistence on the bi-univocity of the relation between term and concept, there has been increased study of the different

kinds of synonymy in terminological systems, and a recognition that some terms are polysemous. More study is being devoted to the collocational and combinatory properties of terms. Moreover, there has been more recognition of terms' complex and inter-related meanings, in contrast to the traditional view of terms as no more than labels for concepts organized in a clear-cut, hierarchical structure.

Thus, although this study is primarily concerned with conceptual relations — that is, relations between the concepts denoted by terms — it is impossible to completely separate many of these from semantic relations, which hold between the meanings or significations of terms, as reflected by their place in the system of signs in a language.

From a terminological perspective, Condamines and Rebeyrolle (2001: 131) discussed the importance of relations between concepts:

> The search for conceptual relationships plays an important role in building a CTKB [corpus-based terminological knowledge base] as long as it is mainly a model of the text content. From this point of view, the most important knowledge within the text is conveyed by conceptual (or semantic) relationships.

As conceptual and semantic relations are of course closely linked, there is a certain variation in the use of these terms. In analyzing previous research, it is common to see variations in the terminology used to describe the same project, even occasionally within the same work. This complexity, and the reasons it must be confronted, are reflected in the description by Ahmad and Rogers (1997: 749):

> Since terminology management and terminology research emphasizes the conceptual organization of subject fields, the semantic relations between terms assume considerable importance. According to the traditional terminological view, the knowledge of the domain is represented by concepts and the relations between them. However, these relations cannot be directly accessed but must be conveyed by largely linguistic means. The relations between terms, [sic] as labels for concepts are therefore a means of accessing this knowledge (and its structure) through text.

This description conveys the inextricable links between the linguistic and conceptual levels in texts, and the need to study not only the links between concepts in their conceptual system — the knowledge structure of the domain — but also the terms that are used to denote these concepts, which provide access to this conceptual system while participating in relations of their own.

The type of approach used in this work, inspired by Meyer and oriented towards identifying knowledge-rich contexts to assist in conceptual analysis, was originally associated with the more traditional view of terminology. However, as Ahmad and Rogers note, it is terms (and other linguistic items) and the relations between them in texts that provide access to this conceptual information.

Moreover, it would be impossible to ignore the fact that terms' linguistic nature will also influence the ways in which they are used in texts. Thus, the analysis here will provide knowledge not only about the concepts denoted by the terms found in texts, but also about the place of these terms in the linguistic system.

1.4 Relation classifications in terminology

While there is general agreement among scholars about the importance of GENERIC–SPECIFIC (or HYPERONYMY) and PART–WHOLE (or MERONYMY) relations in conceptual and terminological systems, there is no widely accepted list of possible relations or a system for classifying them (cf. Chaffin and Hermann 1988).[4] However, relation classifications have been constructed by various authors; below, those presented in two general classifications in terminology — by Sager (1990) and Nuopponen (2005) — and two specific to medicine and related fields — by Feliu (2004) and the Unified Medical Language System (UMLS) (2005) — will be described. Before these classifications are presented, however, some criteria useful for their development may be examined.

---

[4] Moreover, as Feliu noted (2004: 27), there is little agreement between relations and their denominations used in the field of terminology and those in lexical semantics.

1.4.1 Some criteria for classifying conceptual relations

A number of criteria may be used to characterize and classify semantic and conceptual relations in order to facilitate their differentiation, definition and description (cf. Section 1.4.4 on Feliu (2000, 2004), who used some of these in developing and describing her relation typology).

The first of these is the distinction between hierarchical and non-hierarchical relations. The first type, hierarchical relations, involve relations that hold between superordinate and subordinate concepts and include the GENERIC–SPECIFIC and PART–WHOLE relations; these may be used as the basis for creating tree-like concept structures. Non-hierarchical relations, in contrast, link concepts in ways that do not permit this kind of structuring; these relations include those of FUNCTION, CAUSE–EFFECT and ASSOCIATION.

Another criterion for relation classification is the distinction between what Cruse (1986: 113) calls *symmetric* and *asymmetric* relations. In symmetric relations such as SYNONYMY and ASSOCIATION, the link is bi-directional: if A is a synonym of B, then B is a synonym of A, and if A is associated with B, then B is associated with A. In asymmetric relations, such as GENERIC–SPECIFIC and CAUSE–EFFECT, the relation is uni-directional: if A is the generic of B, then B cannot be the generic of A; by the same token, if A causes B this does not imply that B causes A.[5]

Finally, there is the criterion of transitivity. As Cruse (1986: 114) stated, "A relation is said to be transitive if the fact that it holds between two elements A and B, and also between B and a third element C, guarantees that it holds between A and C." He gives the example of the relation IS LONGER THAN to illustrate this, since if A is longer than B and B is longer than C, then A must be longer than C. Conversely Cruse

---

[5] In certain specific cases of cyclical processes, this might in fact be the case, but the expression of the relation is concerned only with the relation of X leading to Y, and in order to describe such a cycle it would be necessary to state separately that Y in turn is the cause of X.

states (114) that intransitive relations are those in which the fact that the relation holds between A and B and between B and C entails that it does not hold between A and C. Since if A is the father of B and B is the father of C, A cannot be the father of C, Cruse identifies the relation FATHER OF as intransitive.

1.4.2 Sager

In his textbook *A Practical Course in Terminology Processing*, Sager (1990) describes the cognitive importance of relations in concept systems. He presents a list of possible relations (Table 1), and also notes (1990: 35) that these relations may be further sub-divided by placing the concepts involved into conceptual reference classes (e.g., objects, methods, properties, qualities, states, processes), or into more general classes (e.g., entities, activities, qualities, relations (cf. 1990: 26–28)). Sager states (1990: 35) that:

> The relationship between two concepts is bound by the conceptual class of each. For example, relationships of product or material can only exist between material entities; in this way a pattern emerges which shows restrictions on the nature of the relationships between concepts by virtue of their categories. Examining concepts in this way may lead to greater insight into ways of establishing conceptual relations.

Sager presents some complexities of relation classification, observing (1990: 29) that limiting the study of conceptual relations to GENERIC–SPECIFIC, PART–WHOLE and the generic "other" is not sufficient in terminology, and that there are exceptions to the seeming simplicity of the GENERIC–SPECIFIC relation (1990: 31–32): the existence of facets, and of quasi-GENERIC relationships. He notes that concepts are classified into types on the basis of a given criterion, which may be only one of many possibilities. It is necessary to specify the criterion used for classification in order to properly classify concepts. Sager uses the term *facetted* classification to denote classification of concepts on the basis of a particular characteristic (e.g., by parts, by process, by method, by function). He also notes that a given concept may be subdivided according to different facets. In addition, some assignments of concepts to types may be more "solid" (1990:

32) than others. For this reason, quasi-GENERIC relationships may also be identified, using the test below (1990: 32). Sager applied this text to the classification of the concept "dandelion," considered to be a weed by some, a medicinal plant by others, and a vegetable by still others, but consistently classified in the same botanical family:

> Generic relationship:
> - All dandelions are members of the family of Compositae.
> - Some members of the family of Compositae are dandelions.

> Quasi-generic relationship
> - Some people consider that dandelions are vegetables.
> - Some vegetables are dandelions.

Sager (1990: 33–4) also notes the existence of polyvalent relationships, in which a concept may have several possible places in a conceptual system (i.e., may be part of more than one hierarchy in a given subject field).

This classification, although not detailed in its description of the nature and possibilities of different relations, presents an initial portrait of the possibilities that may be envisaged for classifying concepts. In addition, Sager points out (1990: 29) that the most useful classification for a given project will be determined in part by the context in which it will be used, i.e., the subject field being studied, and the type of research in which the classification is to be applied.

Table 1. Summary of Sager's conceptual relations (adapted from Sager 1990: 29–37)

Relation	Description	Example
GENERIC (29–32)	A hierarchical relationship in which concepts belong to the same category, and in which the broader (generic) concept is said to be the superordinate of the narrower (specific) concept(s). Formulae: X is a type of A; X, Y and Z are types of A; A has the specific concepts X, Y and Z; A has the subtype X	periodical publications – newsletter, journal, magazine
PARTITIVE (32–33)	A hierarchical relationship that serves to indicate the connection between concepts consisting of more than one part and their constituent parts. Formulae: X is a constituent part of Y; X, Y and Z are constituent parts of A; A consists of X; A consists of X, Y and Z	wheel – hub, spokes, rim
COMPLEX	Complex interrelations between concepts that cannot be conveniently captured by straightforward generic or partitive structures. These may be not only as important in a conceptual system as the hierarchical relationships, but also more revealing about the nature of the concepts they involve. There are many possible complex relations:	
CAUSE–EFFECT	Formula: Y is caused by X	fallout – nuclear explosion
MATERIAL–PRODUCT	Formula: Y is a product of X	steel – girder paper – wood pulp
MATERIAL–PROPERTY	Formula: Y is a property of X	glass – brittle
MATERIAL–STATE	(no formulae indicated)	iron – corrosion
PROCESS–PRODUCT	Formula: Y is a product of X	weaving – cloth petrol – oil refining
PROCESS–INSTRUMENT	Formula: Y is an instrument for X	incision – scalpel data processing – computer
PROCESS–METHOD	Formula: Y is a method of X	storage – freeze-dry

PROCESS–PATIENT	(no formulae indicated)	dyeing – textile
PHENOMENON–MEASUREMENT	Formula: Y is a [quantitative/qualitative] measure of X	light – Watt heat – temperature
OBJECT–COUNTERAGENT	Formula: Y is a counteragent of X	poison – antidote insects – insecticide
OBJECT–CONTAINER	Formula: Y is a container for X	tool – tool box
OBJECT–MATERIAL	(no formulae indicated)	bridge – iron
OBJECT–QUALITY	(no formulae indicated)	petrol – high octane
OBJECT–OPERATION	(no formulae indicated)	drill bit – drilling
OBJECT–CHARACTERISTIC	(no formulae indicated)	fuel – smokeless
OBJECT–FORM	(no formulae indicated)	book – paperback
ACTIVITY–PLACE	Formula: Y is a place for X	coalmining – coal mine

1.4.3 Nuopponen

As recently as 2005, Nuopponen observed (2005: 127) that (despite previous observations by Sager (1990), among others) standard applications in terminology still make use of a limited number of relations, primarily GENERIC–SPECIFIC and PART–WHOLE, although the often vague category of "association relations" — which will be further discussed below in Sections 1.4.4 and 1.4.5 — may also be used in some cases. She also pointed out that for many applications in terminology, including Semantic Web applications and concept modelling, there is a need for a much wider and finer typology of relations. For this reason, she updated and enlarged a classification originally developed in 1994.[6]

Nuopponen recalled several observations also made by a number of other researchers (once again including Sager (1990)). She noted that the relations that are pertinent in the context of a given research project or domain may vary, and that users may choose among the proposed relations those that are most useful for their purposes, and may choose to draw finer distinctions or to add new relations as necessary (2005: 128).[7] Nuopponen also noted (2005: 130) close links between categories of concepts (e.g., entities, activities, processes, methods, properties) and the ontological relations in which they may participate.

Nuopponen's typology of relations is rooted in the distinction made by Wüster (1974, 1985) between logical (i.e., GENERIC–SPECIFIC) and ontological relations, on the basis of these relations' directness (logical) or indirectness (ontological) and the definition of ontological relations as simplifications of relations observed between individual objects in reality. Her relation classifications for logical and ontological

---

[6] In carrying out this task, Nuopponen referred frequently to the relation hierarchy used by Madsen et al. (2001, 2002, 2002a) in their OntoQuery project, as a basis for comparison and for expansion of her own hierarchy.

[7] This comment was intended in part to address some criticisms of her previous research, in which the usefulness of such a detailed relation classification for practical terminology work was questioned.

relations are shown in Table 2 and Table 3 respectively. The author attempted to provide a unique place in the hierarchy for each relation, but also noted that some relation types could belong to several different classes (2005: 134).

Nuopponen noted that logical relations can be classified according to two dimensions, the relative positions of concepts in a concept system or hierarchy (as in the case of SUPERORDINATION, SUBORDINATION and COORDINATION) or by the comparison of concepts according to their intensions or extensions (as in the case of relations of IDENTITY, INCLUSION, OVERLAPPING and DISJUNCTION).

Table 2. Nuopponen's logical concept relations (2005: 129–130)

1. LOGICAL CONCEPT RELATIONS	1.1 SUPERORDINATION	1.1.1 DIRECT SUPERORDINATION
		1.1.2 INDIRECT SUPERORDINATION
	1.2 SUBORDINATION	1.2.1 DIRECT SUBORDINATION
		1.2.2 INDIRECT SUBORDINATION
	1.3 COORDINATION	1.3.1 DIRECT COORDINATION
		1.3.2 INDIRECT COORDINATION
	1.4 DIAGONAL RELATION	
	1.5 INTENSIONAL RELATION	1.5.1 INTENSIONAL IDENTITY
		1.5.2 INTENSIONAL INCLUSION
		1.5.3 INTENSIONAL OVERLAPPING
		1.5.4 INTENSIONAL DISJUNCTION
	1.6 EXTENSIONAL RELATION	1.6.1 EXTENSIONAL IDENTITY
		1.6.2 EXTENSIONAL INCLUSION
		1.6.3 EXTENSIONAL OVERLAPPING
		1.6.4 EXTENSIONAL DISJUNCTION

Causal concept relations appear in this typology under the classification of ONTOLOGICAL INFLUENCE relations (a category defined by the presence of some kind of causal component in the relation, i.e., a one-sided or mutual influence). (See Section 1.5.2.5 for a detailed description of Nuopponen's analysis of the CAUSE–EFFECT relation in 1994.) The CORRELATION sub-type of INTERACTIONAL relations identified in Nuopponen (2005) can be considered to be a type of ASSOCIATION relation (see Section 1.5.1 for a discussion of ASSOCIATION and CORRELATION).

Table 3. Nuopponen's ontological concept relations (2005: 130–135)

<table>
<tr><td rowspan="15">2. ONTOLOGICAL CONCEPT RELATIONS</td><td rowspan="15">2.1 CONCEPT RELATIONS OF CONTIGUITY</td><td rowspan="9">2.1.1 PARTITIVE RELATION</td><td rowspan="6">COMPOUND RELATION[8]</td><td rowspan="2">2.1.1.1 PARTITIVE SUPERORDINATION</td><td>2.1.1.1.1 CANONICAL SUPERORDINATION</td></tr>
<tr><td>2.1.1.1.2 FACULTATIVE SUPERORDINATION</td></tr>
<tr><td rowspan="2">2.1.1.2 PARTITIVE SUBORDINATION</td><td>2.1.1.2.1 CANONICAL SUBORDINATION</td></tr>
<tr><td>2.1.1.2.2 FACULTATIVE SUBORDINATION</td></tr>
<tr><td rowspan="2">2.1.1.3 PARTITIVE COORDINATION</td><td>2.1.1.3.1 DIRECT PARTITIVE COORDINATION</td></tr>
<tr><td>2.1.1.3.2 INDIRECT PARTITIVE COORDINATION</td></tr>
<tr><td colspan="3">PARTITION RELATION</td></tr>
<tr><td rowspan="2">SET RELATION</td><td colspan="2">SET–ELEMENT RELATION</td></tr>
<tr><td colspan="2">ELEMENT–ELEMENT RELATION</td></tr>
<tr><td colspan="4">2.1.2 ENHANCEMENT RELATION</td></tr>
<tr><td colspan="4">2.1.3 LOCATIVE RELATION</td></tr>
<tr><td colspan="4">2.1.4 MATERIAL COMPONENT RELATION</td></tr>
<tr><td colspan="4">2.1.5 PROPERTY RELATION</td></tr>
<tr><td colspan="4">OWNERSHIP RELATION</td></tr>
<tr><td rowspan="2">2.1.6 RANK RELATION</td><td colspan="3">2.1.6.1 RELATION OF ORDER</td></tr>
<tr><td colspan="3">2.1.6.2 RELATION OF EQUIVALENCE</td></tr>
<tr><td rowspan="4">2.1.7 TEMPORAL RELATION</td><td colspan="3">2.1.7.1EVENT RELATION</td></tr>
<tr><td colspan="3">2.1.7.2. SUCCESSION RELATION</td></tr>
<tr><td colspan="3">2.1.7.3 SIMULTANEOUS RELATION</td></tr>
<tr><td colspan="3">2.1.7.4 CONSECUTIVE RELATION</td></tr>
</table>

[8] Relations indicated in italics without numbering are additions to the concept relation typology since the 1994 version.

	2.2 Concept relations of influence	2.2.1 Causal relations	2.2.1.1 Causal sequence	Producing cause – effect
				Explanatory cause – effect
				Causal agent – effect
				Cause – resulting event
				Cause – resulting state
				Cause – resulting product
			2.2.1.2 Causal coordination	2.2.1.2.1 Multicausality
				2.2.1.2.2 Multiple effect relation
		2.2.2 Development relations	2.2.2.1 Phylogenetic relation	
			2.2.2.2 Ontogenetic relation	
			2.2.2.3 Genealogic relation	
			2.2.2.4 Material development relation	
			2.2.2.5 Role change	
		2.2.3 Functional relations	2.2.3.1 Activity relations	2.2.3.1.1 Agent relation
				2.2.3.1.2 Object relation
				2.2.3.1.3 Tool relation
				2.2.3.1.4 Locational relation
				2.2.3.1.5 Temporal action relation
				Teleological relation (action – purpose)
				Resultative relation (2.2.3.2.3)
			2.2.3.2 Origination relation	2.2.3.2.1 Originator relation
				2.2.3.2.2 Product – instrument relation
				2.2.3.2.3 Resultative relation
				2.2.3.2.4 Ingredient relation
				2.2.3.2.5 Origination place relation
				2.2.3.2.6 Origination time relation
				Product – purpose of creation
			Instrumental relations	*Agent – instrument*
				Function relation (entity – way or working)
				Tool relation (2.2.3.1.3; activity – tool)
				Product – instrument relation (2.2.3.2.2)
				Locative relation (2.1.3 object – location)

<table>
<tr><td rowspan="10"></td><td rowspan="10"></td><td rowspan="10">2.2.4 INTERACTIONAL RELATIONS</td><td rowspan="8">2.2.4.1 TRANSMISSION RELATION</td><td colspan="2">2.2.4.1.1 DIRECT TRANSMISSION RELATION (SENDER – RECEIVER)</td></tr>
<tr><td rowspan="2">2.2.4.1.2 SEQUENTIAL TRANSMISSION RELATION</td><td>2.2.4.1.2.1 SENDER – INTERMEDIARY</td></tr>
<tr><td>2.2.4.1.2.2 INTERMEDIARY – RECEIVER</td></tr>
<tr><td rowspan="2">2.2.4.1.3 SOURCE RELATION</td><td>2.2.4.1.3.1 SENDER / PLACE OF DEPARTURE – OBJECT</td></tr>
<tr><td>2.2.4.1.3.2 INTERMEDIARY – OBJECT</td></tr>
<tr><td rowspan="2">2.2.4.1.4 TARGET RELATION</td><td>2.2.4.1.4.1 OBJECT – RECEIVER/DESTINATION</td></tr>
<tr><td>2.2.4.1.4.2 OBJECT – INTERMEDIARY</td></tr>
<tr><td colspan="2">2.2.4.2 DEPENDENCY RELATION</td></tr>
<tr><td colspan="2">2.2.4.3 CORRELATION RELATION</td></tr>
<tr><td colspan="2">2.2.4.4 REPRESENTATIONAL RELATION</td></tr>
</table>

1.4.4 Feliu

In research projects that focused on analyzing Catalan corpora in the fields of heart disease and genomics, Feliu (2000, 2004) addressed the lack of a unified relation typology in terminology. She noted that although lexical semanticians such as Cruse and Lyons and terminologists such as Wüster had identified a certain number of relations (between lexical units in the case of Cruse and Lyons, and between concepts in the case of Wüster), their lists were not exhaustive, and most studies had focused largely on the hierarchical relations of HYPERONYMY and MERONYMY. In order to fill this gap as far as possible, Feliu identified relations described in the literature, and complemented these with observations from a specialized corpus as required. She produced the relation typology presented in Table 4 (adapted from Feliu 2004: 51).

Feliu (2004: 25–7) described relations between concepts as one of the fundamentals of human perception and cognition, and — referring to Otman (1996: 55–6) — noted on the subject of conceptual relations that:

- A conceptual relation is a conceptual link between concepts;
- In a relational model, a concept is defined by the relations that hold between it and other concepts;
- A conceptual relation consists of:
  - A name or identifier specifying the type of relation;
  - The specification of the types of objects the relation admits;
  - The attribution of specific properties to these objects; and
  - Sometimes, conditions of validity.

Feliu also noted (2004: 27) that relations are at least binary, involving two or more concepts, and that a given relationship could conceivably be described using more than one relation name (e.g., ELEMENT–QUALIFICATION or CHARACTERISTIC–ACTIVITY), although she reduced her list of relations as much as possible to eliminate redundancies (2004: 31).

Table 4. Summary of Feliu's relations[9]

Relation		Elements	Description
RESEMBLANCE	POSITIVE RESEMBLANCE		A symmetric and transitive relation. Marker: *ser semblant a*
	TOTAL EQUIVALENCE, SYNONYMY		
	PARTIAL EQUIVALENCE, RESEMBLANCE		
	NEGATIVE RESEMBLANCE		A symmetric and transitive relation. Marker: *ser diferent de*
	OPPOSITENESS		
	PARTIAL OPPOSITENESS, CONTRAST		
INCLUSION	HYPONYMY	GENERIC–SPECIFIC	An asymmetric and transitive relation. Marker: *ser (un tipus) de*
SEQUENCE	SPATIAL		An asymmetric relation. Markers: *ser en*; *ser davant*; *ser darrere*; *anar de x a y*
	LOCATION		
	DIRECTION		
	TEMPORAL		An asymmetric and transitive relation. Marker: *ser simultani/anterior/ posterior a*
	SIMULTANEITY		
	ANTERIORITY–POSTERIORITY		
CAUSALITY	CAUSAL		An asymmetric relation.
	CAUSE–EFFECT		An explicit cause gives rise to a given effect. (33–4) Markers: *causar*; *ser la causa de*; *ser l'efecte de*
	PROCESS–RESULT[10]		A process produces a result, although it may not be seen as a true cause. (33–4) Marker: *produir*; *fer que*

[9] Translations from the Catalan are mine.

[10] Feliu also noted in a previous version of the typology (2004: 44–45) that the manifestation of PROCESS–RESULT relations may vary at the surface level, depending on the specificity with which the type of change that occurs in the process is indicated.

INSTRUMENT	INSTRUMENT–FUNCTION		An asymmetric and intransitive relation. Markers: *servir per a*; *fer-se amb*
MERONYMY	PART–WHOLE		Markers: *ser una part/element de*; *tenir + SN*; *estar format / fet per*; *incloure*; *constar de*; *pertànyer a*
		COMPONENT–OBJECT	An asymmetric relation.
		MEMBER–COLLECTION	An asymmetric relation.
		PORTION–MASS	An asymmetric relation.
		MATERIAL–OBJECT	An asymmetric relation.
		STAGE–PROCESS	An asymmetric relation.
		CHARACTERISTIC–ACTIVITY	An asymmetric relation.
		LOCATION–AREA	An asymmetric relation.
ASSOCIATION	GENERAL		
	SPECIALIZED[11]		Marker: *correlacionar-se amb*

Unsurprisingly, Feliu identified the CAUSE–EFFECT relation as important in her domains. She also distinguished different types of this relation depending on the nature of the causes that participate in them, citing in particular processes and their results. The presence of the ASSOCIATION relation in Feliu's typology, defined as a "relació que s'estableix per la corelació entre dos o més elements" [Eng. a relation of correlation between two or more elements] (39), may also be noted. (See Section 1.5.1 for a discussion of this relation.)

1.4.5 UMLS

The goal of the Unified Medical Language System (UMLS) Semantic Network (2005) is to represent concepts in the medical field, and to link them together using a set of relations that are pertinent in this domain. The relations used and their definitions are

[11] See Section 1.5.1.2 for a discussion of the distinction made in the case of this relation.

illustrated below in Figure 1 and Table 5. The list of possible relations and their organization may be observed to be distinctly different from those illustrated in Sections 1.4.2 to 1.4.4.

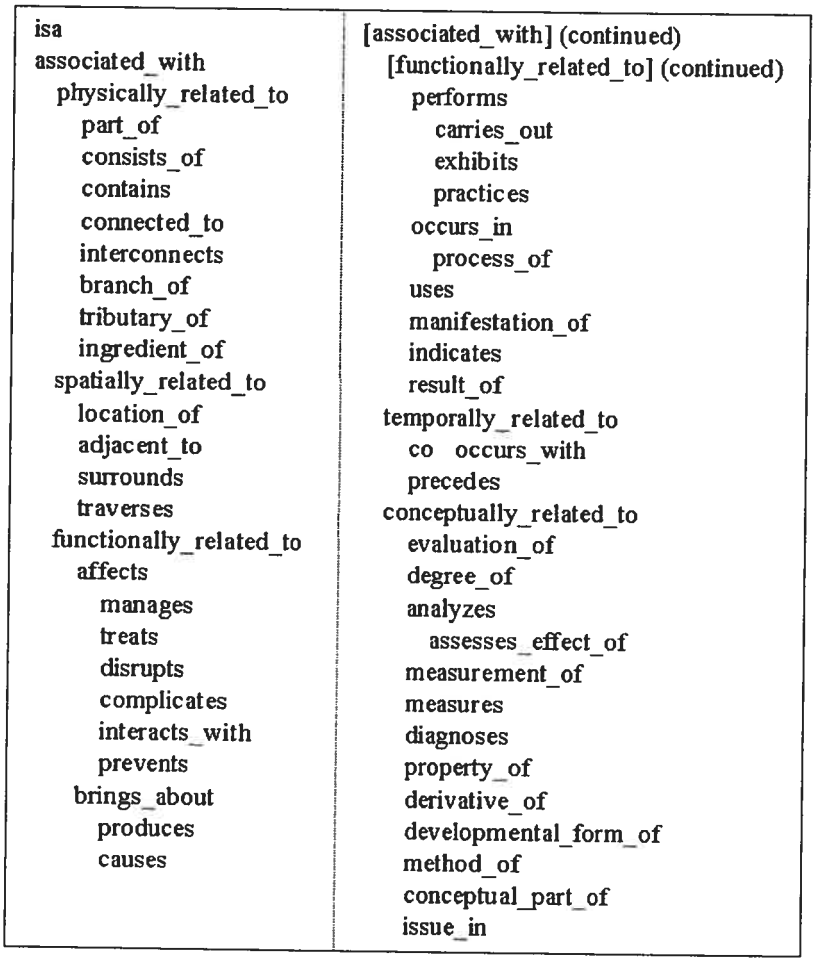


Figure 1. Relations in the UMLS Semantic Network (UMLS 2004)

Table 5. UMLS semantic relations and definitions (UMLS 2005)

Relation	Definition
ISA	The basic hierarchical link in the Network. If one item "isa" another item then the first item is more specific in meaning than the second item.
ASSOCIATED_WITH	Has a significant or salient relationship to.
PHYSICALLY_RELATED_TO	Related by virtue of some physical attribute or characteristic.
PART_OF	Composes, with one or more other physical units, some larger whole. This includes component of, division of, portion of, fragment of, section of, and layer of.
CONTAINS	Holds or is the receptacle for fluids or other substances. This includes is filled with, holds, and is occupied by.
CONSISTS OF	Is structurally made up of in whole or in part of some material or matter. This includes composed of, made of, and formed of.
CONNECTED_TO	Directly attached to another physical unit as tendons are connected to muscles. This includes attached to and anchored to.
INTERCONNECTS	Serves to link or join together two or more other physical units. This includes joins, links, conjoins, articulates, separates, and bridges.
BRANCH_OF	Arises from the division of. For example, the arborization of arteries.
TRIBUTARY_OF	Merges with. For example, the confluence of veins.
INGREDIENT_OF	Is a component of, as in a constituent of a preparation.
TEMPORALLY_RELATED_TO	Related in time by preceding, co-occurring with, or following.
CO-OCCURS_WITH	Occurs at the same time as, together with, or jointly. This includes is co-incident with, is concurrent with, is contemporaneous with, accompanies, coexists with, and is concomitant with.
PRECEDES	Occurs earlier in time. This includes antedates, comes before, is in advance of, predates, and is prior to.
FUNCTIONALLY_RELATED_TO	Related by the carrying out of some function or activity.
MANIFESTATION_OF	That part of a phenomenon which is directly observable or concretely or visibly expressed, or which gives evidence to the underlying process. This includes expression of, display of, and exhibition of.
AFFECTS	Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences,

	predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.
INTERACTS_WITH	Acts, functions, or operates together with.
DISRUPTS	Alters or influences an already existing condition, state, or situation. Produces a negative effect on.
PREVENTS	Stops, hinders or eliminates an action or condition.
COMPLICATES	Causes to become more severe or complex or results in adverse effects.
MANAGES	Administers, or contributes to the care of an individual or group of individuals.
TREATS	Applies a remedy with the object of effecting a cure or managing a condition
OCCURS_IN	Takes place in or happens under given conditions, circumstances, or time periods, or in a given location or population. This includes appears in, transpires, comes about, is present in, and exists in.
PROCESS_OF	Action, function, or state of.
USES	Employs in the carrying out of some activity. This includes applies, utilizes, employs, and avails.
INDICATES	Gives evidence for the presence at some time of an entity or process.
RESULT_OF	The condition, product, or state occurring as a consequence, effect, or conclusion of an activity or process. This includes product of, effect of, sequel of, outcome of, culmination of, and completion of.
BRINGS ABOUT	Acts on or influences an entity.
PRODUCES	Brings forth, generates or creates. This includes yields, secretes, emits, biosynthesizes, generates, releases, discharges, and creates.
CAUSES	Brings about a condition or an effect. Implied here is that an agent, such as for example, a pharmacologic substance or an organism, has brought about the effect. This includes induces, effects, evokes, and etiology.
PERFORMS	Executes, accomplishes, or achieves an activity.
CARRIES_OUT	Executes a function or performs a procedure or activity. This includes transacts, operates on, handles, and executes.
PRACTICES	Performs habitually or customarily.
EXHIBITS	Shows or demonstrates.
CONCEPTUALLY_RELATED_TO	Related by some abstract concept, thought, or idea.

PROPERTY_OF	Characteristic of, or quality of.
CONCEPTUAL_PART_OF	Conceptually a portion, division, or component of some larger whole.
EVALUATION_OF	Judgment of the value or degree of some attribute or process.
MEASURES	Ascertains or marks the dimensions, quantity, degree, or capacity of.
DIAGNOSES	Distinguishes or identifies the nature or characteristics of.
ISSUE_IN	Is an issue in or a point of discussion, study, debate, or dispute.
DERIVATIVE_OF	In chemistry, a substance structurally related to another or that can be made from the other substance. This is used only for structural relationships. This does not include functional relationships such as metabolite of, by product of, nor analog of.
DEVELOPMENTAL_FORM_OF	An earlier stage in the individual maturation of.
DEGREE_OF	The relative intensity of a process or the relative intensity or amount of a quality or attribute.
MEASUREMENT_OF	The dimension, quantity, or capacity determined by measuring.
METHOD_OF	The manner and sequence of events in performing an act or procedure.
ANALYZES	Studies or examines using established quantitative or qualitative methods.
ASSESSES_EFFECT_OF	Analyzes the influence or consequences of the function or action of.
SPATIALLY_RELATED_TO	Related by place or region.
LOCATION_OF	The position, site, or region of an entity or the site of a process.
ADJACENT_TO	Close to, near or abutting another physical unit with no other structure of the same kind intervening. This includes adjoins, abuts, is contiguous to, is juxtaposed, and is close to.
SURROUNDS	Establishes the boundaries for, or defines the limits of another physical structure. This includes limits, bounds, confines, encloses, and circumscribes.
TRAVERSES	Crosses or extends across another physical structure or area. This includes crosses over and crosses through.

The division of all of the relations into two main categories, the GENERIC relations and all others (which thus fall under the general heading of the ASSOCIATION_WITH relations), parallels the division established by Wüster (cf. Nuopponen's use of this distinction as described in Section 1.4.3, specifically Figure 2) and illustrates the emphasis that is placed on this former relation. Every other type of relation is considered to be a kind of ASSOCIATION; this classification includes temporal, spatial, functional and causal relations. In addition, a number of domain-specific relations related to the medical field, such as TREATS and DIAGNOSES, which did not appear in Feliu's classification, are also identified.

Within the relation sub-category FUNCTIONALLY_RELATED_TO, several relations including an element of causation may be identified, including:

- AFFECTS, and its sub-types
  - DISRUPTS
  - PREVENTS
  - COMPLICATES, and
  - MANAGES
- RESULT_OF, and
- BRINGS_ABOUT, with its sub-types
  - CAUSES, and
  - PRODUCES.

However, these CAUSE–EFFECT relations are not clearly defined as such, and do not constitute their own category. Moreover, among sub-types of the AFFECTS relation, it may be difficult to determine the presence of a causal element (e.g., INTERACTS_WITH). This seems to indicate that although CAUSE–EFFECT relations are important enough in the domain to be identified and distinguished in large number, the causal element is not one that is a priority in the classification.

One interesting aspect of the relation definitions given in the UMLS is the reliance on linguistic indicators (essentially lexical markers) in order to explain the nature of the various relations (e.g., citing *composed of, made of*, and *formed of* for the CONSISTS_OF relation, and *component of, division of, portion of, fragment of, section of,*

and *layer of* for the PART_OF relation). This illustrates that although the perspective of this resource is far more conceptually than linguistically oriented, there is nevertheless a close connection between concept and language that cannot be severed, and that this link is a valuable tool for identifying relations as described in texts.[12]

However, it should be noted that the distinctions made in this classification (e.g., the one above, between the relations CONSISTS_OF and PART_OF, which would both generally be subsumed under the heading of MERONYMY, though perhaps in different sub-types, but here are separated) may be too fine for some semi-automatic knowledge extraction applications. In addition, these relations are specifically adapted to the medical domain and as such would not be appropriate for use in many other fields.

1.4.6 Comparison

Two important observations can be made in analyzing these classification systems — and particularly the UMLS. First, the various relations identified are closely linked. Second, a given pair of concepts may be considered to be related in many different ways. Thus, in classifying a given occurrence of a relation, it may be difficult to precisely determine which relation is the best fit. Moreover, by comparing a system such as the UMLS to the others presented in this Section, various levels of granularity with which different relations may be characterized may be observed. This reinforces the statements made, for example, by Nuopponen (2005: 128) and Sager (1990), observing that the subject field being studied may influence the choice of relations considered.

The classifications described above reveal how intended applications can affect the relations identified and how they are organized. Those dealing with the medical domain, and particularly the UMLS, identify a certain number of sub-relations that are

---

[12] Feliu (2004: 232), however, identifies the lack of distinction between relations at a semantic or conceptual level and the markers of these relations at a textual level as a source of fundamental difficulties in the development of relation typologies, stressing the necessity of appreciating the difference between the two levels.

missing from Sager's (1990) and/or Nuopponen's (2005) more general classifications. Feliu's and Sager's more terminologically oriented classifications are far simpler and less atomized; from the point of view of semi-automatic identification and/or classification of relation occurrences in texts, these provide a more reasonable level of detail — and certainly of organization at an intuitive level — than the UMLS.[13] All four sources refer to varying extents to the elements that may participate in the various relations identified, reflecting the importance of this aspect in relation identification and classification on a fine-grained level.

Some distinctions and relations are common to all or most of the classifications identified; the distinction between the GENERIC–SPECIFIC and other relations is observable in all of the classifications, and the PART–WHOLE relation is also clearly identified. All of the classifications except for Sager's identify relations of SPATIAL and TEMPORAL CONTIGUITY as important. In terms of the relations considered in this research, the centrality of the CAUSE–EFFECT relation is made evident by its inclusion (in some form) in all four classifications. The relation of ASSOCIATION is also relatively common in the classifications, with Nuopponen, Feliu and the UMLS identifying relations that correspond at least in part to the definition used for this research.

1.5 Important conceptual relations in medicine

The importance of a wide range of relations for properly, precisely and comprehensively representing knowledge structures in the field of medicine is illustrated, for example, by the long and complex list of relations used in the UMLS (Section 1.4.5). While GENERIC–SPECIFIC and MERONYMY relations are clearly central, others are also critical.

Researchers such as Nuopponen (1994) have recognized the importance of the CAUSE–EFFECT relation in scientific and technical fields, and in particular in medicine.

---

[13] However, as noted in the description of the UMLS's relation definitions, linguistic markers are used as an aid to understanding, and these might serve as a starting point for semi-automatic applications.

In fact, in many ways, the CAUSE–EFFECT relation can be considered to be the central one in this domain: medicine is, after all, the study of the causes of disease and health, the effects diseases have, and the intended and side effects of treatments. All of these can be represented — in a simplified manner, of course — by the CAUSE–EFFECT relation. This relation is not only critical, but also complex, and has been studied in the context of several research projects (Nuopponen 1994; Cabré et al. 1996, 2001; Garcia 1996, 1997; Barrière 2001, 2002; Marshman 2002, 2002a, 2004, 2004a; Feliu 2004; Bodson 2005). However, compared to the relations of HYPERONYMY and MERONYMY, the CAUSE–EFFECT relation has not received as much attention in the field of terminology.

Although the ultimate goal of much medical and epidemiological research is the identification of CAUSE–EFFECT relations, in a field with such critical implications for human health and welfare, conclusions about the existence of CAUSE–EFFECT relations (which always rely to some degree on the judgment of those interpreting data) must be drawn cautiously and on the basis of large amounts of data — and moreover specific kinds of data (i.e., data obtained using specific study designs). Thus, the relation of ASSOCIATION (including CORRELATION) is particularly important in the field. This relation, involving a significant co-occurrence of factors, is relatively frequently expressed in medical texts (and particularly research articles). While important in itself (for example, in identifying risk factors for particular illnesses), it may also become the basis for hypotheses of a CAUSE–EFFECT relationship between the two elements it links. It is important to stress, however, that this kind of relationship is not one of CAUSE–EFFECT in itself, but rather a potential precursor of it (a fact that is illustrated in Hill's criteria for CAUSE–EFFECT relations in medicine, reproduced below in Section 1.5.2.1, which begin with an analysis of observed co-occurrence — i.e., ASSOCIATION — and then analyze this co-occurrence according to various criteria in order to determine if a conclusion of a CAUSE–EFFECT relation on the basis of these observations is justified). Thus, identifying such relations in medical texts may be very important in the information-gathering process.

The ASSOCIATION and CAUSE–EFFECT relations may thus be identified as important in the medical field, and provide a pertinent context for the study of knowledge patterns and their use for terminological knowledge extraction in this domain.

These relations can of course be characterized according to the criteria described in Section 1.4.1. Both the CAUSE–EFFECT and ASSOCIATION relations are non-hierarchical. In contrast, ASSOCIATION is symmetric, while the CAUSE–EFFECT relation is asymmetric. Finally, it can be argued that the CAUSE–EFFECT relation may be considered to be transitive (making reference, for example, to the concept of the causal pathway often mentioned in the context of disease etiology and development (e.g., Friedman 1994: 209), in which each link in a causal chain contributes to the eventual outcome). The ASSOCIATION relation, however, is less easily classified according to this criterion; given that the presumptions that A is associated with B and B is associated with C can guarantee neither that A is associated with C nor that it is not, ASSOCIATION cannot be considered to be either transitive or intransitive.

With these relations selected, they can be defined and described in more detail in the context of the research. Sections 1.5.1 and 1.5.2 will present more detailed information about various analyses and possible classifications for the ASSOCIATION and CAUSE–EFFECT relations respectively.

1.5.1 Association

It is common knowledge that statistical associations do not necessarily imply causation.
Friedman (1994: 213)

Given the delicacy of identifying CAUSE–EFFECT relations, as described above, medical texts and especially research articles generally contain many references to ASSOCIATIONS

between factors or variables in addition to statements of CAUSE–EFFECT relations. Two examples of this are shown below:

1. First, with the notable exception of cardiac toxicity **associated with** anthracyclines, unexpected toxicities to combination therapy have not been identified. (Burstein 2003)
2. The first **association between** CRP and cardiovascular disease was in the context of acute myocardial infarction. (Shah and Newby 2003)

However, although the distinction between ASSOCIATION and CAUSE–EFFECT relations may be "common knowledge" in epidemiological circles, for non-specialists the term *association* may appear general and vague. Moreover, despite the importance of this relation in scientific domains, few research projects have attempted to identify and define the ASSOCIATION relation from a terminological perspective. As a result, the relation is most precisely defined in the context of specialized resources in medical and scientific domains such as epidemiology. Below, the definition of the relation used for the purposes of this project (Section 1.5.1.1) — based primarily on descriptions from Hennekens and Buring (1987), who define the relation as identified in epidemiology — will be discussed, followed by a brief review of references to ASSOCIATION in terminology from Nuopponen (2005) and Feliu (2004) (Section 1.5.1.2).

1.5.1.1 Definition of ASSOCIATION

In this project, ASSOCIATION will be defined as a significant co-occurrence between two factors or variables. The definition given by Hennekens and Buring (1987: 30) — although expressed using terms specific to the calculation of incidence or prevalence of diseases in populations with a given exposure — illustrates this view:

> Association refers to the statistical dependence between two variables, that is, the degree to which the rate of disease in persons with a specific

> exposure is either higher or lower than the rate of disease among those without that exposure.[14]

This kind of relationship, however, is not necessarily indicative of a CAUSE–EFFECT relation. Hennekens and Buring (1987: 30) stress this, stating that "The presence of an association... in no way implies that the observed relationship is one of cause and effect," and continuing on to say that "[a] **causal** association is one in which a change in the frequency or quality of an exposure or characteristic results in a corresponding change in the frequency of the disease or outcome of interest." [emphasis added][15]

The observation of relationships such as the causal one mentioned in the quotation above may generally be recognized as the ultimate goal of studying ASSOCIATION (although ASSOCIATION must nevertheless be distinguished from CAUSE–EFFECT relations). Given this goal, specific types of these relationships that are clearly relations of temporal or spatial contiguity (cf. e.g., Nuopponen 2005) are not included in the definition, since here the focus is on the possibility that a CAUSE–EFFECT relationship of some type will ultimately be identified (and, clearly, strictly temporal or spatial relationships are not likely to lead to such observations).

Determining ASSOCIATIONS between variables is often the goal of various types of medical — and especially epidemiological — research, including identifying risk

---

[14] As in the case of the citation describing causal ASSOCIATIONS, it seems perfectly acceptable to replace *disease* with *outcome* and *exposure* with *characteristic* to reflect other possible contexts of ASSOCIATION.

[15] To this discussion, the observations of Garcia (1997: 10) may be added. In her study of CAUSE–EFFECT relations in the specialized domain of electricity, Garcia describes what she calls — in an allusion to the terminology of Aristotle (Physics II) — *relations causales formelles* (Engl. formal cause relations), characterizing these relations as those in which:

> ... la notion de cause et d'effet sont abandonnées au profit d'une régularité mise en évidence entre des actions (*Le niveau de la puissance produite par l'usine varie dans le temps en fonction de l'hydraulicité...*) (1997: 10)

She identifies four different sub-types of these relations (1997: 11). On the basis of this description, however, these are considered here to be types of ASSOCIATIONS. The propensity to interpret such relationships as causal observed here is also reflected in the observations of Nazarenko (2000: 47; see Section 1.5.2.3) in her lexico-semantic analysis of CAUSALITY in French.

factors for disease and death, and evaluating the effectiveness of treatments. Essentially, the results of research exploring the connection between two variables can be represented in their simplest form in a four-cell contingency table such as Table 6.

By convention, in this table, each of the cells is assigned a letter for reference, *a*, *b*, *c* or *d*. For the purposes of this description, *V1* and *V2* represent the variables being studied, + the presence of a variable, and - its absence. *x*, *y*, *z* and *w* represent the numbers of cases in which these specific combinations of variables were observed (with *x*, appearing in cell *a*, thus being the number of cases in which V1 and V2 were both present, *y* in cell *b* being the number of cases in which V1 was absent but V2 was present, and so on).

Table 6: Example of an epidemiological 2 x 2 contingency table

	V1+	V1-
V2+	x (a)	y (b)
V2-	z (c)	w (d)

This kind of table then forms the basis for calculations of various measures of ASSOCIATION between the variables (e.g., incidence, prevalence, relative risk) according to specific formulae (see Streiner and Norman 1998: 83–107 for more details). To state the case very simply, a positive ASSOCIATION between the two variables V1 and V2 exists when the value of *x* — that is, the number of cases in which both variables are present — is significantly greater than would be expected given a random distribution; a negative ASSOCIATION is present when this value is significantly lower than would be expected for a random distribution.[16]

[16] In this sentence we use *significantly* in the sense of statistical significance, usually represented by a *p* value (expressing the probability of obtaining a given result by chance) of less than $p = 0.05$ (i.e., less than five chances in 100 that the findings could be the result of chance).

Expressions of ASSOCIATION are important in the field for many reasons. First, since they are often used to express hypotheses that may later be confirmed on the strength of more data, they may allow terminologists to identify items that participate in a potential CAUSE–EFFECT relationship, to use this knowledge to monitor new data connecting a given pair of elements, and thus to identify more easily when a causal connection is considered to be proven.[17] Second, as it is relatively difficult to determine the precise point at which ASSOCIATION is accepted as a CAUSE–EFFECT relation, some authors may continue to refer to a connection as an ASSOCIATION between two variables when a causal connection has been accepted by others. Accessing these occurrences will help to obtain additional data about the pair. Third, even in the absence of an established, direct CAUSE–EFFECT relation ASSOCIATIONS are of interest, since they could be used to link terms denoting concepts that are evidently related in some way (for example, that may be shown in light of further data to share a common underlying cause, rather than being involved in a direct CAUSE–EFFECT relation with one another).

Thus, to recap the distinction between ASSOCIATION and CAUSE–EFFECT relations in medicine, a finding of ASSOCIATION between two variables often motivates further research that establishes that one of these is a cause of the other; however, an ASSOCIATION between two variables does not share all of the characteristics of CAUSE–EFFECT relationships. ASSOCIATION involves the co-occurrence of two variables, but not *necessarily* a direct causal connection between them; rather, both may result from a third factor, or they may prove to be related in a non-causal way. (Of course, their observed co-occurrence may be coincidental and they may not in fact be related at all; however, tests for statistical significance and efforts to ensure the reproducibility of results aim to minimize these cases.) As revealed, for example, in Hill's criteria for identifying causal relations in epidemiology, described in Section 1.5.1, more information about the strength and nature of ASSOCIATIONS is necessary before a CAUSE–EFFECT relation can

---

[17] This may be clearly observed in the description of the criteria for conclusions of CAUSE–EFFECT relationships cited in Section 1.5.2.1, as the observation of an ASSOCIATION between the two factors is the starting point for at least the first four, most significant criteria.

be established. To put it another way, ASSOCIATION is a necessary condition for CAUSE–EFFECT relations, but not a sufficient one.

In addition, ASSOCIATION is symmetric (i.e., if A is associated with B, then B is associated with A),[18] while CAUSE–EFFECT relations are asymmetric (i.e., if A is the cause of B, we cannot say that B is the cause of A). If *only* the association between A and B is observed, there is no way of knowing if A is the cause of B, if B is the cause of A, or if both result from another factor altogether. (Of course, most experiments begin with a hypothesis of a CAUSE–EFFECT relationship and therefore an expected directionality in the relationship between the two variables, but it remains just that — a hypothesis — until enough experimental evidence obtained through studies meeting specific design criteria allows for a conclusion to be drawn.[19])

It is nevertheless important to note that a few more complex cases in the expression of this relation may be observed. These will be described below, and include some specific types of ASSOCIATIONS that differ somewhat from the basic definition of the relation chosen for this project. These special cases of the ASSOCIATION relation include CORRELATION and RISK; these will be described below in Sections 1.5.1.3 and 1.5.1.4. These distinctions are also introduced in the discussions of ASSOCIATION in the terminological relation classifications of Nuopponen (2005; cf. Section 1.4.3) and Feliu (2004; cf. Section 1.4.4).

1.5.1.2 ASSOCIATION in terminology

Although terminological discussion of ASSOCIATION may be found in the relation typologies of Feliu (2004) and Nuopponen (2005), both of these correspond most closely to a specific case of the ASSOCIATION as described in the medical literature.

---

[18] There are some exceptions in special cases, which will be discussed in Section 1.5.1.4.

[19] For further discussion of this point, see the descriptions of *dependent* and *independent* variables in Section 1.5.1.4.

Nuopponen's (2005) definition of INTERACTIONAL concept relations — said to be "based on the interplay of referent phenomena" and divided into relations of TRANSMISSION, DEPENDENCY and REPRESENTATION in addition to CORRELATION — may be compared to that of the ASSOCIATION relation as defined by Feliu (2004).

The description of the ASSOCIATION relation in Feliu's revised typology, as a "relació que s'estableix per la corelació entre dos o més elements" [a relation of correlation between two or more elements] (2004: 39), is based not on any resemblance between the two connected elements (2004: 34), but rather on the presence of a point of contact between them. While this description is more general, as are many of the markers indicated as potential indicators of this relation (2004: 47), the definition and the marker of this relation indicated in Table 4 correspond to what will be considered here as a specific sub-type of this relation, CORRELATION.

Feliu notes (2004: 48, 133) the difficulty of distinguishing between ASSOCIATION relation and others such as the CAUSE–EFFECT relation or relations of SIMULTANEITY in the case of some markers; she also notes that her class of ASSOCIATION relation markers includes some that are indicative of symmetrical relations, while others are not; this suggests that there is some variability within this relation as she considers it on the basis of this criterion. She concludes (2004: 50) that there may be two sub-types of ASSOCIATION, one which is more general, and a second, indicated by markers such as *correlacionar-se amb*, which corresponds to a specific relationship in specialized domains.

It may thus be concluded that Nuopponen's (2005) and Feliu's (2004) relation classifications refer to similar relationships, although their insertion within wider classes is not equivalent, given the far larger scope of the INTERACTIONAL concept relations (as may be observed in the descriptions reproduced in Table 3). Moreover, both refer most particularly to relations of CORRELATION, further discussed below in Section 1.5.1.3.

1.5.1.3 CORRELATION

As noted above, Feliu's (2004) observations of the potential for observing variation within the class of ASSOCIATION relations reflects in large part the possibility of observing a CORRELATION between two variables. Moreover, the CORRELATION sub-type of INTERACTIONAL relations identified in Nuopponen (2005) can also be considered from this perspective. Nuopponen classifies the CORRELATION relation under the heading of INTERACTIONAL relations (a subtype of the INFLUENCE relations), and defines it as one in which there is some kind of causal connection between "entities" (i.e., "variables") (2005: 136) that have a reciprocal relationship, in the sense that as one variable changes, the other is likely to change in a corresponding way (2005: 136).

This closely reflects the definition that will be used in this research. For the purposes of this work, CORRELATION is defined as a type of ASSOCIATION involving the systematic variation of two variables in relation to one another, indicating interdependence. In this case, the values of the two variables are dynamic — that is to say, rather than being *categoric*, having one of a set number of discrete values and presenting a static dichotomy as illustrated above in (Table 6), they are *continuous*, with values that fall along a graded scale — and can be compared over a series of changing values.[20] This can be observed in the definition of *correlation* given in the *Oxford English Dictionary* (OED Online 2006):

> *correlation*, n.: In *Statistics*, an interdependence of two or more variable quantities such that a **change** in the value of one is associated with a **change** in the value or the expectation of the others... [emphasis added]

Thus, in the case of CORRELATION, as the value of one variable moves along the scale,

---

[20] For an explanation of the distinction between categoric and continuous variables, see Streiner and Norman (1998: 82). Examples of categoric variables that might be studied in epidemiological research include male/female and married/single/divorced/widowed; examples of continuous variables might include blood pressure readings and serum levels of a particular molecule. See Streiner and Norman for more on ASSOCIATIONS involving categoric variables (1998: 83-107) and continuous variables (1998: 107-117).

the value of the other will also vary proportionally.

The correlation coefficient "indicates the degree to which a set of observations fits a linear relationship" (Friedman 1994: 195). As Friedman states (1994: 195):

> Plotted on a graph showing the relationship between two variables, data points would follow a slanted straight line if the correlation coefficient is +1 or -1. Where there is some, but not complete, correlation, the data points would appear to cluster about a line. If there is no correlation at all, data points would form a regular or irregular clump with no underlying slanted line apparent.

This constitutes a more specific type of relation than that described above for ASSOCIATION, but nevertheless can be considered as belonging to this category of relations because it shares its essential characteristics.

This relation is described by Nuopponen as "rare" (2005: 136), a statement that would likely be controversial in the domains of medicine and epidemiology. (However, it is possible that this conclusion was based on an analysis of concepts belonging to categories or domains in which this kind of relationship is less frequent.) Feliu's (2004) identification of the relation among those most central in medicine and related fields provides an interesting contrast to Nuopponen's statement.

In texts, CORRELATION can be observed in contexts such as Example 3:

3. … **pour** chaque augmentation du rapport albumine/créatinine urinaire de 0,4 mg/mmol, ce risque augmentait de 5,9 %. (Fredenrich et al. 2004)

However, it should be noted that the usage of vocabulary does not always reflect the distinction identified here between CORRELATION and ASSOCIATION, as may be observed in Examples 4 and 5, which indicate a more general type of ASSOCIATION between the two items identified as being related but uses *to correlate* or *corréler*.

4. Cell adhesion molecules **have also been correlated with** CHD. (Rackley 2004)

5. … ses changements peuvent être **corrélés avec** une activation ou une répression de la transcription. (Chailleux et al. 2000)

In other cases, such as Example 6, it may be difficult to tell which of the two is intended:

6. … increased circulating IGF-1 concentrations **correlate** very closely **with** the relative risk for the development of several common cancers, including breast, prostate, colon, and lung. (McCance and Jones 2003)

Given this variability, this sub-type of the relation will be analyzed as part of the set of ASSOCIATION relations.

1.5.1.4 RISK

In the second special case of the ASSOCIATION relation considered here, a variable is identified as contributing to the RISK of a disease or outcome, as in the following sentence:

7. Nous avons recherché chez tous les patients les **facteurs de risque** d'athérosclérose (diabète, hypertension artérielle, tabagisme, hormonothérapie, intoxication alcoolique, dyslipidémie, hérédité)... (Desauw et al. 2002)

As illustrated in the definition of *etiology* shown below, taken from the U.S. National Library of Medicine's Medical Subject Headings (MeSH) (2006; cited in the UMLS Metathesaurus 2006), not only causes (necessary, sufficient, or otherwise), but also some types of ASSOCIATIONS (e.g., predisposing factors,[21] risk factors) may be central in the study of disease:

> *etiology*: The relating of causes to the effects they produce. Causes are termed necessary when they must always precede an effect and sufficient when they initiate or produce an effect. Any of several factors may be associated with the potential disease causation or outcome, including predisposing factors, enabling factors, precipitating factors, reinforcing factors, and risk factors.

---

[21] Cf. Nuopponen's (1994) explanatory causes, described in Section 1.5.2.5.

As Streiner and Norman note (1998: 95–96), RISK is a measure of the likelihood of occurrence of a given event. Accordingly, using the values in Table 6, for the population with the characteristic V2, the RISK of developing V1 is calculated using the formula:

$$\frac{x}{(x+y)}$$

Relative risk measures the strength of an association (Friedman 1994: 214). It is the "ratio of the disease rate in those with the factor to the rate in those without" (Friedman 1994: 214),[22] and can be calculated using the formula:

$$\frac{\frac{x}{(x+y)}}{\frac{z}{(z+w)}}$$

Thus when relative risk is greater than 1, the RISK associated with the presence of a given variable is considered to be elevated as compared to the RISK in the absence of that variable.

RISK is then essentially a measure of the probability of ASSOCIATION of two factors, and thus fits into this relation category. However, from the perspective of this work, RISK is different from some other kinds of ASSOCIATION. One difference lies in the directionality of this relationship, a postulated precondition and effect or outcome (e.g., in Example 7, *diabete*, *hypertension artérielle* or *tabagisme* and *athérosclérose* respectively), rather than a symmetric ASSOCIATION between thetwo variables. This corresponds to the description in Friedman (1994: 55), in which he explains that "[i]n a two-variable relationship one is usually considered the *independent* variable which

[22] That is to say, it is a measure of the likelihood of occurrence of a given disease or outcome in individuals with the presence of a variable versus those with the absence of that variable.

affects the other, or *dependent* variable." (That is, the link expressed in Example 7 is that smoking, for example — the independent variable — appears to increase the chances of developing atherosclerosis — the dependent variable — and not that atherosclerosis increases the chances of smoking.) Thus, in the vast majority of cases, research involving two variables will involve the investigation of a relationship between the variables based on a hypothesis in which one is presumed to be independent, and the other dependent. Data may be presented using lexical markers that make this distinction clear (as in the case of *risk*) or not (as in the case of *association*). Whether the choice is being made on a conceptual level (i.e., based on knowledge that, for example, one factor precedes the other) or a linguistic one (i.e., a function of marker choice alone) may not always be clear. Regardless, this kind of distinction is pertinent for the end user of a KRC, in that when markers clearly corresponding to RISK are used, the postulated cause and outcome are clearly identified, while in the case of markers of ASSOCIATION they are not.

1.5.2 CAUSE–EFFECT relation

Perhaps due to its fundamental role in human perception, the CAUSE–EFFECT relation is generally easier to recognize than to decompose, define and classify. Readers can easily identify one or more such relations in sentences such as those below:

8. Taken alone and without interruption, however, estrogen **causes** cell division in the uterus, which in many women **leads to** uterine cancer. (Watkins 2003)
9. … trials may shed light on the mechanisms by which influenza **triggers** cardiovascular complications. (Madjid et al. 2003)
10. Asymmetrical dimethylarginine (ADMA) is an inhibitor of nitric oxide synthase and thereby **causes** vasoconstriction and hypertension, and **increases** atherogenesis. (Stevens and Levin 2003)

However, there is very little agreement as to the definition of what it is to cause something, and what different kinds of causing may exist. This is in large part due to the many complexities and variants of relationships between causes and effects.

Many possible methods of analyzing CAUSE–EFFECT relations exist; these have been addressed from many points of view, among them those of philosophy, lexical semantics, and terminology. Each of these studies has focused on aspects of this relation that are considered pertinent for a given goal.

The literature on causal relations is abundant and it would be futile to attempt to provide broad coverage of the reflections on the subject here. Rather, the focus in this discussion will be placed largely on some works that outline criteria useful for identifying and characterizing different sub-types of the relation in research projects such as this one.

As described in Section 1.3, various points of view on relations may be pertinent for text-based approaches to identifying conceptual information. The relations studied in this research are indicated by lexical units of a language — often of general language, although some are specialized or have acquired specialized meanings — and these units thus have their own lexical meanings and places in the semantic system of the language. These meanings may provide a basis for the classification of the relationships these units express. Given that these lexical units are the access points through which conceptual relations may be identified in text-based — and particularly pattern-based — applications, it is interesting to consider the ways in which relations have been seen from the perspective of lexical semantics, to observe how the semantics of the language may reflect the conceptual constructs that are in turn the product of human perception and cognition.

Thus, the conceptual perspective of this research may be complemented by and contrasted with cognitive and semantic analyses of the relations in question. Interestingly — but not surprisingly, given the process of evolution described above —

analyses from these different points of view often present significant similarities. This Section will present some analyses of the CAUSE–EFFECT relation in medicine and epidemiology (Hill), lexical semantics (Lyons, Nazarenko, Mel'čuk et al.), and terminology (Nuopponen).

1.5.2.1 Hill

Before addressing the study of the CAUSE–EFFECT relation as described by linguists and terminologists, a more conceptual and empirical approach to identifying this relation may be presented. As several philosophers (e.g., Hume 1739/1985) have noted, the relation between a cause and an effect is perceived, rather than known objectively. In the field of medicine, much thought has been given to criteria that can be used in order to confirm intuitions about CAUSE–EFFECT relations.

Sir A. Bradford Hill (cited in Streiner and Norman 1998: 121–8; cf. also Hennekens and Buring (1987: 39–43) and Greenhalgh (2001: 87)) identified nine criteria that are widely used in medical research in order to justify a conclusion that there is a CAUSE–EFFECT relationship between two variables (here identified as a postulated cause and an outcome). These criteria, presented in descending order of importance, are:

1. **Strength of association**: how closely the postulated cause and outcome are associated;
2. **Consistency of association**: whether the association has been observed in numerous studies carried out by different researchers, in different circumstances;
3. **Specificity of association**: how closely the observed relationship comes to the ideal of one postulated cause being associated with one outcome, and that outcome with a single postulated cause;
4. **Temporality of association**: whether the (exposure to) the postulated cause precedes the outcome;
5. **Biologic gradient**: how direct the correlation is between changes (increases, decreases) in the postulated cause and changes in the outcome;
6. **Biologic plausibility**: how plausible the postulated mechanism for causation is from a biological perspective;

7. **Coherence**: whether there are any conflicts between the postulated cause and existing knowledge;
8. **Experimental evidence**: whether there is evidence from in vitro and/or in vivo experiments that support the presumed causal relationship;
9. **Analogy**: whether there are similarities with known causal relationships.

Clearly, the perception of CAUSE–EFFECT relations begins with an observation of an ASSOCIATION between two variables. On the basis of a sufficient amount of data from appropriately designed studies (the gold standard being randomized controlled trials), researchers attempt to determine whether the two are connected in a causal relationship. While not all of the criteria indicated above must be met in order for the existence of a CAUSE–EFFECT relation to be accepted, the more of them that are met — and the higher their rank in the list — the more certain the existence of a CAUSE–EFFECT relationship.

Having observed these domain-specific criteria, more linguistic and terminological perspectives on CAUSE–EFFECT relations that will be useful for identifying these kinds of conclusions in texts using linguistic cues may be examined.

1.5.2.2 Lyons

Lyons (1977: 490) describes CAUSALITY by stating that "agents are seen as the causes of situations which, by their actions, they bring into existence," and also goes on to state that in another type of CAUSALITY, a situation can also lead to another.

He notes that this portrait is compatible with the description of agency, which he describes (1977: 483) as follows: "animate entity, X, intentionally and responsibly uses its own force or energy, to bring about an event or to initiate a process." He also states that "the paradigm instance of an event or a process in which agency is most obviously involved will be one that results in a change in the physical condition or location of X or some other entity, Y." (1977: 483)

Lyons differentiates between CAUSALITY and CAUSATIVITY by stating that the latter involves both CAUSALITY and agency (1977: 490). Lyons also notes that due to

this interrelation, it may be possible to identify different elements as causes in a given situation, for example, either the agent or the agent's action.

1.5.2.3 Nazarenko

In her portrait of CAUSALITY in the French lexicon, Nazarenko (2000) highlights not only the fundamental nature of this kind of relationship in human reasoning, but also the limited nature of previous descriptions of how it is expressed in language. While she notes (2000: 13) that its prototypical markers (i.e., causal connectors such as *parce que* and *à cause de*) are relatively few compared to those of some other relations (such as temporal relations), Nazarenko observes that the linguistic expression of CAUSALITY can go beyond these simple causal connectors and can also include other lexical units, syntactic means, and what she refers to as *interprétation causale* [Eng. causal interpretation], in which CAUSALITY is not directly asserted, but is inferred by the receptor, generally on the basis of statements that explicitly indicate other relations.

In a description of CAUSALITY on a conceptual level, Nazarenko notes (2000: 3) that definitions are commonly circular, defined by the relationship between a cause and an effect, which are often defined in terms of one another. Thus, she prefers to analyze causal relationships according to their properties, citing (2000: 5–6) five major elements: 1) CAUSALITY is subject to temporal requirements, in that the cause must precede the effect; 2) there is a general applicability of the law of causation, which may itself be hard to define, but is nevertheless intuitively understood to cover a range of specific cases; 3) it is possible to understand this relation as a function of deductions and reasoning based not only on events that did occur, but also on those that did not (e.g., to deduce that if a given event was the cause of another event, that if this event had not occurred, the resulting event also would not have occurred); 4) conclusions of CAUSALITY are based on approximations, with an identified cause constituting only one of multiple factors that may contribute (by their presence or absence) to the occurrence of an effect; and 5) the perception of CAUSALITY is subjective and dependent on the

interpretation of the perceiver. Nazarenko thus defines "causality" as a relation of cause to effect, or causal relation (2000: 10). Causes and effects are viewed as roles in such relationships, rather than entities in their own right; these roles may be played by various types of events, including situations and processes (2000: 10).

Although the recognition of causal relations is a fundamental element in the reasoning that allows humans to interact with and affect the world around them, and therefore often constitutes a guide for intentional human action, Nazarenko notes (2000: 6, also citing Russell 1914: 227) that, contrary to some definitions of causation, the question of volition is not a necessary component of the concept "cause," a reflection that Russell notes is particularly obvious in scientific fields such as physics.[23]

Nazarenko also analyzes some of the criteria that may be used to identify sub-types of CAUSE–EFFECT relations, citing for example (2000: 123–124) a distinction that may be made between direct and indirect causes, punctual and durable causes, and voluntary and involuntary causes (which bring to bear the criteria of awareness and intention). She notes that these kinds of distinctions are essentially based on the type of cause that is present and on the nature of the causal relationship. These distinctions are often linked to differences in the linguistic manifestations of the relationships; she provides more details about possible nuances of causal relations in her discussion of lexical markers (see below).

Nazarenko notes (2000: 8) the close links that exist between the conceptual and linguistic levels in the comprehension of CAUSALITY, noting that one of the few ways of identifying and describing the notion of "cause" is by using a linguistic test: a cause may be identified by its possible function as the answer to a question introduced by *pourquoi...?* [Eng. *why...?*]. Utterances that contain such an element are thus considered to provide causal information (2000: 10).

---

[23] In our opinion, this is also true of medicine, as in the case of the corpora analyzed in this research.

In a discussion of causal connectors and their role in causal utterances, Nazarenko observes, citing the example of *comme* (2000: 79), that some causal connectors may be quite polysemous, introducing ambiguities for the interpretation of the relation underlying a given utterance. She additionally argues (2000: 86–91) that in some interpretations, connectors may indicate the presence of causal information without explicitly expressing this relation (this may be the case in utterances containing markers such as *sans que*), while others may regularly present causal information applying to causes that are either negated or uncertain (e.g., as indicated by *non que*, which rejects a cause or an explanation of a given event, or *soit que... soit que*, which presents two alternative causes or explanations). Finally in her discussion of causal connectors, Nazarenko stresses (2000: 91) that the use of these connectors can be idiosyncratic, and may also vary by text type.

Nazarenko stresses (2000: 124–140, 145) that although causal connectors are relatively limited in number, many other lexical items may also indicate that a causal relation is present; these indicators may be nouns (e.g., *cause, raison, rôle, facteur*, and *origine*), adjectives (e.g., *nécessaire, efficace, responsable*) and verbs (e.g., *causer, provoquer, occasionner*). (She also notes (2000: 125) that in many cases a given item may form the basis of series of derived forms belonging to other part of speech categories but conveying the same notion of cause, e.g., *responsable, responsabilité, être responsable de*). She observes (2000: 137) that among these categories, verbs are the most productive indicators of causation; however, she also notes that a causal relationship cannot be observed in a verb in isolation — it is a function of both the verb and its arguments.

Within each category, various lexical indicators of causation may convey information about the specific type of causal relation present, detailing the type of cause or effect involved or the nature of the causal process. For example, the nouns *rôle, facteur*, and *origine* specify the way that the cause participates in the production of the effect; adjectives such as *nécessaire*, while not necessarily directly expressing causation,

indicate that causes are present and may be characterized using traditionally identified attributes such as "necessary" or "sufficient" (while in contrast adjectives such as *spontané* or *fortuit* may explicitly deny the presence of a cause). Verbs may emphasize various aspects of the relation: the process of causation (as in the case of the examples above); the types of effects that result (e.g., indicators of quantitative variation as in the case of *augmenter* and *renforcer* for increases, *réduire* for decreases, *créer* and *engendrer* for appearance (i.e., coming into being) and *annuler* and *supprimer* for disappearance (i.e., ceasing to be); indicators of qualitative variation as in the case of *améliorer* or *détériorer* in general senses, or *agrandir*, *limiter* or *assurer*, which indicate the involvement of a particular characteristic of the effect (in these cases size, space and certainty respectively));[24] or the role of the cause. This classification by causal role may be reflected in the use of different markers depending on: the orientation of the causal relation (i.e., with a focus on the cause, as in the case of *entraîner* and *provoquer*, or the effect, as in *provenir de, être du à*); the degree of causation, which may be complete (e.g., *causer, conduire à*) or partial (e.g., *influencer*, *contribuer à*, *favoriser*, *intervenir dans*, *aider à*, *participer à*); the value of the causation, which may be positive (e.g., *causer, encourager*) or negative (e.g., *empêcher*, *gêner*); or the temporal relationship between the cause and effect, in which the cause may intervene early in the production of the effect (e.g., *susciter*, *être à l'origine de*) or may be prolonged by the effect (e.g., *aboutir à*). These types of nuances may be pertinent not only for the expression of the causal relationship, but also for the analysis of the relation itself.

In terms of the expression of the causes and effects involved in causal relationships, Nazarenko notes (2000: 147–148) that while at a conceptual level a cause or effect must be an event, situation or process, at a linguistic level these elements may be realized in various forms, including propositions, nouns or noun phrases, and so on. The choice of indicator of the relation may influence the manner in which one of these elements is expressed. For example, the marker *parce que* often introduces a

---

[24] Cf. Feliu (2004) and her distinction between different types of PROCESS–RESULT relations.

propositional cause, while nominal causes are often linked to markers such as *à cause de* (2000: 59–61). The expression of the cause or effect present may also affect the amount of information conveyed: propositional expressions of causes or effects tend to involve the explicit expression of many more elements of the event than nominal ones. Moreover, causal relations may be expressed within a single proposition or between propositions.

Nazarenko is careful to note (2000: 125, 143–144) that given the diversity of the possibilities for expressing CAUSALITY, and the subjective nature of the interpretation of this relation, the list of possible expressions and the classification of the relations present are necessarily incomplete and may be viewed differently by different individuals. In taking a broad view of what may constitute CAUSALITY (including any occurrence which may be considered to provide an answer to a question introduced by *pourquoi*), she chose to present an inclusive description of the possibilities of expressing this relation.

Nazarenko also discusses (2000: 13–49) some cases in which CAUSALITY is not explicitly stated but may nevertheless be interpreted by a receptor from statements that explicitly provide information about other types of relationships, including temporal relations (such as anteriority and simultaneity) and correlation.[25] She observes that there is a strong tendency to interpret statements of these relations as causal, to the point that it may often be necessary to explicitly deny the existence of a causal relationship in utterances involving such relations in order to block a causal interpretation (2000: 47).

---

[25] Nazarenko notes first (2000: 43) that there is an important difference between correlation at a conceptual level (in which two situations or events vary in relation to one another), and linguistic correlation (also called in French *systèmes corrélatifs*), in which two propositions are involved in a reciprocal relationship of implication; she notes that while linguistic correlation may express conceptual correlation, it may also indicate other types of relations, while a conceptual correlation may be expressed by means other than those of linguistic correlation. Nazarenko is concerned solely with conceptual correlation in her discussion, because this kind of relationship may often be interpreted as causal.

1.5.2.4 Mel'čuk et al.

In the context of Meaning ↔ Text Theory, Mel'čuk and his colleagues (Mel'čuk et al. 1995) have formalized various types of relations between the meanings of lexical units that involve causation, using lexical functions (LFs). Some of these LFs describe syntagmatic relationships between lexical units at least one of whose meanings includes a causal component (e.g., `Caus`, `Liqu`, `Perm`), while others describe paradigmatic relationships between lexical units that are linked by a semantic relationship that includes an element of causation (e.g., S_{res}, `Result`).

In associating different types of causation with the LFs `Caus`, `Liqu` and `Perm`, the authors have shown that accurately representing relationships between meanings of lexical units requires a breakdown of different types of causation.

In addition, LFs representing these relationships include an indication of the aspect of the effect, represented by `Incep` (designating beginning), `Fin` (designating ending) or `Cont` (designating continuation). (In the case of causal LFs, `Incep` is seen as a default value, and thus is understood if neither of the other aspectual functions is specified.)

`Caus` represents the basis of the semantic relationship, what would prototypically be thought of as causation and can be paraphrased as 'cause' or 'do something so that a situation begins occurring' (Mel'čuk in preparation: 53). `Liqu` can also be represented as `Anti(Caus)` (Mel'čuk in preparation: 25), that is, the negation of `Caus` or of one element of its meaning, or `CausFin` (Mel'čuk in preparation: 22). This can be paraphrased as 'liquidate' or 'do something so that a situation stops occurring' (Mel'čuk in preparation: 53). Finally, the third function of this type is `Perm`, paraphrased as 'permit' or 'allow,' that is, 'do nothing that would cause that a situation stops occurring.' It can also be seen as the negation of `Liqu` (Mel'čuk in preparation: 53), or as a double negation of `Caus`, that is, `Perm`(P) = `NonCaus`(`NonP`).

These relationships are often seen in complex lexical functions involving support verbs, such as Oper and Func, as in the following examples (Mel'čuk in preparation):

CausFunc_0(*crisis*) = *bring about* [ART ~]
PermFunc_0(*aggression*) = *condone* [ART ~]
LiquFunc_0(*traces*) = *wipe out* [ART ~]
CausFunc_1($hope_N$) = *raise* [~ in N_X]

As a supplement to the LFs mentioned above, it is also worth noting that others may provide additional modifying information, representing some fundamental meanings in quantitative (e.g., Plus, 'more', and Minus, 'less') or qualitative (e.g., Bon, 'good,' and AntiBon, 'bad') modification. Such LFs are commonly combined with causal LFs (often linked by the LF Pred, which can be paraphrased as 'be an (L)') in order to provide additional information about the type of change that occurs.

In addition to these syntagmatic lexical functions, which have as their values lexical units (often verbs) that express different types of causation that are relatively regular and useful for grouping together various occurrences of lexical relations with a causal component, a number of paradigmatic lexical functions that represent links between the meanings of lexical units that can be seen as a "cause" and an "effect" may be observed.

S_{res} represents a circumstantial noun designating the standard result of a situation described by a meaning (generally the meaning associated with a noun or a verb) (Mel'čuk in preparation: 35). Examples are found in the pairs below, taken from the DiCo (Mel'čuk and Polguère 2005):

S_{res} (*coup de foudre*) = *amour*
S_{res} (*labeur*) = *fruit*

`Result` represents verbs meaning 'the expected result of L' (Mel'čuk in preparation: 39). Examples include the following:

`Result` (*buy*) = own_V

`Result` (*have learnt*) = *know*[26]

The identification of the semantic relationships represented by these LFs as some of the most fundamental in languages emphasizes the importance of the sense of 'cause' in lexical semantics — as representative of the equally fundamental relation between causes and effects that may be perceived in reality. In addition, the choice of the criteria used to distinguish between the various causal sense relationships that may exist identifies a number of aspects of causation that are important. First, a number of different types of causation may be identified: causing something to exist, causing something not to exist, and allowing something to exist. Second, another important distinction involves whether this "something" begins or continues to exist, or whether it stops existing. In order to gain a complete picture of causation in language it is necessary to take into account the cause, the effect, the aspect of the effect, and the type of relation between the cause and effect. It may furthermore be argued that these linguistic distinctions reflect important ones at a conceptual level as well.

In a forthcoming article, Kahane and Mel'čuk describe causation (once more from the point of view of lexical semantics) and evaluate the conditions that apply to a representation of a real-world situation in order that this situation can be described using the linguistic sense 'cause,' as well as the conditions that must be satisfied in order to describe a situation in terms of specific linguistic means. They thus very clearly differentiate their work from that of philosophers and logicians, who attempt rather to identify the real-world situations that correspond to the concept of "causation," or what

---

[26] While the infinitive form *learn* is more usual in lexical functions, the perfect form used here reflects the fact that 'knowing' is a result of a finished process of 'learning,' rather than of the process as it is occurring (Mel'čuk 2006, personal communication).

mental constructions and knowledge must be present in order for an individual to use the concept "causation" in reasoning (Kahane and Mel'čuk forthcoming: 2). Nevertheless, they note that these different levels of analysis are inseparable links in a chain from real-world situation to a conceptual representation of that situation, which is then transformed into a semantic representation of sentences in a given language, and finally represented formally in these sentences (Kahane and Mel'čuk forthcoming: 3).

In their analysis, Kahane and Mel'čuk identify a number of pertinent elements in a causal relationship, as well as two separate senses representing causation, and thus two senses of the French verb CAUSER.[27] The first, 'causer1,' is non-agentive causation, which can be paraphrased by 'être la cause de.' The second, 'causer2' is agentive causation, paraphrased by 'être le causateur de.' In this latter case, there is both volition and a given goal for the causation (Kahane and Mel'čuk forthcoming: 25). Thus, the major differentiating criterion used in this analysis is agency (which links to the observations made in Lyons (1977) above, and also reflects observations made by Nazarenko (2000)).

In the case of 'causer1,' the participants in the causal situation can be classified as an effect, a cause, and an elaboration of the cause (cf. Lyons' (1977) action of the agent). This third, less obvious element in the situation represents an event involving the cause (specifically, it is a predicate with the cause as its first actant). The elaboration of the cause may or may not be realized at the surface level, but is always present at some level. The effect must also be an event. Various representations of a situation involving 'causer1' may be observed in Examples 11 to 13 (Kahane and Mel'čuk forthcoming: 6):

11. Les voitures causent l'irritation de Zoé.

---

[27] The authors also mention briefly other complicating aspects of causation that they did not address in this article (Kahane and Mel'čuk forthcoming: 4), notably the difference between direct and indirect causation, the distinctions between different types of causation — including 'cause,' 'make possible,' 'permit' and 'prevent' — and the case of internal causation (in which, to use a primitive decomposition of the meaning, an agent causes itself to do something).

12. Les voitures causent l'irritation de Zoé par leur va-et-vient incessant.
13. Le va-et-vient incessant des voitures cause l'irritation de Zoé.

Thus either two or three of the elements, the cause (*les voitures*), the elaboration of the cause (*le va-et-vient*) and the effect (*l'irritation de Zoé*), are realized in descriptions of the situation, in different surface syntactic structures. This means that there are two possible surface realizations of the actantial structure of 'causer1,' one bi-actantial, the other tri-actantial (Kahane and Mel'čuk forthcoming: 6): *X cause Y* et *X cause Y par Z(X)*, where Z is a predicate that has X as its first actant (e.g., an attribute or action of X, as in the *va-et-vient des voitures*).[28] Moreover, this variation is recursive — that is, can be expanded almost infinitely — so that it is possible to identify many different interconnected "causes" (e.g., *Le bruit du va-et-vient des voitures cause l'irritation de Zoé*). The third actant is characterized as *escamotable* [Eng. syntactically optional], i.e., compulsory in the semantic representation of the situation, but not necessarily realized at the surface level. The authors note (Kahane and Mel'čuk forthcoming: 8) that this reflects the fundamental nature of causation, which relies on human (subjective) judgment to identify not only a single cause among a multitude of contributing factors (e.g., necessary conditions for an event, the absence of other factors that could prevent the event, other contributors to the situation, etc.), but also the correct level of detail to use when describing the cause of an event (e.g., the choice not to express the situation described in the examples above using a sentence such as *L'activité des synapses du cerveau de Zoé en réponse à la stimulation de ses nerfs auditifs par le bruit du va-et-vient des voitures a causé l'irritation de Zoé.*). In addition, there are several possible combinations of semantic classes to which these surface actants may belong. This and other types of variation (e.g., metonymy) ensure that there is likely to be a need for a certain amount of flexibility in describing relations between lexical units participating in causal variations.

---

[28] The semantic structure underlying these surface representations of course remains the same and always has three actants; the form *X cause Y [par Z(X)]* accounts for both possible surface forms.

In the case of agentive causation, 'causer2,' four possible actants may be identified: the causal agent (which might be designated in French by the terms *auteur* or *responsable*), the cause (the action taken by the causal agent), the effect, and the instrument used in the action. The causal agent in this case is either a person or an entity or event that is seen in a personified way (e.g., an animal, an "intelligent" machine, a natural disaster, or even a disease). The authors stress (Kahane and Mel'čuk forthcoming: 26–7) that the cause here is not the same as the elaboration of the cause in 'causer1,' although there may be a superficial resemblance.

Example 14 illustrates the use of 'causer2' (Kahane and Mel'čuk forthcoming: 26):

14. Zoé a causé**2** la mort de la grenouille avec une fourchette en la lui enfonçant dans l'œil.

The authors go on to expand on this analysis by describing several transitive verbs that have one of these two causal senses as their communicatively dominant semantic component (i.e., when represented as a semantic network, can be minimally paraphrased by this component); these include NETTOYER, ÉLIMINER, EXPLIQUER, IRRITER and TUER. Moreover, some of these items also have (at least) two senses (corresponding to two separate acceptions), one that contains 'causer1' and another than contains 'causer2'; these thus parallel the distinction made between the two senses of CAUSER.

In discussing these lexical units, the authors have also found it useful to make another distinction, this one between *verbs of causation* and *causative verbs* (Kahane and Mel'čuk forthcoming: 29–35). The first group, such as DÉCLENCHER and ENTRAÎNER, express causation alone, while the second, such as TUER and CONSTRUIRE, also include in their meaning the effect of the causation. This distinction reflects a number of differences that may be seen on both a semantic and formal level.

1.5.2.5 Nuopponen

As observed in Section 1.4.3, in Nuopponen's 2005 relation classification CAUSE–EFFECT conceptual relations are identified as a type of ONTOLOGICAL INFLUENCE relation (a category defined by the presence of some kind of causal component in the relation, i.e., a one-sided or mutual influence). Although the classification surrounding them has evolved and the terminology used changed slightly, that of the CAUSE–EFFECT relations themselves is not obviously different from Nuopponen's 1994 classification of the relation. (In fact, very little description of these relations is given in the article.)

In 1994, Nuopponen highlighted the importance of the CAUSE–EFFECT relation between concepts from a terminological perspective, particularly in the domains of science and medicine.[29] She developed a classification of the CAUSE–EFFECT relation (1994: 37–8) that, as illustrated in Figure 2, may be grafted onto Wüster's classification of ontological concept relations (which appears above the dotted line), and then introduces a classification of the CAUSE–EFFECT relation that she characterizes as largely based on Mackie (1974) (below the line). Figure 3 provides another representation of the relationships between various types of causes and effects, while Figure 4 gives an example of some CAUSE–EFFECT relationships involving the concept "measles."

Nuopponen recognizes that not all CAUSE–EFFECT relations involve a single cause and a single effect. In fact, many effects — perhaps especially in medicine — may be associated with a number of causes, and many causes with a number of effects.

Nuopponen thus chose to define a category of relations involving more than one cause or effect (causal concept coordination), to subdivide this category into cases of multiple cause and multiple effect, and then to subdivide it further in the case of multiple causes into relations involving alternative causes (with either one or the other of the possible causes producing the effect), and those involving co-operating causes

---

[29] Similar classifications have also been used by other authors (e.g., Cabré et al. 1996).

(which come together to produce the effect). In the case of multiple effects, Nuopponen identifies those that are alternative (in which one of the possible effects occurs) or co-occurring (in which two or more effects occur as a result of the cause).

In addition, the author identifies three different components of cause that may be identified (1994: 39–40). The first is the causative agent, "substances, materials or other elements that cause an effect" (1994: 39), as in the case of allergens causing allergies. The second is the producing cause, commonly seen in philosophy as an event that causes another event, as in the case of the action of an agent or the exposure to an agent that causes disease. Producing causes can be further classified into causative events, causative actions and causative processes. Nuopponen notes (1994: 40) that the patient in a causal event (e.g., the metal in the case of corrosion) is also a pertinent element in the description of the causal system. The third component is the explanatory cause, a fact or a state. The example given by Nuopponen (1994: 40) is the case of existing allergies, which are triggered by the producing cause of exposure to an allergen. Finally, Nuopponen cites counteracting causes, i.e., agents, events, states or facts that counteract the causal process and prevent the effect, as taking allergy medications may interrupt the reaction of a person with allergies to exposure to an allergen. She also notes that the absence of such counteracting causes may in itself be considered to be a causal factor.

In a discussion of effects, Nuopponen also identifies several components (1994: 40–41). These include resulting states (e.g., disease or damage in the case of pathological functions), resulting products (e.g., rust in the case of corrosion) and resulting events (e.g., immunization in the case of vaccination).

Figure 2. Conceptual relations (Wüster 1974; Nuopponen 1994)

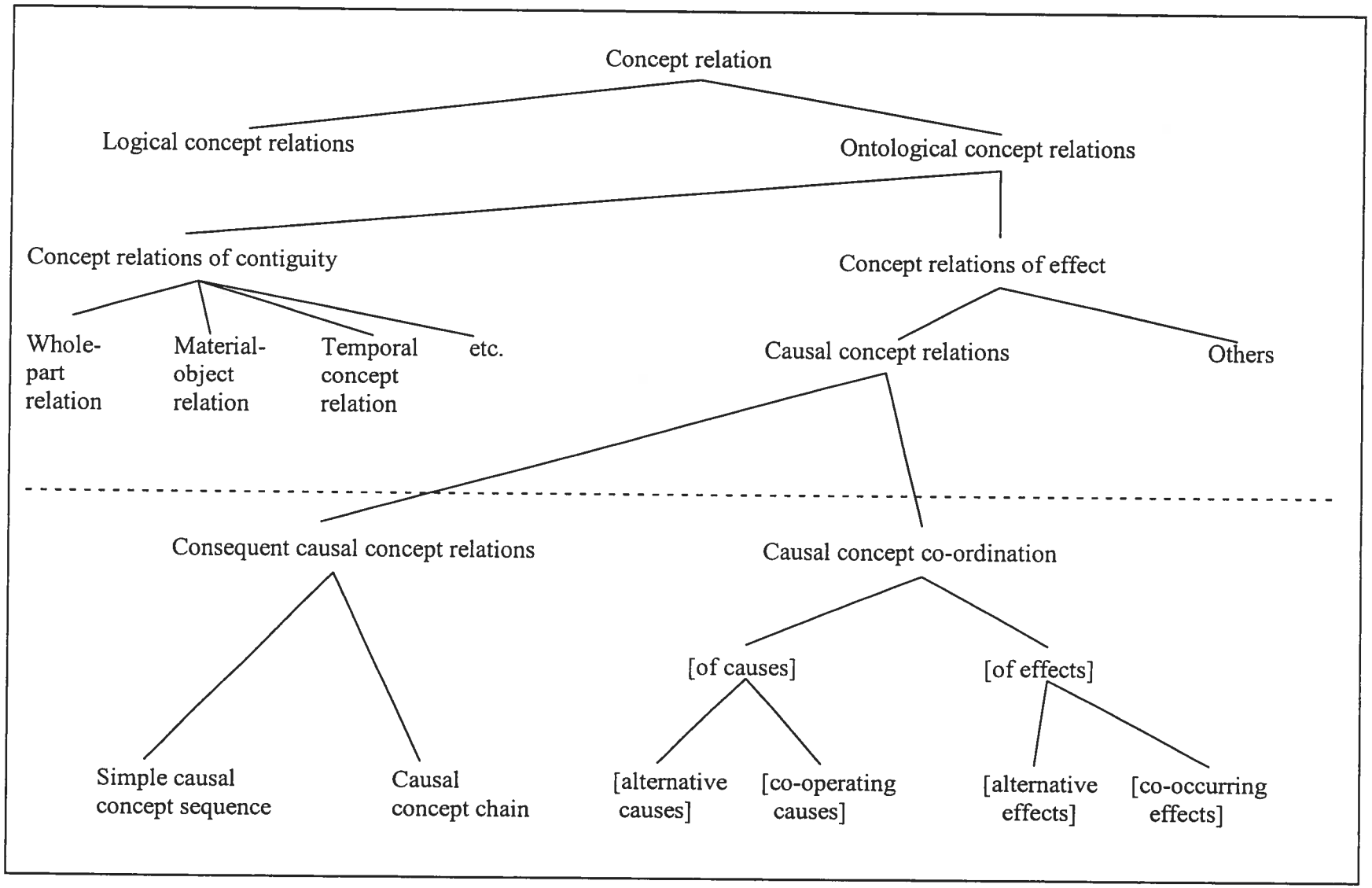


Figure 3: Nuopponen's causes and effects (1994: 41)

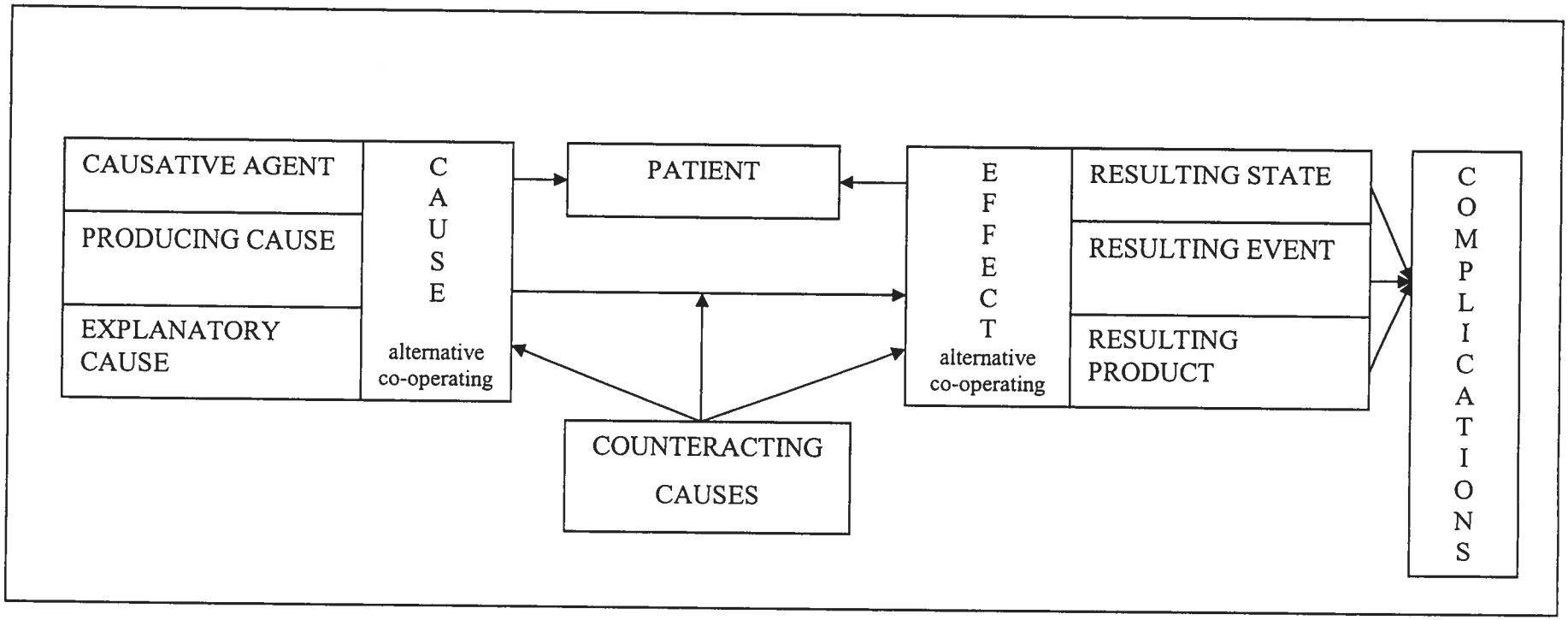
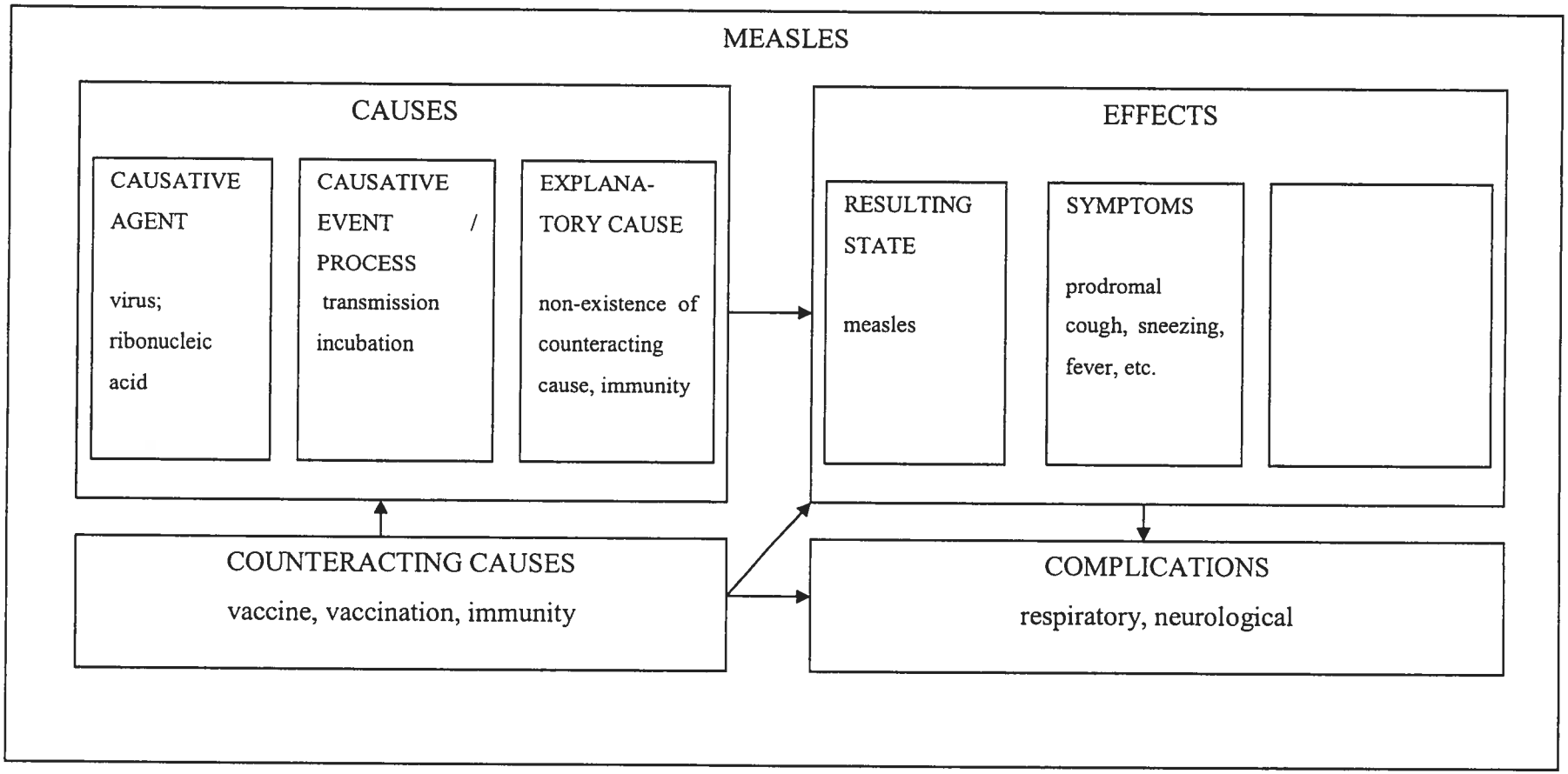


Figure 4: Nuopponen's diagram of the causal concept system for the concept "measles" (1994: 42)



Nuopponen also identifies what she calls *complications*, indirect effects of a first effect (as in the case of respiratory and neurological complications from measles). In the same vein, she notes that often a given relation participates in a chain of CAUSE–EFFECT relations, with the effect in one relation becoming the cause in another, and so on. Nuopponen thus differentiates between simple causal concept sequences and causal concept chains.

1.5.2.6 Synthesis

A number of criteria used by these authors to describe different kinds of relationships involving causation can thus be identified. In her conceptually-oriented analysis, Nuopponen (1994) classifies relations in part according to the number of causes and the number of effects involved in the relation, and on whether these occur together (in the case of co-operating causes and co-occurring effects) or separately (in the case of alternative effects and causes).

These classifications may be compared to other commonly identified classifications of causes, which deal with causes that are necessary and/or sufficient. For example, in the MeSH definition of *etiology* cited in the UMLS Metathesaurus (see Section 1.5.1.4), it becomes apparent that cases of multiple causes, as well as necessary and sufficient causes, are all considered pertinent in the medical domain. Their absence from other classifications noted here (except for brief mentions, for example in Nazarenko 2000) may be linked to the rarity with which these characteristics are reflected in the meanings of lexical units expressing causal relationships, the perspective of most of these studies.

Types of elements that may occur in the roles of causes and effects are also mentioned both from a conceptual perspective (by Nuopponen (1994)) and a lexico-semantic one (by Lyons (1977), Nazarenko (2000), and Kahane and Mel'čuk

(forthcoming)).[30] The granularity of the classification ranges from very precise, as in the case of Kahane and Mel'čuk — who describe various possible combinations of semantic classes in CAUSE–EFFECT relations — to relatively general in the case of Nazarenko (2000) and of Nuopponen (1994), who limits the discussion to relatively general classes considered individually. Particularly important in many contexts is the distinction made between causal agents and causing events; as noted by several authors, either or both of these types of causes may be indicated in contexts expressing CAUSE–EFFECT relations.

Another pertinent characteristic of the description of CAUSE–EFFECT relations is the possibility of a cause potentially — but not necessarily — acting consciously and voluntarily. Kahane and Mel'čuk's distinction of agentive and non-agentive causation is pertinent here, demonstrating that from at least some perspectives, both possibilities may be valid. Analyses by Nuopponen (1994) and by Nazarenko (2000) reflect an acceptance of both types of relationships as causal. As noted in the citation from Russell (1914: 227) provided by Nazarenko (2000: 7), the role of voluntary causation may be very limited in some scientific fields, such as physics (and, we may argue, medicine).

Perhaps the most pertinent criterion for classification, however, is by the nature of the change that occurs to the effect. From a lexico-semantic perspective, this aspect is the central one used in Mel'čuk's lexical functions (Mel'čuk et al. 1995; Mel'čuk in preparation), and is also one of those identified by Nazarenko (2000) that is reflected in lexical markers of CAUSALITY. This is also reflected in Nuopponen's (1994) description of counteracting causes at a conceptual level, which lead to the non-existence or non-occurrence of an effect. While these classifications differ in their specific distinctions of different types of relationships, general commonalities are observed in the

---

[30] Nazarenko, for example, notes that causes may be events, situations or processes (2000: 147–148), but that in some expressions of CAUSALITY agents of processes may also be viewed as causes of those processes' results (2000: 140).

differentiation between, for example, cases in which events occur or cease to occur, in which they increase or decrease, and so on.

These diverse points of view on various ways of analyzing CAUSE–EFFECT relations provide an opportunity to discuss the perspective used in this research, before the classification of the relation used for the work is presented in Section 1.5.2.8.5.

1.5.2.7 Definition of the CAUSE–EFFECT relation in this research

For the purposes of this research, the CAUSE–EFFECT conceptual relation will be defined as a relation between two concepts, i.e., a cause and an effect, in which the cause exerts an influence that determines the existence or occurrence of an effect or changes the existence or occurrence of that effect. These may include relations in which the influence exerted by a cause leads to either the existence or the non-existence of an entity effect or the occurrence or non-occurrence of an event effect (e.g., coming to be or happen, as indicated for example by *produce* or *produire,* or ceasing to be or happen, as in *destroy* or *détruire*), in addition to cases in which the qualitative or quantitative nature of a cause's influence on the effect identified in the context (e.g., modifying as in *change* or *modifier*, increasing as in *increase* or *augmenter*, and decreasing as in *reduce* and *diminuer*) are specified.

In this study, a choice was made not to limit the study of CAUSE–EFFECT relations to those involving the voluntary action of a causal agent, for a number of reasons. First, in the subject field and text types used in the work, the proportion of relations in which there is both volition and a specific goal, and in which this agency is clearly manifested, is likely to be relatively small, and interest in other types of causes widespread (cf. Nuopponen 1994; Nazarenko 2000). Secondly, this volition is most likely to be associated with human causal agents, who are not as likely to be named by terms included in standard terminological resources, and thus for which the extraction of CAUSE–EFFECT relations is probably of limited usefulness in terminology work. Moreover, the involvement of human subjects has been observed to be downplayed in

scientific and medical texts, making the likelihood of observing such contexts minimal. Thus, while the validity of these distinctions for fine-grained semantic analysis is recognized, they will not be made in this work.

Following this decision, it was also not considered to be necessary to limit consideration to specific types of causes, elaborations of cause, and causal agents; all of these were considered pertinent in the research.

In contrast, the analysis in this research was restricted to what may be referred to as basic or "core" CAUSE–EFFECT relations, and the markers that indicate them.[31] Nazarenko (2000) and Kahane and Mel'čuk (forthcoming), in their analyses of causation in lexical semantics, observed that the meanings of a large number of lexical units (for example, a very large number of transitive verbs) contain a component of causation. Kahane and Mel'čuk, for instance, differentiate between *verbs of causation* and *causative verbs*, which distinguishes between items whose senses include "pure" causation from those that include components corresponding both to causation and to its effect. As such, at a semantic level it would be possible to identify very specific sub-types of causation conveyed by a vast range of lexical units, potentially accompanied by expansion of corresponding relations at a conceptual level.

However, the complexity and specificity of these relationships also compromises the usefulness of these markers for extracting KRCs indicating conceptual relations of use in terminology work. The goal of using pattern-based tools in this field — principally the identification of conceptual relations in extracted contexts — involves the identification of links between two elements realized in the text. The complex causal relations in the cases described above do not meet these criteria, as they often involve relations that hold not between the items realized in the context but between one of these items and another element not realized separately or explicitly in the context.

---

[31] A discussion of this choice and some of its effects may be found in Section 5.5.3.1.

In addition, at a practical level, the investment of time and effort in developing marker forms for use in pattern-based tools for such a wide range of lexical units indicating these more complex causal relations would be prohibitive, as would that required to develop strategies for evaluating the information retrieved using such markers.

Finally, restrictions on the usefulness associated with more complex relations are also clear when one considers the range of contexts in which a lexical item such as CAUSER may be found, as compared to one such as NETTOYER or IRRITER, and the informative value of systematically including information about the former type of connection between concepts described in a terminological resource, as compared to that of the latter types.

All of these observations led to the decision to limit the evaluation of CAUSE–EFFECT in this research to the basic "core" varieties that will be described below.

1.5.2.8 Classification of CAUSE–EFFECT relations

In this Section, two classifications of CAUSE–EFFECT relations, by Garcia (1997) (Section 1.5.2.8.2) and Barrière (2002) (Section 1.5.2.8.3), will be presented. These will be preceded by a brief description of the analysis used by Talmy (1985), called upon by both of these classifications (Section 1.5.2.8.1). Then, in Section 1.5.2.8.4, the two classifications will be compared and contrasted with one another and with the analyses of CAUSE–EFFECT relations presented in Section 1.5.2. The motivations for the choice of classification used in this research will then be discussed in Section 1.5.2.8.5.

1.5.2.8.1 Talmy

In developing classifications of CAUSE–EFFECT relations for use in pattern-based tools, both Garcia (1997) and Barrière (2002) refer to an analysis developed in the context of

the theory of force dynamics, as described by Talmy (1985). This analysis is based on a model in which CAUSE–EFFECT relations involve the interaction of two forces, an agonist and an antagonist, which — as their names suggest — are in opposition. The type of effect that results depends on the initial state of the element affected, the relative strength of the forces, and the final state of the element affected given this opposition of forces (i.e., rest or motion). Talmy outlines two possible sets of situations, cases in which the strengths of the opposing forces remain constant (steady-state dynamics), and cases in which the strength of one of these forces changes (shifting-state dynamics).

Talmy (1985) asserts that this kind of classification forms the conceptual basis of many of the distinctions reflected at a semantic level in languages. This makes the classification particularly interesting for use as the basis of relation classifications in pattern-based applications, as these should reflect important conceptual distinctions in a way that mirrors those manifested in the choice of markers themselves as closely as possible, maximizing the possibilities for identifying relations at a conceptual level using their expression in texts (e.g., through the markers of these relations).

This conception of CAUSE–EFFECT relations thus permits the evaluation of relationships that result not only in the occurrence of a situation (cf. motion), but also of its non-occurrence (cf. rest), as well as the onset, end or continuation of these situations. Moreover, it may allow for varying degrees of influence on a situation (as illustrated, for example, by the discussion of *hindrance*). As such it provides a relatively broad portrait of various types of causal relationships.

1.5.2.8.2 Garcia

Garcia (1996, 1997) discusses relations involving efficient causes (cf. Aristotelian terminology for CAUSE–EFFECT relations, *Physics II*, illustrated in Appendix A). Her perspective is that of computer-assisted terminology, and more specifically the development of a tool, called COATIS, for the extraction of CAUSE–EFFECT relations from text corpora in the field of electricity. Garcia's approach relies on the observation

of verbs in a corpus, which she calls *indicateurs linguistiques* of CAUSE–EFFECT relations (1997: 10), *verbes indicateurs de causalité* or simply *indicateurs* (1997: 12). She notes (1997: 10) that these markers may simply indicate the presence of a CAUSE–EFFECT relation (essentially of the semantic primitive 'cause,' as in the case of verbs such as *causer, provoquer,* and *résulter*), or may also give information about the nature of the effect produced or of the causal action that produced it.

Garcia's classification of the efficient CAUSE–EFFECT relation (shown in Figure 5 and in more detail in Table 7), is largely based on that of Talmy (1985), and also reflects many of the aspects of Nuopponen's (1994) analysis (Section 1.5.2.5). In the case of relations in which a marker gives added information about a cause, she identifies a sub-category of relations involving contributing causes (indicated by markers such as *participer dans* and *contribuer à*), and within it another of collaborating causes (indicated by markers such as *coopérer à* and *collaborer à*). Garcia thus separates CONTRIBUTION from other types of CAUSE–EFFECT relations, in which the cause is (presumably) sufficient to lead to the effect.[32]

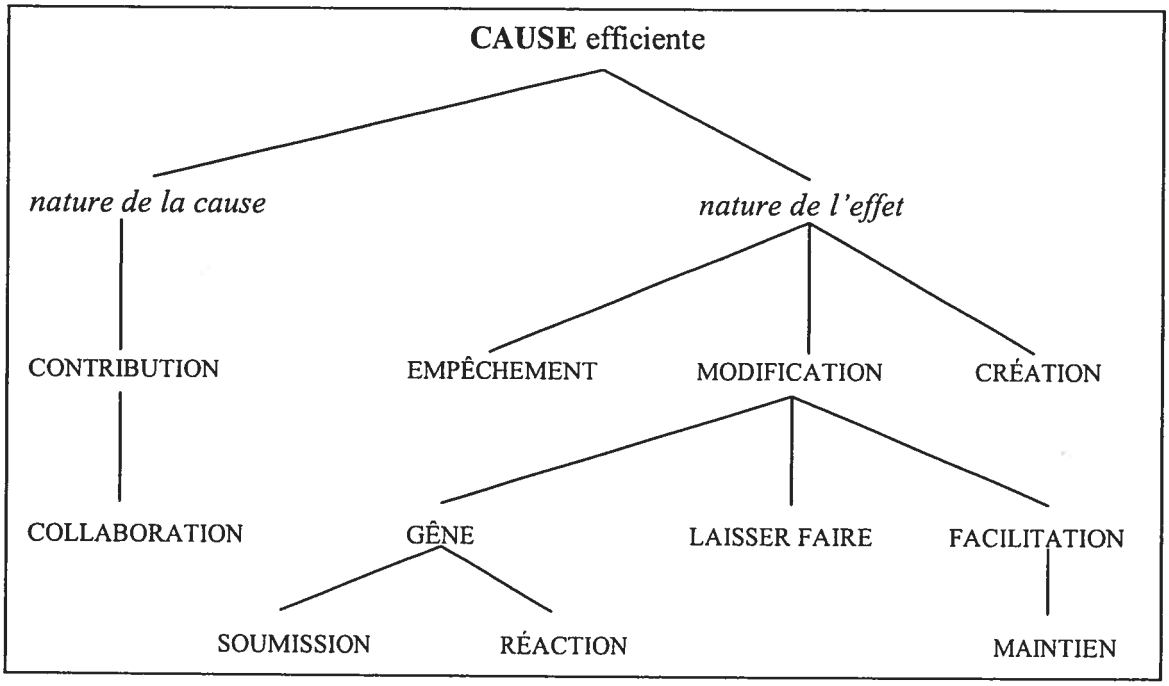
For efficient causes in which additional information about the nature of the effect produced is indicated, Garcia defines (1997: 11–12) a classification of CAUSE–EFFECT relations between two actions, associated with some typical verbal markers:

- CRÉATION (e.g., *déclencher, produire*);
- EMPÊCHEMENT (e.g., *empêcher, bloquer*);
- MODIFICATION (e.g., *changer, influencer*);
  - LAISSER-FAIRE (e.g., *permettre, autoriser*), a neutral modification;
  - FACILITATION (e.g., *favoriser, faciliter)*, a relatively positive modification;
    - MAINTIEN (e.g., *conserver, garder*), an extreme case of facilitation in which the cause not only facilitates but also is necessary to the effect;
  - GÊNE (e.g., *entraver, gêner*), a relatively negative modification;

[32] Nuopponen's (1994) co-operating causes can also be considered to form a category of contributing causes (although her alternative causes cannot, as each of these individual causes could be sufficient to bring about the effect).

- SOUMISSION (e.g., *accepter, souffrir de*), in which the effect submits to the influence of the cause; and
- RÉACTION (*réagir à, résister à*), in which the effect resists the influence of the cause.

Figure 5. Garcia's efficient causes (adapted from Garcia 1997: 11)



Garcia chose to limit her discussion to cases of CAUSE–EFFECT relations, i.e., INFLUENCE or INTERACTION, between actions in order to separate the discussion of CAUSALITY from that of agency, although she noted that in discussions of verbal polysemy these aspects are both important and closely connected (1997: 13). She noted that various types of causes may nevertheless be realized in a given utterance, including the action itself, its result, and the agent of the action, although rarely are all of these expressed in a single utterance; rather, the focus is usually placed on one of the three.

Again evoking terminology used by Aristotle, Garcia also mentioned formal causes (12) (cf. footnote 15 in Section 1.5.1.1).

Table 7. Garcia's classification of CAUSE–EFFECT relations (1997)

Type de causalité					**Indicateurs lexicaux**
Cause efficiente					*causer, provoquer, résulter*
	nature de la cause	CONTRIBUTION			*participer dans, intervenir dans, contribuer à*
			COLLABORATION		*coopérer à, concourir à, collaborer à*
	nature de l'effet	EMPÊCHEMENT			*éviter, empêcher, bloquer*
		MODIFICATION			*changer, métamorphoser, influencer*
			GÊNE		*bouleverser, entraver, gêner*
				SOUMISSION	*accepter, pâtir, souffrir de*
				RÉACTION	*réagir à, résister à, s'opposer à*
			LAISSER-FAIRE		*permettre, laisser, autoriser*
			FACILITATION		*aider, favoriser, faciliter*
				MAINTIEN	*maintenir, conserver, garder*
		CRÉATION			*déclencher, faire naître, produire*

1.5.2.8.3 Barrière

Barrière's (2001, 2002) classification (Figure 6) is quite similar in many ways to those used by Garcia (1997) and Talmy (1985), owing much to both of these. However, Barrière further develops Garcia's classification by emphasizing and developing a larger hierarchical structure in which the types of relations identified may be organized. This structure highlights some of the important distinctions in the types of CAUSE–EFFECT relations that may be identified in corpora using lexical markers, and offers possibilities for classifying these types of relations automatically and for dealing with some cases of ambiguity of markers.

This classification, developed for use in a marker-based application for extracting CAUSE–EFFECT relations from text corpora (see Section 2.1.8), reflects not only a need to reflect conceptual realities and important distinctions between different types of conceptual CAUSE–EFFECT relations, but also the goal of automatic identification of different sub-relations through these markers, a task that may also be informed by more semantic aspects of marker meaning.

Barrière's classification begins with the identification of two categories, depending on the type of element affected by this interaction of forces. In the Existence dependency, the interaction affects the existence or non-existence of an entity, or the occurrence or non-occurrence of an event, while in the Influence dependency, it affects a given feature or property of an entity or an event.

In the Existence dependency, Barrière establishes parallels between the different interactions of forces identified by Talmy in the context of steady-state dynamics, associating Talmy's rest with non-existence of an entity or non-occurrence of an event, and motion with the existence of an entity or occurrence of an event. Correspondingly, in the Influence dependency, referring to the set of situations identified in shifting-state dynamics, Barrière creates a classification based on the change in a feature of an event or entity.

The Existence dependency is divided into four sub-types, CREATION, DESTRUCTION, MAINTENANCE, and PREVENTION. CREATION occurs when the interaction between the opposing forces brings into being an entity that did not previously exist or causes an event that was not previously occurring to take place.

The opposite of CREATION, DESTRUCTION occurs when the interaction between opposing forces causes something that previously existed to cease to exist, or an event that was previously taking place to stop.

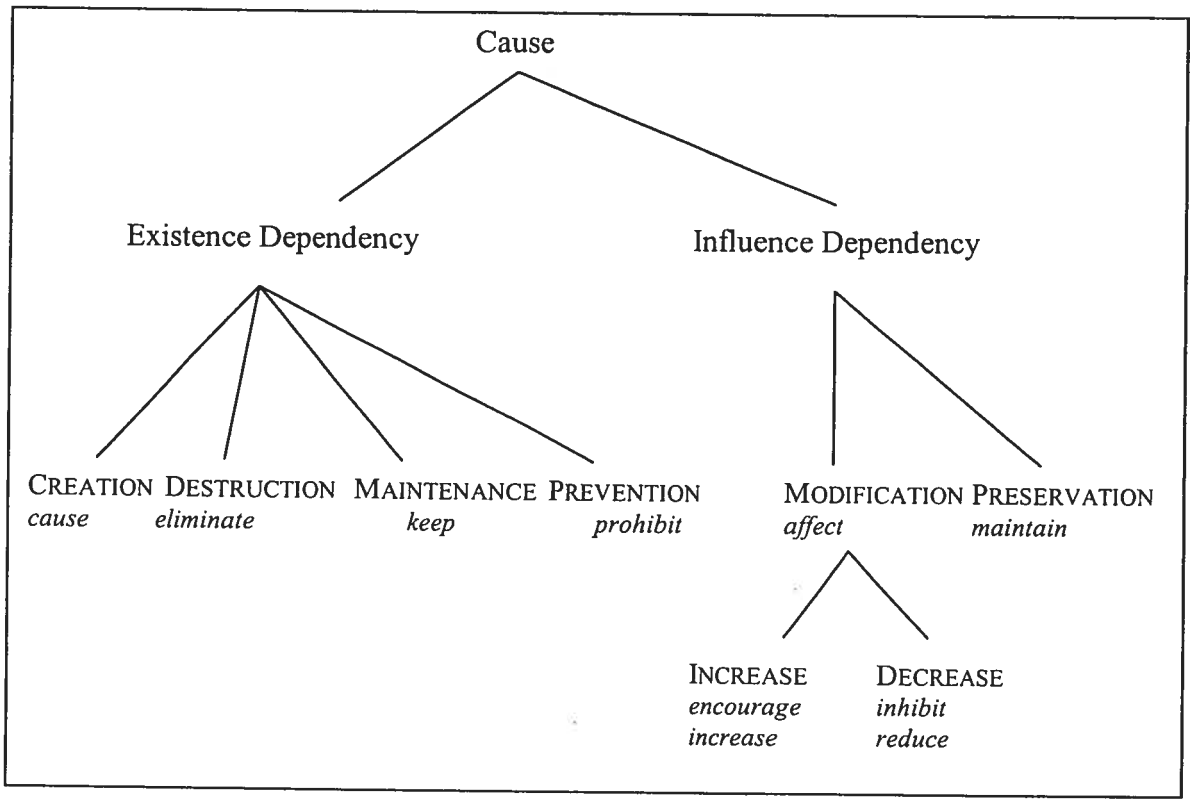


Figure 6. Barrière's classification of the CAUSE–EFFECT relation (Barrière 2002)

MAINTENANCE designates a situation in which an entity or event existed or was occurring before the interaction of the opposing forces and continues to exist or occur thereafter. It is worth noting Barrière classes in this category the relations denoted by verbs such as *allow* and *permit*.

In PREVENTION, an entity or event did not exist or occur before the interaction between opposing forces, and continues not to exist or occur.

Barrière's classification of the Influence dependency also includes four sub-types: MODIFICATION, INCREASE, DECREASE, and PRESERVATION. MODIFICATION is a sub-category that in turn includes those of INCREASE and DECREASE. It groups together all types of the relation in which the interaction between forces causes a change in a feature or property of an entity or event. However, this category is not limited to INCREASE and DECREASE, but also constitutes its own sub-type of the Influence dependency. This sub-group includes cases in which the forces' interaction have an effect on a characteristic or feature of an event or entity, but in which the kind of MODIFICATION may not be specified, or may be qualitative rather than quantitative (i.e., not easily identified as either an increase or a decrease). Examples are found in the pattern markers *influence* and *change*: without further explanation there is no way to know what form this influence may take or what kind of change occurs.

The category of INCREASE covers situations in which the feature of the entity or event is intensified or augmented by the interaction between the opposing forces.

DECREASE is the mirror image of the INCREASE sub-type: a feature of the entity or event is lessened or reduced by the interaction between the opposing forces.

PRESERVATION is analogous to the existence dependency's MAINTENANCE sub-category, since the feature of the entity or event exists before the interaction of the forces and continues to exist unchanged afterwards. However, perhaps less intuitively, it also includes those instances in which a feature or property is not present and continues not to be present after the interaction between the forces.

Table 8 presents Barrière's hierarchical classification of the relation and the effects of the force interaction, in addition to some of the relation markers for each sub-type. The left-hand column indicates the dependency; the column to its right

indicates the sub-categories. The third column indicates the effect of the interaction between the opposing forces. *E* designates an entity or event, *~E* the non-existence of an entity or the non-occurrence of an event, and f_i a feature of an entity or event, both before and after the interaction of forces. The right-hand column lists some English relation markers discovered in the course of Barrière's research (2001, 2002).

Table 8. Barrière's classification of the CAUSE–EFFECT relation (2002)

Existence dependency	CREATION		~E → E	*create* *generate* *produce*
	DESTRUCTION		E → ~E	*kill* *eliminate* *destroy*
	MAINTENANCE		E → E	*allow* *keep* *maintain*
	PREVENTION		~E → ~E	*prevent* *discourage* *control*
Influence dependency	MODIFICATION		$E:f_i <> E:f_i$	*influence* *change* *modify*
		INCREASE	$E:f_i < E:f_i$	*increase* *improve* *promote* *enhance*
		DECREASE	$E:f_i > E:f_i$	*reduce* *decrease* *shorten* *slow down* *deter* *discourage*
	PRESERVATION		$E:f_i = E:f_i$	*maintain* *keep* *retain*

1.5.2.8.4 Synthesis

Both of these classifications have been used successfully in research on pattern-based applications. Clearly — and unsurprisingly, given their common links to Talmy's analysis (1985) and Barrière's consultation of Garcia's classification in her work — they present strong similarities. Moreover, the development of Garcia's system for use in

French and Barrière's in English indicates that such systems — and particularly their commonalities — are promising for bilingual use.

In addition, these classifications reflect many aspects of the analyses of CAUSE–EFFECT relations in lexical semantics and terminology described above in Section 1.5.2, highlighting features of the relations that are important from both perspectives. This reflects the hybrid nature of the application itself, specifically its goal of using textual items (i.e., relation markers) to access information for use in conceptual analysis.

Both of these classifications — and of course the analysis developed by Talmy (1985), to which both of these refer — are heavily based on the nature of the change that occurs (or does not occur, in the case of sub-relations such as PREVENTION). This establishes commonalities with the lexical semantic analyses represented in Mel'čuk's lexical functions (Mel'čuk et al. 1995; Mel'čuk in preparation), and in Nazarenko (2000).

Similarities to the criterion of aspect as included in Mel'čuk's verbal lexical functions (described in Section 1.5.2.4) may also be observed, since sub-relations are distinguished according to whether a given entity's existence or an event's occurrence begins, stops or continues.

The classifications diverge somewhat in the presence of some additional aspects in Garcia's (1997) classification. Like Nuopponen (1994), Garcia identifies sub-types of the CAUSE–EFFECT relation involving multiple causes. She also identifies specific ways in which the elements affected in the relations characterized as GÊNE may react to the influence exerted on them (i.e., by submitting or reacting to this influence). Moreover, like Nuopponen (1994) from a conceptual perspective and Kahane and Mel'čuk (forthcoming) from a lexical semantic point of view, Garcia (1996, 1997) also discusses types of elements that may occur in texts in the roles of causes and effects as criteria for classifying CAUSE–EFFECT relations.

1.5.2.8.5 Choice of CAUSE–EFFECT relation classification

Barrière's classification of the CAUSE–EFFECT relation (2002; cf. Section 1.5.2.8.3) was chosen for use in this research, for a number of reasons discussed below.

In the medical field, differentiating between sub-types of the CAUSE–EFFECT relation — and in particular the types of effects that may be expected — was considered to be essential: obviously, it is important in medicine to be able to distinguish factors that prevent an event from those that cause it, those that reduce an effect from those that increase it, and so on. Both Garcia's and Barrière's classifications (as well as the analysis by Talmy (1985), to which both authors refer) reflect these criteria, and thus important aspects of CAUSE–EFFECT relations in the medical domain that would suit the needs of terminologists attempting to describe and link concepts and terms terms in terminological resources. (The classifications thus parallel counteracting causes as identified by Nuopponen (1994), as well as important lexico-semantic distinctions made by Mel'čuk et al. (1996) (in lexical functions such as `Liqu`), and Nazarenko (2000).)

Moreover, in previous research (Garcia 1996, 1997; Barrière 2001, 2002; Marshman 2002) it has been shown that the classifications discussed here are adequate for use in pattern-based applications, and that individual knowledge patterns can often (although not always or always exclusively (cf. Barrière 2002: 102–3)) be linked to the sub-types of CAUSE–EFFECT relations they include. Moreover, the two classifications reflect a level of granularity that is realistic in semi-automatic applications. Both were thus considered to be good preliminary candidates for use in this research.

While in the field of medicine other aspects of CAUSE–EFFECT relations are clearly important, it was felt that the nature of the field and the texts in the corpus, as well as the goal of semi-automatic extraction of relations, precluded the consideration of a certain number of factors considered in the analyses presented above.

Given the fact that many of the corpus texts deal with current research in the domain, they often describe knowledge that is developing, and may not yet be complete.

In such cases, it can be very difficult to determine definitively whether a given cause is necessary and/or sufficient to produce an effect, and thus explicit indications of these characteristics in texts are likely to be infrequent. For this reason, and from observations of the corpus, contexts that clearly indicate whether a given cause is necessary or sufficient were considered likely to be too few for this distinction to be taken into account in the classification used.[33] In addition, given the evolving character of knowledge in the domain and the complex and multi-criterial evaluations that are necessary to attribute these characteristics to a given relation in many cases, the definitive attribution of these characteristics may not be advisable in future applications of the data extracted for terminology work, making these criteria marginal in their value for the evaluation and/or sorting of contexts.

In addition, while in some cases (e.g., when two or more items representing causes or effects are found in association with a relation marker) it could be possible to identify cases of causal concept co-ordination, it may be extremely difficult to determine whether these causes are alternative or co-operating, and the effects alternative or co-occurring. Thus it was not considered to be advisable to use these criteria in the classification of CAUSE–EFFECT relations extracted in the research, as was the case for example in Nuopponen's (1994) analysis, reflected in a less detailed form in Garcia (1997).[34]

---

[33] However, this does not preclude subsequent identification of indicators of these phenomena in contexts extracted, for example using a technique that would search for lexical indicators (e.g., *necessary, sufficient, always, only*) of these specific types of causes in proximity to markers of CAUSE–EFFECT relations.

[34] However, the presence of multiple elements sharing a role in a relation was noted in the applicable relation occurrences (cf. Section 2.6.2); this kind of processing may allow for pertinent contexts to be identified for a user, who can then interpret the kind of phenomenon observed. Moreover, the analysis of the type of link that exists between these related elements at a surface level may serve as a cue to aid the user in determining whether causes are co-operating or alternative, or effects are co-occurring or alternative. (For example, observations of lexical indicators of the presence of these specific sub-types of links (e.g., *both... and*, *either... or*) may be observed in proximity to more general causal markers. See Section 4.9.1.2.1 for a discussion of some of these markers, and Section 5.5.3.3 for a discussion of some of these possibilities.)

While the existence of causal chains, again reflected in Nuopponen's (1994) analysis of the relation, is clearly important in the field of medicine, the identification of such chains is somewhat outside the scope of conceptual relation identification in the kinds of applications discussed in this research, which are likely to involve identifying links between specific pairs of objects (e.g., between term records in a terminological resource or nodes in an ontology), or the formulation of brief descriptions of relations (e.g., in the case of definitions), rather than the comprehensive description of complex interactions and chains of causation. Finally — but most importantly — semi-automatic, pattern-based knowledge extraction methods are simply not adapted for locating chains of causation in their entirety, since a given marker is most likely to indicate only one instance of a relation at a time (although this relation may hold between different pairs of elements), while additional links in chains of CAUSE–EFFECT relations will probably be associated with separate markers. Thus, in this research it was deemed preferable to treat individual relations separately.[35]

Another distinction included in Nuopponen's (1994) analysis and mentioned briefly in Garcia (1997) — also discussed by Nazarenko (2000) and Kahane and Mel'čuk (forthcoming) — but not included in Garcia's or Barrière's classifications as such, is that between different classes of causes and effects, including causal agents and events. However, it may be observed in the descriptions by several authors that the expressions of participants in a relation may vary significantly at a textual level from occurrence to occurrence (with one or more types of causes or effects potentially indicated in a single occurrence), without necessarily reflecting an underlying difference at a conceptual level. While certainly these types of distinctions may be pertinent in

---

[35] In thematic research that attempts to identify all of the relations present in a given text collection, additional relations in a chain of causation should ideally be identified through their own markers. If they are not, or if more targeted research is being carried out, the approach favoured for applying knowledge patterns — which allows the user to consult original contexts as needed — would give textual information about causal chains on demand. In addition, where specifically required (for example, for the acquisition of domain knowledge, in which more comprehensive information may be required), causal chains could be located in searches for a given term occurring as either a cause or an effect, or using multiple searches.

many analyses, in terminology work focusing on the relations between concepts as denoted in texts, and the identification of relation occurrences using lexical markers, this criterion was not considered to be a necessary one in the initial classification of relations.[36]

For the above reasons, aspects of classifications such as those of Nuopponen (1994) and Garcia (1997) were not considered to be particularly well suited for the kind of semi-automatic application intended here, and thus were not used in this research. It is our belief that many such fine distinctions are certainly valid and important in terminology work — and especially in medicine, to which Nuopponen's classification was applied — but are best left to an expert or terminologist, and not attempted by an automatic or semi-automatic application. While Barrière's classification may still pose challenges for semi-automatic applications (see the discussion in Section 5.5.3.2), this level of granularity should allow a preliminary sorting of contexts according to criteria important in the field.

Some research in the domain — and how these projects may be compared and contrasted with this work — will be presented in Chapter 2.

---

[36] It may, however, be taken into account in applications such as the sorting of contexts or refinement of pattern forms using semantic classes of actants, as described in Section 2.2. A brief discussion of some cases in which variation in the expression of related elements may be observed is also included in Section 5.5.4.

2 The state of the art

In this chapter, key points in previous research projects will be identified and summarized. Section 2.1 describes a number of research projects that have explored knowledge-pattern-type approaches to extracting information from texts. Section 2.2 gives a description of some research on refining pattern forms using the semantic classes of related elements. Section 2.3 outlines characteristics of knowledge patterns and their markers that have been identified as pertinent in the development and use of pattern-based semi-automatic knowledge extraction systems, and Section 2.4 describes additional textual elements that may present challenges in identifying and using knowledge patterns for extracting information from text, and in re-using the contexts identified using these kinds of tools for various applications. In light of this discussion, Section 2.5 will then present the objectives of this work, and Section 2.6 its originality as compared to the other projects described in this chapter, including a description of a number of factors evaluated from a new perspective in the course of the research.

2.1 Research in pattern-based knowledge extraction

Many researchers, including Hearst (1992), Ahmad and Fulford (1992), Pearson (1998) and Condamines and Rebeyrolle (2000), have worked on developing lists of patterns indicating semantic and conceptual relations and methodologies for discovering and applying these patterns. (For a summary of these research projects, see Appendix B.) A number of projects that have evaluated and implemented knowledge patterns are discussed below.

2.1.1 Hearst

Hearst (1992, 1998) has been recognized as one of the first researchers to use patterns for the extraction of semantic relations from text corpora. At COLING in 1992, she presented research focusing on the automatic detection of the relation of HYPONYMY

from free text. Hearst focused on identifying patterns that were frequent (i.e., that would locate a large number of relation occurrences), domain-independent, reliable (i.e., precise, almost always indicating the desired relation), and that did not require annotation of the texts or previously-encoded knowledge. Such an approach, in Hearst's view, complements statistical methods of semantic relation identification, specifically because it does not require previously encoded (domain-specific) knowledge, and can identify relationships that occur only once in a corpus, while statistical methods require several occurrences.

Hearst's approach focused on discovering relations between noun phrases, and her pattern forms were designed accordingly. They target noun phrases linked by lexical markers, e.g., *such as*, *including*, and *especially*. Some examples are shown below:

NP *such as* {NP_1, NP_2 ...., (*and*|*or*)} NP_n
NP {,} *including* {NP,}* {*or*|*and*} NP
NP {,} *especially* {NP,}* {*or*|*and*} NP

The form indicated is quite restrictive, allowing for only minimal modification by adjectives (including quantification by *some*, *many*, *certain*, and *other*), and for the possibility of multi-item lists of hyponyms (as long as these are not interrupted by any external elements), and the last item is preceded by *and* (indicating conjunction) or *or* (indicating disjunction)).

In a later publication in the context of the WordNet project (1998), Hearst described efforts to simply and effectively identify cases of the relation of HYPONYMY from general language (journalistic) corpora with minimal annotation, following a similar approach.

Hearst's pattern discovery method — similar to that later applied by many other researchers, such as Feliu (2004) (see Section 2.1.11) — begins with the choice of a relation of interest, and with a word (or term) pair (taken from WordNet in Hearst's case) that illustrates this relation. Sentences containing the pair are extracted from a

corpus and regularities — possible patterns — are observed. This kind of approach, given its deliberate simplicity, should be relatively easy to transpose into new, specialized subject fields.

2.1.2 Ahmad et al.

Ahmad and Fulford (1992) and later Ahmad and Rogers (1997) have also considered and discussed knowledge patterns (or *knowledge probes*) for information retrieval in the context of terminological description.

Ahmad and Fulford studied an English-language corpus on automotive engineering in order to identify knowledge probes that could be used to identify semantic relations including HYPONYMY, MERONYMY, CAUSE, MATERIAL and SYNONYMY to assist in term identification and term system development. Markers were initially identified manually in the corpus, and the pattern sets were then expanded using markers' synonyms as given in a synonym dictionary and/or thesaurus. The markers used belonged to several different grammatical categories, including verbs, nouns, adjectives and adverbs, and were applied in a character-string-based approach that integrated wildcard characters as needed. In the application of the markers to a corpus for relation identification, the precision results obtained varied fairly widely; while some patterns were highly (or even perfectly) precise in the evaluation, others produced few — if any — useful contexts.

Ahmad and Rogers (1997: 749–750) mention briefly some of the paradigmatic sense relationships into which terms can enter, including SYNONYMY, ANTONYMY, HYPONYMY, HYPERONYMY, and MERONYMY. The indicators of these relationships as described include paralinguistic markers — such as parentheses following a domain term (which may indicate SYNONYMY) — as indicators of defining or explaining information. Lexical markers, including *is a* and *a kind of* (presumably for HYPONYMY) and *is composed of, consists of, has a* (MERONYMY) are also mentioned. The authors go on to note that "[w]hile such textual cues are language-dependent, the principles are

valid across languages" (1997: 750). However, this point is mentioned only in passing, and there are no examples from other languages given, much less any discussion of the effectiveness of the approach in other languages.

2.1.3 Meyer et al.

Of the approaches described here, Ingrid Meyer's is closest to that used in this research. In 1999, Meyer et al. described the possibilities of using corpus-analysis tools to perform conceptual sampling, targeting a subset of the contexts in corpora that illustrate important conceptual relations for the term being researched (i.e., knowledge-rich contexts in the updated definition of the term (1999: 256)). The authors note that terminologists could use knowledge-rich contexts intact in definitions, as starting points for constructing definitions, or as resources for their own knowledge acquisition and conceptual analysis (1999: 256–257). The conceptual sampling approach developed in the research was based on knowledge patterns, described as "predictable, recurring patterns in text" that manifest the relations in which the terminologist is interested (1999: 257). These knowledge patterns are classified into three categories: lexical, grammatical and paralinguistic.[37]

This research differed from many projects in the field at the time because it dealt not only with the more conventionally investigated relations of HYPERONYMY/HYPONYMY and MERONYMY/HOLONYMY, but also with relations such as FUNCTION (i.e., the relation between an object and the function it fulfills) and CAUSE–EFFECT. (One of the co-authors was Barrière, who went on to work extensively on the CAUSE–EFFECT relation (2001, 2002).) The research, which focused on the analysis of a corpus on childbirth, identified knowledge patterns in the form of character strings (with possible truncation) and could be used in combination with the term being researched in order to identify knowledge-rich contexts.

---

[37] See Section 1.2 for further description of Meyer's classes of knowledge patterns.

Several difficulties of the knowledge pattern approach were noted; among them anaphora (see Section 2.3.1.5.2.1), marker polysemy (see Section 2.3.1.2) and expressions of uncertainty (see Section 2.4.2). The research also identified some knowledge patterns that were domain-linked, in this case to the medical domain, such as *risk*, *exposed to* and *complication*. The article also mentions the possibility of patterns that are linked to a particular semantic class.[38] To explain this phenomenon, Meyer et al. (1999: 262) used the example of MERONYMY in relation to processes: as processes have temporal rather than physical parts, lexical markers that are useful may not be (only) the conventional ones (e.g., *part*, *contain*), but rather lexical items such as *stage*, *phase*, *during*, and *throughout*.

2.1.4 Pearson

Pearson (1998, 1999), used relatively strictly defined lexico-syntactic patterns, which she called *defining expositives*, in order to identify contexts that would be pertinent for terminologists for the construction of definitions. While not explicitly separating different relations that may underlie this kind of information, she nevertheless identified patterns that can be associated with HYPERONYMY, MERONYMY, and FUNCTION. Pearson worked on a series of sub-corpora, one of English specialized texts in the field of science (in an expert-to-expert communicative situation), one of English, French and Spanish specialized texts (in an expert-to-semi-expert communicative situation) in the field of telecommunications, and a third, of didactic texts (in a teacher-to-student communicative situation) used in preparing for the General Certificate of Secondary Education (GCSE) examinations (given in Britain at the end of the secondary school program in preparation for university studies). The patterns were composed of relation markers (what Pearson called *hinges* or *connective verbs*) in relatively specific forms (e.g., *is/are defined as*, *denote(s)*) which linked specific parts of speech. As discussed in

---

[38] On this point cf. also Pearson (1998: 209) cited in Bodson (2005: 86) and described in Section 2.1.4, and other projects outlined in Section 2.2.

the principal section of Pearson (1999), dealing with identifying formal definitions. Pearson's defining expositives are subject to further restrictions on the related elements (e.g., one of the elements linked by the hinge being a domain term, the other belonging to a set of labels that can be equated with semantic classes). She also imposed specific restrictions on the composition of the sentence containing the defining information (e.g., the principal clause containing the defining expositive, there being very little possibility for separation between the hinge and the terms linked by it).

The author does however recognize that not all definitions are formal ones, and that these rules must be modified if these less formal definitions are to be located in texts. She describes (1998: 168–190) what she calls *connectives* or *connective phrases* (e.g., *called*, *known as*), which may introduce relations such as SYNONYMY, EQUIVALENCE and GENERIC–SPECIFIC in less formal structures.

2.1.5 Garcia

Garcia (1996, 1997) worked on developing a tool, COATIS, for automatically analyzing causal relationships in texts; this tool aims not only to identify contexts containing relation occurrences, but also the elements linked by the relation as expressed in these contexts. Using linguistic markers (i.e., verbs that often express CAUSE–EFFECT relations), COATIS identifies contexts in parsed texts that appear to express this relation, analyzes these contexts in order to confirm whether this initial interpretation stands up to further tests, and finally identifies the cause (causal agent or action) involved. The final result of the analysis is a network consisting of nodes representing the causes identified, as expressed in the text, which are linked by arcs tagged with the linguistic marker of CAUSE–EFFECT relations identified (which is associated with a given sub-type of the CAUSE–EFFECT relation).

Thus, this tool provides a fine-grained analysis of the contexts, with a higher level of automation than many applications in semi-automatic knowledge extraction.

2.1.6 Séguéla

Séguéla (1999) also developed an automated system, Caméléon, which uses collections of lexico-syntactic knowledge patterns for the extraction of relations of HYPONYMY and MERONYMY from a variety of corpora. Calling upon observations by authors such as Jouis (1993; Jouis et al. 1997), he discussed the use of an approach integrating a base of general, widely applicable patterns with more domain-specific patterns for application in each specific corpus.[39] The combination of general and domain-specific patterns is an interesting approach, because it not only allows for the reusing of patterns where possible, maximizing return on investment of time and resources, but also allows for the addition or substitution of patterns which give increased or more precise access to the information likely to be contained in a corpus.

Another interesting aspect of some of the pattern forms described by Séguéla (1999: 55) is their representation in the form of regular expressions reflecting each inflected form of the marker that was considered likely to be pertinent.[40] This kind of approach allows the precise identification of desired pattern forms, but also of course requires detailed preliminary analysis of the pattern forms that are likely to be useful (which may vary from corpus to corpus — for example, depending on the level of specialization — and will of course vary by the part of speech of the pattern element in question), and could also reduce recall if not all pertinent pattern forms were included. Other patterns (1999: 57–58) included classes of markers or marker elements (e.g., indefinite articles) that may comprise a number of different members, or a list of two or more possible options (e.g., *dans/sur*; *plus/moins*). This strategy would also require the

---

[39] Probable domain-specific patterns have since been noted in many works and in different fields, among them Meyer et al. (1999) in medicine, Morgan (2000) and Marshman, Morgan and Meyer (2002), in computing and genetics, Condamines and Rebeyrolle (2001) in software engineering, Meyer (2001) in computing, and Bodson (2005) in computing and medicine. Cf. also Bowker (2003) on this subject, and Marshman and L'Homme (2006a) for an evaluation of the portability of markers of CAUSE–EFFECT relations between the fields of medicine and computing.

[40] Some of the more complex forms were represented as regular expressions in a separate resource, and a placeholder representing the class of possible values in this resource used in pattern forms (1999: 55).

description of various possible alternatives, but allows for some regularities in pattern structure to be exploited, reducing the need for specific pattern forms corresponding to each unit.

2.1.7 Condamines and Rebeyrolle

In her doctoral thesis (2000, cf. also 2000a), Rebeyrolle characterized various forms (and structures) of definitions (*énoncés définitoires directs* and *indirects*) in corpora representing different communicative situations in the domains of science and technology (including electricity, geomorphology, knowledge engineering and software development), in order to study potential differences in the distribution and expression of definitions indicated by linguistic markers in these situations. She identified a number of lexico-syntactic and paralinguistic knowledge patterns which she then represented as regular expressions.

Condamines and Rebeyrolle (2001) discussed the construction of a corpus-based terminological knowledge base, a form of text modelling. From the perspective of producing corpus-based terminological knowledge bases for corporations, they aimed to develop an approach for creating a conceptual representation of a text (in this case, a French-language text on software engineering) by automatically identifying candidate terms and the relations between the concepts they denote. Once the candidate terms were identified, they began by studying paradigmatic relationships, constructing a series of hierarchies of concepts denoted by terms. These hierarchies were based on the inheritance of features and shared heads of candidate terms (CTs). This work was carried out using a previously identified core set of knowledge patterns for the relations of HYPERONYMY and MERONYMY (e.g., *a* CT1 *is a* CT2 *which*, *a* CT1 *is split into* CT2 *and* CT3). The authors classified the results obtained into two categories: contexts that show the expected relationship with a generic point of view (i.e., useful contexts); and contexts that either do not show the expected relationship, or show it from a specific or subjective point of view (i.e., less useful contexts). They then went on to discover

syntagmatic relationships by searching for combinations of candidate terms from different hierarchies.

The approach used was iterative: contexts including two candidate terms were observed for regularities that were associated with a particular conceptual relationship. The common linguistic elements (patterns) and their morphological, combinatory and syntactic restrictions were observed. These patterns were then used as search terms in combination with the previously identified candidate terms, in order to locate new patterns. This cycle was repeated as necessary.

In the testing phase of the approach described above, the authors used very strictly morphologically, syntactically and combinatorially defined patterns linking two candidate terms, in order to minimize noise produced. Below are two examples for the relation of HYPERONYMY (138–9):

> def_det + CT1 + Vtobe (present) + undef_det + {kind, type, etc. of} CT2 + {relative clause, past participle, present participle, adjective}
>
> def_det + CT1 + Vtobe (present) + undef_det + CT2 + {relative clause, past participle, present participle, adjective}

Several syntagmatic relationships were identified, each associated with a particular combination of classes of nouns (activities, documents, humans, and time periods). These relations included the relations "is composed of" (document/document), "precedes" (activity/time period, time period/time period), "starts during," "ends during," "occurs during," "conditions the start of" and "conditions the end of" (activity/time period). For these relations, a series of verbs that may indicate the relation were identified. It should be noted that since the goal of the research was to create a model of this particular corpus (although in this case it was for testing purposes), there was no reference to previously established lists of relations; rather the relations were taken from the texts themselves.

Among the difficulties encountered in the research was that of ellipsis in the expression of the items participating in a given relationship (2001: 142). In addition to potentially causing problems with the identification of potential patterns using an approach that depends on the presence of candidate terms, ellipsis can also cause difficulties in the identification of classes of nouns involved occurrences of a relation (e.g., in the sentence *The project leader is responsible for the development plan* [document], which would be more precisely rendered as *The project leader is responsible for the writing* [activity] *of the development plan.*)

The authors noted (2001: 135) that the observed linguistic patterns could be stored and re-used in further research projects in various domains. However, they also observed that some patterns are likely to be linked to a particular domain or text type, and that an iterative approach like the one used above could be beneficial in locating domain-specific patterns in new corpora. Moreover, they stressed (2001: 136) that tools for automatic extraction of information do not present perfect results, and still rely on subsequent human analysis.

In another research project, Condamines (2000–2) examined the case of *chez*, a particularly polysemous candidate marker for the relation of MERONYMY (or more exactly — to make a seldom-used distinction — HOLONYMY) in French. Using corpora in the field of the natural sciences, she attempted to formalize some of the possible meanings of this marker, and the contexts in which it can be found, using internal restrictions (syntactic structure, the semantic classes of co-occurrents) and external restrictions (domain and text type). By taking into account the classes of verbs that can co-occur with the marker, the semantic class of the noun that follows it, and the presence and type of a determiner (definite, indefinite, quantifying) in this noun phrase, Condamines performed a fine-grained analysis of the contexts in order to define cases in which the context provides an example of a relation, and more particularly in which the relation is the one that is sought, that of part to whole. She concluded that this marker

was particularly frequently and reliably associated with the relation of HOLONYMY in corpora in the domain of natural sciences, and especially in didactic texts.

In addition, Condamines (2002) noted a number of phenomena that can complicate the prototypical but somewhat simplistic view of the knowledge pattern. One observation made (2002: 146) was that the lexical markers generally described as "denoting" a given relation may not play this role in all knowledge patterns; some markers may be very reliably associated with a given relation — i.e., may consistently occur in contexts that indicate a particular relation — while not actually expressing it themselves.[41] She went on to give the example (151) of the case described above — of the marker *chez* for the relation of MERONYMY in French natural science texts — noting that in contexts containing this marker, the relation is not being asserted, but rather simply mentioned in passing. The reader is expected to know this information already, but from a terminological perspective, it may be new and valuable data that is worth extracting.

The researcher also noted (Condamines 2002: 146, 153–5) that patterns may not always take the binary form mentioned above, linking two and only two elements. Rather, markers may participate in structures that involve two, three, or more arguments, some or all of which may be pertinent for the identification of various relations. Moreover, these structures may vary in their surface realizations.

Finally, Condamines observed that knowledge pattern occurrences may not always be complete in a given sentence (or in some cases even in a larger context).[42] One of the elements involved in a relation may be presented in a text a sentence or more

---

[41] This observation is also clearly true of many — if not all — paralinguistic knowledge patterns, although these are of course often not as reliably associated with a given relation as some of their lexical counterparts.

[42] This rejoins Pearson's (1998) description of what she calls *complex defining expositives*.

before the occurrence of a marker that indicates the relation; while such contexts are valid, they do not provide complete information.

2.1.8 Barrière

Barrière (2001, 2002) used an English-language corpus in the field of composting to identify and test some lexical knowledge patterns for the CAUSE–EFFECT relation. Her initial approach to finding candidate knowledge patterns was slightly different from the others described above, as the corpus was read line-by-line and possible markers identified manually (in contrast to the more common method using terms or term pairs to locate occurrences of relations and the markers that indicate them).

Like Garcia (1997), Barrière chose to concentrate primarily on the refinement and application of verbal patterns (although in this case there were some exceptions).[43] The focus on verb forms in this research was largely motivated by the precision of pattern markers belonging to various part of speech (POS) categories as observed in the initial results of pattern identification and testing (Table 9); patterns containing verbal markers were found to be approximately twice as precise as those containing nouns, adjectives and adverbs, which in turn showed precision approximately twice as high as those containing conjunctions. This variation in pattern precision led Barrière to consider verbs as the most promising possibilities for research.

Table 9. Pattern precision by marker POS (adapted from Barrière 2001: 145)

POS	Occurrences	Noise	Precision
Conjunction	1671	1387	0.17
Verb	1217	389	0.68
Noun	172	108	0.37
Adjective/Adverb	58	36	0.38
Total	3118	1911	0.39

[43] A similar focus on verbal relation markers can also be observed in Garcia (1997) in French, Pearson (1999) and Condamines and Rebeyrolle (2001) in English, Feliu (2004) in Catalan and Weilgaard (2004) in Danish. See the individual sections on these research projects for more details of the work.

2.1.9 Marshman et al.

The goal of my research at the M.A. level (Marshman 2002, cf. also Marshman 2002a, Marshman 2004) was to discover lexical knowledge patterns that could be used to find knowledge-rich contexts indicating the CAUSE–EFFECT relation in the field of biopharmaceuticals (i.e., pharmaceutical products that are produced, purified or activated using biological organisms, systems or processes), in both English and French. The research also included a description of some of the difficulties encountered in the context of the research, and a preliminary comparison of the patterns and difficulties encountered in English and French.

The methodology used was term-based, using domain terms (individually) to discover simple character strings representing markers that could be used to search unannotated corpora and extract contexts that (in a liberal interpretation similar to that of Meyer et al. (1999)) could be used by terminologists in their work. A character-string-based approach was chosen mainly because of the simplicity of its implementation: these markers could be put to use with minimal resources of time, software, and technological expertise. This is certainly an advantage for a terminologist who is carrying out specific rather than thematic research or who does not have access to many technological resources.

This research led to the identification of lists of potential lexical knowledge pattern markers in English and French, many of which showed excellent precision in the corpora. Moreover, as these markers were classified using Barrière's classification of the CAUSE–EFFECT relation (see Section 1.5.2.8.1), the research offered an opportunity to test the classification not only in a new domain and new corpus and in another language (French) in addition to English, but also in a term-based approach that differed from Barrière's manual one. On all counts, the classification was found to be largely satisfactory. The research also offered an opportunity to consider markers belonging to various part of speech classes.

However, inherent in such a character-string-based approach are many challenges, which a number of researchers have attempted to minimize using more sophisticated approaches, such as more restricted pattern forms and linguistic analyses. The results of the work — while positive — confirmed the usefulness of further refinement of many of the pattern markers observed in order to improve the product of the extraction.

The research also provided an opportunity to carry out a brief comparison of the results in English and French, taking into account not only the patterns themselves, but also some of the challenges encountered. This comparison highlighted the need to evaluate the performance of pattern-based approaches in the two languages, in order to determine the effectiveness that may be expected of this kind of application and the challenges that will be encountered, and how these will affect pattern-based tool development and use.

In a preliminary interlinguistic comparison carried out in Marshman (2002), some interesting phenomena were observed. Many (e.g., the relative prevalence of verbs occurring as pattern markers in English, and of nouns in French) seemed to draw parallels with observations already made in the field of comparative stylistics (e.g., by Vinay and Darbelnet (1958)). Some other typical stylistic issues in scientific and technical language (e.g., the backgrounding of the observer, often by using the passive voice) can also be compared and contrasted in light of unilingual descriptions, e.g., by Sager et al. (1980), for English, and Kocourek (1991), for French.

More discussion of difficulties confronted in the use of pattern-based approaches, including difficulties related to language, was included in Marshman et al. (2002). This article described the results of a study (Morgan 2000) focusing on patterns indicating the relation of HYPERONYMY in French, using corpora in the fields of computing and genetics that were built from texts and TERMIUM® term bank record definitions. Some of the interlingual differences observed were pattern-specific (e.g., lack of easily identified equivalents in the other language, differences in frequency

and/or precision, and differences in pattern variation), while others were more generalized (e.g., more flexible word order, emphatic forms, more highly inflected pattern forms, elision, more complex pattern structures in French).

However, this description remained at the level of observations — and often of observations of individual markers or small sets of markers — rather than formal evaluation of these phenomena. The lack of large-scale, more formal evaluations of differences and similarities of patterns and difficulties of identifying and using them in different languages leaves a gap in existing knowledge of how knowledge patterns may be useful in a bilingual or multilingual context, and the need for such research is evident.

2.1.10 Bowker

Bowker (2003) describes one of very few comparative studies of lexical knowledge patterns in different language varieties. In comparing occurrences of previously identified knowledge patterns for HYPERONYMY, MERONYMY, FUNCTION and CAUSE–EFFECT relations (adapted from Davidson 1998, Morgan 2000 and Marshman 2002) in corpora of French-language popular science texts from France and Quebec, Bowker identified some significant differences.

Overall, more contexts were located in the French than in the Québécois corpus. Moreover, some specific markers showed significantly different productivity in the two corpora. Bowker notes that these differences — while they were observed in a pilot project only and would benefit from further investigation on a larger scale under more controlled conditions — may be indicative of a need to adapt pattern lists to the language variety of the corpus being analyzed, either by selecting appropriate patterns or by ranking patterns according to their usefulness in a given variety.

2.1.11 Feliu

In her doctoral thesis, Feliu (2004) — a member of the IULATERM research team (IULA, Universitat Pompeu Fabra) working on the Genoma-KB project — investigated the primary relations in the field of genetics (specifically, the human genome). Her main goals were the construction of a prototype system for the semiautomatic extraction of conceptual relations from specialized texts, and the development and evaluation of a typology of conceptual relations using the data observed in the course of the research (cf. Section 1.4.4).

Feliu used lexical markers — specifically verbs — to search texts for what she identified as the most central conceptual relations in that field. In the phase of pattern discovery, she used an approach similar to Hearst's (1992), adapted for use in specialized corpora (involving the use of pairs of terms previously identified as being linked by the relation of interest).

Feliu then explored a number of contexts containing these verbal markers to evaluate their usefulness — with the tool Mercedes (Vivaldi 2003), which extracts sentences that contain one of a series of verbal markers in addition to at least one previously identified domain term — and to identify possibilities for refining pattern forms.

In her description of the markers' usefulness for extracting contexts containing relation occurrences, Feliu noted (2004: 126–127, 137–138) that the structures in which these verbal markers occurred, and especially the prepositions that occurred with them, were often particularly useful in identifying the specifc conceptual relation present in a given context. However, Feliu also noted (2004: 169) that the implementation of these markers using Mercedes was complicated by the possibility of interruptions of these complex structures, and thus that the inclusion of these other elements in marker forms for character-string-based applications may lead to significant numbers of silences in the results of extraction.

In the evaluation of the usability of contexts retrieved by Mercedes, Feliu (2004: 173–183) identified several phenomena that call into question the presence or usefulness of the conceptual relation observed and require the disqualification of these contexts from further analysis: negation, expressions of possibility (i.e., uncertainty), anaphora, impersonal verb forms, and the failure to express one or more of the entities linked by the relation in the context.

Feliu then went on to discuss possibilities for future applications of these markers in syntactically annotated corpora, describing fairly precisely pattern structures involving the presence of elements linked by the marker (presumably domain terms) directly preceding and following these markers for application (2004: 192–202).

2.1.12 Weilgaard

Weilgaard (2004), in a research project using a Danish hydraulics corpus and a popular science corpus, studied the use of the Pronominal Approach (PA) in order to analyze the application of verbs' valency structures for the retrieval of information (definitions, synonyms). For the simple reason that corpus-annotation tools for Danish are not widely available, she worked with corpora that were not syntactically or semantically annotated, using data on verb valency structures drawn from the Odense Valency Dictionary (Daugaard and Kirchmeier-Andersen 1995).

The pronominal approach involves standardizing contexts found in corpora by describing the argument structures of potential knowledge pattern markers in verb form (including both compulsory and optional arguments) and by using pronouns, which carry information about these arguments such as some of their most basic semantic classes (human, animate, inanimate, etc.). (Further details of Weilgaard's use of semantic classes in the characterization of the verbs' arguments are found in Section 2.2.2.) This information was used for the disambiguation of potentially polysemous verbs, as well as for the identification of pertinent information for definitions (e.g., differentiating characteristics). The verbs studied fell into three categories:

metalinguistic concept-related verbs (e.g., *define*, *characterize*), metalinguistic term-related verbs (e.g., *call*, *denote*), and relational verbs (e.g., *consist of*, *belong to*). Results showed that the more promising verbs belonged to the first and third groups, those that dealt more with the conceptual than the term level. One of the factors contributing to this increased success was the presence of prepositions linking certain types of arguments to the verb marker, which could be used to identify the role of a given element in a context. In contrast, term-related metalinguistic verbs often took direct objects with no linking preposition, which precluded this kind of analysis.

In light of her results, Weilgaard suggested using a structured system of corpus analysis, using the verb patterns that showed the most regular valency patterns and class associations in a first pass and adding less regular patterns to complement the information observed in subsequent passes. She also observed that expanding the pattern list (with this type of analysis) to include nominalizations of verbs, other lexical elements, and paralinguistic markers could also be productive and should be explored.

2.1.13 Rodríguez Penagos

In developing MOP (for *Metalinguistic Operator Processing*), Rodríguez (2004, 2004a) aimed to create an information extraction system that could help specialists in specialized lexicography and terminology to keep lexical resources up-to-date. The tool is designed as an aid to processing free texts in a variety of domains, in order to extract metalinguistic information — what the author calls *Explicit Metalinguistic Operations*, primarily information about terminological creation and modification — and help in entering it into what he terms a *Metalinguistic Information Database*. This resource, a kind of intermediary between raw corpus data and a terminological knowledge base (TKB), contains semi-structured information that specialists can use to update and add to lexical resources. While this information rarely constitutes a complete definition, it is often extremely useful, providing pragmatic information about a term's value,

acceptance, or usage constraints. Such a database is seen both as a complement to and as a tool for enriching TKBs.

The system uses previously identified markers of metalinguistic information similar to those noted in definitional contexts (e.g., by Pearson (1998) and Meyer et al. (1999)), observed in a corpus of sociology and refined using concordances from the British National Corpus. These markers may be lexical (e.g., *is called, termed, coined*) or paralinguistic (e.g., quotation marks, text layout), as well as pragmatic. However, unlike Meyer et al.'s punctual, term-based approach, Rodríguez Penagos takes a more thematic angle, beginning with the markers themselves. (This can be compared to Condamines and Rebeyrolle's goal of text modelling — an effort to collect the majority of the pertinent information in a text collection — and to applications in automatic ontology construction, in contrast to approaches such as Meyer et al.'s in which users are generally expected to target a specific term.) These markers were studied in part-of-speech tagged corpora, and were subsequently refined to reduce noise, using contextual information and applying restrictions by POS tag and by string in the context of the marker (up to three words before or after the marker).

A second phase of study using these markers involved the use of machine-learning techniques to automatically discover patterns. These algorithms were developed using POS-tagged examples.

Once identified, contexts containing metalinguistic information were automatically processed (using POS-tagging, shallow parsing, and other analyses) and inserted into a database template containing slots for the term described, the information about the term that was provided, and the marker of the relation.

2.1.14 Gillam et al.

Gillam et al. (2005) carried out research on a corpus in the field of nanotechnology with the aim of ontology construction. To achieve their goal of identifying candidate terms

and the GENERIC–SPECIFIC and SYNONYMY relations that hold between them — and ultimately of creating a concept hierarchy or ontology — they began by using statistical methods to create initial concept sub-hierarchies, which were then enriched using linguistic analysis. Their approach was iterative, using candidate terms discovered in the first phases of research, coupled with phrase patterns, to identify new candidates for a similar analysis.[44]

In the first stage of the process, candidate terms were identified using their "weirdness" in the specialized corpus as compared to the British National Corpus, and a distributional analysis of the contexts in which they occurred was carried out. This analysis produced a series of "trees" of candidate terms and their collocates,[45] which were then used as the starting point for a new series of analyses, and so on until the list of collocates was exhausted.

In the pattern-based phase of research, candidate terms and their collocates were used to identify linguistic patterns containing specific parts of speech that could indicate a relation such as that between a generic and one of its specifics. The markers in these patterns included ADJECTIVE + PREPOSITION (e.g., *such as*) and ADJECTIVAL PRONOMINAL (e.g., *and other* and *or other*) (2005: 70). These patterns were then represented formally using regular expressions and applied to the part-of-speech tagged corpus to identify corresponding occurrences, which were parsed to identify the terms involved. The trees thus created were then unified to create an integrated concept hierarchy.

The research as described in the article appears to use a rather unusual approach to the use of patterns, defining pattern markers using their parts of speech. These

---

[44] This iterative approach is somewhat similar to that used by Condamines and Rebeyrolle (2001).

[45] The term *collocate* is used here to denote a unit that co-occurs with another in a text or texts. In some schools (e.g., associated with Meaning-Text Theory), the term may be used specifically to refer to cases in which a link exists between co-occurring elements that involves a change in the meaning of one of the elements in that context. In this case, to refer to simple co-occurrence in a context, the term *co-occurrent* may be preferred in order to avoid confusion.

patterns thus have characteristics both of lexical or lexico-syntactic patterns (including a marker of the relation that joins two candidate terms) and of grammatical patterns (which are defined using their parts of speech). This approach can be contrasted with the approaches used by other researchers mentioned above, who in defining lexico-syntactic patterns generally associated syntactic information with specific lexical units, rather than using any combination corresponding to a given part of speech or combination of parts of speech.[46] The technique is also distinguished from grammatical patterns as identified by Meyer (2001), in which the parts of speech composing the pattern (e.g., NOUN + VERB for the FUNCTION relation) correspond to the elements linked by the relation, and not to the marker of this relation (e.g., as in the case of *a scalpel cuts* or *a monitor displays*).

In addition to this difference, this research can also be distinguished from those described above because of its integration of both statistical methods (weirdness, distributional analysis) and linguistic analysis (patterns).

2.1.15 Malaisé et al.

Malaisé et al. (2005), described research focusing on the identification of semantic relations pertinent for the construction of differential ontologies — that is, ontologies containing information differentiating parent nodes from children and sibling nodes from one another. Specifically, the tasks involved were the selection of terms for an ontology, and then the structuring of the ontology both vertically (with links between hyperonyms and hyponyms) and horizontally (by differentiating between elements at the

---

[46] Some of the more complex pattern forms described by Séguéla (1999) more strongly resembled those described by Gillam et al., specifically in their use in the pattern forms of part of speech classes of marker elements. However, these classes also often indicated information about the semantic characteristics of the members of these classes (e.g., verbs of decomposition), and lists of specific units corresponding to each class were specified in a separate resource (1999: 55). Where possible, however, Séguéla's pattern forms included specific strings.

In terms of the relations observed, Bodson noted (2005: 275) that the most frequently observed relations in this study were not those that are most often studied in the field of terminology; HYPERONYMY and MERONYMY were not as common as FUNCTION in these corpora, and CAUSE–EFFECT relations were quite common as well. She went on to observe (2005: 276) that the use of a collection of relations such as the ones observed in her research and the links that they have with specific semantic types of terms would allow a user to predict the type of definitional information that is most likely to be used (and useful) for a given term.

2.2.4 Marshman and L'Homme

In Marshman and L'Homme (2006), a series of 14 polysemous verbal markers of CAUSE–EFFECT relations in English medical texts were evaluated, on the basis of a manual analysis of their actantial structures and the classes of their actants. The goal of the work was to determine what possibilities these characteristics might offer for disambiguation of markers in a medical corpus and for the distinction between their causal and non-causal senses.

The WordNet and UMLS classifications (and particularly their upper levels) were consulted in order to evaluate possibilities for assigning semantic classes to actants of these verbal markers. Generalization based on UMLS classes where available — with recourse to WordNet as necessary — was chosen as the most functional solution (in light of some of the challenges in using these classifications, described below).

This analysis provided promising results. In the first stage of disambiguation, efforts were made to distinguish contexts containing non-causal senses of the markers from those containing causal ones. Using the actantial structures of the markers, 9 non-causal senses among the total of 46 senses identified for the set of markers could be completely excluded, and some occurrences of another 8 senses, for a total of approximately half of the contexts containing non-causal senses of the markers. In the next step, using the semantic classes of the actants of the markers, 2 senses were

eliminated completely and some contexts containing another 7 senses were excluded, accounting for approximately $1/7^{th}$ of the remaining non-causal contexts. This step thus eliminated a total of approximately two-thirds of the original number of non-causal contexts in the sample. In the second main stage of disambiguation involving the sorting of various causal senses, nine senses, representing approximately half of the causal contexts, required no sorting because they were the only causal senses for a given marker. Among the occurrences of markers with two or more causal senses, actantial structures allowed for the sorting of all occurrences of five senses and some occurrences of another 4, approximately a third of the total number of contexts remaining. Semantic classes allowed for the sorting of all occurrences of 3 senses and some occurrences of 5 more, accounting for approximately three quarters of the remaining contexts. While there were a number of senses and contexts that could not be disambiguated using these techniques, the approach showed promise for disambiguation, especially if technical difficulties could be overcome.

Difficulties were noted in the use of both classifications evaluated. The coverage offered by WordNet was not comprehensive for this task, because of its general-language orientation and the emphasis placed on single-word units rather than the complex noun phrases that were most commonly observed in the contexts analyzed. The UMLS with its specialized orientation offered more comprehensive domain coverage, but the level of granularity of its classification often provided several possibilities for each unit found, which required significant user intervention in class choice. In addition, some apparent inconsistencies in the classification of some units (at least, for the purposes of this project) posed challenges. The need for a resource more adapted to this kind of task was thus clearly apparent, especially if automated approaches to disambiguation are envisaged. As a result, neither resource was found to be particularly well suited to this task.

2.3 Pattern characteristics

This Section will review and discuss some of the characteristics of knowledge patterns that have been addressed in some of the research projects described above and identified as being pertinent for the identification and application of knowledge patterns for various purposes. These include the frequency of marker and/or pattern occurrences and the types of pattern markers that are found (specifically the part of speech classes to which they belong), marker precision and polysemy, as well as the number and form of the elements linked by the patterns.

2.3.1.1 Number of occurrences of each marker

The impact of the number of marker occurrences in a given corpus is significant in terms of a pattern-based approach's productivity, determining the number of potentially valid contexts that may be retrieved using that marker. Logically, the fewer contexts identified per pattern, the more patterns are likely to be required in order to obtain the same number of potentially useful contexts from a corpus. While pattern precision is of course also a factor in the ultimate productivity of markers (since frequent but imprecise patterns may not provide any more useful information than rarer but more precise ones), marker frequency is the starting point for evaluating potential productivity.

Bowker (2003) addressed differences in numbers of occurrences of markers in a comparison between two varieties of French. Variation in the productivity of a pattern-based approach using knowledge patterns was considered to be likely, as a general trend towards more occurrences of patterns was noted in the texts from France.

In Marshman (2004) it was noted that mean pattern frequency for the English patterns identified was significantly higher than that of the French patterns in corpora of approximately the same number of tokens as calculated by WordSmith Tools (225,000 words in English and 224,000 words in French); mean frequency in English was 25, and in French 20, a difference of 20%. This would seem to reflect conventional wisdom that

French is less tolerant of repetition than English, and that patterns may be used less frequently in order to vary expression. This difference in frequency could be indicative of lower productivity of the French patterns identified as compared to the English markers observed in corpora containing comparable numbers of tokens.

2.3.1.2 Types of pattern markers observed

In pattern-based applications, the types of patterns used may have a significant effect on the productivity of a tool and on the difficulties associated with its use. In this Section, the focus will be placed specifically on the part of speech classes to which markers may belong, and how these have been considered in various projects. This factor may be pertinent in a number of ways in the development and use of pattern-based tools, for example because of its pertinence in guiding the choice of markers for use, and in its potential link with the recall and precision of the results obtained.

As noted in Section 2.1, in the course of research projects on pattern-based identification of conceptual relationships, researchers have used different forms of knowledge patterns, but in many cases (e.g., Garcia 1997; Barrière 2002; Feliu 2004) ultimately concentrated on those containing verbal markers. In Barrière's case, this decision was supported by observations that the verbal markers she identified were observed in her corpus to be significantly more precise than those belonging to other part of speech classes (Barrière 2001: 145).

However, both Garcia and Barrière were focused on markers of the CAUSE–EFFECT relation exclusively, and Feliu also addressed this relation. For other relations, e.g., in the case of definitions, which often contain GENERIC–SPECIFIC relations, very productive patterns belonging to other grammatical categories have been observed (e.g., Pearson 1999; Meyer et al. 1999; Condamines and Rebeyrolle 2001; Meyer 2001; Marshman et al. 2002); an excellent example is one of the most commonly cited examples of patterns for the GENERIC–SPECIFIC relation, *type of*.

In Marshman (2002, 2004), a difference in the part of speech classes of the lexical units that corresponded to the markers identified in English and French corpora was observed.[51] When only the three main categories of the patterns (nouns, verbs, and adjectives) were considered, there were proportionally more nouns in French than in English, and fewer verbs. Verb forms (including past and present participles) accounted for 54% of pattern occurrences in the English corpus, noun forms (including nominalizations of verbs) for 32%, and adjective forms for 14%. In the French corpus, verb forms accounted for 49% of the pattern occurrences, noun forms for 44%, and adjective forms for 7%. These figures reflect conventional wisdom on the preference of French for nominal forms, and of English for verb forms (e.g., Vinay and Darbelnet 1958: 102–104).

While verbs may be among the more precise and thus effective markers for use in pattern-based applications, it is nevertheless obvious that in making the choice to limit patterns used to those containing a particular type of marker, the potential of a tool for retrieving relation occurrences will be significantly diminished.

In terms of the usefulness of various types of markers for the application of patterns for knowledge extraction, one difficulty immediately suggested by differences in the distribution of markers in part of speech classes in French and English observed in Marshman (2002) — specifically the higher proportion of nouns than verbs in French and the opposite in English — is reliability with which the concepts involved in a relation can be identified. For example, in a preliminary analysis of several patterns in the previous project, it seemed that both concepts were clearly specified more frequently in contexts with verb forms of the pattern than with noun forms.

---

[51] As the markers identified in the project were in character-string form, some could be used to retrieve occurrences of multiple units belonging to different part of speech classes. The comparison took into account the POS-class distribution of the range of lexical units potentially retrieved, in light of the forms observed in the corpora.

2.3.1.3 Marker precision

All researchers in the field agree that one of the major criteria determining the effectiveness of pattern-based resources is that of the precision with which markers and/or patterns identify the relations of interest in research. Clearly, then, it is essential to evaluate this aspect of marker performance in any project that studies knowledge patterns for KRC extraction. This measurement is indicative of the efficiency of an approach using knowledge patterns may offer, and the degree to which it succeeds in reducing the time and effort required to identify a particular type of information in corpora.

Precision is generally evaluated (e.g., Meyer et al. 1999; Séguéla 1999; Meyer 2001; Barrière 2001, 2002; Marshman et al. 2002; Marshman 2002, 2004) by calculating the proportions of results retrieved using a given tool, marker or pattern that are considered to be valid (i.e., in pattern-based tools for KRC extraction, generally the proportion of contexts retrieved that contain useful, complete occurrences of the desired relation).

However, some authors (e.g., Meyer et al. 1999) have noted that not all contexts that do not meet the criteria described above are without value. Rather, these often convey other types of information that may assist in the process of concept analysis and terminological description, though from another perspective. These have been characterized as *good noise* (Meyer et al. 1999: 261).

2.3.1.4 Marker polysemy

Many researchers (e.g., Meyer et al. 1999; Séguéla 1999; Meyer 2001; Condamines 2000–2, 2002; Bowker 2003; Feliu 2004; Weilgaard 2004; Bodson 2005; Malaisé et al. 2005; Marshman and L'Homme 2006) have observed that one of the major difficulties — if not the major difficulty — of lexical-pattern-based approaches is the polysemy of pattern markers. This polysemy is closely related to the issue of pattern precision.

Meyer et al. (1999), for example, discussed cases in which patterns may represent more than one relation (e.g., *consist* of*, which may indicate MERONYMY, but also HYPERONYMY or SYMPTOM (i.e., the link between a disease or other disorder and one of its manifestations)).[52] Again, although the relation specifically targeted may not be present in these cases, these contexts may constitute "good noise."

Condamines (2000–2), among others, noted that the usefulness of a given, polysemous marker for identifying a specific relation may be linked to a particular domain or text type (e.g., the marker of MERONYMY *chez* in didactic natural science texts). The use of these types of markers outside these parameters would be likely to produce more noise in semi-automatic extraction.

Studies of verbal markers of CAUSE–EFFECT relations in English and French (Marshman and L'Homme 2006, 2006a) — which focused on distinguishing different senses of these verbs and evaluating the proportions of their occurrences that conveyed causal and non-causal senses, and in the former case on the evaluation of strategies for distinguishing these senses automatically — identified a high level of polysemy in these verbal markers, and the presence of both causal and non-causal senses for many of these. These results indicate a need for disambiguation of marker senses in order to decrease noise levels in the results of marker-based KRC extraction. Moreover, the presence of distinct causal senses for many markers also indicates a possibility of further refining the processing of occurrences of these markers to provide a more fine-grained analysis of the relationships present in KRCs. Another phenomenon noted involved the existence of not only relatively basic, "core" causal senses of markers, but also senses that included a causal component accompanied by additional components that increase the complexity of the relationship expressed. As discussed in Section 1.5.2.7, the practical value of these occurrences for specific applications — and thus the precision of markers as it applies to these different applications — may vary.

---

[52] Similar phenomena were observed by Feliu (e.g., 2004: 119).

As described in Section 2.2, some of the approaches used in order to deal with polysemy include the analysis of contexts in order to identify the structures in which pattern markers may participate and the semantic classes to which the elements they connect belong, and to differentiate between senses of markers on this basis.[53]

Evaluations of the polysemy of pattern markers in different languages, text types or domains and of the performance of disambiguation techniques (e.g., Marshman and L'Homme 2006, 2006a), provide information about some of the differences that may affect the efficiency of pattern-based tools, and the problems that may be confronted in developing knowledge extraction applications.

2.3.1.5 Number and form of elements linked by patterns

In designing pattern forms for use in pattern-based tools that specify the structures in which markers may occur — and specifically the forms that the elements linked by these markers may take — or that attempt to identify these related elements automatically, it is necessary to evaluate both the number and form of these related elements. The occurrence of structures that are not taken into account in these activities may lead to difficulties in the identification of potentially useful contexts or in the automatic analysis of these contexts to identify pertinent information.

2.3.1.5.1 Number of elements linked by patterns

The pattern forms used by some researchers — such as those described by Malaisé et al. (2005) — would not allow for the appearance of more than one term in a given role. However, as has been observed in many research projects (e.g., Hearst 1992; Feliu 2004), the occurrence of multiple terms in the same role in a given context is not unusual and should be taken into account when developing pattern forms. Malaisé et al. do in fact note (2005: 28) that difficulties are encountered when more than one term is

---

[53] The strategies explored in some research projects are described in Section 2.2.4.

associated with a hyperonym in a given context. This observation indicates the importance of developing pattern forms and strategies for analysis that reflect the kinds of structures likely to be encountered in texts. (This issue will be discussed in more detail in Section 2.6.3.)

2.3.1.5.2 Form of elements linked by patterns

In a common view of pattern-based knowledge extraction, most elements connected by knowledge patterns are either intended or assumed to be terms — and thus usually nouns and noun phrases. This assumption is reflected in approaches and pattern forms used in numerous research projects (e.g., Hearst 1992; Pearson 1998; Condamines and Rebeyrolle 2001; Feliu 2004). Researchers often either specify the nominal form of the elements connected to pattern markers, or use previously identified candidate terms that are likely to take noun forms, given the traditional noun-centred view of the term.

However, approaches that take for granted that related elements will be terms, and these terms will occur in nominal form, may not find all occurrences of relations that may be useful, particularly for the formulation of definitions and for domain knowledge acquisition. Moreover, applications that attempt to identify the related elements in a given context must also be adapted to accommodate these variations if contexts containing non-nominal items are to be located and properly analyzed.

2.3.1.5.2.1 Anaphora

For any application in natural language processing, anaphora are among the most difficult elements to process, as automatically or semi-automatically identifying the all of the information that a human would understand in reading a text as a whole presents a number of challenges. The most general factor, of course, involves the human's ability to make logical inferences from the use of anaphoric expressions, which is a difficult task for automated processes. (For descriptions of attempts to develop algorithms and other techniques for resolving anaphora, see Boudreau (2005), discussing Hobbs (1978),

Lappin and Leass (1994) and Mitkov (1998), as well as Boudreau and Kittredge (2006).) Thus, in any computer application it may be difficult to identify the relationship between an anaphoric expression and its antecedent.

The importance of this issue in knowledge extraction using knowledge patterns has been noted by several researchers, including Pearson (1998), Meyer et al. (1999), Meyer (2001), Marshman et al. (2002), Bowker (2003), Feliu (2004) and Malaisé et al. (2005). At a formal level, Meyer et al. (1999), for example, noted that anaphora may interfere with term-based approaches in pattern-based applications (i.e., tools that search for a term in connection with a pattern) if the term is replaced by a pronoun or other anaphoric expression. The impact of the phenomenon on the usefulness of extracted contexts was noted, for example, by Feliu (2004), who chose in her evaluation of contexts containing markers to exclude those that contained anaphoric expressions, because they provided incomplete information for the intended application of her research.

In addition to these pattern characteristics, the frequency and nature of several difficulties involving items external to pattern forms may affect the development, use and usefulness of pattern-based tools. These are described in Section 2.4.

2.4 Challenges in using knowledge patterns and extracted contexts

All researchers in the field agree that several issues must be addressed before effective (semi-)automatic extraction of knowledge-rich contexts and of information related to conceptual relations can be achieved. In many research projects (e.g., Hearst 1992; Pearson 1998; Séguéla 1999; Feliu 2004), attempts were made to minimize these problems by using restricted sets of markers and/or pattern forms or by excluding from consideration contexts containing particular types of additional elements.

Two challenges in using lexical knowledge patterns for knowledge extraction — interruptions of pattern forms and expressions of uncertainty — are described below.

2.4.1 Pattern interruptions

The variation of knowledge pattern structures from their prototypical form is widely recognized as a challenge for many types of knowledge pattern-based applications. Interruptions that occur within pattern forms or their elements of course may interfere with the recognition of KRCs in texts, and dealing with this phenomenon often involves adapting pattern forms to allow for such divergences from the norm.

Different types of interruptions may pose distinct challenges for pattern design and tool performance. One type of interruption involves the non-contiguity of relation markers and the elements that they link; this phenomenon was noted, for example, in Marshman (2002) and Bowker (2003) and poses difficulties particularly for applications that specify the context in which markers may occur (either in the recognition or the analysis of potential KRCs) and that attempt to identify the elements linked by markers automatically. Interruptions of markers (e.g., as mentioned by Séguéla (1999), Bowker (2003) and Feliu (2004: 169)) and/or of the elements they link may also be expected to interfere with the recognition of KRCs, because of the risk that these will not be recognized in their modified form. (Such interruptions of related elements are of course particularly pertinent for applications that search for previously identified terms and/or candidate terms.) If high recall is to be maintained, the potential for interruptions should be taken into account when developing pattern forms for use and choosing approaches for implementing markers.

Some projects (e.g., Hearst 1992; Pearson 1998) have described fairly restrictive pattern forms that do not allow for the insertion of additional elements within pattern structures. Such an approach may be effective for targeting the most immediately and certainly useful contexts available, and also for ensuring that further analysis of contexts identified is as straightforward as possible.

However, this inevitably leads to the exclusion of a certain proportion of potentially useful contexts, and alternatives may be sought. Séguéla (1999: 55, 58) allowed in some pattern forms for minimal interruptions of pattern structures and/or markers (e.g., the occurrence of a maximum of two words between a verbal marker and a preposition with which it is used). Another approach, used for example by Meyer et al. (1999), involves the representation of markers alone as patterns, and the search for these markers in proximity to (but not necessarily contiguous with) a specified term. Feliu (2004) also chose at least in some applications (i.e., in her use of the tool Mercedes, for the gathering of contexts containing markers for the description of pattern forms) to use simple marker forms to retrieve any context that also contained a pertinent term, in order to reduce silences in the results. (However, Feliu also noted an associated reduction in the precision with which specific conceptual relations could be identified.)

Thus, the choice of approach for dealing with potential interruptions may vary depending on the context in which patterns are to be used and the needs of the users (including the user's goals, the level of automation desired in the processing and use of extracted contexts, the size of the corpus to be analyzed, and the requirements for precision and recall that result from these factors): in cases in which a more automatic approach to establishing links between elements (e.g., in the construction of ontologies) is intended (e.g., Hearst 1992), more restrictive and therefore more precise forms may be chosen; in cases in which a substantial human participation in the evaluation of contexts (e.g., for conceptual analysis) is envisaged (e.g., in Meyer et al. 1999), more permissive forms may be appropriate and may allow for greater recall.

In a very specific potential source of interruption, discussed in Section 2.1.14 in the context of work by Malaisé et al. (2005), more than one relation marker and/or knowledge pattern may be observed in a single context. This phenomenon was also discussed in Marshman et al. (2002: 9–10) and Marshman (2002: 103–104). Malaisé et al. (2005) cited the presence of multiple markers in a context as a positive predictor of the pertinence of a given context for knowledge extraction. However, while the presence

of more than one pattern or pattern marker certainly does indicate that a context is very likely be pertinent for analysis in some way (as it increases the chances that a relation will indeed be present), recognizing such a context's form as corresponding to a pattern may become considerably more difficult in the presence of multiple markers because of the interruption of knowledge pattern structures that may result.

Moreover, at a conceptual level the presence of multiple markers may pose challenges for identifying the relation present in a given context if these markers are generally associated with different relations or sub-relations, particularly if these co-occurring markers link the same two elements.

2.4.2 Expressions of uncertainty

In an article that focused on the intersection between fuzzy logic and linguistic semantics, Lakoff (1975: 221) observed:

> "[N]atural language concepts have vague boundaries and fuzzy edges and [...] consequently, natural language sentences will very often be neither true nor false, nor nonsensical, but rather true to a certain extent and false to a certain extent, true in certain respects and false in other respects."

These fuzzy boundaries make it difficult to reduce the truth value of many natural language statements to a form that meets formal logical principles; similarly, the information contained in many such statements may be equally difficult to process for applications that attempt to identify clear-cut, universal assertions — such as those of the existence of relations between concepts — in natural language texts (whether they do so using knowledge patterns or other techniques). A sort of continuum between the extremes of complete certainty and complete uncertainty of an assertion (or rather, the certainty of the untruth of an assertion) may be established, and indicators of the place of a statement along this continuum must often be taken into account in the development of strategies for information extraction from texts.

For example, in an analysis of information appearing in a children's dictionary, and calling upon a classification set out in Cruse (1986), Barrière (1996) identified five levels of certainty that may be identified in automatic processing of texts using textual cues. Cruse (1986: 16–20) presented these five "statuses" in the context of the analysis of semantic traits that are part of word meanings and the linguistic tests that can be used to evaluate them: criterial (i.e., necessarily included in a word's meaning), expected (i.e., normally assumed to be included in the meaning of a word), possible (i.e., neither assumed to be included in nor assumed not to be included in the meaning of a word), unexpected (i.e., normally assumed not to be included in a word's meaning) and excluded (i.e., necessarily not included in a word's meaning). Barrière applied these criteria in her analysis of definitions and examples given in the dictionary, and identified some markers of the status of information about a word's meaning.

The presence of expressions of the uncertainty of a statement may of course affect the usefulness of contexts for a given application, depending on the requirements of the task at hand in terms of the universality, strength and reliability of the information identified. In some cases, automatic strategies for identifying levels of certainty using textual cues may be developed (as in the case of Barrière (1996)); these may allow tools to present only the most valid contexts to a user, or to sort contexts according to their potential usefulness, presenting a user with the most promising contexts first.

Moreover, in addition to their impact on the interpretation of contexts' content and their usefulness for various applications in terminology work, expressions of uncertainty can also affect the structure of KRCs by interrupting pattern forms. Such interruptions, if unaccounted for in pattern design, may interfere with the recognition of KRCs in texts. These choices impact not only the form of patterns, but also the recall that can be expected.

2.4.2.1 Quantification of related elements

Barrière (1996) identified quantification as one of the primary methods of expressing either certainty or uncertainty in a statement of a relationship between two elements. In examples given in a children's dictionary, she identified quantifiers that indicated various levels of certainty, including *all* (criterial), *most*, *many* (expected), *some* (possible), *few* (unexpected) and *no* (excluded)) (1996: 187).

Another observation of this phenomenon — although from a different perspective — may be found in Sager's (1990: 32) discussion of tests for distinguishing "true" GENERIC–SPECIFIC from quasi-GENERIC relations, which introduce quantifiers into statements designed to test the solidity of relationships (Section 1.4.2).

This phenomenon is thus important to take into account, given that it can affect the value of a context for future use (e.g., particularly if the element present is excluded from participation in a relation); however, the evaluation of the impact of quantification requires that the specific indicator of quantification occurring in a context be taken into account, as some quantifiers (e.g., *all*, *tout*) may indicate certainty rather than uncertainty.

2.4.2.2 Hedging

Hedging, i.e., the use of linguistic markers to express uncertainty in regard to or to attenuate a statement, is extremely common in scientific texts and is often found in contexts that are potentially knowledge-rich.

Lakoff described (1975: 234) what he calls *hedges* as "words whose meaning implicitly involves fuzziness — words whose job is to make things fuzzier or less fuzzy," i.e., intensifiers and deintensifiers. He cites (1975: 235) examples of these (and related phenomena), including *more or less*, *roughly*, *somewhat*, *mostly*, *essentially*, *very*, *especially*, *exceptionally*, *often*, *almost*, *practically*, *actually*, and *really*. These hedges (1975: 248–250) may affect the interpretation of various aspects of meaning

(e.g., may indicate that a statement is true to a certain degree, or in a certain respect); they may also apply not only to the predicate they modify, but to the value of a statement as a whole. Lakoff notes (1975: 249) that hedges may indicate different degrees of hedging in a kind of continuum. Moreover, he also notes (1975: 247) that the context in which a hedge is used has an important effect on its meaning, and on the types of elements to which the intensification or attenuation can apply.

Pearson (1998: 115) discussed the presence of hedged definitions in her corpora, identifying two types of hedging. The first involves indicators of tentativeness, which indicate that "an author is being tentative about his/her claims" and is reserving the right to return to, refine and/or revise a statement at a later stage. The second involves indications of scope, which call into question the general applicability of a statement, and is used to avoid controversy and guard against challenges from others. Pearson discussed this issue primarily in terms of what she called *focusing adverbs*.

While Pearson (1998: 142–144) chose, in her work, to consider contexts that contain focusing adverbs such as *commonly*, *usually* and *generally* to be valid, presenting a widely accepted definition of a term, she chose to reject those containing markers such as *chiefly, mostly*, *frequently* and *often*, because she considered that these restricted the applicability of a statement (i.e., if a statement often applies, it does not always apply). However, she also noted (1998: 143) that many of the focusing adverbs she considered to justify the elimination of potential definitions from consideration in other contexts might be considered to be an assertion of the general applicability of a statement, rather than the reverse. This indicates, therefore, that the evaluation of the validity or general applicability of contexts containing such expressions of hedging may be extremely complex, and may be difficult to implement automatically.

However, adverbs are not the only available methods of hedging; other examples include the verbs *seem*, *appear*, *suppose* and *consider*, as noted for example by Lysvåg (1975: 125), which convey the belief and/or interpretation of a human in regard to the truth value of a given statement. Furthermore, Aijmer (1986), observed a number of

English hedges involving nouns, adjectives, conjunctions, interjections, adverbial or prepositional constructions, and clauses, showing that the means of hedging are extremely varied.

In contrast to approaches such as Pearson's (1998),[54] researchers including Meyer et al. (1999) and Barrière (Barrière 1996; Barrière and Hermet 2002) did not consider that expressions of uncertainty justified excluding contexts from consideration, although they stressed the need to account for indications of certainty and uncertainty when extracting knowledge from texts. In a section on assessing hits (1999: 265), Meyer et al. note the prevalence of "attenuating phrases" (including expressions of hedging, e.g., *is thought to*, *appear to be*, and modal verbs) in medical texts, and the fact that some potentially knowledge-rich contexts also contain negation, both of which may call into question the usefulness of these contexts. However, given their semi-automatic approach to information extraction, the authors chose to retain these potentially useful contexts and leave the decision as to their validity to the terminographer.

This perspective is reflected in the portrait provided by Barrière (2002: 105–107), focusing on indications of what the author referred to as the *probability* of a given relationship existing. She identified various means of expressing different levels of certainty, and a selection of adjectival, adverbial and other markers of these levels that may occur within contexts expressing relations.[55] She also noted the importance of this phenomenon and its evaluation in the processing of CAUSE–EFFECT relations in particular, citing the many nuances of certainty that may affect the usefulness of various statements for extracting information. The author stressed that the representation of these levels of certainty can be challenging for formal representation and evaluation.

---

[54] A similar approach was taken by Feliu (2004), who excluded from consideration in the evaluation of pattern forms contexts containing indications of what she referred to as *posibilitat* (i.e., possibility), on the basis that the information they contained was thus not universally valid or applicable.

[55] Such expressions may include *always* (criterial), *often* (expected), *sometimes* (possible) and *never* (unexpected/excluded) (Barrière 1996: 188).

2.4.2.3 Modal verbs

The presence of modal verbs in a given context can, like the use of quantification and hedging, express doubt about the certainty of a statement, and may thus call into question the validity of that context for subsequent applications in terminology work (and thus the usefulness of extracting that context), as observed in Pearson (1998: 115), Marshman (2002) and Bowker (2003), among many others.

The polysemy of modal verbs has been widely recognized, for example by Swan (1995: 334–336), who identifies potential uses including indicating various degrees of certainty (such as complete certainty, possibility and probability (be it strong, weak, theoretical, habitual, or conditional)), as well as ability, permission, obligation and necessity. In the context of specialized language, Sager et al. (1980: 210–212) also noted that modal verbs may be used in various ways, including indicating possibilities, predictions, generally applicable statements and logical expectations.

The impact that these verbs may have on the interpretation of a relation may thus vary from context to context; however, as mentioned by Sager et al. (1980: 210), there may be a preference for some of the possible meanings over others in specialized discourse, which may reduce the difficulties posed by their interpretation to more manageable levels. The authors noted, for example, that modal verbs such as *may*, *might* and *can* generally indicate possibilities in specialized language.

In the context of KRC extraction, Barrière (2002: 107) also discussed the levels of certainty that may be expressed by these verbs in the context of relation occurrences in English, also identifying *can*, *may* and *might* as indicators of possibility, as well as *must* as an indicator of criterial certainty.

2.4.2.4 Negation

Perhaps even more than the presence of hedging and modal verbs, negation can call into question the validity of a context for knowledge extraction and subsequent use. This issue has been discussed by several researchers (including Bowker (2003) and Feliu (2004)); for example, Bowden at al. (1996) in their research used negation as a *negative trigger*, eliminating from consideration contexts containing negation.

While excluding contexts containing negation is certainly the most conservative and thus probably "safest" approach in many (especially more highly automated) applications — and was also used, for example, by Feliu (2004) — there is always in these cases the possibility of eliminating valid contexts.[56]

Moreover, it is worth considering whether negated contexts may in fact constitute "good noise" (Meyer et al. 1999: 261), i.e., whether for some applications it is as important to know that a given relation does not hold between a given pair of items as that it does between another pair. This is a question that can only be answered by individual users in the context of their particular research projects. In such cases, an application might be designed to indicate the presence of negation to a user and/or to sort results according to the presence or absence of negation.

At a formal level, pattern forms may thus need to take into account a variety of different forms of negation in order to identify contexts in which negation is present and further — a far more complex task — to distinguish between cases in which negation affects the validity of a context from those in which it does not.

As the above discussion has illustrated, a number of characteristics of knowledge patterns can affect the development and performance of knowledge-pattern-based applications, and these tools can confront numerous and varied difficulties. Few

---

[56] This may be particularly significant when the data available for analysis are limited.

systematic studies of these issues have been carried out, and fewer still have compared them between languages. The analysis of these factors will constitute an important part of this research. The methodology adopted in the evaluation will be presented in Chapter 3. First, however, the objectives and originality of this research will be presented.

2.5 Objectives

In light of the results of previous research, we have developed a series of research questions and a methodology to be used to attempt to answer them.

2.5.1 Research questions

As a number of characteristics of knowledge patterns and their markers, as well as a number of factors external to the forms of patterns, can affect the development and performance of pattern-based tools, a number of questions can be asked about the potential for developing and using such tools in a bilingual environment: What are the differences in knowledge patterns (and their components) and how they occur in English and French? Are factors external to the patterns likely to affect these tasks in the two languages differently? Will these (potential) differences affect the possibilities and difficulties of identifying and designing patterns, of using them to extract knowledge-rich contexts containing information about conceptual relations in corpora, and/or of using the extracted contexts in terminology work? If so, what may their impact be? How can these differences be taken into account in developing pattern-based applications, in expectations of application performance, and the use of the contexts retrieved?

2.5.2 Hypothesis

We hypothesize that differences exist in a number of the characteristics of knowledge patterns and pattern markers and of the contexts in which they occur in English and French medical texts, and that these indicate a need for adaptation of methodologies and

expectations for knowledge pattern identification, development, and use in the two languages if comparable results are to be obtained.

2.5.3 General objectives

The general objectives of this research were thus to identify, analyze and evaluate a selection of characteristics of knowledge patterns and markers indicating conceptual relations of CAUSE–EFFECT and ASSOCIATION — and some aspects of the contexts in which they appear — that are pertinent in the development and use of pattern-based tools. The next goal of the work was to compare these results and observations of the process to observe similarities and differences in these patterns and markers and the challenges in their identification and potential use in English and French, in order to evaluate the impact that these similarities and differences might have on the design and implementation of knowledge patterns for the extraction of information about conceptual relations for the purposes of concept analysis and terminological description in a bilingual context.

2.5.4 Specific objectives

The specific goals of this research were thus:

- To identify in English- and French-language corpora of medical texts lexico-syntactic knowledge patterns indicating CAUSE–EFFECT and ASSOCIATION relations involving concepts denoted by domain terms;
- To observe and classify the characteristics of the patterns identified — including the nature, form and characteristics of the markers observed and/or of the elements that they link — that may affect pattern-based tool development and performance and the subsequent use of the candidate KRCs identified;
- To observe and classify phenomena related to the presence of elements external to the pattern structures — including pattern interruptions and expressions of uncertainty — that may pose challenges in pattern-based tool development and performance, and in the subsequent use of candidate KRCs identified;
- To compare the patterns, their markers and characteristics and the challenges observed in English and French; and

- To evaluate the similarities and differences observed between the results in the two languages, in order to study their implications for KRC extraction using these patterns in a bilingual context.

2.6 Originality of this research

In pursuing the objectives outlined above, this work will differ on several fronts from others in the area of extraction of conceptual relations in the medical domain.

First, the descriptive and comparative orientation of the study, and thus its general approach, set it apart from other research projects that have focused on developing functioning knowledge-extraction systems, terminological knowledge bases, or ontologies. Automatic knowledge structure or term base development — as in research projects such as MENELAS (e.g., Bouaud et al. 1995; Nazarenko et al. 1997, 2001; Zweigenbaum 1994; Zweigenbaum et al. 1995), GENIA (Ohta et al. 2001, 2002; Tateisi et al. 2000) and Genoma-KB (Cabré et al. 2004; Feliu et al. 2004), for example — often requires conservative choices, in order to ensure highest-quality output and little noise. However, given its goals, the aim in this study is to start with a wide range of potentially useful data (i.e., KRCs) that could be found using some kind of lexical pattern-based technique.

In fact, these choices are inspired not only from a desire to consider a wide range of data for the purposes of interlinguistic comparison, but also from a liberal perspective on what constitutes useful information in text corpora, and how this information may be used. The needs of terminologists — the primary users envisioned in this research, and those for whom the approach based on the extraction of knowledge-rich contexts was developed — are extremely varied. Meeting those needs in as many ways as possible requires flexibility, and thus may favour the consideration of an interactive approach, which presumes a certain amount of human intervention and judgment (i.e., *semi-automatic knowledge extraction* (Meyer et al. 1999: 258; Meyer 2001)). Semi-automatic KRC tools may, however, implement additional stages of context analysis, attempting to provide the user with the most pertinent data in the most efficient way possible and thus

to reduce the time and effort required to locate this information. Using the idea of semi-automatic KRC extraction as a starting point and then considering various ways extracted contexts might be processed to identify useful information provides an opportunity not only to observe phenomena that may be pertinent in a range of different situations and applications, but also to consider how potential interlinguistic differences may come into play at each of these levels.

This perspective thus determines aspects of the methodology such as: 1) the use of a single term rather than a term pair to generate the first set of concordances, which allows the retrieval of contexts that would otherwise be excluded, for example due to anaphora or "non-standard" forms of related elements; 2) in the wide range of candidate pattern markers and forms identified and evaluated, which is intended to provide maximum access to data on knowledge patterns and the ways in which they may appear in texts, rather than restricting occurrences to a certain type of marker or specific, standard forms and structures; and 3) in the choices made when analyzing the results, notably in the decisions on the inclusion and evaluation of contexts presenting difficulties such as interruptions or expressions of uncertainty.

The evaluation in this work of a wide range of occurrences of knowledge patterns according to a number of characteristics affecting their design and use (some discussed by other researchers, others added or further developed as a function of the observations made in this corpus and the comparative orientation of this project), thus presents an opportunity for structured observation of how and to what degree various factors may influence the semi-automatic knowledge extraction process and the results of this extraction.

In addition, the systematic analysis and comparison of challenges related to elements external to the knowledge patterns but occurring within contexts constitutes a new contribution to knowledge pattern research, as the gathering of this data will permit not only the evaluation of the relative frequencies of these issues and the forms in which they occur, but also the interlinguistic comparison of these frequencies and forms. This

will allow for further, more targeted analysis of the impact of these difficulties and for making informed choices in developing strategies for dealing with them.

Another difference from many previous projects is observed in the relations studied. Most research has tended to deal with the most widely recognized, hierarchical GENERIC–SPECIFIC and PART–WHOLE relations, particularly in relation to their use in definitions. While there has been increasing interest in the FUNCTION and CAUSE–EFFECT relations, others have still been largely neglected. However, their importance in the field (observable for example in their inclusion in the UMLS) shows that other relations such as ASSOCIATION are important in medicine. The study of these relations may fill some gaps in the information about useful knowledge patterns in the field, which may be particularly important given the volume of available data in text form.

Finally, perhaps the most significant points here are the bilingual nature of the research and its comparative approach. While research on knowledge patterns has been carried out in many languages (e.g., English, French, Spanish, Danish, Catalan), few studies have evaluated the possibility of using parallel techniques in two or more languages. Moreover, we are not aware of other research that has systematically compared the process, results, and difficulties of a pattern-based approach in two or more languages. This reveals a significant gap in knowledge about pattern-based approaches, especially as so much terminology work, particularly in the Canadian context, is carried out bilingually or multilingually.

The development of tools that may be used — equally effectively — in both English and French, and that thus offer users a way to identify and evaluate occurrences of relations from a similar perspective (for example, a similar search approach and analysis of relation sub-types) in the two languages, would be a valuable contribution to the field of terminology. However, it cannot be assumed that a knowledge-extraction approach will be equally effective, or will meet exactly the same challenges and successes, in the two languages. A truly bilingual approach must begin with an evaluation of the characteristics of knowledge patterns and the contexts in which they

appear in each language, and bilingual tools will likely need to be adapted to take any differences into account, in order to optimize results in both languages. This research will begin to fill the gap in our knowledge by gathering data that will improve our understanding of the approach and how it may perform in English and French and suggesting strategies for future development.

In addition to the aspects of the methodology and goals described above that set this research apart from other projects, the approach used in this work allowed for the evaluation of a certain number of criteria not explicitly described or not described in detail in many previous studies. These — and the motivations for examining them — are described below.

2.6.1 Evaluation of pattern marker types observed: Simple and complex

Different pattern types may be identified according to the form of the lexical marker they contain, and one of the distinctions that can be made is between patterns containing a simple marker (i.e., a marker that is in the form of a single lexical unit, such as the verbs *induce* or *induire*, as seen in the patterns X *induces* Y or X *induit* Y) and those containing a complex marker (i.e., a marker composed of two or more lexical units, such as *induction of*, as in X *induction of* Y, or *induction of... by* in *induction of* Y *by* X).

These variations are often simply different surface manifestations of the same lexical unit, but the fact remains that if surface-structure-based methods of identifying contexts in corpora (e.g., character strings, regular expressions) are used, different pattern forms must be used in order to locate these contexts (since, for example, a pattern that requires that *by* be present would not retrieve contexts in which other variants are present, and the placement of the related elements differs from one structure to another).

The form of pattern markers clearly affects the complexity of identifying (and then applying) potential pattern forms because it vastly increases the possibilities for variation in marker form. Pattern markers containing several different lexical units may show not only morphological variation (potentially for each lexical unit included in the marker), but may also vary in the order of the elements, which increases the number of possible pattern forms required to locate contexts containing these markers (as discussed in Section 2.6.2). Over and above the order of elements, complex markers may also be interrupted by external elements, again indicating a need for adjustments in pattern forms (as discussed in Section 2.4.1).[57] Because of this increased complexity in pattern design and KRC identification, this criterion was considered to be worthy of evaluation.

2.6.2 Evaluation of pattern variation

As mentioned in Bowker (2003), while pattern-based tools rely on the representation of forms of knowledge patterns identified as recurring and therefore promising for information retrieval, occurrences in texts may vary (in the form of pattern markers and/or of the structures in which they appear). This variation must be taken into account when designing patterns for use in knowledge-extraction applications; this may require a significant investment of time and effort when high levels of variation are present. Additionally, several pattern forms corresponding to a single marker may be required in order to implement these different structures. Finally, the adaptation of pattern forms to deal with variation may have consequences for the productivity of patterns; adapting patterns to allow high levels of variation may introduce noise in the results of knowledge extraction, and conversely, not taking into account certain kinds of variation may reduce recall.

---

[57] This phenomenon and its impact was mentioned, although not formally evaluated, by Feliu (2004: 169), who chose in the extraction of contexts for the description of potential pattern forms to use verbs alone as markers, although she acknowledged the advantage of including additional elements such as prepositions in marker forms for more precisely identifying relation types and occurrences.

The significance of different kinds of variations in marker form and in pattern structure will be addressed below.

2.6.2.1 Variation in marker form

Variation in marker forms occurring in the corpus can correspond to a difference in the number of pattern forms (e.g., character strings or regular expressions) needed to extract pertinent occurrences of knowledge patterns. Differences in marker variation can be significant both in terms of the time and effort required to develop pattern forms, and in the possibilities for precision and recall offered by the patterns. As discussed, for example, in Bowker (2003), the form of markers of relations can vary in various ways. In addition to morphological variation (which will not be discussed here, although as noted in Marshman (2002, 2004) and Weilgaard (2004), it can present difficulties in cases when pre-established lists of inflected forms of markers are not available or not used), these include variation in marker elements and variation in the use of active and passive voices.

Markers may vary in the addition or change of auxiliary elements associated with a "principal," open-class marker; these additional elements are often, for example, function words (particularly prepositions or conjunctions) that combine with nouns or verbs (e.g., *result from*, *result in*, *suppression of*, *suppression of... by*, *association of ... with, correlation between... and*).

These supplementary elements may be very important to take into account in pattern forms, as they can not only clarify the structure of a given context and the directionality of asymmetric relations,[58] but may also constitute formal indicators of the elements that are involved in it and be good indicators of the completeness of a given context (e.g., in the case of *suppression of... by*, which reliably indicates that both

---

[58] One example of this is the verbal marker *result*. In the pattern forms [CAUSE] *results in* [EFFECT] and [EFFECT] *results from* [CAUSE], the directionality of the relation changes depending on the preposition used with the verb.

elements of interest are present in a context, while *suppression of* alone may be observed in contexts in which only one of the related elements is realized).

Variation in the voice of verbal markers may affect not only the markers' form, but also the structure of patterns (including the relative positions of the participants in asymmetric relations), as observed in Marshman (2002) and Bowker (2003: 159–160).

In scientific texts, it has been widely observed (e.g., Sager et al. 1980: 226; Ouellet 1984, 1985; Kocourek 1991: 70, 83–85) that mentions of the observing subject tend to be minimal and backgrounded, most likely in an effort to increase the appearance of objectivity. One symptom of this effort is the frequent use of the passive voice, which is common in English scientific texts. While conventional wisdom asserts that French is less tolerant of the passive than English, it has been observed (e.g., Kocourek 1991: 84) that the passive is nevertheless used, particularly in scientific and technical writing, in an effort to give such an impression of objectivity. Given this motivation, in cases in which the passive voice is used, often a context will not clearly indicate the agent of an action, creating challenges for knowledge extraction (in cases in which this information is pertinent). The relative frequencies of the use of the passive in English and French may thus have an impact on the usefulness of the contexts extracted and therefore the precision of the patterns, particularly in patterns that may involve the presence of linguistic elements denoting human participants in a situation.

An alternative also exists in French: the use of the impersonal pronoun *on*, which can fulfill the same function as the passive, while maintaining the backgrounding of the subject. This is generally far more common than the use of *on*'s English counterpart, *one*. This phenomenon involves less variation from standard pattern structures than the use of the passive (although it nevertheless constitutes a variation from standard patterns involving related elements in noun form). Such constructions are also less informative than a clear indication of the entity represented by the pronoun, but more so than a passive involving no indication at all of the agent, as they at least indicate that the entity involved is animate and human.

As different techniques may be preferred in the two languages, the effects on various types of applications may be expected to differ as well, and distinct types of difficulties and possibilities for dealing with them may also be observed. An observation of these variations is thus of interest in this type of research.

2.6.2.2 Variation in pattern structures

One of the most widespread challenges of pattern-based applications is the inherent variability of language and the resulting reality that pattern forms in use are rarely invariable. For example, the number and order of elements within a pattern structure may vary: different numbers of arguments of markers may be realized (e.g., as in *X has been correlated with Y* versus *Z has correlated X with Y*), the elements of patterns may appear in different orders (e.g., X *plays a role in* Y versus *the role played by* X *in* Y), and additional elements may appear within pattern structures (e.g., a copula may be present in X *is associated with* Y but not in X *associated with* Y).

All of these variants would need to be represented in one way or another in pattern forms, in order to achieve maximum recall, and this would often require the use of either multiple or complex pattern forms (e.g., that include indications of optional elements within marker forms, or allow for the possibility of marker elements appearing in varying places relative to one another). Given its pertinence for developing pattern-based tools, the level of variation in pattern structures was evaluated and compared in the research.

2.6.3 Evaluation of the presence of and relationships between multiple elements sharing a role in a relation

Although the CAUSE–EFFECT and ASSOCIATION relations are considered to be binary, corresponding to links between two concepts, in many cases in texts more than one instantiation of a given role in a relation (i.e., more than one item filling a single "slot" in a knowledge pattern) may be observed. At a formal level, this type of variation must

be taken into account in planning for semi-automatic knowledge extraction, because pattern forms must represent the possible structures in which complex elements participating in relation may appear, as was the case, for example, in Hearst (1992).

The analysis of the frequency with which multiple elements may share roles in a relation allows for the evaluation of the proportion of relation occurrences that may require adaptations of pattern forms to accommodate this phenomenon, in applications that attempt to identify related elements automatically or that impose restrictions on the form and/or placement of related elements.

The analysis of the structures in which multiple elements occur provides an indication of the complexity of defining such structures (e.g., the numbers of different markers that can link the various elements), as well as the impact this variation may have for the development of pattern forms that can recognize different forms and ensure that all elements linked in a given relation are included in an extracted context (and/or are identified in further automatic processing of those contexts, as required).

In addition, the presence of multiple elements may also be associated with additional phenomena, including those of ellipsis of part of multiple, complex elements that share either a head or an expansion, and the repetition of part of complex markers in connection with multiple related elements.

The difficulties of the former phenomenon have been noted by researchers including Lauriston (1994: 164) and Ahmad and Rogers (1997: 753) (cf. also Daille 2005). It may, for example, affect the possibilities for automatic identification of related elements and the immediate usefulness of a given context for applications such as linking term records and nodes in ontologies. For example, tools that attempt to identify related elements, or that target specific items for research, may not be able to identify these (correctly and/or completely) if their forms are modified by ellipsis. The search for specific terms adjacent to pattern markers will encounter problems if the two items are separated by an elliptical form of another element, or if the form of the term itself has

been interrupted. Pattern forms that specify POS classes of related elements (generally restricting these to nouns and noun phrases) may also not correspond to forms observed in texts if ellipsis involves the absence of a noun directly adjacent to a pattern marker.

Moreover, the ellipsis of different portions of related elements (i.e., of heads or expansions) clearly poses different types of difficulties for automatic applications. Both the form of the element(s) and the value of the individual elements for information extraction may vary depending on whether the head or expansion is omitted in a given context. For example, expansions may more often take non-nominal form, potentially interfering with the recognition of KRCs or of related elements by pattern forms that search for elements in noun form, and in many cases, even if an expansion is identified as a related element, it may not be productive for knowledge extraction without the head to which it should be attached.[59] Analysis of a considerable amount of data would be necessary to develop formal representations of this phenomenon that can automatically and reliably process the various types of ellipsis observed; an alternative approach might involve the design of pattern forms that allow for the extraction of the entire related element structure for human analysis. Regardless, the phenomenon must be accounted for in developing pattern forms and choosing strategies for pattern-based extraction.

Repetition of part of a complex marker may affect the requirements for pattern design in such cases, as pattern forms that do not allow for this repetition may encounter problems in identifying contexts and/or their components (as the repetition of the marker would thus constitute an interruption of the pattern form).

In addition, the involvement of multiple elements may have an impact on the interpretation of a relation at a conceptual level (e.g., in conclusions that may be drawn about the type of relation present and the necessity or sufficiency of a given element's

---

[59] Conversely, in some cases the expansion of a complex item may contain the essential information being expressed. It may be extremely difficult to distinguish these cases formally.

involvement in a relation). The relation classifications used by Nuopponen (1994) and Garcia (1996, 1997) reflect the distinctions between types of CAUSE–EFFECT relations that may be identified in cases in which multiple causes and effects participate in a relationship. Barrière and Hermet (2002) also described the impact of this phenomenon on the creation of conceptual graphs representing CAUSE–EFFECT relations.

From this perspective, the semantic analysis of the links between multiple related elements can provide information that is useful for assisting in sorting and/or using the contexts extracted. The nature of the connection between multiple elements may indicate that a given context describes a relation that holds between two or more separate element pairs (e.g., in the case of conjunction of related elements), one that may hold between only one of two possible pairs included in a context, or one that may hold between a given pair of elements in only some cases (e.g., in the case of disjunction of elements). Alternatively, some types of connections may indicate that a relation holds between a single pair of concepts that may be denoted by different linguistic units, thus providing not only data on the principal relation observed but also information about additional relations (e.g., in the case of variants and abbreviations). Finally, some connections between related elements may not only indicate that an additional relation is present in a given context, but also that the scope of the relation indicated by the marker may be larger than that of a pair of contexts mentioned explicitly in the context, extending to additional pairs through inheritance (e.g., in the case of hierarchical relations between related elements).

Because of their importance for the interpretation of conceptual relations and the development of pattern forms, these factors were evaluated in both of the languages, and then compared in order to estimate their relative impacts in English and French.

2.6.4 Identification and evaluation of types of anaphoric expressions

Various authors (cf. Section 2.3.1.5.2.1) have observed the challenges inherent in the presence of anaphora within potential KRCs. However, the impact of anaphora in semi-

automatic extraction of KRCs is likely to vary according to the type of expression found in context. In particular the replacement of a given element participating in a relation by a pronoun or generic term may be challenging to deal with, although in different ways.

The replacement of a related element by an anaphoric expression obviously reduces the usefulness of the context in which it occurs, particularly when the antecedent of that expression occurs at some distance from the occurrence identified using the pattern: the inability to identify (with an acceptable level of specificity) one or more of the related elements in a context may make this context only marginally useful or even useless for some applications. In addition, if such occurrences — particularly involving replacement using pronouns — are to be considered for extraction, patterns must be designed to permit this possibility (allowing, for example, for a related element to take the form of a pronoun in addition to a noun or noun phrase).

For applications that attempt to process contexts according to the semantic classes of the actants involved, anaphoric expressions — and particularly those in pronoun or possessive adjective form — are also problematic, as they may make contexts very difficult to classify automatically by the class of the antecedent.[60]

In contrast, the use of a generic term in the place of a more specific one may pose fewer problems at a formal level, and may also be helpful for the identification of semantic classes. However, the less obvious nature of the anaphora in these cases may pose problems in the potential for identifying the anaphoric nature of the expression that is present and the need to identify an antecedent to determine with precision and specificity what concept is involved in the relationship The analysis of the forms in which this phenomenon occurs may help to suggest strategies for ensuring that accurate information may be identified.

---

[60] However, Weilgaard (2004) describes another perspective on this phenomenon.

Because of these factors, the types of anaphoric expressions found in the research were analyzed to evaluate the ways in which pattern-based tools and the usefulness of the contexts they extract may be affected in each language.

2.6.5 Text-related issues

In this research, occurrences of potential difficulties related to the form of individual contexts were also annotated, in order to estimate the prevalence of these phenomena. By its very nature, all corpus-based work is closely dependent on the texts that form the corpus, and thus indirectly on the authors of these texts. Knowledge extraction applications rely heavily on the presupposition that corpus texts will present correct information in ways that are relatively clear, easily interpretable and unambiguous. In carrying out this kind of research, it is necessary to make the fundamental assumption that when authors use the relation markers identified, they are generally doing so with at least a minimal respect for the generally accepted meanings of these markers and for the interpretation that a reader is likely to make of their use. Clearly, this is far from guaranteed, however, and pattern-based applications can be affected by various types of problems in corpus texts; this is one of the major vulnerabilities of such an approach.

In addition to the fundamental issues of choices in expression made by authors, in some cases, issues in the original text (or possibly introduced in a text conversion process necessary for preparing texts for analysis using tools) may pose problems for semi-automatic and automatic approaches to knowledge extraction, interfering with correspondences between patterns and contexts. These problems may range from unusual structures that are not accounted for in pattern forms to stylistic variations that involve changes in these pattern forms, to actual mistakes in the text (typographical errors, incorrect punctuation, misspellings, etc.).

In addition to the impact these phenomena may have on pattern identification (affecting the recognition of the marker, of the pattern structure, or of the related elements) and the interpretation of individual contexts (including the identification of

the relation present and of how various elements are involved in it), at a more general level they may also be indicative of the authors' skill and care in the expression of the content; problems in writing at this level may be indicative of problems on other levels, which may call into question the reliability and ultimate reusability of information expressed.

However, it is our opinion that this kind of analysis is best dealt with by the end-user of the information, taking into account the intended application. Moreover, the very nature of the phenomenon entails a high degree of variability, and the construction of a comprehensive and generally applicable typology of such problems would be almost unthinkably complex. Certainly such a task is beyond the scope of this research.

Therefore, an evaluation of the frequency of such text-related issues was carried out solely in order to determine whether similar proportions of contexts in English and French were affected. This evaluation focused solely on the prevalence of phenomena that may interfere with the identification and interpretation of candidate KRCs.

The methodology developed for carrying out the evaluations described above will be presented in Chapter 3.

3 Methodology

This chapter outlines the methodology used in the research: Section 3.1 describes the corpus-building process, Section 3.2 the generation of the initial, term-based series of concordances, Section 3.3 the manual analysis of patterns and difficulties, and Section 3.4 the interlinguistic comparison of the results.

3.1 Corpus-building

Two corpora were constructed, one in English and one in French. A full list of the corpus texts appears in Appendix C.[61] The issues of languages and language varieties, domains, corpus size, the dates of texts and the text types selected are described below.

3.1.1 Languages and language varieties

In order to maintain a relatively broad linguistic representation, efforts were made to include texts from a variety of sources and geographical areas. This geographical variety was identified using the affiliations of the authors of the corpus texts, as well as the place of publication of the sources (cf. Vandaele 2001). English sources from the United States, the United Kingdom, Canada, Ireland and Australia were included. French sources originated in France, Quebec, Belgium and Switzerland.

Author affiliation was also used as an indication of the language communities to which authors likely belonged. Texts were retained if at least one author was affiliated with an institution in a Francophone or Anglophone country (depending on the corpus).

---

[61] We will not describe in detail the theoretical framework surrounding corpus-building in terminology (for a detailed review of the literature and some issues in corpus-building, see e.g., Meyer and Mackintosh (1996), Pearson (1998) and Bodson (2005: 13–33)).

3.1.2 The domain and sub-domains

Within the medical domain chosen as the subject of the corpora for this project, the sub-domains selected were those of breast cancer and heart disease (with a specific focus on atherosclerosis). Texts were located using terms that were identified as representing 1) concepts that denoted central concepts in these fields (e.g., the diseases themselves), or 2) other concepts of interest in current research in the domain, identified through background reading on the fields. In English these terms included *breast cancer*, *breast neoplasm*, *p53*, *BRCA1* and *BRCA2* for the breast cancer corpus, and *atherosclerosis*, *arteriosclerosis*, *cholesterol*, and *cardiovascular disease* for the heart disease sub-corpus. In French, the keywords used included *cancer du sein*, *BRCA1*, *BRCA2* and *p53* for the cancer sub-corpus, and *athérosclérose*, *artériosclérose*, *cholestérol* and *maladies cardiovasculaires* for the heart disease sub-corpus.[62] The corpora in each sub-domain included articles discussing the etiology of the disease, its development, its diagnosis, its treatment, its effects, and its prevention.

3.1.3 Corpus size

The English corpus contains approximately 573,000 tokens in total (approximately 305,000 in the breast cancer sub-corpus and 268,000 in the heart disease sub-corpus). The French corpus contains 692,000 tokens (478,000 in the breast cancer corpus and 214,000 in the heart disease corpus). To minimize variation in results of corpus analysis due to differences in numbers of words and permit comparison of frequencies of occurrence of various units in the two corpora, where applicable measures of frequency were expressed and compared in occurrences per thousand corpus tokens.

---

[62] Where this option was available (e.g., in MEDLINE), articles classified according to subject-headings corresponding to these terms were used; in other cases, these terms occurred as keywords, or in journal or article titles or abstracts.

3.1.4 Dates of corpus texts

The corpus documents were published between 1997 and 2004.[63]

3.1.5 Text types

The corpora for the research were built using texts available in electronic form from various databases including MEDLINE and Repère,[64] as well as online sources such as the *Canadian Medical Association Journal*, *Le Médecin du Québec* and *Le Clinicien*.[65] The texts were primarily specialized texts — including review and research articles, with a few additional types, including continuing education articles, case reviews and clinical cases — although a small proportion of popularized texts (i.e., specialized journalism) were included in each corpus.[66]

The classification of the corpus texts by level of specialization was based on criteria similar to those used by Pearson (1998) in defining communicative situations; these focused particularly on the level of expertise of the author and audience. Articles were considered to be specialized if their authors were experts in the medical domain (e.g., physicians, specialists and/or researchers) and to be popularized if their authors were knowledgeable in the field but not experts (e.g., medical journalists). The intended audience also contributed to the classification: texts intended for the general public or an interested general public (e.g., in the case of specialized science and medical journalism) were considered to be part of the popularized sub-corpus, while texts intended for

---

[63] Most dated from between 2000 and 2004. The English breast cancer corpus also contained approximately 15 older documents, published between 1988 and 1997.

[64] Cf. Vandaele 2001 for an overview of the usage of such databases in the medical field.

[65] The exceptions are texts from *La Recherche*, which were scanned and processed using optical character recognition software.

[66] These popularized texts complement more specialized texts and moreover are often extremely useful for terminologists and terminographers, especially those who are beginning work in a new domain and acquiring domain knowledge, or who are preparing resources for users who are not themselves subject-field specialists. For this reason, it is important for pattern-based tools — even those intended for use in specialized subject areas — to be able to process these documents as well as more specialized texts.

experts (e.g., physicians, specialists and researchers) were considered to be specialized.[67]

3.2 Initial concordances

Initial concordances were generated using candidate terms and manually analyzed to identify occurrences of the relations of interest and the candidate patterns that indicated them. This reflects approaches used by many researchers in the field of markers of semantic and conceptual relations, including Hearst (1992), Meyer et al. (1999), Meyer (2001), Condamines and Rebeyrolle (2001), Marshman (2002, 2002a) and Bodson (2005).

In terminology, this widely used technique generally aims to identify markers that may occur in association with one or more terms that terminologists and/or terminographers might seek to include in a resource, and thus that may be of assistance in the analysis and description (e.g., definition) of the concepts denoted by these items.

The methodology chosen thus reflects not only the practice in the domain, but also the principal needs of typical end-users of a pattern-based tool — in this case terminologists and terminographers carrying out conceptual analysis for terminological description — while still allowing for the evaluation of a wide range of data that may be helpful to these users, and the challenges in exploiting it. Moreover, it allows for the observation of some situations and challenges such an approach may encounter.

---

[67] It is possible to subdivide these sections even more specifically if desired, by evaluating the relative levels of knowledge of the author and receptor of texts. For example, a text on a very specialized subject written by a specialist in that area for readers who have background knowledge in the field but not the same level of specialization (e.g., general practitioners) may show some characteristics that texts written by specialists for other specialists in the same field may not. However, this fine-grained level of classification was not used in this study.

3.2.1 Choice of the terms for initial concordances

The choice of terms used to generate the initial concordances raises a certain number of questions: What kinds of terms would end-users be likely to research using a semi-automatic tool? How can appropriate terms be identified? What effects may term choice have on the observations?

In regard to the first issue, more and more applications for the automatic extraction of candidate terms are being developed, and interest is growing in these tools. Several different techniques have been proposed, and can be grouped into three general categories: 1) approaches based on statistics (repetition of forms; cf. Gillam and Ahmad (2002) and Ahmad and Davies (1994) on the concept of "weirdness"); 2) approaches based on common patterns of term formation (e.g., common forms of noun phrases); and 3) hybrid approaches integrating both statistical and term formation pattern approaches. This is a technique that is being more and more frequently used in the field of term extraction, and has been recognized to provide good results. (See also Drouin (2002: 53–114) for an overview of various term extraction methods.) Promising results have been observed for all of these approaches, and they seem to present good starting points for corpus analysis, providing terminologists with a list of candidate terms that may be important in a given corpus, and by extension in a given domain.

Moreover, automatically identified candidate terms have been used as a starting point in several research projects focusing on pattern discovery and application (e.g., Pearson 1998; Condamines and Rebeyrolle 2001; Gillam et al. 2005).

The use of automatic term extraction software to identify candidate terms for the initial concordances thus offers both a starting point for the research that does not require the involvement of a terminologist to analyze the corpus manually (which is impractical in this context), and a relatively realistic picture of the kinds of terms a terminologist might begin to research in a real-world project.

In regard to the last question, Bodson (2005), working in French, observed in her corpora that different types of semantic relations were strongly associated with specific semantic classes of terms. Moreover, although the associations were far less pronounced, she did consider the possibility that some patterns could be linked to a particular class of terms (i.e., that the patterns indicating a specific semantic relation between entities might not be the same as those that express the same relation between activities, and so on). Given these observations, it was considered important to use terms corresponding to a variety of classes in order to observe a range of results.

Finally, for the purposes of this project, focusing on an interlinguistic comparison of results, it was important that a certain parallelism be maintained between the terms chosen in the English and French corpora. Comparable numbers of terms in each corpus, as well as of terms associated with each semantic class, were required.

Three steps were thus identified for the selection of terms for the initial concordances: 1) the identification of candidate terms in the corpus using a term-extraction tool; 2) the identification of terms representing different semantic classes and each of the sub-domains, in order to ensure variety; and finally 3) the selection among a set of these candidates of terms that had similar characteristics in the two languages, in order to maintain a level of interlinguistic parallelism.

3.2.1.1 First criterion: Specificity

The terms used were identified using the term extractor TermoStat (Drouin 2002, 2003) (see Appendix D for a sample of the candidate terms proposed). This tool uses a hybrid (i.e., linguistic plus statistical) approach: it identifies simple and complex terms —

generally, nouns and noun phrases — in a corpus using a statistical measure of their specificity, i.e., frequency in the corpus as compared to a reference corpus.[68]

TermoStat offers advantages over other tools that use statistical or linguistic approaches alone: it proposes an integrated list of both simple and complex candidate terms; it carries out its calculations on data that has been part-of-speech tagged and lemmatized using TreeTagger (Schmid 1994), to provide more accurate indications of frequency in the corpora and thus of relative specificity; and it can use typical patterns of term formation to reduce some noise in the results.

The threshold set for considering lexical units as candidate terms was a score of +3.09 (Drouin 2003: 148). TermoStat identified a total of 14,536 candidate terms in the English corpus and 14,058 candidate terms in the French corpus that met this criterion.

A first, superficial sort of the candidate terms proposed was carried out to eliminate those that were not considered to be appropriate for the task of generating the first set of concordances to search for patterns. Exclusions were made in the case of:

- candidate terms that were judged not to be domain terms (e.g., *fois* and *mois* in the French heart disease corpus, *fig*, the abbreviation for *figure* in the English breast cancer corpus);
- candidate terms that were judged unlikely to occur in relations of interest in this research (e.g., *mg*, the abbreviation for *milligram*, in the English heart disease corpus);
- candidate terms that very often occurred as modifiers in larger complex terms (e.g., *LDL* in the English heart disease corpus, which occurred often in compounds such as *LDL-C*, the abbreviation for *low-density lipoprotein cholesterol*), or which occurred in other expressions (e.g., *cours* in the French heart disease corpus, which often occurred in the expressions *en cours*, *au cours de*, and *au long cours*);
- candidate terms that were considered likely to occur as relation markers (e.g., *risk* in the English corpus, *risque* and *augmentation* in the French corpus); and

---

[68] In this case, the English reference corpus was taken from the Montreal newspaper *The Gazette* (articles published between March and May 1989) and the French from *Le Monde* (articles published in 2002). More details on the English corpus are available in Drouin (2002: 122).

- candidate terms that would provide too much overlap with other terms of interest (e.g., *cellule endothéliale*, excluded from the candidates from the French heart disease corpus because of the overlap with *cellule*).

3.2.1.2 Second criterion: Representation

3.2.1.2.1 Semantic classes

In order to ensure that a variety of term classes were used in generating the initial concordances and later to permit interlinguistic balancing of the semantic classes represented, the roughly sorted candidate terms identified using TermoStat were associated with classes including entities, activities and processes.

Two classification systems, WordNet and the Unified Medical Language System, were consulted in order to identify the high-level semantic classes with which candidate terms might be associated.[69] The two resources use a different approach to classification, and each presents a certain number of advantages and disadvantages for the purposes of this kind of research.

WordNet is an English-only system based on the classification of sets of synonyms and quasi-synonyms (*synsets*) according to their semantic characteristics.[70] It covers a large range of general language words, and also some specialized terms from various domains (e.g., medicine, computing; cf. Bodson 2005). However, the entries are largely single words, and few complex items are included.

The UMLS Semantic Network is specialized for the field of medicine, and classifies concepts according to a hierarchy based on the harmonization of a number of medical ontologies, terminologies, thesauri and other resources. This classification can be accessed using a large number of complex (and some simple) terms through the

---

[69] Only the higher levels of the hierarchies were used, in order to remain as neutral as possible to the inter-system variations in classifications, as there is far more variation at the lower levels of the classification systems than in the upper ones.

[70] While WordNet is an English-language resource, many projects are underway to expand the classification system used to other languages.

Metathesaurus, which offers far more complete coverage of terms in the medical domain than that of WordNet and links these terms to the concepts of the Semantic Network. A multilingual resource, the UMLS Metathesaurus offers searching in many languages, and synthesizes results in a single entry. It is freely available for research purposes.

The UMLS was chosen as the primary basis for classification in this research, and its classes of entities (including physical objects and conceptual entities) and events (including activities and phenomena/processes) assigned to the candidate terms.[71,72]

A choice was made, however, to exclude from the list of candidate terms some that were especially difficult to classify (e.g., *data*, *study*). In the case of processes, because the classification of these types of terms was challenging (largely due to the fact that the UMLS identifies a class of phenomena or processes rather than processes alone), terms that were described as denoting processes in both the UMLS and WordNet were chosen, in an effort to ensure as precise a classification as possible.[73]

One problem encountered was difficult to overcome, and yet extremely important in the corpora being analyzed. This was the classification of pathologies (i.e., diseases and other disorders), which were frequently represented in the list of candidate terms. These may be considered to have multiple inheritance, possessing semantic properties characteristic of both states and processes. This fluidity in the classification

---

[71] In the rare cases that candidate terms (particularly in French, e.g., *traitement*, *cellule*) were not present as independent forms in the UMLS, their classification was deduced from the classification of more specific concepts. In one case, *oxydation*, the French term was not present, but its equivalence with the English term *oxidation* was established and the class corresponding to the English term was used.

[72] However, it is clear that neither of the systems was perfectly suited to the purposes of this project: the classifications used by these systems for particular terms were often not immediately clear or intuitively understood, and were certainly not always coherent between systems.

[73] In the case of the French terms, English equivalents were used to consult WordNet. As it was relatively difficult to find terms that were classified in both systems as processes, the specificity of these terms tends to be lower than that of the other candidate terms used. However, their TermoStat specificity scores still far exceeded the threshold set for validity.

of disease terms was easily observed when comparing the classifications given in WordNet and in the UMLS. WordNet classified all of these terms as states, while the UMLS attached these terms to concepts characterized as phenomena or processes (specifically, diseases or syndromes, or pathologic functions).[74] Given both the centrality of terms denoting disease in the medical field, and of the high number of such terms among the candidates proposed by TermoStat, it was not considered advisable to exclude these terms because of the challenges of their classification. However, neither were they considered to be reliable examples of terms denoting phenomena or processes. As such, they were included as a special category of phenomena or processes (pathologies). (Further discussion of this decision may be found in Section 5.5.2.)[75]

3.2.1.2.2 Sub-domains

In order to ensure that candidate terms representative of each sub-corpus were chosen in the two languages, the specificities of the terms selected were also evaluated in the two sub-corpora, and these specificities were also considered in the selection of terms from among the candidates identified using the previous criteria.

3.2.1.3 Third criterion: Interlinguistic similarity

A number of pairs of candidate terms evaluated using the above criteria were identified as equivalents in the two languages, and were given preference in the final selection.

---

[74] One exception was the French term *récidive*, which was not classified at a more specific level than that of Phenomenon or process, but which on the basis of this classification and its definition was considered to be most acceptably included in this category.

[75] Somewhat less problematic, but nevertheless present, were challenges linked to the classification of terms associated with treatments (*chemotherapy*, *hormone replacement therapy*, *chimiothérapie*, *traitement*) which may present some variation in usage, denoting either an activity (as reflected in the classification used) or, for example, the agent used in this activity. Once again, as the focus in the use of these classifications was to ensure variety in the types of terms chosen and parallelism in the two languages, and the potential for variation was observed in the two corpora, this variation was not considered to preclude the use of these kinds of central domain terms in the research.

3.2.2 Selected terms

These criteria were used to select 15 terms in each language (Table 10).

Table 10. Terms for initial concordances

Term	UMLS Semantic Type	Frequency in the corpus
English		
chemotherapy	Event/activity	540
hormone replacement therapy	Event/activity	516
patient	Entity/conceptual entity[76]	3992
cell	Entity/physical object	2143
C-reactive protein	Entity/physical object	562
atherosclerosis	Event/phenomenon or process	410
breast cancer	Event/phenomenon or process	2533
diabetes	Event/phenomenon or process	425
coronary heart disease	Event/phenomenon or process	373
tumour	Event/phenomenon or process	1325
activation	Event/phenomenon or process	266
development	Event/phenomenon or process	367
expression	Event/phenomenon or process	592
oxidation	Event/phenomenon or process	84
pathogenesis	Event/phenomenon or process	61
French		
chimiothérapie	Event/activity	738
traitement	Event/activity	2357
patient	Entity/conceptual entity	3504
cellule	Entity/physical object	1678
cholestérol	Entity/physical object	359
athérosclérose	Event/phenomenon or process	392
cancer du sein	Event/phenomenon or process	2092
diabète	Event/phenomenon or process	233
récidive	Event/phenomenon or process	272
tumeur	Event/phenomenon or process	1481
activation	Event/phenomenon or process	237
coagulation	Event/phenomenon or process	41
oxydation	Event/phenomenon or process	54
prolifération	Event/phenomenon or process	138
transcription	Event/phenomenon or process	101

Two terms classified as denoting activities and three as denoting entities were chosen. Given the prevalence and importance of terms denoting pathologies (e.g.,

[76] The terms *patient* in English and in French are linked to a concept that is further classified as a Group, explaining the somewhat surprising characterization of the term as denoting a conceptual entity.

diseases), these were more numerous in the terms chosen, with five representatives of the class in each language. Processes were also considered not only to be of central importance in the understanding of the field and particularly the aspects targeted in the corpus-building process (e.g., disease development, effects, treatments), but also to be likely to participate in the relations of interest in the research, and thus were also numerous in the list of candidates retained. Five terms representing this class were retained. Nine pairs of these terms were considered to be equivalents. Six additional non-equivalent terms in each language that provided relatively comparable distributions between classes and sub-corpora as well as numbers of occurrences for analysis were retained to complement these pairs. (More data about the terms are available in Appendix E, which presents the terms chosen in each language, their semantic classes, their frequencies, and their specificity in the corpora. In addition, Appendix F presents definitions of each of the candidates selected.)

3.2.3 Generation of initial concordances

Once the lists were finalized, the concordancer WordSmith Tools was used to generate concordances for each of the selected terms. A random sample of up to approximately 100 occurrences of each term — for a total of approximately 1,400 contexts in each language — was then added to a Microsoft Access database, to be analyzed and annotated if a relation was present.[77]

3.2.3.1 English terms

The 15 selected terms, shown in Table 11, were used to generate a sample of 1,412 concordance lines. The second column in this table, identifying the terms' class as

[77] Sampling was carried out using a function of WordSmith Tools that allows for the random selection of a certain proportion of occurrences of a search string located using the tool, up to a desired number of occurrences. Some variation in numbers of occurrences retained resulted from the need to ensure that the maximum was not consistently attained before the end of the corpus was reached, which could have introduced bias in the results. In cases in which an approach using automatic random selection produced an excessively large number of occurrences, a manual random selection was made within these.

retained for the purposes of this project, reflects the choices made (cf. Section 3.2.1.2.1) to consider terms denoting diseases and disorders (in addition to tumours) as a separate class of terms, and to choose terms classified as denoting processes in both the UMLS and WordNet.

Table 11. English terms used to generate the initial concordances

English candidate term	Class	Number of contexts
chemotherapy	Activity	100
hormone replacement therapy	Activity	101
patient	Entity (conceptual entity)	100
cell	Entity (physical object)	106
c-reactive protein	Entity (physical object)	101
activation	Process	107
development	Process	99
expression	Process	100
oxidation	Process	84
pathogenesis	Process	61
atherosclerosis	Phenomenon or process (pathology)	85
breast cancer	Phenomenon or process (pathology)	99
diabetes	Phenomenon or process (pathology)	92
coronary heart disease	Phenomenon or process (pathology)	77
tumour	Phenomenon or process (pathology)	100
Total		**1412**

3.2.3.2 French terms

The 15 French terms, shown in Table 12, were used to generate a sample of 1,392 concordance lines. The distribution of contexts by term class is shown in Table 13. These contexts were then analyzed manually and annotated as described in Section 3.3.

Table 12. French terms used to generate the initial concordances

French candidate term	Class	Number of contexts
chimiothérapie	Activity	100
traitement	Activity	100
patient	Entity (conceptual entity)	100
cellule	Entity (physical object)	100

cholestérol	Entity (physical object)	100
activation	Process	100
transcription	Process	101
coagulation	Process	41
oxydation	Process	54
prolifération	Process	101
athérosclérose	Phenomenon or process (pathology)	100
cancer du sein	Phenomenon or process (pathology)	96
diabète	Phenomenon or process (pathology)	100
récidive	Phenomenon or process (pathology)	100
tumeur	Phenomenon or process (pathology)	99
Total		**1392**

Table 13. Contexts analyzed by term class

Class	Number of contexts — English	Number of contexts — French
Activity	201	200
Entity	307	300
Process	451	397
Phenomenon or process (pathology)	453	495
	1412	**1392**

3.3 Manual identification of relation occurrences and candidate patterns

The analysis of the term-based concordances began with the identification of contexts of interest for identifying and evaluating relation occurrences and the candidate patterns indicating them. Contexts were retained and annotated if:

- the context expressed a relation of interest in this project;
- both of the elements participating in the relation were expressed in the context;[78]
- one of the elements linked in this relation corresponded to the term used to generate the concordance, or to a complex term with this term as its head;
- a candidate lexical knowledge pattern indicating the relation (i.e., a knowledge pattern containing a lexical marker — a verb (e.g., *induce*), noun (e.g.,

[78] If a complex related element was incomplete in the context extracted, the context was retained if the head of the element was at least partially present in the context. In cases in which multiple elements shared a role in a relation and at least one of these occurred within the extracted context, the occurrence was also retained.

prevention), adjective (e.g., *dependent*), participial adjective (e.g., *associated*), conjunction (e.g., *and*), or preposition (e.g., *with*), a prefix (e.g., *anti-*), or a combination of such units) could be identified in the context.

3.3.1 Annotation

For each context retained for annotation according to the criteria described above, a certain number of elements were noted in separate database fields: the relation (and sub-relation in the case of the CAUSE–EFFECT relation) and the base form of the pattern marker observed, as well as the related elements (as observed in the context).[79] The form of the pattern as observed in the context was also noted, with the related elements and other intervening items replaced by placeholders (related elements by variables and intervening items by parts of speech).[80] A sample of an analyzed concordance is shown in Table 14 (page 175).

If two or more distinct relations meeting the criteria were present, a separate entry was created for each relation and the elements linked by it. Several characteristics of the markers identified and their contexts were evaluated, analyzed and compared. Characteristics of the pattern marker and the related elements (e.g., part of speech classes) and challenges for pattern-based tool use were noted in fields designed for this purpose (not shown in Table 14). More details of each aspect of the annotation are provided below in Sections 3.3.1.2 to 3.3.1.5. However, before entering into this discussion, a note on one aspect of the annotation is presented in Section 3.3.1.1.

---

[79] Some markers were ambiguous in the sub-relation they represented. Occurrences of these markers were classified according to the sub-relation that best represented the relationship as interpreted in the specific context being annotated, and thus markers could be associated with different sub-relations in different contexts. In addition, a small number of markers were difficult to interpret within a single context; these were classified according to the best interpretation possible of the context. A discussion of this phenomenon and some strategies for dealing with it may be found in Section 5.5.3.2.

[80] As described in Section 3.3.1.5.1.1, an exception was made for cases in which multiple markers were observed in a given context.

3.3.1.1 Special case in the annotation

Although terms (e.g., *risk* in English and *risque* and *augmentation* in French) that were initially suspected to be potential markers of relations were excluded, in some cases — of *activation* in both English and French and of *expression* in English — it became evident that a term identified as specific to the corpus and used to generate one of the initial concordances could also play the part of a lexical marker in a knowledge pattern of interest in this project (e.g., the expression of a molecule by a cell involves a CREATION relation). This issue was dealt with by excluding the term used to generate concordances from consideration as a pattern in those concordance lines only. This was chosen as the best solution, since: 1) these candidate patterns could also be found in other contexts and thus were not *a priori* excluded from study; 2) this method prevented the introduction of a serious bias in the observations of candidate pattern frequency at this stage and gave these patterns the same likelihood of being observed as others that were not used to generate term-based concordances; 3) the approach allowed for the study of candidate terms that may be particularly interesting foci for evaluation and description using information such as that offered by pattern-based tools. (Some further discussion of this decision is found in Section 5.5.2.) It must nevertheless be acknowledged that this may lead to some inconsistencies in annotation, since the same context, if identified using another term, would likely have been annotated differently.

3.3.1.2 Relations

For each context, the relation present — as described in Section 1.5 — was determined. The criteria for relation classifications are described below in Sections 3.3.1.2.1 and 3.3.1.2.2.

Table 14. Sample of annotated relation occurrences

Initial term	Context	1st element (X)	Marker	2nd element (Y)	Relation	Pattern Form
atherosclerosis	… surgery--is performed 500,000 times a year in the U.S. to treat coronary arteries that are becoming blocked as a result of atherosclerosis. … (Beardsley 2000)	coronary arteries that are becoming blocked	as a result of	atherosclerosis	EFFECT–CAUSE (creation)	X as a result of Y
atherosclerosis	… The "response to injury" hypothesis developed by Russell Ross in the late 1970s suggested that atherosclerosis, at least, resulted from an initial injury to endothelial cells … (Griendling and FitzGerald 2003a)	atherosclerosis	result from	injury to endothelial cells	EFFECT–CAUSE (creation)	X, [adverb phrase], resulted from [article] [adjective] Y
atherosclerosis	… MMPs have been broadly implicated in a number of cardiovascular diseases, including atherosclerosis, 90,94 aortic aneurysms, 95 and heart failure, 96 … (Jaffer and Weissleder 2004)	MMPs	implicate	cardiovascular diseases, atherosclerosis, aortic aneurysms, heart failure	CAUSE–EFFECT (creation)	X have been [adverb] implicated in [quantifier] Y1, including Y2 [ref], Y3, [ref] [conjunction] Y4
atherosclerosis	... Recently, an exciting report provided evidence for a new pathway by which hepatic lipase may modulate atherosclerosis. … (Zambon et al. 2003)	hepatic lipase	modulate	atherosclerosis	CAUSE–EFFECT (modification)	X [modal] modulate Y

3.3.1.2.1 Criteria for classification of ASSOCIATION relations

As described in Section 1.5.1, within the larger framework of the ASSOCIATION relation, some specific sub-types may be identified. While it would be possible — and in some cases beneficial to a user — to separate these sub-types of ASSOCIATION (according to the criteria of variable type or symmetry, for example), for the purposes of this research only the main relation of ASSOCIATION was used. This decision was made because of the relatively low number of occurrences of this relation as compared to the CAUSE-EFFECT relation, of the nature of the relation itself and of its primary function for the purposes of this kind of research (i.e., indicating the advisability of a surveillance of connections between two elements in order to determine whether a causal link is present, and if so what its nature may be), and of the difficulty in many cases of differentiating between the different potential sub-types of this relation. (However, a brief discussion of the possibilities of refining the classification, in light of the results of the project, appears in Section 5.5.3.4.)

3.3.1.2.2 Criteria for classification of CAUSE–EFFECT relations

In the first stage of the analysis of CAUSE–EFFECT relations, the relation occurrences were annotated using Barrière's classification (2002; cf. Section 1.5.2.8.3) to assign a sub-relation (CREATION, DESTRUCTION, MAINTENANCE, PREVENTION, MODIFICATION, INCREASE, DECREASE or PRESERVATION) to each occurrence.

One minor modification in the sub-relation names assigned by Barrière was made: given that the inclusion in the MAINTENANCE sub-relation of cases in which one element allows or permits the other to exist or to occur may be somewhat confusing, this sub-relation was identified as MAINTENANCE (PERMISSION) to improve clarity.

3.3.1.3 Number of relation occurrences observed

Once the relation (and, where applicable, sub-relation) present in each context was identified, the number of occurrences of each type was tallied, in order to provide an idea of the density of relation occurrences meeting the criteria set out in the project, as compared to the number of contexts analyzed. This measure is indicative of the raw potential of this kind of approach for locating relation occurrences in the two languages.

Once the number of relation occurrences present was determined, the analysis of these occurrences themselves, and of the markers and patterns identified in them, could proceed. This evaluation focused on two distinct aspects of the relation occurrences identified: the characteristics of the patterns themselves (including the markers identified, their characteristics, the structures in which they participated, and the elements that they linked); and some challenges in the identification and use of these patterns or the information conveyed by the contexts in which they occur due to items external to the patterns themselves (including interruptions of the patterns and the presence of expressions of uncertainty in the contexts evaluated).

3.3.1.4 Annotation and analysis of pattern characteristics

The various pattern characteristics that were annotated and evaluated are described in more detail below.

3.3.1.4.1 Candidate markers

The first step in analyzing the term-based concordances was the identification of candidate markers, which required a preliminary definition of what kinds of markers were of interest in the project, and what information was needed about them.

The candidate knowledge patterns identified in the course of this research are lexical, containing a relation marker that is a lexical unit or series of lexical units (or, as discussed in Section 1.2, in rare cases, a derivational affix). However, these may then be

refined by the addition of syntactic information about these forms and the elements they link (cf. lexico-syntactic knowledge patterns).

Given the different relations being studied, the comparative focus of this project, and specifically the goal of comparing various kinds of patterns in English and French, the patterns studied were not restricted to those containing a marker belonging to a specific part of speech class (as was the case of Barrière (2002) and Garcia (1996, 1997) with verbal markers). Patterns containing markers that were verbs (e.g., *to prevent*), nouns (e.g., *prevention*), adjectives (e.g., *preventive*), adverbs (e.g., *preventatively)* and prepositions (*X from Y*), as well as conjunctions (e.g., *X and Y*) were considered.

Thus, in pattern discovery, word forms that were associated with the expression of a pertinent relation (e.g., for the CAUSE–EFFECT relation of PREVENTION, *prevent, prevents, prevented,* and *preventing*) were noted. These individual forms were then associated with a base form of a lexical unit (in this case, the verb *to prevent*).

Both simple (e.g., *prevent*, *prevention*) and complex markers (e.g., *prevention of... by*) were considered. The most complete possible form of lexical marker was identified in each relation occurrence (e.g., *prevention of... by* was identified as the marker form present in structures such as *prevention of* X *by* Y, *prevention* in structures such as X Y *prevention*), although forms such as these containing the same open-class marker that differed only in the presence of additional closed-class elements such as prepositions were considered to be occurrences of a single base marker. (The analysis of these variations is discussed separately in Section 3.3.1.4.4.2.)

Through this annotation, the candidate markers identified were associated with a specific (sub-)relation (or potentially more than one in the case of polysemous markers).

3.3.1.4.2 Number of markers observed

Once the lexical markers present were identified, these were counted in order to obtain an idea of the number of distinct markers that were present in the relation occurrences

analyzed, and for each relation and sub-relation. This measure permits the estimation of the number and variety of distinct markers required to retrieve the number of relation occurrences observed in the contexts analyzed, and thus of the variety of markers that may be needed in pattern-based applications in each language.

3.3.1.4.3 Number of occurrences of markers

Because the productivity of a pattern-based tool depends largely on the number of potentially useful contexts accessible using pattern sets, the numbers of occurrences of markers are an important component in evaluating the potential of this kind of approach. This analysis was based on two separate measures: the proportions of relation occurrences observed that corresponded to the most frequent markers in the sample analyzed (which is indicative of the productivity of individual markers for retrieving relation occurrences), and the frequency of the sets of markers in the corpus as a whole (indicative of the overall number of potentially useful contexts that may be located using these sets).

In the first case, the evaluation focused on the numbers of markers observed that would be required to identify a given proportion of the relation occurrences identified (in this evaluation, 50% and 75%).

The evaluation of the frequencies of sets of markers in the corpora as a whole was generally determined using concordances generated for the markers using character strings and WordSmith Tools, in a version of the corpora that had been part-of-speech tagged and lemmatized using TreeTagger (IMS Textcorpora and Lexicon Group 1994; Schmid 1994). This allowed for the evaluation of marker frequency in such a way as to take into account inflected forms of the pattern markers and eliminate (within the limits of the performance of the part of speech tagger and lemmatizer) categorial ambiguities presented by the markers. In the case of markers that could not be effectively evaluated using this approach — for instance those that were complex and interrupted or that shared elements (for example, in the case of complex markers that contained elements

that may serve as markers in their own right, such as *risk* and *risk factor* in English) — occurrences were sorted manually to distinguish between forms and these figures were retained for the purposes of the analysis.[81] To permit marker frequency comparisons between the two data sets, a measure of frequencies per 1,000 tokens in the corpus was calculated.[82]

Combined with observations of the numbers of markers observed for each relation, these measures may indicate the variety of markers used to express a given relation and the potential of the marker sets to identify contexts, and thus help to estimate and compare the number of markers required in pattern-based applications to retrieve relation occurrences.

3.3.1.4.4 Types of markers observed

In order to evaluate the types of markers observed, each occurrence of a marker was associated with an indication of its part of speech class and form, as described below.

3.3.1.4.4.1 Part of speech class of markers

Given the possibilities of targeting specific part of speech classes of markers for use in pattern-based applications (e.g., Garcia 1997; Barrière 2001; Feliu 2004), as well as possible links with marker performance, the part of speech class of each pattern marker was noted in a database field included for this purpose, to allow for the evaluation of the proportions of individual markers and marker occurrences belonging to each category.

---

[81] In addition, when technical restrictions (e.g., the maximum frequency identifiable by WordSmith Tools) did not permit evaluation in the tagged corpus, the untagged corpus texts were used. As this was an issue only in the case of prepositions and conjunctions in these results, the possibilities of morphological variation and categorial ambiguity were not considered to be problematic in these cases.

[82] This conversion was carried out simply by dividing the total number of occurrences by the number of tokens in the corpus (as calculated by WordSmith Tools) and then multiplying the result by 1,000. The choice of occurrences per 1,000 tokens as a basis for comparison was made because the resulting figures were easily interpretable (i.e., were generally not too large or too small) and could be easily manipulated to estimate expected numbers of occurrences in corpora of different sizes, if desired.

In the case of complex markers, the part of speech class of each marker element was noted, and classification into pattern marker classes was based on the part of speech of the base, open-class element (e.g., in the case of NOUN + PREPOSITION combinations, the marker was considered to be nominal; in the case of VERB + PREPOSITION combinations, as verbal). This annotation allowed for the evaluation of different types of markers as coherent groups. To facilitate some aspects of this evaluation and comparison, some categories were grouped together for analysis; for instance, PARTICIPIAL ADJECTIVES (occurring both independently and in association with a preposition) were included with verb forms.[83] Moreover, as noted in the Introduction, derivational affixes — while not strictly speaking lexical units — were also considered in this research, as they may clearly mark relations at a textual level; these were considered as a separate category.

Analyzing the part of speech class distribution of markers allows for the evaluation of the types of markers that may be considered for pattern set development, and the impact that choices of specific classes of markers may have on pattern-based tool performance in the two languages.

3.3.1.4.4.2 Complex and simple marker forms

In order to allow for the evaluation of the proportions of complex and simple marker occurrences observed — and thus of the potential prevalence of the difficulties associated with these characteristics (e.g., interruptions of the forms or variation in the order of marker elements) — each occurrence of a marker form was classified as simple (in the case of single lexical items) or complex (in the case of marker forms consisting of multiple open-class lexical items or of an open-class and one or more closed-class items such as NOUN + PREPOSITION combinations).

---

[83] The choice to include these with the verbs rather than adjectives was made because these items were formally very closely related to verbs, and moreover were often difficult to differentiate from verbal forms appearing in elliptical structures.

3.3.1.4.5 Marker precision

Once the list of candidate markers was established, a sample of 13 of these patterns was used to evaluate the productivity of these markers for the identification of relation occurrences. For the purposes of this evaluation, the markers that were observed most frequently in the initial series of concordances were considered to be most interesting, as they are likely to be among the most promising for developing pattern sets. The sample was designed to include markers of the two relations evaluated, as well as a range of different part of speech classes of markers.

The selected markers were then used to generate a second set of concordances, similar to the one below in Figure 7; a random sample of 100 occurrences (generated in untagged corpora using character strings and the random sampling feature of WordSmith Tools) was extracted from the corpora. The character strings used were designed to allow for the retrieval of inflected forms of the markers while excluding where possible sources of noise from similar forms (e.g., belonging to other part of speech categories). The longest standard form of the marker observed was used in the two languages; for example, the form *risk* in English was used to find occurrences of *risk*, *risk from*, *risk of... from*, and so on; in French the form *risque de* was used, as this was the standard form observed in the initial concordances.

This second set of concordances was then manually analyzed, in order to identify contexts that contained a pertinent conceptual (sub-)relation, those that indicated another type of relationship (including more complex relations involving a causal component, as discussed in Section 1.5.2.7), those that involved forms corresponding to other lexical items (i.e., categorial ambiguities), and those that appeared to indicate the relation in question, but in which the two concepts linked were not clearly denoted using linguistic

means within the context extracted (i.e., that were incomplete), as well as non-pertinent uses of the items in question (i.e., noise).[84]

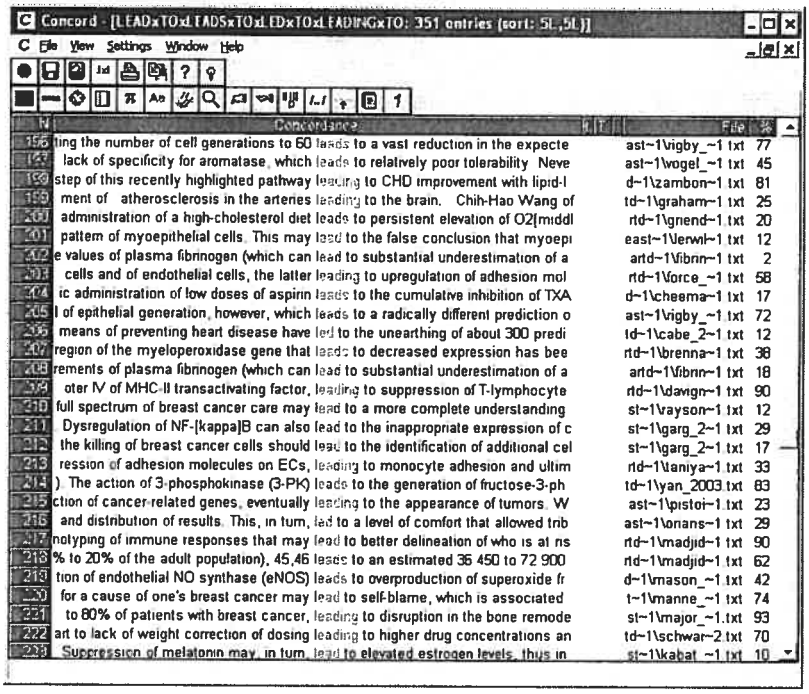


Figure 7. Pattern-based concordance for the marker *lead to*

Using the classification of the patterns' occurrences, pattern precision in the sample was calculated by comparing the total numbers of occurrences evaluated and those that were useful for identifying the desired relation. The evaluation of contexts' validity in this research was more liberal than those used by many other researchers, as the goal of the research was to study a wider range of occurrences of conceptual

---

[84] A comment on the classification of categorial ambiguities in this evaluation may clarify the decision made to distinguish these contexts systematically from valid hits. Certainly, character strings that correspond to more than one lexical item can be used to identify occurrences of specific relations (cf. the approach used in Marshman 2002), and may often provide an efficient means of retrieving a large number of candidate KRCs using a limited number of marker forms. However, in light of the methodology used in this project, which coupled markers with indications of their part of speech category and distinguished between formally similar markers on this basis (e.g., distinguishing the verb *cause* from the noun *cause*), the distinction was considered necessary in evaluating marker precision. Moreover, this also allowed for preliminary observation of some differences in precision linked to the part of speech category of markers, which can help to provide a basis for gathering data to guide strategies for choosing markers for use in tools.

relations, in order to better understand how they are represented in texts and thus how they can be located semi-automatically. Thus, not only contexts conforming to the TERM + [MARKER] + TERM structure with no expressions of uncertainty or negation were accepted, but rather any context in which the relation and the concepts involved were clearly expressed.

In cases in which the results obtained for a specific marker suggested possibilities for refinements in the use of markers involving the use of lexico-syntactic pattern forms in part-of-speech tagged corpora, samples of 50 contexts containing that marker and its counterpart in the other language were extracted from the corpora as processed using the tool Syntex (Bourigault et al. 2005), which analyzes syntactic dependencies in versions of the corpora that have been part-of-speech tagged and lemmatized using TreeTagger (Schmid 1994) and allows for contexts containing occurrences of specific lemmas to be identified and extracted. These contexts were once again analyzed manually in order to classify the occurrences as indicated above. This allowed for the potential for improving results of extraction using more developed approaches to be evaluated for specific markers that appeared to confront difficulties in character-string-based techniques.

In addition to these evaluations of the set of markers as a whole, a sub-set of ten markers in each language — two of ASSOCIATION and eight of CAUSE–EFFECT — that had similar distributions among the relations and part of speech classes was selected and evaluated in order to permit the evaluation of precision without potential bias due to differences in the performance of different classes of markers.

3.3.1.4.6 Polysemy of pattern markers

As discussed above in Section 2.3.1.4, pattern-based applications are very vulnerable to problems of ambiguity. This ambiguity can be identified in several ways. The first of these involves the observation of cases in which markers may denote more than one (sub-)relation among those considered pertinent in the research. A second involves the

occurrence of marker forms in contexts that do not indicate the relations of interest in this research, a major contributor to noise in analyses of marker precision. Finally, markers may be observed to denote not only one or more CAUSE–EFFECT sub-relations of interest, but also a more complex relationship that also includes an element of causation.

In the annotation of the relation occurrences analyzed in this research, the relation (and for the CAUSE–EFFECT relation, the sub-relation) present in each context was identified, as well as the pattern marker and pattern form observed. As markers were associated with a relation and sub-relation in each context individually, those that may indicate more than one sub-relation (according to the criteria set out for this project) could be identified and the nature of their polysemy evaluated.

In order to evaluate this polysemy, the lists of markers observed in each language were analyzed, and markers that were observed to indicate two or more (sub-)types of relations identified. In these cases, the variety of relations associated with the markers was compared, and the contexts in which they occurred were analyzed, in order to evaluate the impact that these factors may have on the usefulness of the results and the possibilities for differentiating between the different types of relations indicated.

The two remaining types of evaluation were carried out on the data from the evaluation of marker precision (as described in Section 3.3.1.4.5 and reported in Section 4.6). The prevalence and nature of phenomena such as noise and complex relationships observed in the contexts extracted using candidate markers may indicate additional types of challenges that may be encountered in pattern-based applications. The impact of these types of polysemy may be significantly different from that described above.

In the case of noise, the markers may not indicate relations of interest (or indeed, any relation at all) and therefore may introduce non-pertinent context into results. Ideally, strategies should be developed to eliminate these contexts from the results of KRC extraction. A brief analysis of this kind of polysemy was carried out in the context of the evaluation of marker precision (Section 4.6).

In the case of complex relations, while the markers may indicate relations other than the "pure" relations analyzed in this research, the contexts in which this kind of polysemy is observed may constitute *good noise*, in that the information they convey — while complex and therefore not immediately useful for some applications — may indeed ultimately help a user to better understand a concept.

The possibility that markers may indicate not only the type of "core" CAUSE–EFFECT relations considered in this research, but also more complex relationships with a causal component was also evaluated to some extent in the results of the study of precision. This kind of polysemy indicates that contexts retrieved by these markers, while useful, may pose difficulties for the identification of the specific relations present. Applications that attempt to sort or otherwise process these contexts automatically according to the relations present will confront significant challenges. The evaluation of the frequency of this phenomenon may help to determine its impact on pattern-based applications; moreover, the analysis of the types of polysemy observed and the forms in which polysemous markers occur may indicate possible avenues for dealing with the issue.

3.3.1.4.7 Pattern variation

As discussed in Section 2.6.2, pattern variation may involve not only morphological variation in marker form, but also variations in the number and nature of elements associated with a primary marker (e.g., the presence of prepositions or conjunctions in association with an open-class lexical marker), or variations associated with the voice of a verbal marker, as well as variations in pattern structure (e.g., in the location of the marker and related elements relative to one another, or the presence of additional, regular elements within a pattern structure). As these variations must be taken into account when developing pattern sets, and can also influence the recall of pattern-based tools, they were considered to be important in this research.

In order to observe these variations, a two-level annotation of the relation occurrence and marker present was used in the database, the first reflecting the base form of the marker, and the second indicating the structure of the context as it was observed and including the pattern marker and the related elements in the order and form in which they appeared. Additional characteristics, e.g., the voice of verbal markers, were also described in database fields designed to receive this information. Table 15 presents a sample of this annotation. The evaluation of marker and pattern variation is described below.

Table 15. Sample of annotation accounting for pattern variation

Context	Marker	Marker POS	Voice	Pattern Form
… LDL-C remains the primary target of lipid-lowering therapy based on a robust database of studies linking LDL-C to atherosclerosis and cardiovascular events … (Bittner 2003)	link… to	v. + prep.	active	[studies] linking X to Y1 [conjunction] Y2
… these findings, together with those in chronic atherosclerosis, importantly link ligand-RAGE interaction to the pathogenesis of exaggerated neointimal expansion … (Yan et al. 2003)	link… to	v. + prep.	active	[findings] … [adverb] link X to [article] Y
… Oxidative stress has been linked to the activation of both NF-[kappa]B and AP-1. … (Granger et al. 2004)	link... to	v. + prep.	passive	X has been linked to [article] Y1 [quantifier] Y2a [conjunction] Y2b
… homocysteine, dyslipidaemia, malnutrition and inflammation [1*,2,3*], some of which have also been linked to the pathogenesis of anaemia itself. … (Stevens and Levin 2003)	link… to	v. + prep.	passive	X1, X2, X3 [conjunction] X4 [ref], [quantifier] have been linked to [article] Y

3.3.1.4.7.1 Variation in marker form

The methodology used in this project, which identified the most comprehensive form of markers present, allowed for variants in marker form to be grouped together and analyzed. The data thus obtained were then studied in order to evaluate the numbers of marker variants and the types of pattern structures in which each marker participated.

The data for the markers that were observed two or more times in the samples annotated were analyzed to identify the level of variation in markers (including the presence of additional elements such as prepositions or conjunctions in addition to the principal marker elements and the change in the order in which marker elements appeared). The level of variation was evaluated using a simple ratio of the number of forms observed relative to the numbers of markers overall and for each relation.

However, the variation that can be observed per marker is influenced by the number of times a given marker was observed. In order to present the level of variation more accurately, the mean number of forms observed was calculated for groups of markers observed a specific number of times in the sample analyzed.[85]

Variation in the voice of verbal pattern markers was also considered specifically, as it affects not only the form of the marker but also that of the pattern in which the marker participates, including, for example, the inversion of causes and effects in the pattern structure.

3.3.1.4.7.2 Variation in pattern structures

The annotation of pattern structures observed in relation occurrences allowed for the identification of candidate pattern forms that may subsequently be evaluated and refined for use in pattern-based tools. Moreover, the potential for variation in these structures can be observed in the case of markers that occurred two or more times in these contexts.

Variations considered involved the relative placement of pattern elements (i.e., markers and related elements) and the presence of additional but regular items within pattern structures (e.g., the presence of a copula before adjectival or participial adjective markers). These variations were evaluated as a ratio of pattern structures relative to the

---

[85] The range of frequencies that were observed in both languages, and thus could provide a basis for comparison, was between 2 and 8.

number of markers overall and for each relation, and for groups of markers that were observed comparable numbers of times in the sample analyzed.

An additional specific case of pattern variation evaluated involved that observed when a marker occurred within a structure involving a relative clause introduced by a relative pronoun. This kind of variation in structures may be one of the most interesting to take into account in pattern forms, given its relative regularity.

3.3.1.4.8 Number and form of the elements linked by patterns

As markers are only one part of the knowledge pattern, it was also important to consider characteristics of the elements they linked. As described in Sections 2.3.1.5 and 2.6.3, variation from the prototypical pattern structure involving non-nominal forms of related elements or the presence of multiple elements sharing a "slot" in a knowledge pattern affects not only the forms of patterns required to extract complete KRCs, but also the value of the information they contain. Thus, for each relation occurrence observed, the number and form of the related elements were noted.[86]

3.3.1.4.8.1 Number of related elements

Cases in which more than two elements were linked by a given marker — generally when two or more elements shared a given role in a relation (e.g., two or more possible causes or effects were indicated in connection with a single occurrence of a marker) — were noted and their structures analyzed in terms of the relationship present between the elements sharing a role and the form of the occurrences.[87]

---

[86] As noted in Section 3.3, to be retained in this analysis contexts were required to include at least two participants in a relation, e.g., a cause and an effect, or two associated elements. Contexts that did not include one of the elements involved in a given relation because the context extracted was too short were also excluded from study.

[87] This was not considered in the evaluation of pattern variation, however, as it was considered to be distinct from other types of pattern variation (primarily because it is not linked to specific markers but rather may be observed in a wide range of relation occurrences).

Each context containing multiple related elements sharing a role in a relation was classified according to the relationship between these elements. These relationships included the appearance of two or more variant expressions of a single concept (including a full form of a linguistic item (generally a term) denoting a concept accompanied by an abbreviation or symbol representing this same concept), the conjunction (e.g., X and Y cause Z), disjunction (e.g., X causes Y or Z), and conjunction/disjunction (e.g., X causes Y and/or Z)) of linguistic expressions denoting multiple related elements, and finally the case of co-occurrence of expressions denoting concepts participating in GENERIC–SPECIFIC relationships with one another.

As part of the analysis of the form these occurrences may take, the lexical or paralinguistic indicators (e.g., punctuation) of the relationship that existed between the two elements were noted and their distribution analyzed; this permitted an evaluation of the complexity of the task of developing pattern forms that could allow for complete identification and analysis of contexts containing these types of structures and of the information these contexts offer.

The presence of the associated phenomena of ellipsis of part of complex elements and repetition of a pattern marker or part of a marker in association with the phenomenon were also noted and evaluated. In the case of ellipsis, a further distinction was made between cases in which the head of a complex item was omitted, and those in which it was an expansion that was omitted, as the impact of these phenomena for pattern design and the potential usefulness of extracted information may differ.

3.3.1.4.8.2 Form of the related elements

As observed in the Introduction and described in Section 2.1, some pattern-based tools attempt to identify contexts in which relations link specific types of elements (e.g., previously identified terms or candidate terms, or elements that appear in a form typical of terms (generally NOUNS and NOUN PHRASES)). This raises questions about the

proportion of relation occurrences that involve such forms — and more importantly, of those that do not. This aspect of the relation occurrences was thus also evaluated.

Any occurrence of a related element that did not occur in NOUN or NOUN PHRASE form was thus annotated with an indication of its part of speech class (including ADJECTIVES, PRONOUNS, VERBS, and propositions). The numbers and proportions of these items were then evaluated in order to determine the proportions of occurrences that diverged from the "standard" nominal forms.

3.3.1.4.8.2.1 Anaphora

One specific type of variation in the form of related elements — and one that may be particularly problematic for the analysis of KRCs and the identification of the information they convey (as discussed in Section 2.3.1.5.2.1) — is anaphora. The database used for the annotation of the contexts contained a field that allowed for each occurrence of an anaphoric expression that replaced a related element (or some part thereof) to be identified, and for the form of this expression to be identified and subsequently analyzed.

Moreover, other types of anaphoric expressions that appeared in candidate KRCs were also noted, as they may affect the form of pattern markers or of the patterns themselves, and certainly indicate a need for additional information to evaluate the information conveyed in a given context fully.

Within the category of anaphoric expressions, a distinction was made between the various types of part-of-speech classes identified (including PRONOUNS, POSSESSIVE ADJECTIVES and generics introduced by a DEMONSTRATIVE ADJECTIVE or DEFINITE ARTICLE), given the differences in the characteristics of these classes of elements (such as the location of the anaphoric expressions relative to their antecedents), the ways in

which these could be integrated in pattern set design, and the usefulness of the anaphoric expressions themselves for knowledge extraction.

This data was then used not only to determine the proportion of potentially useful contexts that might be lost if such contexts were excluded and to evaluate the need for access to a larger context for knowledge acquisition, but also to identify the types of anaphoric expressions that were used in the contexts in each language and the ways in which their use might be taken into account in the creation of pattern forms that can be used to identify contexts containing anaphora.

In order to facilitate the evaluation of the most problematic forms of anaphora for KRC identification and processing, the proportion of occurrences of the phenomenon in which an entire related element or the head of a complex related element was replaced by an anaphoric element was also evaluated specifically, and the types of anaphoric elements observed in these cases analyzed.

3.3.1.5 Annotation of challenges for pattern-based tool use

As observed in Section 2.4, identifying knowledge-rich contexts using knowledge patterns — and even more so further processing these contexts automatically — is not always as simple as it might first appear. The difficulties of pattern-based information extraction from corpora must thus be taken into account when working in any language, and it is also important to take into account any differences in their nature and frequency when adapting the pattern-based approach for bilingual use.

Throughout this project, a record was kept of challenges pertinent to the identification and application of the patterns that were observed in the analyzed contexts. The database structure used contained fields corresponding to the majority of the types of difficulties described in Section 2.4, and for each annotated context the appropriate information was entered in these fields, allowing the identification of all cases of a given phenomenon as well as specific details of the occurrence in a given

context. The annotation of these difficulties was carried out within the structure illustrated in Figure 8. In the following sections, some of the details of the decisions made for the annotation of specific difficulties will be presented.

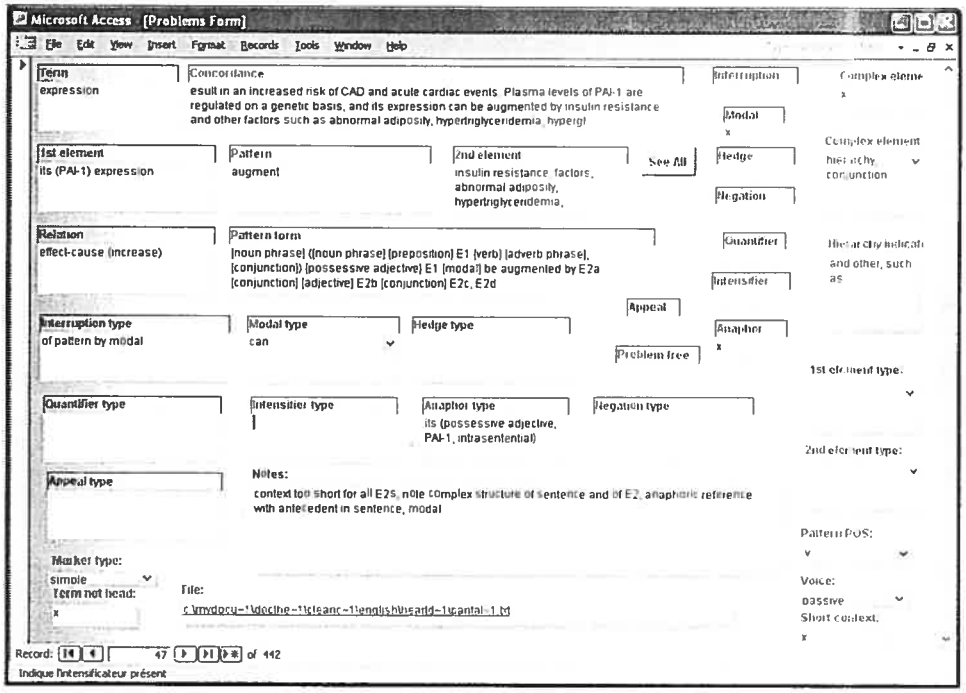


Figure 8. Annotation of challenges for pattern-based applications

3.3.1.5.1 Pattern interruptions

As discussed in Section 2.4.1, the challenges posed by the interruption of pattern structures and/or elements in the design of pattern forms and the recognition of KRCs are significant, and thus the occurrences of such interruptions were noted and analyzed.

The two-level annotation described in Section 3.3.1.4.7 allowed any such interruptions of patterns to be taken into account. In addition, fields were included in the database for identifying contexts in which patterns were interrupted by external elements, and for describing the location of the interruption and the interrupting item. Interruptions were evaluated in general, and in addition three main types of interruptions were identified that may have different impacts on the development of pattern forms and

the recognition of relation occurrences in corpora: those of patterns by other patterns (discussed below in Section 3.3.1.5.1.1), of complex markers, and of related elements.

The annotation of these phenomena allows for the calculation of statistics of the number of pattern occurrences that were interrupted (and thus of the proportions of potentially useful contexts that would be affected by the phenomenon), and the evaluation of the frequency of different types of interruptions.

3.3.1.5.1.1 Multiple markers and interruptions by other patterns

The presence of multiple markers in a single context can pose significant challenges for the identification of relations present in a candidate KRC, as discussed in Section 2.4.1.

Annotating the cases in which multiple markers and/or patterns were present required the evaluation and processing of occurrences on various levels. Relation occurrences may be observed to contain multiple markers corresponding to separate relations between distinct pairs of concepts or to a single relation between a pair of concepts (or potentially more, in some cases of multiple elements sharing a role in a relation). These required distinct forms of evaluation.

When multiple markers denoted distinct relationships, some chains of relations were observed. Since in this annotation, one of the conditions for retaining and annotating pattern occurrences was that the term used to generate the concordance realize one of the concepts in a relation, the question of annotating chains of relations within a single context was somewhat simplified. If the term in question was linked to only one of the markers indicating relations expressed in the context, only the relation indicated by that marker was annotated. If, however, the term was linked to more than one marker and thus denoted a participant in more than one separate relation, the context was duplicated and each relation annotated separately. If the pattern structure identified in the relation occurrence was interrupted by another pattern (e.g., occurring in a relative clause inserted within the main clause), this interruption was noted.

In contexts containing multiple markers describing a single relationship between a pair of items, only the marker that was most decisive in characterizing this relationship was retained in the annotation (and used in the calculation of pattern statistics); the presence of another marker in the context was nevertheless noted and included in the pattern form identified.[88] This is illustrated in Table 16.

Table 16. Annotation of contexts containing multiple markers

Concordance	Relation	1st element	Pattern	2nd element	Pattern form
Strenuous PA was generally associated with a reduced breast cancer risk. (Dorn et al. 2003)	ASSOCIATION	strenuous PA	associated with	breast cancer	X [copula] [adverb] associated with [article] [reduced] Y [risk]
Receptor-mediated leukocyte activation leads to conformational changes in LFA-1 structure… (Granger et al. 2004)	CAUSE–EFFECT (modification)	receptor-mediated leukocyte activation	change in	LFA-1 structure	X [leads to] [adjective] changes in Y
L'activation de récepteurs endothéliaux produit une augmentation de [Ca]i dans les cellules endothéliales… (Feletou et al. 2003)	CAUSE–EFFECT (increase)	activation de récepteurs endothéliaux	augmenta-tion de	[Ca]i	X [produire] [article] augmentation de Y
Cette activation directe permet d'engendrer une réponse cellulaire cytotoxique protectrice. (Catros-Quemener et al. 2003)	CAUSE–EFFECT (creation)	cette activation directe	engendrer	réponse cellulaire cytotoxique protectrice	X [permettre de] engendrer [article] Y

[88] The presence of an additional marker of ASSOCIATION or CAUSE–EFFECT relations was also noted in the canonical pattern form(s) ultimately identified for a pattern, if this occurred regularly. However, the specific marker was not indicated as in the pattern form illustrated in Table 16; an indication of the presence of some marker of the appropriate relation was considered to be sufficient for these purposes. This can be observed in the results presented in Appendix H.

"Decisive" patterns in these contexts were identified based on the evaluation of the relation that was identified: if the markers present generally indicated (and were interpreted by the annotator as denoting) different relations or sub-relations, the marker that was associated with the relation that was judged to hold between the two items in a general interpretation of the context was chosen.

Thus, for example, in the structure X *leads to changes in* Y, *lead to* generally indicates CREATION, *change in* generally indicates MODIFICATION, and in a general evaluation of the context, the overall relation was judged to be MODIFICATION, since the overall effect is a change in a feature of Y; as such, *change in* was retained as the marker and *lead to* was indicated only as part of the pattern form. When markers that generally expressed ASSOCIATION and CAUSE–EFFECT relations co-occurred, as in the case of structures such as X *is associated with reduced* Y, the ASSOCIATION relation was considered to be dominant, and thus the marker *associated with* was indicated as the principal marker, and *reduced* was included in the pattern form. This decision was made because of the unconfirmed nature of potentially causal relations inherent in ASSOCIATION; thus, it was considered to be premature (and potentially misleading to an end-user of the results produced by a pattern-based tool) to classify such occurrences as CAUSE–EFFECT relations.

In the rare cases in which two markers of the same relation (and sub-relation) occurred connecting the same pair of elements, the marker that was identified as the 1) clearest and 2) most prevalent indicator of the relation was retained.

This approach allows for counting each relation between a given pair of elements only once, rather than several times (once for each marker present). It also reflects the most effective strategy for presenting contexts to an end user, with each context ideally appearing only once for each related element pair it contains. Moreover, it avoids potential problems of mis-classifications of contexts (as in the cases of co-occurrence of markers of ASSOCIATION and CAUSE–EFFECT relations). In addition, it ensures that each marker can be analyzed separately, and that no false indications of polysemy of a given

marker are drawn on the basis of occurrences containing multiple markers. The tagging of each context with an indication of the presence of multiple markers and the inclusion of these markers in the pattern form identifies these cases for further analysis and study, and ensures that the presence of additional markers is taken into account in the analysis of pattern forms. (However, the approach does have an impact on the results obtained; see the discussion of this decision in Section 5.5.3.5).

3.3.1.5.1.2 Interruptions of complex markers and of related elements

In addition to interruptions of pattern structures, in some cases external elements (e.g., modifiers, quantifiers) occurred between elements of complex markers or related elements or between multiple related elements. As these interruptions can interfere with the identification of KRCs and the information they convey by pattern-based tools, and should also be taken into account when developing many types of pattern forms, these cases were noted and their frequency evaluated.

In the case of interruptions of complex markers, a distinction was made between cases in which the interruption was systematic (in the case of marker forms that surround one of the elements that they link, e.g., *prevention of* X *by* Y) and those in which the interruption was irregular (e.g., in the case of interruptions by modifiers). This distinction allows for the differentiation between types of interruptions that are relatively predictable and thus may be accounted for in pattern forms, and those that are not and thus present more serious challenges for pattern-based tools.

3.3.1.5.2 Expressions of uncertainty

As discussed in Section 2.4.2, the presence of expressions of uncertainty in candidate KRCs may affect not only the recognition of these contexts but also the value of the information they convey for subsequent use, and it is thus important to identify the types of indicators that may be used, and to develop strategies for dealing with the

phenomenon, for example by classifying contexts containing these markers semi-automatically according to the level of "reliability" that they indicate.

In this research, the presence of expressions of uncertainty in the relation occurrences identified was noted where applicable. In addition, these expressions of uncertainty were classified into four types: quantification of related elements, hedging, modal verbs and negation. Each of these types was annotated and evaluated separately, in order to permit the evaluation of the various characteristics of the different types of expressions and their impact on pattern form and the strategies that may be developed for dealing with the uncertainty indicated.

The potential for these expressions to interrupt pattern forms was taken into account in the annotation of pattern interruptions. In addition, each type of expression of uncertainty observed in the analyzed contexts was noted in a database field intended for this purpose. This allowed for the description of the types of expressions used to indicate degrees of uncertainty both at a formal level (including the part of speech classes to which they belonged, for the purposes of their inclusion in pattern forms as required) and at a semantic level (in order to describe the ways in which relations may be characterized and to determine the effect on the reliability and reusability of contexts for future applications).

3.3.1.5.3 Text-related issues

For the purposes of this research, text-related issues related to individual texts or contexts were identified only when they might interfere with the identification or interpretation of a relation occurrence. Contexts containing writing problems or that were written in complex style such that it was impossible to identify (using the manual approach used in this research) the relation present, the element involved in a relation, or a candidate pattern marker and/or form in the manual analysis were excluded from study, as the possibilities for automating the identification of these occurrences were considered to be limited. Cases of writing problems that did not interfere with manual

identification of these elements were retained, although the presence of potential problems for pattern-based tools was noted.

3.3.1.5.4 Difficulties overall

While, for organizational purposes, the discussion of the relation occurrences has been divided according to the nature of the elements in question (i.e., on whether the discussion focuses on an element is a part of the pattern itself — pattern characteristics — or an element that is external to the pattern form — challenges for pattern-based applications), it is clear that a number of difficulties may affect the performance of tools for extraction of KRCs and the ultimate usability of the contexts extracted.

These include the form of related elements (particularly elements that are non-nominal in form), anaphoric expressions occurring within patterns, interruptions of pattern forms (including complex markers and related elements), expressions of uncertainty, and text-related issues. The proportions of the contexts containing any one of these factors is indicative of the frequency with which relation occurrences in the two data sets diverge from prototypical knowledge pattern structures indicating a certain and reliable relation, and therefore in turn may reveal the proportion of relation occurrences that would be missed by the most conservative approaches (e.g., that rely on very restricted pattern forms and that exclude contexts containing anaphora and expressions of uncertainty).

For this reason, the proportion of relation occurrences containing one or more of these phenomena was evaluated.[89] This analysis indicates the importance of considering such factors in the context of pattern-based applications, and may help to determine how profitable investments in time and effort in developing strategies for dealing with them

---

[89] However, the evaluation excludes the interruption of complex markers by related elements, which is not generally considered as a difficulty as such, although it adds to the complexity of developing pattern forms.

may be. Moreover, it may provide data to support decision-making as to the selection of pattern-based approaches that are appropriate for a given situation.

3.4 Interlinguistic comparison

We researchers use statistics the way a drunkard uses a lamp post: more for support than illumination.
Winifred Castle, Statistician

After the annotation of the pattern characteristics and challenges for pattern-based applications was completed, an interlinguistic comparison between the results of the analysis for each of the criteria described above in Section 3.3 was carried out, in order to evaluate the similarities and differences in the product of this methodology in the English and French. This comparison focused on both quantitative data (e.g., the number of occurrences of relations, markers of relations and of challenges observed) and qualitative data (e.g., the forms in which these phenomena were observed, and the potential impact of these for pattern-based applications).

The quantitative results were analyzed where appropriate using statistical tests in order to evaluate the statistical significance of any observed differences. Qualitative data were compared through general observations of the parallels observed and any differences that became apparent in the course of the analysis.

Details of the methods used for the interlinguistic comparison of quantitative data, specifically of the statistical tests used to determine the significance of differences observed, are presented in Section 3.4.1, focusing on the comparison of the overall numbers of relation occurrences observed, in Section 3.4.2, focusing on the comparison of patterns and their characteristics, and in Section 3.4.3, describing the comparison of the challenges observed for pattern-based applications.

3.4.1 Comparison of numbers of relation occurrences observed

The overall proportions of the contexts evaluated that produced relation occurrences meeting the criteria for evaluation in this study were calculated from the frequencies of relation occurrences identified; this calculation was also done for each relation individually.

Differences between the data samples in proportions of relation occurrences to contexts analyzed were evaluated using the Chi-square (χ^2) test (Muller 1973: 109–127; Oakes 1998: 24–29; Norman and Streiner 2003: 86–88), used to compare rates and proportions and evaluate the probability that a variation at least as large as the one observed could occur strictly by chance. More precisely, this statistical evaluation allows for testing of what is generally referred to as a *null hypothesis*, i.e., in this case that there is no difference between the samples in two languages in regard to the criterion evaluated, and specifically for estimating the probability that, if this null hypothesis is true, any difference observed in the results can be entirely accounted for by chance. This probability is generally expressed as a p value, which can vary between 0 and 1. A high probability value (i.e., a p value approaching 1) indicates that any variation is very likely to be the result of chance alone. A low p value (i.e., approaching 0) indicates that chance is unlikely to be entirely responsible for the variation observed, and thus suggests that the null hypothesis should in all likelihood be rejected, and that there is a statistically significant difference between the two samples in regard to the criterion being tested.[90]

The Chi-square test is generally agreed to be applicable when the expected occurrences of a given phenomenon in a given sample size are 5 or higher; smaller values may not provide valid results. Most commonly, a p value less than or equal to

---

[90] It is important to note that a null hypothesis (e.g., in this case, that there is no difference between the samples in the two languages in terms of the criteria analyzed) can never be proven. When a non-significant difference is present, a statistical test can only suggest the scope of future research that could allow for a statistically significant difference — if one exists — to be observed (e.g., in what size of sample the level of discrepancy observed would be significant).

0.05 is considered as the threshold of significance permitting the rejection of the null hypothesis (e.g., Norman and Streiner 2003: 32); this criterion has been adopted in this thesis. The exact calculations in this research were done in a Microsoft Excel spreadsheet (version 2003). (This test is described in more detail in Appendix G.)

The numbers of relation occurrences identified for the sub-sets of relation occurrences involving terms that were equivalents in the two languages, and those that involved non-equivalent terms were also compared. These sub-analyses permitted the evaluation of the contribution of these sub-sets to the overall data.

3.4.2 Comparison of pattern characteristics

Following the structure set out in the evaluation of the results in each language, in the interlinguistic comparison the numbers of markers observed, the number of occurrences of these markers, the types of markers observed, pattern variations (in marker form and pattern structure), and the number and form of related elements were compared.

3.4.2.1 Number of markers observed

In order to evaluate the variety of pattern markers observed in the research, the numbers of distinct markers observed for each (sub-)relation in each corpus were compared with the total number of contexts analyzed and the number of relation occurrences annotated.

Comparison of the numbers of markers observed relative to the number of contexts evaluated reflects the productivity of the methodology applied in this research for pattern discovery in the two corpora. As the number of distinct markers observed relative to the total number of relation occurrences identified in the sample may suggest how many markers and patterns will be required to obtain a given number of pertinent contexts in the two languages; variations observed between the English and French data sets may suggest discrepancies that should be further evaluated and ultimately taken into account in pattern-based tool development.

The difference in the numbers of markers observed in the two samples is not evaluated from a precise statistical perspective in this study, because of the restrictions imposed by the data (i.e., the fact that the comparison involves the ratios of *distinct* markers to a total number of relation occurrences associated with the set of markers and not a simple evaluation of proportions of marker occurrences, which can be measured as above using the Chi-square test). Rather, the ratios of distinct markers relative to the total numbers of contexts evaluated and relation occurrences observed are simply compared in order to give an indication of the potential for differences between the samples in the two languages that may be worthy of further investigation using a methodology that allows for precise statistical evaluation.

3.4.2.2 Number of occurrences of markers

The counterpart to comparison of the variety of pattern markers observed is comparison of the frequencies of the individual markers and of the sets of markers in the English and French data. In these evaluations, the numbers of occurrences of markers in the sample of annotated contexts (specifically as a proportion of the total numbers of occurrences observed), as well as the marker sets' frequency in the corpus as a whole — using occurrences per 1,000 corpus tokens— were contrasted.[91]

The nature of the data precludes precise statistical comparisons of the English and French data in respect to these factors. Given that the lists of markers are of course different in the two data sets (which consist not only of different numbers of markers observed in different numbers of relation occurrences, but also of markers that are in themselves different in the two languages), it is not possible to compare the numbers of occurrences of each marker in the two corpora directly. The comparison carried out here rather reflects the overall productivity of the markers observed in terms of the proportions of relation occurrences identified in the sample that corresponded to the

---

[91] See Section 3.3.1.4.3, footnote 82 for details of these frequency calculations.

most frequent markers (indicating the differences in the productivity of marker sets for locating the relation occurrences identified in the two corpora, which in turn may reflect the numbers of markers required to locate a given number of pertinent relation occurrences in each language), and of the numbers of occurrences for the marker sets per 1,000 corpus tokens (which may indicate differences in the number of potentially useful contexts that may be retrieved using these marker sets).[92] Discrepancies observed in these measures may thus suggest the need for further study in a context that allows for direct and precise statistical evaluation.

3.4.2.3 Types of pattern markers observed

The part of speech class of markers and marker occurrences and the proportions of marker occurrences in complex or simple form were compared in the samples in the two languages.

Potential differences in the proportions of markers and marker occurrences belonging to the various POS classes were evaluated using the Chi-square test, providing data on possible variations in the types of markers that indicated the relations in the two corpora and the distribution of relation occurrences associated with each type of marker, thus suggesting the potential usefulness of each class for inclusion in pattern-based applications in corpora in the two languages.

The Chi-square test was also used to evaluate potential differences in the proportions of complex and simple marker occurrences, in order to reveal possible interlinguistic differences in the challenges for pattern design and application associated with complex markers.

---

[92] Of course, since these figures indicate the presence of the marker only, and not the proportion of pertinent occurrences, nor of specific pattern forms, they can only predict the total pool of possibly pertinent contexts.

3.4.2.4 Marker precision

An interlinguistic comparison of the results of the precision evaluation for a sample of markers and some of the principal differences observed between the markers' performances in the two data sets was carried out to evaluate potential differences in the two data sets. The Chi-square test was used to evaluate differences in the proportions of marker occurrences assigned to each category.

3.4.2.5 Marker polysemy

In light of the data gathered in the analyses in English and French, the observations of various kinds of polysemy that were noted in the two corpora may be compared. However, given the limited size of the data samples available, statistical testing would not be reliable and certainly does not provide a strong basis for generalizations. A comparison may only be considered as potentially indicating differences that merit further investigation with more data (ideally extracted specifically for this purpose using appropriate criteria). Given this situation, no statistical test was applied for comparing polysemy in the two corpora.

3.4.2.6 Pattern variation

The comparison of levels of variation in marker and pattern forms is a complex task, and one that cannot be undertaken in a strictly accurate way using the data gathered in this project, as the numbers of occurrences of the markers — and thus the potential for observing variation — differed from one marker to another and between the sets of markers observed in the English and French corpora.

For this reason, while the levels of variation of marker forms and of pattern structures observed can be represented roughly by calculating a simple ratio of different forms observed relative to the number of occurrences of each marker, and the mean values calculated for the marker sets in the English and French data and for each of the relations within these sets, these measures cannot be considered to be strictly

comparable. Rather, any discrepancies should be considered as possible foci for future research in a more structured context rather than as statistically significant differences.

The need for further study of interlinguistic differences in levels of variation may be somewhat more accurately evaluated by comparing the mean numbers of marker and pattern forms within groups of markers that were observed the same number of times in the English and French data sets. For this reason, the ratios of marker and pattern forms to markers observed from two to eight times (i.e., the numbers of occurrences that were common to markers observed in the two data sets) were also compared in the English and French data, overall and for the CAUSE–EFFECT and ASSOCIATION relations.

3.4.2.7 Number and form of related elements

The proportions of the relation occurrences in the two data sets involving the various numbers and forms of related elements (as well as the related phenomena) described in Section 3.3.1.4.8 were all compared using the Chi-square test described in Appendix G.

The proportions of occurrences of different types of relationships between multiple elements sharing a role in a relation were also compared using this test. The prevalence of various indicators that identified these relationships was also analyzed, in order to evaluate the challenges of the task of representing these formally in the two languages. In the cases of conjunction and disjunction of related elements and of GENERIC–SPECIFIC relationships holding between such elements, the distribution of the occurrences of each phenomenon among the various lexical indicators observed was also compared in the two data sets, to evaluate the potential for developing pattern forms including these structures in each language. (However, the different numbers of occurrences and markers observed in the two data sets precludes a formal statistical evaluation.)

3.4.3 Comparison of challenges for pattern-based tools

The quantitative data on the proportions of relation occurrences containing the types of challenges for pattern-based applications discussed in Section 3.3.1.5 (interruptions of patterns, complex markers and related elements, presence of multiple markers, expressions of uncertainty), were also compared, and the differences evaluated using the Chi-square test described in Appendix G.

The proportions of the various types of expressions of uncertainty present were also analyzed using the Chi-square test to evaluate variation in the observations of the phenomena in the two data sets. Where applicable, the proportions of occurrences of each type of expression (e.g., specific modal verbs, hedges belonging to different part of speech classes) were also compared in the two data sets, to evaluate the possibilities for developing pattern forms.

The results of the analyses described above will be presented in Chapter 4.

4 Results

This Chapter will present the results of the analysis of the data according to the criteria identified above in Chapter 3, including characteristics of the candidate markers and patterns observed for each relation as well as some challenges for pattern-based applications. The discussion of each characteristic will begin with a brief restatement of its pertinence, followed by the comparison of the results in the two data sets, and finally some specific characteristics observed in each language.

4.1 Number of relation occurrences observed

As this project involved the use of an approach similar to that used in many pattern-discovery applications and the analysis of corpora similar to those that may be exploited using pattern-based tools, the number of relation occurrences provides information about the productivity of the methodology in each language and of the possibilities of identifying relation occurrences that meet the criteria used in the project.

As shown in Table 17, the proportion of contexts observed to contain occurrences of relations was somewhat different in the two data sets, with approximately 31% of the 1,412 contexts analyzed in English producing pertinent relations associated with lexical knowledge patterns, and 25% of the total of 1,392 in French.[93] An evaluation using the Chi-square test confirms that the proportion of relation occurrences to contexts analyzed is significantly higher in the English data than in the French ($p < 0.001$).[94]

---

[93] However, as some contexts produced more than one annotated relation, the proportion of distinct contexts that were retained is actually slightly lower. In the English sample, 413 distinct contexts produced relation occurrences (123 distinct contexts with ASSOCIATION relations and 294 CAUSE–EFFECT), and in the French sample this figure was 325 (70 distinct contexts with ASSOCIATION relations and 258 CAUSE–EFFECT). This slight variation nevertheless does not significantly alter the results of the statistical comparisons reported.

[94] All p values in this thesis refer to the results of Chi-square tests.

Table 17. Comparison of the proportions of ASSOCIATION (A+) and CAUSE–EFFECT (CE+) relation occurrences relative to the total number of contexts analyzed in English and French[95]

	EN	FR	Total
A+	125	70	195
CE+	317	279	596
Total contexts	1412	1392	2804

In both languages, more occurrences of CAUSE–EFFECT relations were observed than ASSOCIATION relations (72% CAUSE–EFFECT relations in English and 80% in French). The proportions of the contexts containing the individual relations nevertheless varied between the languages, with 22% of the contexts in English returning CAUSE–EFFECT relations and 20% in French, and 9% of the contexts analyzed containing ASSOCIATION relations in English and only 5% in French. The Chi-square test confirms that these proportions are significantly different. If the relations are analyzed separately, the difference in the proportions of contexts that returned CAUSE–EFFECT relations is not as evident — and is not statistically significant ($p = 0.119$) — but the proportion of those returning ASSOCIATION relations is significantly different ($p < 0.001$).[96]

The distribution among the sub-relations of CAUSE–EFFECT relations in the two data sets was roughly parallel, with most occurrences found for the CREATION sub-relation, followed by MODIFICATION, INCREASE and DECREASE, as shown in Table 18.

Relative to the terms used to generate the initial concordances (Table 19) the terms from the process class in both languages accounted for a high proportion of the relations analyzed in this project, and the other elements fewer, with the class of

---

[95] In this table, EN indicates the numbers of cases identified in the English, FR the number of cases in French, and + the presence of the criterion evaluated (i.e., in this table, a pertinent relation). These conventions have been retained in the Chi-square tables throughout the thesis.

[96] Within the category of relation occurrences, again a significant difference is observed ($p = 0.008$), with French showing a lower proportion of ASSOCIATION relations and higher proportion of CAUSE–EFFECT relations than English.

pathologies next, followed by the activities and finally the entities. A comparison of the proportions of contexts associated with each class that produced relation occurrences indicates a perfect rank-order correlation in the two data sets in the productivity of the various classes of terms. However, a consistently lower proportion of the contexts provided relations in French than in English, with a difference of 6.4% for the activity class, 6.2% for the class of pathologies, 4.9% for the entity class, and 3.4% for the class of processes, resulting in a 6.1% difference overall.

Table 18. Comparison of distribution of relation occurrences in English and French

Relation	Number of occurrences annotated in English	Percentage of relations annotated in English	Number of occurrences annotated in French	Percentage of relations annotated in French
ASSOCIATION	125	28.3	70	20.1
CAUSE–EFFECT	317	71.7	279	79.9
CREATION	167	37.8	133	38.1
DESTRUCTION	8	1.8	9	2.6
MAINTENANCE/PERMISSION	12	2.7	21	6.0
PREVENTION	20	4.5	18	5.1
MODIFICATION	46	10.4	48	13.8
INCREASE	36	8.1	25	7.1
DECREASE	27	6.1	24	6.9
PRESERVATION	1	0.2	1	0.3
Total	**442**	**100**	**349**	**100**

Table 19. Comparison of proportions of relation occurrences by term class in English and French

	English				French			
Class	**Number of contexts**	**Number of relations**	**% of contexts with relations**	**% of relation occur-rences from class**	**Number of contexts**	**Number of relations**	**% of contexts with relations**	**% of relation occur-rences from class**
Activity	201	41	20.4	9.3	200	28	14.0	8.0
Entity	307	48	15.6	10.9	300	32	10.7	9.2
Process	451	248	55.0	56.1	397	205	51.6	58.7
Process (pathology)	453	105	23.2	23.7	495	84	17.0	24.1
Total	**1412**	**442**	**31.3**		**1392**	**349**	**25.2**	

As shown in Table 20, there are rough parallels between the two data sets in the productivity of each term class for identifying occurrences of the individual relations. In both languages, the category of pathologies is particularly productive for locating occurrences of ASSOCIATION relations. However, the proportion is somewhat higher in French, and those for the classes of activities and entities somewhat lower than in English, while the distribution of occurrences in English between processes and pathologies is much more even. The distribution for the CAUSE–EFFECT relation, however, is very similar in the two data sets, although French shows a very slightly higher productivity of process terms and lower productivity of pathology terms than English.

These observations suggest that that on the whole, terms belonging to the class of processes are both good candidates for observing these relations, and particularly the CAUSE–EFFECT relation, and also may be good candidates for description according to their participation in these kinds of relations. In addition, the class of pathologies is particularly productive for observing ASSOCIATION relations, especially in the case of French.

This general parallelism indicates possibilities for the development and application of this kind of methodology for pattern discovery and use in the two corpora, and shows promise for further use of this kind of approach in corpora in both English and French. Further research could also help to clarify the sources of the variation that was observed between the two languages; possible explanations include variations in the contents of the corpora and the choice of terms used to generate the initial concordances.

Table 20. Comparison of numbers of individual relation occurrences linked to term classes in English and French

Term class	% of English contexts	% of French contexts	Numbers of English ASSOCIA–TION relations	Numbers of French ASSOCIA–TION relations	% of English ASSOCIA–TION relations	% of French ASSOCI–ATION relations	Numbers of English CAUSE–EFFECT relations	Numbers of French CAUSE–EFFECT relations	% of English CAUSE–EFFECT relations	% of French CAUSE–EFFECT relations
Activity	14.2	14.4	13	3	10.4	4.3	28	25	8.8	9.0
Entity	21.7	21.6	19	5	15.2	7.1	29	27	9.2	9.7
Process	31.9	28.5	42	19	33.6	27.1	206	186	65.0	66.7
Process (pathology)	32.1	35.6	51	43	40.8	61.4	54	41	17.0	14.7
			125	**70**			**317**	**279**		

The primary exception to this parallelism is the overall number of relations observed, and particularly that in the ASSOCIATION relation. As mentioned above, one strategy for identifying the source of the variation in the numbers of relations observed is the comparison of the terms used to generate the initial concordances. As observed in Section 3.2.1, in the selection of terms the primary criteria used involved the specificity of terms in the corpora and their representation of semantic classes and the two sub-fields evaluated. As reported in Section 3.2.2, these criteria produced term lists in which 9 of 15 terms were candidate term pairs that can be identified as equivalents, while another 6 were associated with the same semantic classes but were not equivalents. It is thus interesting to compare the results of the analysis of these concordances, in order to determine whether these two groups produced significantly different numbers of relation occurrences.

This is indeed the case. As illustrated in Table 21 and Table 23, the numbers of relation occurrences and the proportions of relations observed for the pairs of equivalents are more equally distributed, as compared to the groups of non-equivalents, shown in Table 22 (with the processes ordered in increasing number of occurrences) and Table 24.

If the figures for the group of 9 equivalent term pairs alone are evaluated, the differences in the numbers of relation occurrences relative to the number of contexts analyzed is still present (with the overall proportion in French slightly lower than in the English), but the difference is much reduced, and is not significant for the two relations together (p = 0.283) or for the relations separately (p = 0.388 for the ASSOCIATION relation, with the proportion of occurrences slightly higher in French, and p = 0.153 for the CAUSE–EFFECT relation, with the proportion of occurrences somewhat higher in English). These figures are illustrated in Table 25.

Table 21. Relation occurrences per term for equivalent pairs in English and French

English						French					
Term	**Class**	**Number of contexts**	**Number of relations**	**ASSOCIATION relations**	**CAUSE–EFFECT relations**	**Term**	**Class**	**Number of contexts**	**Number of relations**	**ASSOCIATION relations**	**CAUSE–EFFECT relations**
chemotherapy	Activity	100	14	1	13	chimiothérapie	Activity	100	13	1	12
patient	Conceptual entity	100	4	3	1	patient	Conceptual entity	100	0	0	0
cell	Entity	106	10	1	9	cellule	Entity	100	22	2	20
activation	Process	107	73	5	68	activation	Process	100	69	6	63
oxidation	Process	84	41	1	40	oxydation	Process	54	23	4	19
atherosclerosis	Process (pathology)	85	35	14	21	athérosclérose	Process (pathology)	100	19	10	9
breast cancer	Process (pathology)	99	17	14	3	cancer du sein	Process (pathology)	96	15	8	7
diabetes	Process (pathology)	92	15	5	10	diabète	Process (pathology)	100	30	20	10
tumour	Process (pathology)	100	18	3	15	tumeur	Process (pathology)	99	15	3	12
Total		**873**	**227**	**47**	**180**	**Total**		**849**	**206**	**54**	**152**

Table 22. Relation occurrences per term for non-equivalents in English and French

English						French					
Term	**Class**	**Number of contexts**	**Number of relations**	**ASSOCIATION relations**	**CAUSE–EFFECT relations**	**Term**	**Class**	**Number of contexts**	**Number of relations**	**ASSOCIATION relations**	**CAUSE–EFFECT relations**
hormone replacement therapy	Activity	101	27	12	15	traitement	Activity	100	15	2	13
c-reactive protein	Entity	101	34	15	19	cholestérol	Entity	100	10	3	7
expression	Process	100	36	17	19	coagulation	Process	41	11	0	11
pathogenesis	Process	61	41	3	38	transcription	Process	101	47	3	44
development	Process	99	57	16	41	prolifération	Process	101	55	6	49
coronary heart disease	Process (pathology)	77	20	15	5	récidive	Process (pathology)	100	5	2	3
Total		**539**	**215**	**78**	**137**			**543**	**143**	**16**	**127**

Table 23. Comparison of relation occurrences for equivalent pairs in English and French

	English					French				
Class	**Number of contexts**	**% of total contexts**	**Number of annotated relations**	**% of contexts with relations**	**% of relations from category**	**Number of contexts**	**% of total contexts**	**Number of annotated relations**	**% of contexts with relations**	**% of relations from category**
Activity	100	11.4	14	14.0	6.2	100	11.8	13	13.0	6.3
Entity	206	23.6	14	6.8	6.2	200	23.6	22	11.0	10.7
Process	191	21.9	114	59.7	50.2	154	18.1	92	69.7	44.7
Process (pathology)	376	43.1	85	22.6	37.4	395	46.5	79	20.0	38.3
	873		**227**	**26.0**		**849**		**206**	**24.3**	

Table 24. Comparison of relation occurrences for non-equivalents in English and French

	English					French				
Class	**Number of contexts**	**% of total contexts**	**Number of annotated relations**	**% of contexts with relations**	**% of relations from category**	**Number of contexts**	**% of total contexts**	**Number of annotated relations**	**% of contexts with relations**	**% of relations from category**
Activity	101	18.7	27	26.7	12.6	100	18.4	15	15.0	10.5
Entity	101	18.7	34	33.7	15.8	100	18.4	10	10.0	7.0
Process	260	48.2	134	51.5	62.3	243	44.8	113	46.5	79.0
Process (pathology)	77	14.3	20	26.0	9.3	100	18.4	5	5.0	3.5
	539		**215**	**39.9**		**543**		**143**	**26.3**	

Table 25. Comparison of the proportions of ASSOCIATION (A+) and CAUSE–EFFECT (CE+) relation occurrences in the contexts with equivalent terms in English and French

	EN	FR	Total
A+	47	54	101
CE+	180	152	332
Total contexts	873	849	1722

Semantic classes show rough parallels in their relation densities in the two groups of terms, but the variability is somewhat higher in the group of equivalents than in the set of terms as a whole, and is much higher between the classes of non-equivalent terms. In addition, there was more interlinguistic variation between the semantic classes in terms of the proportions of contexts that were annotated and the proportions of relations associated with each semantic class among the equivalents than among the group as a whole.

As these data were not gathered specifically in order to compare the productivity of equivalent and non-equivalent terms, their scope and nature is not adequate to evaluate the role of equivalence in the discovery of relation occurrences and markers. While the kind of analysis reported above reveals some interesting results, the variations observed between these two groups may be related to a number of factors.

First, any particularities of individual terms and the relations in which they may be involved — and also potentially their associations with markers of these relations — are of course likely to have a greater impact on the results based on a more restricted term set, as each term accounts for a higher proportion of the total occurrences observed.

Second, the effect of the degrees of resemblance between the concepts denoted by these terms — and the semantic proximity of the terms themselves — cannot be evaluated on the strength of these data. Evaluation of a range of terms representing

differing degrees of resemblance would be necessary in order to evaluate the role of equivalence specifically in the kind of variation observed.

Third, the proportions of contexts analyzed belonging to each class vary substantially between the groups of equivalents and non-equivalents (particularly in the case of the processes, the majority of which were not equivalents, and therefore account for a much smaller proportion of the contexts observed in the category of equivalents). This leads to a decrease in the proportion of occurrences of relations involving processes in the equivalents, although this category still provides the highest proportion of the relation occurrences. Given the results of Bodson (2005), focusing on the links between semantic classes, relations and markers of relations, it is thus to be expected that the relations and also potentially pattern sets associated with the two groups will be different, which is likely in turn to create differences in certain characteristics of these data and possibly in some of the difficulties observed in connection with them.

These factors make it challenging to further evaluate the impact of differences in the data gathered using equivalent and non-equivalent terms on the criteria analyzed in this project, as the comparisons of the groups with the overall data and with one another may be affected not only by term choice but also by these other factors. For these reasons, this kind of comparison was considered to be beyond the scope of this project, but nevertheless to suggest very important subjects for future research. Some subjects of and avenues for this further evaluation are discussed further in Section 5.5.2.1.

4.2 Number of markers observed

In order to evaluate the overall possibilities of an approach based on lexical pattern sets, the factor of marker variety — as well as the impact this may have on potential recall of a knowledge-extraction tool — may be analyzed.

As illustrated in Table 26, relative to the number of contexts analyzed, the data sets include comparable numbers of markers. Therefore, these data suggest that (for these relations at least) a pattern discovery approach that begins with the identification of pattern markers in contexts generated using a methodology similar to that used in this project shows a comparable potential for identifying candidate markers in the two languages.

Table 26. Numbers of markers observed relative to contexts analyzed in English and French

Relation	English			French			Difference
	Contexts	Markers	Ratio of markers to contexts	Contexts	Markers	Ratio of markers to contexts	Ratio of markers to contexts
ASSOCIATION	1,412	33	0.02	1,392	30	0.02	0.00
CAUSE–EFFECT		121	0.09		137	0.10	0.01
Total		154	0.11		167	0.12	0.01

A rough parallel was observed in the two corpora in the numbers of distinct markers identified. However, the relations vary markedly in the number of markers observed, with more markers for the CAUSE–EFFECT relation. In both corpora, the most markers were found for the CREATION and MODIFICATION sub-relations.

The markers of ASSOCIATION show less variety than the CAUSE–EFFECT markers, suggesting they are likely to be somewhat more productive than CAUSE–EFFECT markers according to this criterion (although it is, of course, only one among many that are pertinent in evaluating the value of markers for pattern-based KRC extraction). For pattern-based tool design, the figures suggest that while both relations are promising candidates for automatic extraction, the CAUSE–EFFECT relation may require a more involved pattern set design process (given the larger numbers of markers to be included).

Relative to relation occurrences, marker variety shows more interlinguistic differences, as shown in Table 27. For both relations, the number of different markers

relative to the number of relation occurrences is smaller in English than in French. A more detailed analysis of the markers for the CAUSE–EFFECT sub-relations is provided in Table 28. The differences in the proportions of distinct markers compared to numbers of relation occurrences continue to show a trend towards less variety in English, with the only exceptions the MAINTENANCE/PERMISSION and PRESERVATION sub-relations.[97]

Table 27. Numbers of markers observed relative to relation occurrences in English and French

Relation	English			French			Difference
	Occur-rences	Markers	Ratio of occur-rences to markers	Occur-rences	Markers	Ratio of occur-rences to markers	Difference in ratio of occurrences to markers
ASSOCIA-TION	125	33	3.8	70	30	2.3	1.5
CAUSE–EFFECT	317	121	2.6	279	137	2.0	0.6
Total	442	154	2.9	349	167	2.1	0.8

Table 28. Comparison of number of markers and occurrences in English and French

Relation	Number of English occur-rences	Number of French occur-rences	Number of English markers	Number of French markers	Ratio of English occur-rences to markers	Ratio of French occur-rences to markers
ASSOCIATION	125	70	33	30	3.8	2.3
CAUSE–EFFECT	317	279	121	137	2.6	2.0
CREATION	167	133	51	54	3.3	2.5
DESTRUCTION	8	9	5	7	1.6	1.3
MAINTENANCE/ PERMISSION	12	21	11	10	1.1	2.1
PREVENTION	20	18	6	11	3.3	1.6
MODIFICATION	46	48	20	32	2.3	1.5
INCREASE	36	25	14	10	2.6	2.5
DECREASE	27	24	13	12	2.1	2.0
PRESERVATION	1	1	1	1	1.0	1.0
Total	**442**	**349**	**154**	**167**	**2.9**	**2.1**

[97] Note that the very low occurrences for the PRESERVATION sub-relation make this figure difficult to use for generalization.

Thus, a wider variety of markers are used to denote the relations in the French data analyzed. Additional research could further investigate this apparent difference in other corpora and/or using different methodologies. Such research could permit the evaluation of other potential explanations for this difference (for example, related to the corpora evaluated, the methodology used, or the terms used to generate the initial concordances).

4.3 Markers observed

The more frequent markers observed in the term-based concordances (i.e., those observed twice or more) are shown below, in decreasing order of frequency, in tables that also present the number of occurrences observed in the sample concordances and examples of contexts in which they were observed. Full lists of the markers identified, including those observed only once in the sample, appear in Appendix H.

4.3.1 Markers observed in English

4.3.1.1 ASSOCIATION

The markers identified for this relation included the 18 illustrated below in Table 29.

Table 29. English markers observed for the ASSOCIATION relation

Marker	Occur-rences in sample	Sample contexts
associated (with)	17	Diabetes was **associated with** accelerated atherosclerosis at both 14 and 20 weeks of age… (Yan et al. 2003)
risk (of/for/in relation to)	14	There is good evidence that HRT increases the **risk for** VTE… (Kocjan and Prelevic 2003)
risk factor (for/as a ~ for)	10	Hyperhomocysteinaemia is a **risk factor for** the development of CHD. (Mackness et al. 2004)
marker (of/for/ as a ~ of)	9	As carotid IMT is a good early **marker of** atherosclerosis and risk of cerebrovascular ischemic events… (Zambon et al. 2003)

Marker	Occur-rences in sample	Sample contexts
relationship (between… and)	9	… additional randomized clinical trials are necessary to further elucidate the **relationship between** CRP **and** CHD. (Rackley 2004)
in	8	Moreover, these processes are exaggerated **in** diabetes… (Yan et al. 2003)
association (between… and/of… with)	7	Overall, results of our investigation indicate that the **association between** risk of breast cancer **and** HRT varies by regimen. (Weiss et al. 2002)
and	5	CRP **and** Acute Myocardial Infarction The first association between CRP and cardiovascular disease was in the context of… (Shah and Newby 2003)
link (to/with) [VERB]	5	LDL-C remains the primary target of lipid-lowering therapy based on a robust database of studies **linking** LDL-C **to** atherosclerosis and cardiovascular events… (Bittner 2003)
with	5	A further aspect of the change of atherogenicity of lipoproteins **with** HRT was tackled by Wakatsuki et al. … (Seed and Knopp 2004)
related to	4	… the risk of mortality from breast cancer **related to** HRT could not be determined. (Watkins 2003)
correlate (with/… and)	3	… increased circulating IGF-1 concentrations **correlate** very closely **with** the relative risk for the development of several common cancers, including breast, prostate, colon, and lung. (McCance and Jones 2003)
relevant to	3	… lipid-independent effects of statins on various signaling pathways that are potentially **relevant to** the pathogenesis of atherosclerosis. (Balk et al. 2003)
find… in	2	… strong expression of cyclin D1, p21WAF1/CIP1, and Ki-67 **was found in** a DCIS lesion… (Wang et al. 2003))
link between… and [NOUN]	2	Part 1 will provide a brief overview of the **link between** inflammation, endothelial dysfunction, **and** atherosclerosis… (Szmitko et al. 2003)
predict	2	In addition, baseline renal function **predicted** development of CHF. (Coresh et al. 2004)
prediction of	2	High-sensitivity C-reactive protein and the **prediction of** coronary events among patients with renal disease (Torres and Ridker 2003)
relation (between… and/ of… to)	2	… the exact nature of the **relation between** hepatic lipase **and** atherosclerosis remains controversial (Zambon et al. 2003)

4.3.1.2 CAUSE–EFFECT

Of the 121 lexical markers observed for all of the CAUSE–EFFECT sub-relations combined, 52 occurred twice or more in the sample analyzed.

4.3.1.2.1 CREATION

The markers identified for this sub-relation included the 26 shown in Table 30.

Table 30. English markers observed for the CREATION sub-relation

Marker	Occur-rences in sample	Sample context
role (for… in/of… in/in/ play a r~ in/in which… plays a ~)	33	William Osler 3 was one of the first to propose a major **role for** acute infection **in** the pathogenesis of atherosclerosis. (Madjid et al. 2004)
contribute to	13	By studying the normal function of BRCA2, we can understand how changes in the protein **contribute to** the development of cancer… (Graham 2002)
induce	11	hs-CRP has also been reported to **induce** the expression of plasminogen activator inhibitor-1… (Torres and Ridker 2003)
lead to	9	While the ADH3 [gamma]1 allele **leads to** rapid oxidation of ethanol, the [gamma]2 allele results in slow ethanol oxidation. (Humphries et al. 2004)
involved in [PPL.A.]	8	Recently, accumulating evidence has shown that fractalkine is **involved in** the pathogenesis of various clinical disease states or processes, such as atherosclerosis… (Umehara et al. 2004)
implicate in [VERB]	7	There is a large body of evidence that **implicates** inflammation and adhesion molecules **in** the pathogenesis of CVD, including atherosclerosis, stroke, and myocardial infarction. (Granger et al. 2004)
result (in/from) [VERB]	7	The response to injury hypothesis developed by Russell Ross in the late 1970s suggested that atherosclerosis, at least, **resulted from** an initial injury to endothelial cells… (Griendling and FitzGerald 2003a)
mediated (by) [PPL.A.]	6	Endothelial dysfunction and the subsequent changes in blood flow promote CD40-**mediated** endothelial activation by decreasing the intracellular expression of a CD40 signaling blocker. (Szmitko et al. 2003)
cause [VERB]	5	Preoperative chemotherapy often **caused** shrinkage of the tumour… (Shenkier et al. 2004)
importance of… in	5	Third, researchers increasingly recognize the **importance of** nonlipid factors **in** the pathogenesis of atherosclerosis. (Balk et al. 2003)
important in	5	We now appreciate that the fractalkine/CX3CR1 system is **important in** various clinical diseases, such as atherosclerosis, cardiovascular disease, graft rejection, HIV infection, and inflammatory diseases. (Umehara et al. 2004)
pathway (for/in/ as a ~ of)	4	Endothelial dysfunction is a new **pathway in** cardiovascular disease (CVD) development. (Harris and Matthews 2004)
due to	3	… persons scoring higher on a scale of spirituality or religious participation have lower mortality **due to** CHD… (Haskell 2003)

Marker	Occur-rences in sample	Sample context
mediate (by/through/ via)	3	The chemopreventive effects of retinoic acids might **be mediated** via PKC-[delta] activation. (Schondorf et al. 2004)
produce [VERB]	3	Activation of these receptors **produces** endothelium-dependent relaxation of human coronary arteries. (Harris and Matthews 2004)
cause of [NOUN]	2	Atherosclerosis is the leading **cause of** morbidity and mortality in developed countries. (Jaffer and Weissleder 2004)
drive	2	It is presumed that aberrant cyclin D1 expression **drives** the phosphorylation and functional inactivation of pRB in tumor cells. (Sicinski and Weinberg 1997)
implicated in [PPL.A.]	2	... we recently tested whether statins decrease formation of nitric oxide-derived oxidants in vivo [22**], species **implicated in** development of atherosclerosis. (Brennan and Hazen 2003)
induced (by)	2	As is the case for chemotherapy, radiation-**induced** NF-[kappa]B activation has been reported in a variety of cancer cell types... (Garg et al. 2003)
initiate [VERB]	2	Thus, other triggers--including diabetes, high blood pressure, or chemicals in cigarette smoke--can also **initiate** the signals... (Stix 2003)
key... in	2	Oxidation of LDL is a **key** process **in** atherogenesis. (Mason et al. 2003)
mechanism of	2	Further, recent studies implicating translocation of SK1 to the membrane as a **mechanism of** activation have not been demonstrated for SK2. (Saba and Hla 2004)
participate in	2	... is consistent with this heme protein **participating in** the development of atherosclerosis and its thrombotic complications. (Brennan and Hazen 2003)
product of	2	AGEs, the **products of** nonenzymatic glycation and oxidation of proteins and lipids, accumulate in the vessel wall... (Yan et al. 2003)
trigger [VERB]	2	This enhances retention of the lipoprotein and possibly **triggers**, along with oxidation, the formation of a recognizably foreign substance... (Caslake and Packard 2003)
via	2	Lipid oxidation **via** reactive nitrogen species (Brennan and Hazen 2003)

4.3.1.2.2 DESTRUCTION

The markers of DESTRUCTION observed included the two shown below in Table 31.

Table 31. English markers observed for the DESTRUCTION sub-relation

Marker	Occur-rences in sample	Sample context
anti-	3	Administration of Virulizin showed **anti**-tumor efficacy in the treatment of human pancreatic cancers and melanoma... (Du et al. 2003)
against	2	COX-2 inhibition combined with immune-based therapy that would induce cytotoxic T-lymphocyte activity **against** tumor cells is a novel concept that needs further exploration in preclinical animal models and in clinical settings. (Pockaj et al. 2004)

4.3.1.2.3 *MAINTENANCE (PERMISSION)*

The one marker for this relation identified twice in the sample is shown in Table 32.

Table 32. English markers observed for the MAINTENANCE/PERMISSION sub-relation

Marker	Occur-rences in sample	Sample context
required for	2	Therefore, it is currently suggested that ER[alpha] function may be **required for** maximum activation of IGF-signaling pathways. (McCance and Jones 2003)

4.3.1.2.4 *PREVENTION*

Three markers of PREVENTION are illustrated below in Table 33.

Table 33. English markers observed for the PREVENTION sub-relation

Marker	Occur-rences in sample	Sample context
prevent	6	Normally, HDL **prevents** LDL oxidation. (Cabe 2000)
prevention (as... ~/in ~ of/ for ~ of)	6	HRT is effective **for prevention** or treatment **of** osteoporosis... (Kocjan and Prelevic 2003)
suppressor [NOUN]	4	BRCA1 and BRCA2 in their nonmutated forms function as tumor **suppressor** genes. (Khoury-Collado and Bombard 2004)

4.3.1.2.5 *MODIFICATION*

Six markers occurred twice more in the results and are illustrated below in Table 34.

Table 34. English markers observed for the MODIFICATION sub-relation

Marker	Occur-rences in sample	Sample context
effect (of/on/of... on) [NOUN]	12	Recognition of the **effects of** influenza **on** CHD provides the medical community with a valuable opportunity to further reduce cardiovascular death and morbidity. (Madjid et al. 2004)
affect [VERB]	7	In addition, interactions between dihydropyridines and these pathways **affect** lipid oxidation and cholesterol metabolism and can thereby reduce atherosclerosis development. (Mason et al. 2003)
respond to	6	... among ER-positive tumors, nearly 70% of those that are also progesterone receptor (PR)-positive and 25-30% of PR-negative tumors **will respond to** hormonal therapy. (Vogel 2003)

Marker	Occur-rences in sample	Sample context
response (to/ of... to)	5	The conceptual advantage of in vivo assessment of primary tumor **response to** the selected CTX regimen is another benefit derived from the neoadjuvant CTX approach. (Newman et al. 2003)
influence	2	Emerging data reveals that a large number of additional proteins (i.e., growth factors) **influence** the transcriptional activation of ER[alpha] and possibly ER[beta]. (McCance and Jones 2003)
regulated [PPL.A.]	2	TNF-[alpha]-**regulated** SK activation is likely to be important in nuclear factor-[kappa]B (NF-[kappa]B) activation and inhibition of apoptosis. (Saba and Hla 2004)

4.3.1.2.6 INCREASE

Seven markers of INCREASE, shown in Table 35, occurred twice or more in the sample.

Table 35. English markers observed for the INCREASE sub-relation

Marker	Occur-rences in sample	Sample context
promote	10	IL-18 also **promotes** adhesion molecule expression on the endothelium ... (Szmitko et al. 2003)
increase	9	Several recent reports have demonstrated that estrogen therapy **increases** expression of MMP. (Karas 2004)
enhance	2	Lp(a) also **enhances** oxidation of LDL. (Cabe 2000)
facilitate	2	Other preclinical studies show that CRP may **facilitate** the development of atherosclerosis... (Rackley 2004)
increased [PPL.A.]	2	Receptor-mediated leukocyte activation leads to ... **increased** adhesiveness... (Granger et al. 2004)
stimulate	2	... activation of the B2-kinin receptor **stimulates** NO production... (Mason et al. 2003
upregulate	2	Because LDL **upregulates** angiotensin II receptor type 1 (AT1) receptor expression... (Griendling and FitzGerald 2003)

4.3.1.2.7 DECREASE

The markers identified for this relation included the seven illustrated below in Table 36.

Table 36. English markers observed for the DECREASE sub-relation

Marker	Occur-rences in sample	Sample context
reduce	6	... CRP was recently shown to **reduce** synthesis of the vasodilator nitric oxide in cultured endothelial cells. (Rackley 2004)

Marker	Occur-rences in sample	Sample context
inhibit	5	Hydroxy metabolites of atorvastatin… **inhibit** oxidation of both LDL and very-low-density lipoprotein …. (Davignon 2004)
decrease [VERB]	2	NO is an important vasodilator that **decreases** LDL oxidation and smooth muscle cell proliferation. (Torres and Ridker 2003)
downsizing (for… ~/ with)	2	… breast-conserving surgery after tumor **downsizing with** preoperative chemotherapy… (Meric-Bernstam 2004)
inhibition of	2	… free radical-scavenging abilities that may contribute to **inhibition of** lipoprotein oxidation. (Davignon 2004)
lower [VERB]	2	… studies showed that HRT **lowered** low-density lipoprotein (LDL) cholesterol levels… (Aschenbrenner 2004)
reduced [PPL.A.]	2	Loss of ER[alpha] in MCF-7 cells causes **reduced** expression of IGF-signaling molecules... (McCance and Jones 2003)

4.3.1.2.8 PRESERVATION

No markers of PRESERVATION occurred more than once in the sample analyzed.

4.3.2 Markers observed in French

4.3.2.1 ASSOCIATION

The 13 markers identified for this relation that occurred twice or more in the contexts analyzed are shown below in Table 37.

Table 37. French markers observed for the ASSOCIATION relation

Marker	Occur-rences in sample	Sample context
et	10	Traitement hormonal substitutif **et** risque de cancer du sein (Serin and Escoute 1998)
lié à	7	L'hypertension artérielle exacerbe les complications **liées au** diabète, telles que les complications microvasculaires (néphropathie et rétinopathie)… (Gonzalez and Palardy 2004)
facteur de risque	6	… enfants démontrant d'autres **facteurs de risque** cardiovasculaire (obésité, tabagisme, hypertension, diabète, consommation d'aliments riches en matières grasses, prise de médicaments augmentant les lipides plasmatiques…)… (Lambert 2002)
caractérisé par	5	Dans l'adénose sclérosante, affection bénigne du sein **caractérisée par** une prolifération des cellules épithéliales et myoépithéliales… (Angèle et al. 2001)

Marker	Occur-rences in sample	Sample context
risque de	5	L'obésité, le syndrome métabolique et le diabète accroissent notablement le **risque de** maladies cardiovasculaires. (Lambert 2002)
associé à	4	… la dyslipidémie ou des autres troubles fréquemment **associés à** l'athérosclérose (notamment le diabète et l'hypertension). (Gendreau 2003)
lien entre… et	4	… un suivi attentif permettant d'établir les **liens entre** les anomalies lipidiques, le tabagisme, l'hypertension artérielle, le diabète **et** la maladie coronaire. (Bauduceau et al. 2004)
au cours de	2	La vitesse de l'onde de pouls est significativement altérée **au cours du** vieillissement, de l'hypertension artérielle, du diabète et de l'athérosclérose. (Levenson et al. 2000)
corrélé avec	2	… ses changements peuvent être **corrélés avec** une activation ou une répression de la transcription. (Chailleux et al. 2000)
en cas de	2	**En cas de** diabète équilibré, TG et LDL sont quasi normaux, cependant on peut noter un taux de HDL… (Fredenrich et al. 2004)
observé (dans/au niveau de)	2	Par ailleurs, les anomalies qualitatives des lipoprotéines sont similaires à celles **observées** dans le diabète de type 2. (Fredenrich et al. 2004)
prédisposition (de ~ à)	2	Nous présentons ici une mise au point des connaissances sur les gènes **de prédisposition** héréditaire **au** cancer du sein…. (Bonadona and Lasset 2003)
retrouvé dans	2	Le profil lipidique le plus fréquemment **retrouvé dans** le diabète de type 2 associe une élévation du taux plasmatique des triglycérides (TG)… (Fredenrich et al. 2004) à

4.3.2.2 CAUSE–EFFECT

In total, 52 CAUSE–EFFECT markers occurred twice or more in the sample analyzed.

4.3.2.2.1 CREATION

The markers observed for this relation included the 25 shown below in Table 38.

Table 38. French markers observed for the CREATION sub-relation

Marker	Occur-rences in sample	Sample context
conduire à	8	Cette oxydation **conduit à** la déplétion des LDL en antioxydants, en phosphatidylcholines et en esters de cholestérol… (Bonnefont-Rousselot et al. 2002)
entraîner	7	Cette activation **entraîne** de nombreuses réponses cellulaires avec stimulation de la croissance et de la division cellulaire… (Penault-Llorca et al. 2002)

Marker	Occur-rences in sample	Sample context
induire	7	… l'engagement de Fas **induit** la dénitrosylation de la caspase 3... (Kolb 2001)
induit par [PPL.A.]	7	Une hypothèse est que l'activation des récepteurs TP **induite par** les isoprostanes est responsable des effets indépendants des cyclooxygénases. (Cracowski 2004)
participer à	6	Cette prolifération musculaire lisse **participe à** la constitution de la plaque athéroscléreuse… (Teiger 2001)
résulter de (il en ~)	6	… la formation d'adduits hépatiques **résulte de** l'activation des microsomes hépatiques. (Sasco 2000)
activer	5	En conséquence, la caténine ß n'est plus dégradée… et **active** la transcription sous le contrôle de LEF/Tcf. (Blanchard 2003)
exprimer	5	La cellule transfectée produisant du NO endogène **exprimerait** Fas et produirait du FasL autotoxique. (Gauthier et al. 2004)
facteur de	4	À l'opposé, le **facteur de** transcription c-Jun, en se fixant sur le promoteur de son propre gène, contribue à amplifier sa production. (Blanchard 2003)
impliqué dans [PPL.A.]	4	Les ERO formées par la NADPH oxydase des cellules musculaires lisses sont également **impliquées dans** l'activation par la thrombine du facteur de transcription hypoxia-inducible factor-1… (Bonnefont-Rousselot et al. 2002)
provoquer	4	Le dimère ainsi formé se lie au PPRE et **provoque** l'activation de la transcription du gène cible. (Gervois and Fruchart 2003)
rôle (de… dans/jouer un ~ dans/ jouer un ~ lors de/rôle joué par)	4	Le **rôle des** estrogènes **dans** la prolifération des tumeurs mammaires hormonodépendantes a été montré depuis de nombreuses années [1]. (De Crémoux 2000)
stimuler	4	NO **stimule** l'activation de caspases et l'apoptose …. (Kolb 2001)
conséquence de	3	… les maladies métaboliques qui en découlent, c'est à- dire [sic] le diabète, les dyslipidémies et l'hypertension artérielle, sont les **conséquences du** mode de vie adopté par les humains… (Essiambre 2003)
déclenche-ment de (par)	3	Lorsque la plaque est rompue, le **déclenchement de** la coagulation **par** les cellules inflammatoires aboutit à la thrombose… (Collet et al. 2004)
déclencher	3	L'oxydation exagérée des acides gras de ces lipoprotéines modifiées **déclenche** une réaction in- flammatoire [sic]… (Ferrières 2004)
pour	3	… mastectomies subtotales **pour** tumeur maligne… (Lilliu et al. 2002)
produire	3	… elle est ainsi plus fréquente dans les régions riches en cellules **produisant** des cytokines pro-inflammatoires. (Mallat and Tedgui 2004)
responsable de [ADJ.]	3	L'activation des ostéoclastes est **responsable de** l'hyperrésorption osseuse et de la libération de facteurs de dégradation… (Tubiana-Hulin et al. 2001)
à l'origine de	2	L'athérosclérose est **à l'origine de** la plupart des maladies coronaires. (Ferrières 2004)
important (dans/pour)	2	… les c-jun kinases (JNK), **importantes pour** la croissance et la prolifération cellulaire… (Bonnefont-Rousselot et al. 2002)
intervenir dans	2	… d'autres cellules vasculaires **intervenant dans** la pathologie thrombotique, principalement les monocytes… (Drouet 2004)

Marker	Occur-rences in sample	Sample context
médié par	2	… l'effet vasculaire de la 15- F2t-IsoP **est médié par** une activation du récepteur TP (récepteur commun à la prostaglandine H2 et au thromboxane)… (Cracowski 2004)
par	2	… peut réduire de façon significative la mortalité **par** cancer du sein. (Spyckerelle et al. 2002)
réponse à (en ~ à)	2	L'athérosclérose est considérée actuellement comme une **réponse** inflammatoire **aux** lésions de la paroi artérielle. (Duriez 2004)

4.3.2.2.2 DESTRUCTION

The markers of DESTRUCTION observed include the two shown below in Table 39.

Table 39. French markers observed for the DESTRUCTION sub-relation

Marker	Occur-rences in sample	Sample context
anti-	2	… un nouveau traitement **antitumoral** ... (Tubiana-Hulin et al. 2001)
destruction de	2	... des processus qui conduisent à la **destruction de** la cellule. (Chène 1999)

4.3.2.2.3 MAINTENANCE (PERMISSION)

Three of the markers identified for this relation are shown below in Table 40.

Table 40. French markers observed for the MAINTENANCE/PERMISSION sub-relation

Marker	Occur-rences in sample	Sample context
permettre	5	… l'exercice physique **permet** l'oxydation mitochondriale des acides gras au niveau des muscles… (Ferrières 2004)
nécessaire à	4	… l'activation du protéasome est, au contraire, **nécessaire à** l'accomplissement du processus apoptotique… (Kolb 2001)
dépendant (de)	3	Inhibition de la transcription REα **dépendante** de gènes de la prolifération par BRCA1 (Pujol et al. 2004)

4.3.2.2.4 PREVENTION

Four markers for this relation are shown below in Table 41.

Table 41. French markers observed for the PREVENTION sub-relation

Marker	Occur-rences in sample	Sample context
suppresseur de [ADJ.]	4	Avec les gènes RB p53, WTA ou APC, est apparue une première génération de gènes **suppresseurs de** tumeurs. (Bénard 1997)
prévention (de/de ~)	3	Le THS n'est recommandé qu'en cas d'intolérance à un autre traitement indiqué dans la **prévention de** l'ostéoporose et après une évaluation individuelle précise et soigneuse… (Rozenbaum 2004)
bloquer	2	Le tamoxifène **bloque** la prolifération cellulaire qui est rétablie par l'addition d'estrogènes. (Vinatier and Orazi 2003)
préventif	2	… les cellules dendritiques présentent un pouvoir curatif et **préventif** à l'égard de tumeurs greffées. (Catros-Quemener et al. 2003)

4.3.2.2.5 MODIFICATION

Eight markers of MODIFICATION are shown below in Table 42.

Table 42. French markers observed for the MODIFICATION sub-relation

Marker	Occur-rences in sample	Sample context
effet (sur/ de… sur)	7	… l'exercice physique n'a pas eu d'**effet sur** le cholestérol total ou le LDL cholestérol. (Ferrières 2004)
régulation (de/entre… et)	4	…il existe une **régulation** étroite **entre** apoptose **et** prolifération cellulaire … (Lavelle and Jehanno 1998)
moduler	3	Les molécules qui **modulent** sélectivement l'activation des récepteurs hormonaux (SERM)… (Vinatier and Orazi 2003)
anti-	2	… une hormonothérapie (anti-aromatase) ou une chimiothérapie **anti**tubuline… (Guastalla et al. 2004)
commander	2	L'activation de ces récepteurs **commande** la transcription des gènes insulinosensibles… (Leblond 2001)
complica-tion … de	2	Les interactions entre système rénine-angiotensine et **complications** vasculaires **du** diabète constituent un autre exemple de l'implication du TGF-ß. (Michel 2004)
contrôler	2	La prolifération des cellules cancéreuses mammaires **est contrôlée par** les oestrogènes et les facteurs de croissance… (Chailleux et al. 2000)
nuire à	2	Plus besoin non plus du coeur-poumon artificiel, qui dégrade le sang et **nuit à** sa coagulation. (Simard and Dussault 1997)

4.3.2.2.6 INCREASE

The markers identified for this relation included the five shown below in Table 43.

Table 43. French markers observed for the INCREASE sub-relation

Marker	Occur-rences in sample	Sample context
favoriser	7	...l'expression de Cox2 **favorise** la prolifération tumorale en inhibant l'apoptose... (Guastalla et al. 2004)
augmenta-tion de	4	L'activation de récepteurs endothéliaux produit une **augmentation de** [Ca]i dans les cellules endothéliales... (Feletou et al. 2003)
augmenter	3	...une chimiothérapie d'induction peut **augmenter** les possibilités de chirurgie ... (Lerouge et al. 2004)
accroître	2	Le traitement hormonal substitutif **accroît** l'incidence du cancer du sein. (Noël et al. 1998)
faciliter	2	Si l'on considère que les macrophages peuvent **faciliter** la prolifération des cellules musculaires lisses ... (Caligiuri 2004)

4.3.2.2.7 DECREASE

The markers identified for this relation included the five shown below in Table 44.

Table 44. French markers observed for the DECREASE sub-relation

Marker	Occur-rences in sample	Sample context
inhiber	7	... l'activation de la protéine G Arf par ses GEF à domaine Sec7 peut **être inhibée par** stabilisation de complexes abortifs Arf-GD... (Cherfils and Pacaud 2004)
diminuer	3	..., les inhibiteurs du système rénine-angiotensine **diminuent** la prolifération intimale des cellules musculaires lisses .... (Michel 2004)
réduire	3	Ces médicaments non seulement **réduisent** le cholestérol plasmatique et ses dérivés, mais aussi ont des effets " pléïotropes "... (Asmar et al. 2003)
inhibiteur de [ADJ.]	2	L'activité paraoxanase 1 **inhibitrice de** l'oxydation des LDL est très diminuée chez les patients ayant des antiphospholipides. (Meyer 2001)
inhibition de (... par)	2	**Inhibition de** la transcription REα dépendante de gènes de la prolifération **par** BRCA (Pujol et al. 2004)

4.3.2.2.8 PRESERVATION

None of the markers observed occurred twice or more in the sample analyzed.

4.4 Number of occurrences of markers

In order to evaluate the potential performance of the markers, their frequency as observed in the sample and in the whole corpora was evaluated. These measures are presented below.

4.4.1 Number of occurrences of markers in the samples

Given the numbers of markers observed in the samples analyzed, as discussed above in Section 4.2, it is not surprising that the French markers had lower mean frequencies in the relation occurrences analyzed than the English ones.

Since in the context of designing pattern-based tools a selection of the most promising patterns located may be chosen for inclusion in a pattern set — according to criteria that may include the number or proportion of desired relation occurrences that can be retrieved by a given marker[98] — another way of evaluating the relative frequencies of the markers observed involves the comparison of the numbers of markers required to retrieve a given proportion of the relation occurrences observed in the two data sets.

If the CAUSE–EFFECT markers are ranked from most to least frequent in the sets of relation occurrences evaluated (in order to target those that appear most productive for retrieving occurrences of the relations according to these data), in English the top 17 markers account for 50% of the relation occurrences observed, and the top 46 for 75%. In French, retrieving 50% of the occurrences would require the top 30 markers, and 75% would require 73. The markers required to retrieve 50% of the occurrences are shown in Table 45 and Table 46.

---

[98] Other criteria in pattern choice of course include the marker's potential for recall (i.e., overall frequency in corpora), precision, and the ease with which pattern forms may be developed for that marker.

Table 45. Most frequent CAUSE–EFFECT markers: Markers required to retrieve 50% of English relation occurrences

Marker	Occurrences in sample	% of occurrences
role	33	10.4
contribute to	13	4.1
effect	12	3.8
induce	11	3.5
promote	10	3.2
lead to	9	2.8
increase	9	2.8
involved in	8	2.5
implicate	7	2.2
result from	7	2.2
affect	7	2.2
mediated	6	1.9
prevent	6	1.9
prevention	6	1.9
respond to	6	1.9
reduce	6	1.9
cause	5	1.6
Total	161	139

Table 46. Most frequent CAUSE–EFFECT markers: Markers required to retrieve 50% of French relation occurrences

Marker	Occurrences in sample	% of occurrences
conduire à	8	2.9
entraîner	7	2.5
induire	7	2.5
induit par	7	2.5
effet	7	2.5
favoriser	7	2.5
inhiber	7	2.5
participer à	6	2.2
résulter de	6	2.2
activer	5	1.8
exprimer	5	1.8
permettre	5	1.8
facteur de	4	1.4
impliqué dans	4	1.4
provoquer	4	1.4
rôle	4	1.4
stimuler	4	1.4
nécessaire à	4	1.4
suppresseur de	4	1.4
régulation	4	1.4
augmentation de	4	1.4

conséquence de	3	1.1
déclenchement de	3	1.1
déclencher	3	1.1
pour	3	1.1
produire	3	1.1
responsable de	3	1.1
dépendant	3	1.1
prévention	3	1.1
moduler	3	1.1
Total	140	151

If the ASSOCIATION markers are ranked from most to least frequent in the sets of relation occurrences evaluated, in English the top 6 markers account for 50% of the relation occurrences observed, and the top 12 for 75%. In French, retrieving 50% of the occurrences would also require the top 6 markers, and 75% would require 13. The markers required to retrieve 50% of the occurrences are shown in Table 47 and Table 48.

Table 47. Most frequent markers of ASSOCIATION: Markers required to retrieve 50% of the English relation occurrences

English		
Marker	**Occurrences in sample**	**% of occurrences**
associated	17	13.6
risk	14	11.2
risk factor	10	8.0
marker	9	7.2
relationship	9	7.2
in	8	6.4
Total	67	68

Table 48. Most frequent markers of ASSOCIATION: Markers required to retrieve 50% of the French relation occurrences

French		
Marker	**Occurrences in sample**	**% of occurrences**
et	10	14.3
lié à	7	10.0
facteur de risque	6	8.6
caractérisé par	5	7.1
risque de	5	7.1
associé à	4	5.7
Total	37	69

These data show that for the CAUSE–EFFECT relation, the number of markers required by a pattern-based tool to locate a given proportion of the relation occurrences observed in the corpus would be lower in English than in French. (Although a large part of this discrepancy can be traced to the most frequent marker, *role*, the overall trend is still visible without this marker.) Such a trend would clearly have implications for the creation of pattern sets in the two languages, as the investment of time and energy in creating pattern forms for markers would be increased in French if the choice was made to include more markers. Conversely, the performance of an application could be poorer in French if this choice was not made.

However, the relation of ASSOCIATION shows far less variation, suggesting that differences in pattern variety are less likely to raise questions for pattern design and application performance in the case of this relation.

Further research could evaluate this apparent difference in light of results in other corpora and/or using another methodology, to determine whether this trend is widely observed or whether particularities of the corpus, methodology or data retrieved have contributed to these observations. It would also be interesting to evaluate individual markers more specifically to determine their contribution to the difference observed.

4.4.2 Number of occurrences of markers in the corpora

Evaluating the number of occurrences of the markers in the sets observed in the corpora as a whole can indicate their overall potential for retrieving contexts: markers that occur more frequently in the corpora will give access to more potentially useful contexts. Table 49 presents the total frequencies per 1,000 corpus tokens for the marker sets for

each relation and sub-relation in each of the languages (based on data provided in Appendix H).[99]

Table 49. Comparison of total occurrences of markers in sets per 1,000 corpus tokens in English and French

Relation	English	French	Difference
ASSOCIATION	80.0	47.7	32.3
CAUSE–EFFECT	52.6	50.3	2.3
CREATION	24.6	22.7	1.9
DESTRUCTION	1.9	1.9	0.0
MAINTENANCE/ PERMISSION	1.5	2.5	-1.0
PREVENTION	1.4	0.9	0.5
MODIFICATION	16.3	17.9	-1.6
INCREASE	3.6	2.2	1.4
DECREASE	3.3	2.7	0.6
PRESERVATION	0.05	0.1	-0.05
Total	**132.5**	**98.1**	**34.4**

In these results, the trend towards higher frequencies in English continues overall and for the two relations individually, with only the CAUSE–EFFECT sub-relations of MAINTENANCE/PERMISSION, MODIFICATION and PRESERVATION showing higher values in French.[100] Both far higher frequencies and a substantial difference were observed in the case of the ASSOCIATION relation (likely due to very common prepositional markers, particularly in English, as well as markers such as *risk* and *risk factor* in English and *risque* and *facteur de risque* in French, also extremely frequent in the corpora). Lower frequencies and smaller differences were observed for the CAUSE–EFFECT relation and its sub-relations.

A few comments, however, should be made on this subject. First is that these statistics are based on simple numbers of tokens as calculated by WordSmith Tools, and

---

[99] As noted above in Section 3.3.1.4.3, the evaluation of marker frequency expressed in occurrences per 1,000 corpus tokens allows for comparison in corpora of varying sizes.

[100] Moreover, the single marker observed for PRESERVATION in each corpus does not allow for generalization.

the generally higher prevalence of articles and prepositions in French no doubt affects this measure. Although these observations may be used as a guideline for estimating productivity in corpus size measured using such means, an evaluation of frequency that takes into account this kind of variation would provide a more exact picture of the potential for variation between corpora in English and French. A second observation is that of course these figures do not determine overall productivity: the pattern forms used, for example, affect how many contexts are retrieved by a pattern-based tool, and the precision of each marker how many of these contain the desired relation.[101]

Nevertheless, on the basis of the data gathered in this analysis, and considering the generally more numerous markers, more even distribution of relation occurrences among the markers and lower marker frequency per 1,000 corpus tokens in the French data, it appears that — at least for the CAUSE–EFFECT relation — in order to access the same number of potentially useful contexts in the two languages (i.e., to achieve the same potential for identifying relation occurrences) more French markers may be required. The increased number of markers required would also be accompanied by an increase in the number of pattern forms required to exploit them.[102] The use of more markers would thus be likely to involve a significant investment of time and effort on the part of application developers. A larger corpus could also be used to equalize the numbers of occurrences retrieved, although this would increase silences as well as hits.

The differences observed in the two relations, however, indicate that the ASSOCIATION relation is likely to be less affected by the general differences in marker frequency and the distribution of relation occurrences among the markers. Marker sets that are comparable in number in the two languages are likely to show more similar performance for this relation than for that of CAUSE–EFFECT.

---

[101] See Section 4.6 for a discussion of precision in a sample of markers in the two corpora.

[102] Moreover, given that individual pattern forms containing a given marker may vary, this increase may be not linear, but rather exponential, with multiple pattern forms required to exploit a given marker.

These observations must nevertheless be further investigated, preferably using other corpora, and certainly using methodologies that will allow for more precise statistical evaluation of these criteria, in order to confirm whether the trends observed in this study are widespread. It will be important to evaluate and/or neutralize other potential sources of variation linked, for example, to the corpora used and their content, the sample of data retrieved, or the methodology used to retrieve it (e.g., the choice of terms for retrieving the contexts evaluated).

In developing pattern sets, it could also be productive, for example, to target particular types of markers according to usage observed in the languages. The characteristics of the markers observed, as evaluated in Section 4.5, may provide some data to help in targeting particularly useful types of markers or marker forms.

4.5 Types of markers observed

As discussed above in Section 3.3.1.4.4, the markers identified were characterized in a number of ways, including their part of speech classes and their form.

4.5.1 Part of speech class of markers

Interlinguistic variation was also observed in the independent analysis in the parts of speech of the markers observed. This analysis (presented in Appendix I) provides information about general tendencies, and may also allow for the identification of a potential link with marker precision (discussed below in Section 4.6).

4.5.1.1 Individual markers

As shown in Table 50, the proportions of markers belonging to individual part of speech classes showed basic parallels between the two data sets, with the verbal (and participial

adjective) markers most prevalent, followed by nominal markers, adjectives and adverbs, function words, and finally affixes.[103]

Table 50. Comparison of proportions of markers belonging to various POS classes in English and French

POS	Both relations		ASSOCIATION		CAUSE–EFFECT	
	English	French	English	French	English	French
Nouns, Noun phrases	48 (31%)	54 (31%)	12 (36%)	8 (27%)	36 (30%)	46 (34%)
Verbs, Verb phrases, Participial adjectives, Participial adjective phrases[104]	84 (55%)	77 (47%)	16 (48%)	10 (33%)	68 (56%)	67 (49%)
Adjectives, Adjective phrases, Adverb phrases[105]	11 (7%)	20 (12%)	2 (6%)	4 (13%)	9 (7%)	16 (12%)
Prepositions, Conjunctions	9 (6%)	13 (8%)	3 (9%)	8 (27%)	6 (5%)	5 (4%)
Affixes	2 (1%)	3 (2%)	0	0	2 (2%)	3 (2%)
Total	154	167	33	30	121	137

The dominance of verbal markers in both languages is evident, indicating that these are likely to be promising subjects for developing pattern-based applications for both CAUSE–EFFECT and ASSOCIATION relations in the two languages, and especially promising in English. However, it is apparent that nominal markers are also good candidates for research, particularly for the ASSOCIATION relation. In light of these observations, it is clear that research in pattern identification and pattern set

---

[103] If the verbs and participial adjectives are considered separately, the rank order of the categories is not identical, but strong similarities are noted.

[104] If this group is broken down internally, 67 (44%) of the English markers are verbs and 17 (11%) participial adjectives, and in French these figures are 62 (37%) and 15 (9%). For the ASSOCIATION relation, in English 14 (42%) verbs were observed, and 2 (6%) participial adjectives, and in French the figures were 6 (20%) and 4 (13%). For the CAUSE–EFFECT relation, in English 53 (44%) verbs were observed, and 15 (12%) participial adjectives, and in French the figures were 59 (42%) and 9 (7%).

[105] The categories of adjectives and adverbs are considered together for the purposes of this research. However, adjectives are far more prevalent than adverbs. (See Appendix I for details.)

development should certainly focus on both of these part of speech classes for the two relations studied here, although verbs are likely to be particularly prevalent.

When the distribution across all of the categories is compared overall in the two data sets (with the categories of function words and affixes combined to allow for accurate testing using the Chi-square test), no significant difference is observed (p = 0.299).[106] The percentages of markers for the two relations together that belong to the POS classes of nouns and noun phrases, function words and affixes remain relatively parallel between the two data sets, although a slightly higher proportion of function words was noted in French.

However, differences in the percentages of adjectival or adverbial and verbal markers individually are more apparent, with a higher proportion of the former in French and the latter in English. However, when the Chi-square test is applied, no statistically significant variation is observed in these proportions (p = 0.143 and p = 0.131 respectively). It may nevertheless be interesting to continue to evaluate these potential differences, to determine whether more data could reveal a difference that should be taken into account in pattern and marker discovery projects (i.e., in specifying structures for analysis) and in the design of pattern-based tools (e.g., in planning the types of markers and structures to include in order to maximize recall, and adjusting strategies to deal with any variations that may be associated with differences in marker POS).[107]

Although differences were observed in the proportions of the POS classes of markers in each relation, these variations also showed parallels in the data sets.

---

[106] If the verbs and participial adjectives are considered separately, $p = 0.450$.

[107] Such differences could involve influence of marker POS on the types of structures in which markers occur, the complexity of representing these structures, and possibilities of further processing contexts using additional components of the contexts.

For the ASSOCIATION relation, the rank order of the classes remained similar. When analyzed using the Chi-square test, the proportions of all of the classes considered together (with the function word and affix categories combined) were not significantly different ($p = 0.172$).[108] Further, when the classes are compared individually, none of the categories shows statistically significant differences. Nevertheless, once again in French the proportions of adjectives and function words are slightly higher (with the difference in the case of the function words trending towards significance, $p = 0.066$), and the proportion of verbs is slightly higher in English.[109] Moreover, slightly more nouns were observed in English. These latter results suggest that more data could reveal some interesting variations, and that taking these into account in pattern discovery and pattern set development could be important. French function word markers specifically were prevalent in the ASSOCIATION relation, suggesting that at least this category would be another promising avenue for the identification of markers for this relation in that language. The potential need to take more types of markers into account could complicate pattern set development, and would likely have significant effects on the time and effort required for this task.

Moreover, the fact that these differences are more pronounced than those observed overall suggests that the ASSOCIATION relation is an important contributor to overall differences, and could be targeted for further research.[110]

For the CAUSE–EFFECT relation, a perfect correlation between the ranks of the POS classes of the markers is observed, with the verbal markers most numerous,

---

[108] However, the expected values for the adjective and adverb category were too low for the Chi-square test to be considered strictly valid; if the categories of adjectives/adverbs, function word and affixes are combined and contrasted with the nouns and verbs, this value increases to $p = 0.235$. The expected values for participial adjectives were too low for this category to be considered separately from verbs for this relation.

[109] When the verbs are considered separately from participial adjectives, the higher prevalence of verbal markers in English trends toward significance ($p = 0.056$).

[110] Given the relatively low numbers of markers observed for this relation, more data would certainly be necessary to obtain relevant and reliable conclusions.

followed by the nominal and then adjectival or adverbial markers, function words, and finally affixes. When the Chi-square test is applied, for all categories considered together (with the function words and affixes combined to allow for accurate results), no significant difference was observed ($p = 0.531$).[111,112]

Further, no statistically significant difference was observed for the individual classes.[113] However, the somewhat higher proportion of adjectival and adverbial markers in French for this relation, similar to those observed overall and for the ASSOCIATION relation, suggests that this phenomenon is one that would be particularly interesting to evaluate in light of more data, as it could indicate that markers belonging to this category should be considered in French. The trend towards higher prevalence of verbs in English, although fairly subtle, could also be interesting to investigate, to determine what effect this might have on the productivity of approaches that focus on verbal markers in the two languages.

The relative similarity in the distribution of markers among POS classes indicates that pattern sets that are similar in regard to this characteristic may be located using the kind of pattern discovery approach used in this project, and that in order to reflect usage in the two languages, candidate pattern sets may include similar types of markers in relatively similar proportions. However, the presence of minor variations suggests that some categories (for example adjectival and adverbial markers) could be interesting to investigate further in light of more data.

From another perspective, the choice to target markers of a specific part of speech class for identification and development may have relatively — but certainly not exactly — comparable effects on the potential of a pattern-based tool for retrieving contexts in the two languages. For example, if the choices made in some projects to

---

[111] Including the category of affixes, which are too few for accurate Chi-square testing, $p = 0.645$.

[112] If the verbs and participial adjectives are considered separately, $p = 0.416$.

[113] This is still the case when verbs are considered separately from participial adjectives.

consider only verbal markers of the CAUSE–EFFECT relation are considered, it becomes clear that while this category is the most prevalent, a significant proportion of other types of markers — particularly nouns — are used in both languages, and a number potentially useful relation occurrences would thus be excluded. However, it appears that the impact of such a choice may affect French somewhat (but as far as these data indicate, not significantly) more than English, signalling a potential problem with obtaining comparable results in the two languages using this kind of approach. This is worth evaluating on the strength of more data to determine if a significant difference may be observed.

The study of marker POS distribution in further projects may not only provide more data to assist in confirming whether the variations observed become significant, but may also help to eliminate other potential sources of variation (for example, related to the content of the corpora or the terms chosen to identify the contexts, which may affect the results directly or indirectly).

4.5.1.2 Marker occurrences

The proportions of markers belonging to each class are only part of the equation in the design of pattern sets; the comparison of proportions of marker occurrences complements these observations, reflecting not only the types of markers that may be used but also their potential for productivity in identifying relation occurrences.

The proportions of relation occurrences associated with each POS class of markers (Table 51) shows general parallels with the proportions of the individual markers observed, with verbs most prevalent, followed by nouns. These data indicate even more strongly than the distribution of the individual markers that both the categories of verbs and nouns should be considered in identifying candidate markers for locating both relations. The prevalence of nominal markers for ASSOCIATION in English particularly underlines their importance in this relation and language, but the loss of potentially useful contexts that would occur if this category of markers were excluded

from consideration even in the case of the CAUSE–EFFECT relation is obvious. The prevalence of function word marker occurrences for ASSOCIATION also suggests that this category might be interesting to take into account in developing pattern sets, although it is still far more prevalent in French.

Table 51. Comparison of proportions of occurrences of markers of various POS classes in English and French

POS	Both relations		ASSOCIATION		CAUSE–EFFECT	
	English	French	English	French	English	French
Nouns, Noun phrases	159 (36%)	96 (28%)	59 (47%)	21 (30%)	100 (31.5%)	75 (27%)
Verbs, Verb phrases, Participial adjectives, Participial adjective phrases[114]	232 (52%)	183 (52%)	44 (35%)	27 (39%)	188 (59%)	156 (56%)
Adjectives, Adjective phrases, Adverb phrases	21 (5%)	38 (11%)	4 (3%)	4 (6%)	17 (5%)	34 (12%)
Prepositions, Conjunctions	26 (6%)	27 (8%)	18 (14%)	18 (26%)	8 (2.5%)	9 (3%)
Affixes	4 (1%)	5 (1%)	0	0	4 (1%)	5 (2%)
Total	442	349	125	70	317	279

Overall, the rank order of the POS classes shows a certain degree of positive correlation, with the classes of adjectives/adverbs and function words varying between third and fourth rank in the two data sets. However, in the Chi-square test of the proportions of marker occurrences belonging to each POS category for the two relations together, the languages showed a significant difference in the types of markers observed ($p = 0.001$).[115]

---

[114] If this group is broken down internally, 175 (40%) of the English marker occurrences were of verbs and 57 (13%) of participial adjectives, and in French these figures were 140 (40%) and 43 (12%). For the ASSOCIATION relation, in English 21 (17%) occurrences of verbs were observed, and 23 (18%) of participial adjectives, and in French the figures were 23 (33%) and 4 (6%). For the CAUSE–EFFECT relation, in English 152 (48%) occurrences of verbs were observed, and 36 (11%) of participial adjectives, and in French the figures were 136 (49%) and 20 (7%).

[115] If the classes of verbs and participial adjectives are considered separately, $p = 0.004$.

The most pronounced variation is found in the category of adjectival/adverbial markers, which when evaluated separately using the Chi-square test are much more frequent in French than in English ($p = 0.001$). The proportions of nouns are also different, with the English significantly higher than the French ($p = 0.011$), while no significant difference was observed in the other categories. Interestingly, while the verbal markers were more numerous in English than in French, the difference is reduced in the comparison of occurrences, suggesting that the French verbal markers observed are slightly more productive than their English counterparts.[116]

When compared to the proportions of individual markers observed, these data indicate that the more numerous adjectival/adverbial markers observed in French are even more significant in the proportions of occurrences, and support the argument that taking into account the types of markers observed in a given language (e.g. the higher prevalence of adjectival/adverbial markers in French) is important in the design of pattern sets. Pattern sets that do not include adjectival markers such as those observed in this project would likely miss a higher proportion of relation occurrences in French than in English. Although no significant difference was observed in the numbers of nominal markers, a comparison of nominal marker occurrences indicates that these markers were somewhat more productive in the English data than the French. These observations underline the possibility that even if comparable proportions of the various types of markers are observed in the two corpora, the performance of a given type of marker may differ. Thus, the evaluation of the proportions of occurrences containing different kinds of markers in the (types of) corpora being evaluated may be very important in the choice of pattern-based application approaches.

The rank order of the marker occurrences corresponding to the POS classes for the ASSOCIATION relation showed a weak positive correlation. The categories of verbal and nominal markers varied between first and second rank in the two data sets. In the

---

[116] There is no change in these results when verbs and participial adjectives are considered separately.

distribution of marker occurrences among the POS classes for the ASSOCIATION relation, a trend towards statistically significant variation was observed overall (p = 0.066, although the low expected frequency of the adjectival markers may interfere minimally with the accuracy of the Chi-square test).[117] When the individual classes are compared, a significantly higher proportion of noun markers is found in English (p = 0.019), and there is an extremely strong trend towards significance in the higher prevalence of function word occurrences in French (p = 0.051). The other categories do not show significant differences, indicating that the nominal markers are the primary source of the overall variation, although the function words do contribute.[118] This once again underlines the greater difference seen between the languages in terms of relation occurrences, and the impact that the POS of markers included in pattern sets may have on the performance of tools.

The rank order of the POS classes in the two data sets for the CAUSE–EFFECT relation showed perfect positive correlation, with the verbal markers in first place, followed by nominal and adjectival markers, with function words in fourth place and affixes fifth. Overall, a significant difference in the distribution of marker occurrences among the POS classes for the CAUSE–EFFECT relation was observed (p = 0.019 when the rarer category of affixes was merged with the function words to permit more accurate Chi-square testing).[119] For this relation, most marker categories considered individually do not show significant differences, but the proportion of

---

[117] When the categories of verbs and participial adjectives are considered separately and the smaller values (adjectives, function words, affixes) collapsed to allow for accurate testing using the Chi-square test, a more significant difference is observed (p = 0.001).

[118] When the categories of verbs and participial adjectives indicating ASSOCIATION are considered separately, some differences in distribution may be noted. The proportion of verbs is significantly higher in French (p = 0.010), while the proportion of participial adjectives is significantly higher in English (p = 0.014).

[119] This figure is p = 0.040 if the category of affixes is not merged with that of function words. If the verbs and participial adjectives are considered separately, p = 0.014.

adjectives/adverbs was significantly higher in French ($p = 0.003$).[120] This once again suggests that this class of pattern markers is especially important to take into account when designing pattern sets in French, or the recall of tools is likely to suffer.

In these data, the relation being analyzed was very closely linked to the level of inter-corpus variation observed. Less variation was observed for the CAUSE–EFFECT relation than the ASSOCIATION relation (although in terms of statistical significance, the lower number of markers and of occurrences no doubt influenced the results). In both data sets roughly parallel inter-relational variations in the proportions of marker POS classes were observed, with a higher proportion of verbal markers and verbal marker occurrences in the CAUSE–EFFECT relation, and a more even distribution between the nominal and verbal categories for the ASSOCIATION relation.

Nevertheless, some interlinguistic differences observed indicate the importance of considering the types of markers used in each language and the proportions of relation occurrences that may be indicated by different classes of markers. The results suggest that potential differences (for example, in the use of relation markers in adjective form) may be important to take into account in pattern discovery (as limiting the search for markers to a certain class may affect one language more than another), pattern set development (as the pattern sets retained should reflect the usage in each language as closely as possible to ensure comparable recall) and application performance (as decisions made in the prior two steps are likely to influence recall and even potentially precision, and certain types of markers may be more productive in one language than another). Moreover, the potential for variation between the proportions of markers and of marker occurrences associated with each POS class indicates the importance of evaluating both of these factors in order to more accurately identify the effect of POS distribution on pattern set design and performance.

---

[120] If the verbs and participial adjectives are considered separately, there is also a trend towards higher prevalence of participial adjective markers of the relation in English ($p = 0.080$).

Additional data on the proportions of markers and occurrences in each POS class may reveal further differences, and allow for a more comprehensive look at trends in each language and for each relation.[121] Testing in other corpora and/or using another methodology could also help to eliminate the possibility of additional sources of variation related to these factors.

4.5.2 Simple and complex markers

As discussed in Section 2.6.1, complex marker forms may encounter difficulties not generally seen with simple markers, such as their interruption by external elements (Section 4.10.1.2) and variation in the order of their elements (cf. Section 4.8.1).

Despite the formal challenges posed by complex forms, and particularly in the case of nominal and participial adjective pattern markers, contexts containing such marker forms may be more likely to provide complete occurrences of relations than those containing simple forms. The explanation for this becomes relatively obvious when contexts containing the different types of markers are compared. For nominal markers, variants such as those observed in Examples 15 to 17 may be observed.

15. These were the first findings demonstrating conclusively that heat shock protein **induction** in the intact heart was able to produce a protective effect against subsequent exposure to ischemia and reperfusion… (Gupta et al. 2004)
16. The specific **induction of** cyclin D1 in the mammary epithelium of pregnant animals raised the possibility that… (Sicinski and Weinberg 1997)
17. … **induction of** apoptosis in MCa-35 and A549 tumor cells **by** celecoxib or **by** radiation… (Liu et al. 2003)

---

[121] The advisability of further evaluation is highlighted by the fact that these results reflect the distribution observed in Marshman (2002; cf. Section 2.3.1.2) only to a certain extent (with a slightly higher proportion of individual, verbal markers in English and of nouns in French). Differences in approaches to evaluating the proportions of markers in each class may have contributed to this difference, as Marshman (2002) focused on identification of candidate markers in character string form (which thus could be associated with multiple parts of speech) and thus compared the part of speech classes of the items in the corpora that could be retrieved using these strings, rather than the proportions of markers or occurrences observed in the concordances analyzed to identify the markers.

In Examples 15 and 16, the contexts provide only one of the elements linked by the relation (the element that is created: *heat shock protein* and *cyclin D1*), while Example 17, both elements involved in the induction (i.e., the cause and the effect) are specified. In observations in both languages, occurrences of the most complete marker forms were generally the most promising for observing complete relations, although some individual markers may behave somewhat differently (e.g., in English, the simple, participial adjective markers *stimulated* and *induced* are very precise in the pattern structure X-[MARKER] Y).

Although complex markers are likely to be particularly useful, representing them in pattern design is more challenging: character strings or regular expressions that represent markers, for example, have to take into account phenomena such as the interruption of the marker form by external elements, either in fairly regular form, by one of the elements it links (e.g., by *apoptosis* in Example 17), or by external elements (also seen in Example 17). Problems in accounting for these phenomena (e.g., in not allowing for a long enough interruption of the marker, or for the type of interruption present) may also interfere with KRC recognition in a significant proportion of cases.

The proportions of occurrences of simple and complex marker forms in the English and French data (Table 52) were fairly parallel, with only a 1% variation between the languages (60% complex in English and 59% in French). The Chi-square test confirms that the difference in the two data sets is far from significant ($p = 0.842$).

Table 52. Comparison of proportions of complex and simple marker occurrences in English and French

	EN	FR	Total
Simple	178	143	321
Complex	264	206	470
Total	442	349	791

As shown in Table 53 and Table 54, in neither of the relations did the proportions of simple and complex markers show a statistically significant difference

(75% complex in English versus 77% in French for the ASSOCIATION relation and 54% complex in both English and French for the CAUSE–EFFECT relation; with Chi-square test results of $p = 0.734$ and $p = 0.835$ respectively). Moreover, the inter-relational differences were parallel in the two data sets, suggesting that this aspect of pattern marker form will vary more from relation to relation than from language to language.

Table 53. Comparison of proportions of simple and complex marker occurrences for the ASSOCIATION relation in English and French

	EN	FR	Total
Simple	31	16	47
Complex	94	54	148
Total	125	70	195

Table 54. Comparison of proportions of simple and complex marker occurrences for the CAUSE–EFFECT relation in English and French

	EN	FR	Total
Simple	147	127	274
Complex	170	152	322
Total	317	279	596

The results indicate that the types of pattern markers that may be integrated into pattern-based tools for the two languages are likely to resemble one another in this aspect of their form, and therefore also to involve comparable complexity in the development of pattern forms and to confront some of the same difficulties.

As the distribution of simple and complex markers differed substantially between the two relations (with the proportions in both data sets for the CAUSE–EFFECT relation more even than those for the ASSOCIATION relation, for which complex markers were far more prevalent), the complexity of designing pattern forms for ASSOCIATION and the prevalence of difficulties associated with these forms seem likely to be higher in both languages.

4.6 Marker precision

Measurements of marker precision (i.e., the proportion of contexts retrieved using a marker that express the desired relation) complement the evaluations of markers described above, indicating the efficiency with which these markers retrieve useful contexts. In this project, the precision of a set of 13 of the most frequently observed markers in English and French was evaluated.

The markers retained for initial analysis, as well as the character strings used to generate the concordances, are shown in Table 55 and Table 56. The sample analyzed — a total of 2,549 randomly selected contexts (1,300 in English and 1,249 in French) containing occurrences of 13 distinct markers in each language[122] — can be further sub-divided according to criteria such as the relation indicated and the part of speech class of the marker. It provides data that may identify trends in the performance of markers in the two corpora that are worthy of evaluation on a larger scale.[123] The sample represents approximately 8.5% of the 154 English markers observed and 8% of the 167 French markers. However, as these are the most frequently observed markers, they account for 162 or 37% of the relation occurrences observed in English and 90 or 26% in French.

Table 55. List of English markers used to for evaluating precision

Marker	Character string[124]
ASSOCIATION	
associated	associated
risk	risk/risks
risk factor	risk factor/risk factors
marker	marker/markers

[122] The French marker *résulter de* was less than 100 times in the corpus, and thus the sample for this marker is smaller.

[123] Of course the restricted size of the sample does not allow for broad generalizations. More data from a wider variety of markers will be essential to establish the consistency of the potential variations noted.

[124] In these representations of character strings, / represents alternative forms (the equivalent of the operator OR), * represents a wildcard character replacing zero, one or many characters, and *NOT* represents the exclusion of the forms that follow.

CAUSE–EFFECT	
role	role/roles
contribute to	contribut* to NOT contribution* to/contributor* to
effect	effect/effects
induce	induc* NOT induction*
promote	promot* NOT promotion*/promoter*
increase	increas*
lead to	lead* to/led to
involved in	involved in
result	result*

Table 56. List of French markers used for evaluating precision

Marker	Character string
ASSOCIATION	
et	et
lié à	lié* à/ lié* au/ lié* aux
risque	risque/risques
facteur de risque	facteur de risque*/facteurs de risque*
CAUSE–EFFECT	
conduire à	condui* à/ condui* au/ condui* aux
effet	effet/effets
entraîner	entraîn*
favoriser	favoris*
induire	indui*
induit par	induit* par
inhiber	inhib* NOT inhibition*/inhibiteur*
participer à	particip* à/ particip* au/ particip* aux
résulter de	résult* de/résult* du/résult* des NOT résultat* de/ résultat* des/résultat* du

The results of the evaluations are presented in Table 57 and Table 58, which break down the contexts analyzed into those that were considered valid hits (i.e., that presented complete relations of interest in the research), those that involved complex relationships that nevertheless included a component of the relations considered in this evaluation (cf. Section 1.5.2.7), those that were potentially pertinent but incomplete (i.e., in which one or more of the related elements was not explicitly indicated in a context that might otherwise have been useful), those that presented categorial ambiguities, those that constituted noise for the purposes of this research (including the occurrence of the marker as part of a more complex unit), and finally those that could not be classified (for example, due to problems related to the form of the context, or to ambiguities that could not be resolved).

Table 57. Results of the evaluation of English marker occurrences

Marker	Valid hits	Complex relations	Incomplete	Categorial ambiguities	Noise	Unknown	Total
associated	88	0	0	3	7	2	100
risk	48	0	32	0	20	0	100
risk factor	29	0	69	0	2	0	100
marker	39	0	56	0	3	2	100
role	73	1	21	0	2	3	100
contribute to	95	0	0	0	2	3	100
effect	51	0	44	0	3	2	100
induce	35	4	1	59	0	1	100
promote	83	0	1	13	1	2	100
increase	24	0	0	69	7	0	100
lead to	99	0	0	0	1	0	100
involved in	91	3	1	0	4	1	100
result	29	0	0	71	0	0	100
Total	784	8	225	215	52	16	1300

Table 58. Results of the evaluation of French marker occurrences

Marker	Valid hits	Complex relations	Incomplete	Categorial ambiguities	Noise	Unknown	Total
et	2	0	0	0	98	0	100
lié à	96	0	0	0	3	1	100
facteur de risque	31	0	69	0	0	0	100
risque de	61	0	35	0	3	1	100
conduire à	97	0	1	0	1	1	100
effet	27	0	51	0	21	1	100
entraîner	87	0	0	11	0	2	100
favoriser	80	1	6	7	6	0	100
induire	52	0	1	44	2	1	100
induit par	92	1	0	0	2	5	100
inhiber	62	22	4	9	2	1	100
participer à	58	0	1	6	34	1	100
résulter de	43	0	0	3	0	3	49
Total	788	24	168	80	172	17	1249

A rough parallel is observed in the distribution of marker occurrences overall, with the majority in each language (approximately 60%) identified as valid hits (Figure 9). Much smaller proportions were considered to be incomplete or to constitute cases of categorial ambiguity or noise. However, as the data in the tables above indicate, there was a significant amount of variation between the individual markers. Clearly, each marker's performance must be evaluated individually in order to target the most useful.

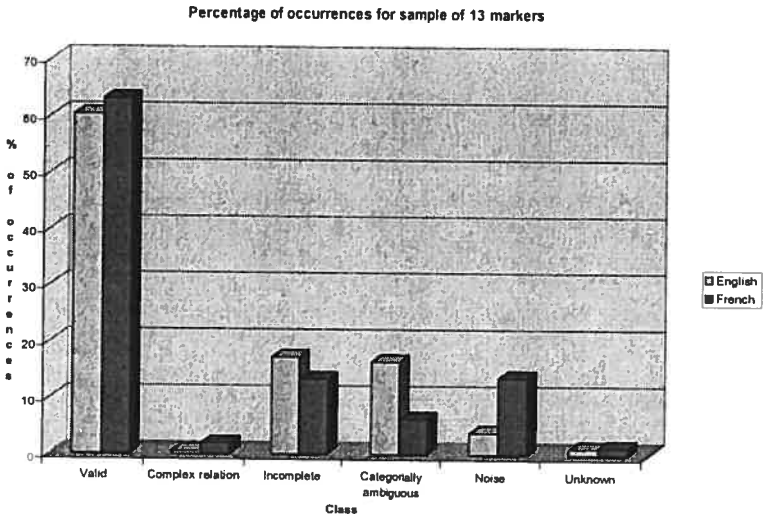


Figure 9. Marker precision for a sample of 13 markers in each language

While the distribution among the relations is consistent in the two groups, it is worth considering that the part of speech distribution of the markers differs. This may be expected — in light of observations in the results and in projects such as that of Barrière (2001; cf. Section 2.1.8) — to have a significant effect on the results observed.

For example, the precision of *et*, a conjunction identified as a marker of ASSOCIATION in French, is extremely low. This indeed reflects Barrière's observations — as well as intuitive expectations — that such markers will tend to be less precise indicators than nouns or verbs. This marker presents interesting examples of some phenomena that may interfere with marker precision.

Not surprisingly, a major source of noise in the case of this marker is the use of *et* to indicate the conjunction of two items, rather than their participation in one of the relations observed in this project. (This will be discussed in detail in Section 4.9.1.2.) Another phenomenon involves the possibility that a single marker may indicate more than one type of relation or sub-relation. Interestingly, although the marker *et* was identified as a candidate marker of ASSOCIATION in the contexts generated using domain terms, in the sample of contexts retrieved using the marker itself, no examples of

ASSOCIATION, but rather two occurrences of CAUSE–EFFECT relations were found. (In both cases, *et* occurred in conjunction with other potential markers, including *ainsi* and *donc*). While both of these relation types would be admissible according to the criteria of this study, they are of course considerably different, and should be distinguished in the presentation of results to a user. However, given the fact that the same marker may indicate ASSOCIATION, CREATION or neither of these, this distinction could be difficult to make automatically. Structural cues may assist in some cases (for example, the propositional form of at least one of the items linked by the marker in the case of CAUSE–EFFECT relations); in others, paralinguistic factors may also provide cues (e.g., the fact that the cases of ASSOCIATION were identified exclusively in headings or sub-headings). The implementation of these techniques for sorting occurrences would nevertheless require both a meticulous evaluation of the contexts in which the marker may occur, and a considerable investment of time and effort to develop effective strategies for exploiting these cues effectively.

Given the variation in the part of speech classes observed in the initial sample, a sub-set of the markers that have the same distribution among both the relations and part of speech classes in the two languages may be chosen for more detailed evaluation, in order to provide a more uniform basis for comparison. This set of ten markers (two for ASSOCIATION and eight for CAUSE–EFFECT) is illustrated in Table 59.

Table 59. List of English and French markers for precision evaluation by relation and part of speech category

	English markers	**French markers**
ASSOCIATION		
NOUNS	risk factor, risk	facteur de risque, risque
CAUSE–EFFECT		
NOUNS	effect	effet
PARTICIPIAL ADJECTIVES	involved in	induit par
VERBS	result, contribute to, increase, induce, promote, lead to	entraîner, conduire à, favoriser, induire, inhiber, participer à

The distribution among the classes identified for the 1,000 randomly selected occurrences in each corpus containing this set of ten markers is illustrated in Figure 10.

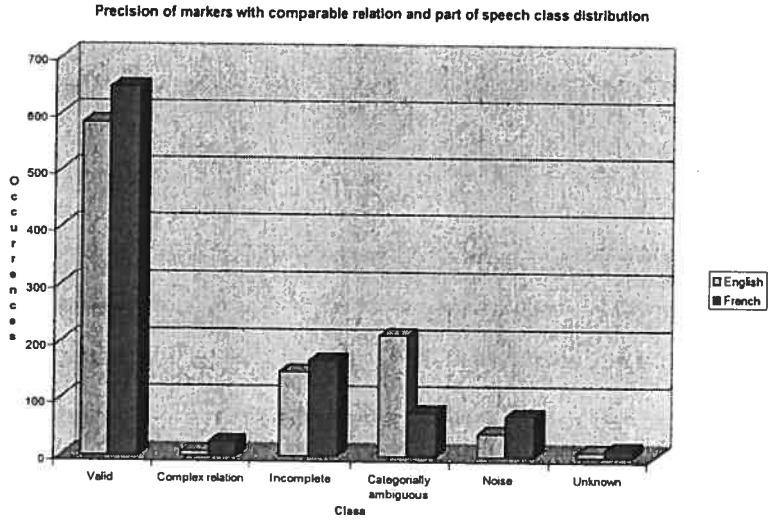


Figure 10. Precision of markers with comparable relation and POS class distribution

The general trends observed in the results of the occurrences of the set of 13 markers remain present, with the majority of the occurrences presenting valid relations. Overall, according to a Chi-square test comparing the proportions of valid occurrences observed in the English and French data sets, significantly more of the contexts retrieved using the French markers were valid ($p = 0.004$), and a similar difference was observed in the case of complex relations ($p = 0.002$). The proportions of incomplete contexts were not significantly different ($p = 0.220$), although once again the proportion was higher in French. Conversely, the proportion of occurrences presenting categorial ambiguities was considerably higher in English ($p < 0.001$).

These results indicate that the markers identified frequently in the study are likely to be effective for identifying relation occurrences in corpora, a very positive result. However, the French markers appear to be even more precise than the English, possibly indicating that the results of KRC extraction using the English markers may require more user intervention to eliminate noise and identify required information.

However, as much of this difference was attributable to categorial ambiguities, an approach using a part-of-speech tagged corpus as input could be particularly beneficial in English, reducing some of this noise and improving the effectiveness of the approach. (In fact, if the occurrences of categorial ambiguities are excluded from the data presented above, the proportion of valid contexts is somewhat higher in the English sample, a difference that trends towards significance ($p = 0.065$).)

Moreover, while occurrences of character strings corresponding to lexical items other than those expressly targeted in this evaluation were distinguished from valid hits here, some of these occurrences may in fact indicate the desired relation. This distinction was made in light of the methodology used in this research and the goals of the comparison, but the observations could in future be reviewed from a more inclusive perspective (e.g., similar to that used in Marshman 2002).

The evaluation of the sample data also allows for preliminary comparison of the precision of markers from different part of speech classes. As illustrated in Figure 11, Figure 12 and Figure 13, some differences between the two data sets are observed.

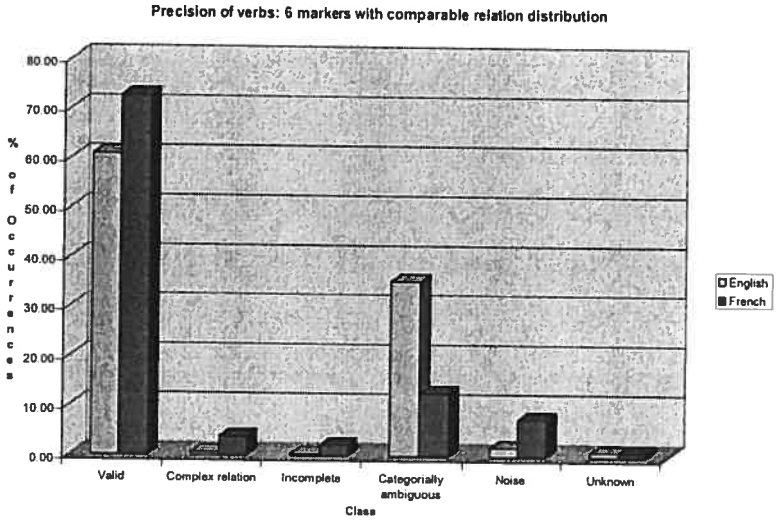


Figure 11. Precision of 6 verbal markers with comparable relation distribution

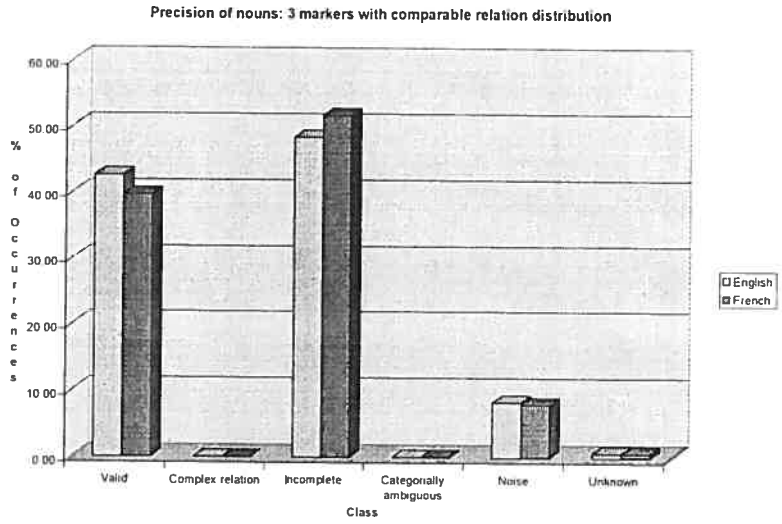


Figure 12. Precision of 3 nominal markers with comparable relation distribution

While these data suggest that both verbal and nominal markers may be productive for identifying KRCs in the two languages, a higher proportion of valid occurrences was obtained using verbal markers than using nominal ones. The difference was far more pronounced in the French markers evaluated, as the French verbs were more precise than the English, but the English nouns were more precise than the French. This potential for variation could be an interesting subject for future work in order to evaluate whether the difference is observed in an evaluation based on more data.

It is also obvious that a much higher proportion of the occurrences of nominal markers evaluated in the two data sets were found to be incomplete (i.e., did not include an explicit indication of one or more of the elements linked in a potentially valid relation), although the proportions were higher in French for both verbal and nominal markers. The presence of a high proportion of categorially ambiguous verbal forms in English is likely to have contributed to the decreased precision in this language.

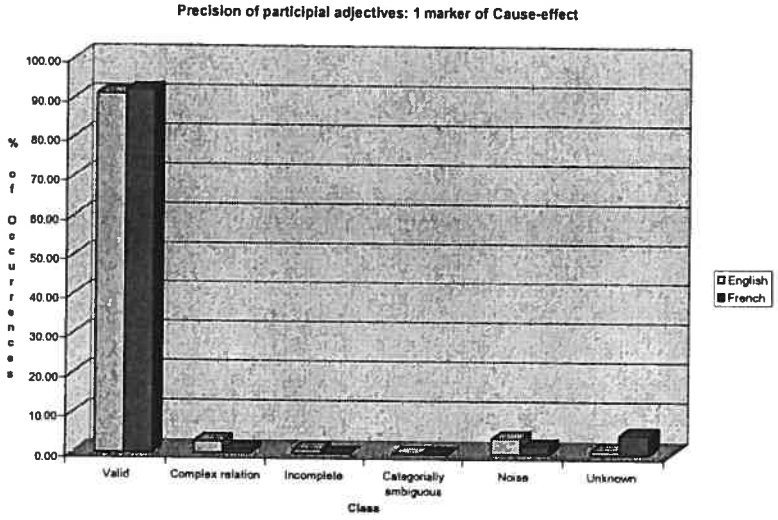


Figure 13. Precision of participial adjective CAUSE–EFFECT marker

The fact that a single participial adjective was evaluated in each language makes generalization impossible; however, the two markers showed good precision and fairly close results in the two data sets.

Perhaps the most important information gained from this evaluation is the proportion of contexts containing nominal markers that were found to be incomplete. This clearly reduces the value of these contexts for knowledge extraction and introduces noise in the results of KRC extraction (although of course the partial information provided by such contexts may be useful to some extent in some contexts). These results parallel the observations of Barrière (2001), who noted that the precision of verbal markers in English was significantly higher than that of nominal ones, and also suggest that this tendency is also present in French.

These data may be discussed further in light of the distribution of the markers identified in this research between POS classes in the samples analyzed, as discussed in Section 4.5.1. Overall, the proportions of nominal markers were approximately equal in the two data sets, although English showed a higher proportion of nominal markers for the ASSOCIATION relation and French for the CAUSE–EFFECT relation. There may thus be

a potential for variation in the levels of noise observed for specific relations in the two languages if marker sets that reflect the types of markers identified in this project are used.

This information may also affect the choice of markers for inclusion in the pattern sets, for example encouraging a focus on verbal rather than nominal markers, as in several previous research projects on CAUSE–EFFECT relations. If this apparent trend is observed in larger samples of data and does encourage such a decision, however, it will be necessary to consider the potential for silences in the results if nominal markers are excluded. Moreover, the difficulties posed by the fact that nominal markers of ASSOCIATION are quite numerous but also — if the two evaluated in this analysis are any indication — quite likely to be incomplete will be important to consider and further analyze in the development of pattern sets for this relation. Additional difficulties linked to the prevalence of function word markers (as illustrated by the case of *et*) may also greatly increase the complexity of creating pattern sets for identifying this relation. Regardless, from the proportions of markers observed in this study, it is clear that excluding these markers of ASSOCIATION is not a valid option if a certain level of recall is to be maintained. Other strategies will need to be developed, which may require a considerable investment of time and effort.

It is also possible to compare the precision of markers for the two relations in the samples (Figure 14 and Figure 15). Once again, parallels between the performances of markers in the corpora may be observed. (However, it is important to note that the markers of ASSOCIATION are exclusively nominal while the markers of CAUSE–EFFECT are primarily verbal, which is also likely to contribute to the differences observed.)

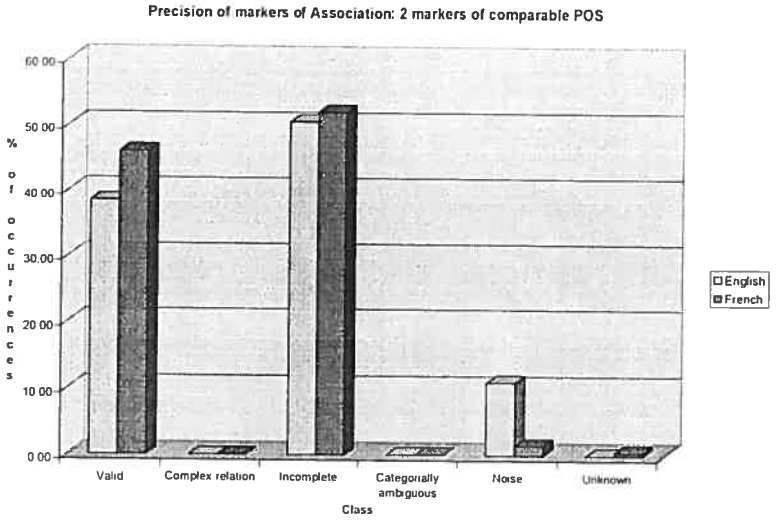


Figure 14. Precision of 2 nominal ASSOCIATION markers

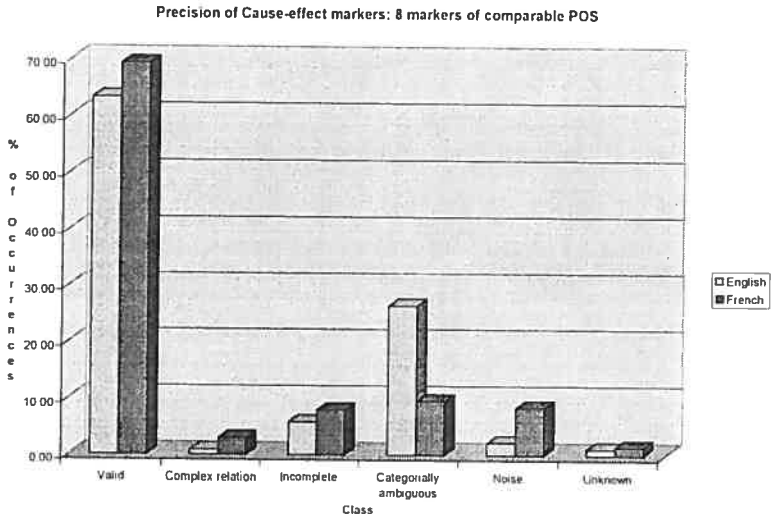


Figure 15. Precision of 8 CAUSE–EFFECT markers with comparable POS class distribution

In both cases, the proportions of valid occurrences are slightly higher in French. The proportions of noise also differ, with more observed in English for the ASSOCIATION relation and in French for the CAUSE–EFFECT relation. Categorial ambiguity, observed

A first possibility for consideration is that the form of the markers *résulter de* and *risque de* may play a role in the higher precision of these markers in French as compared to *risk* and *result* in English, as a large amount of noise may be eliminated by the specification of these additional elements. In fact, in a sample of contexts retrieved using the marker *risque* alone, only 31 of 100 occurrences were identified as valid, indicating that the additional element did improve precision. In a sample for the marker *résulter* alone, however, 93 of 100 occurrences were found to be valid, which actually constitutes an increase in the precision of this marker. This may be due to the fact that the marker may occur in forms that were not observed in the relation occurrences initially analyzed but that nevertheless indicate the presence of a CAUSE–EFFECT relation (e.g., X *résulte en* Y).

It is clear that the difference in the distribution of occurrences of the markers *result* and *résulter de*, with the French marker showing a much higher proportion of valid occurrences than the English, is likely to be closely linked to the proportion of categorially ambiguous occurrences in English. This can be traced to the form of the associated nouns in the two languages, *result* and *résultat*. The categorially ambiguous form of the English verb and noun forms (i.e., *result*, *results*) causes serious difficulties in a character-string-based approach: while the French noun *résultat* can be explicitly excluded from the results of extraction without eliminating verb forms, this is not the case in English. Moreover, the fact that the noun form *results* is very commonly used in scientific tests such as those included in the corpus for this research (e.g., as a heading introducing the observations in an experiment in research articles) produces an extremely high proportion of character-string occurrences associated with the noun rather than the verb (of which a large number are not used to indicate a relationship between two elements).

These kinds of differences pose significant challenges for bilingual, character-string-based approaches. The performance of even very similar markers may differ considerably; moreover, possibilities for this kind of variation in performance must be

evaluated for each marker individually, involving a detailed analysis in the two languages.

The differences between the data sets in the proportions of contexts involving categorial ambiguities, significant in the sample ($p < 0.001$), can be traced to markers such as *result* and *increase*, which shows a similar ambiguity. For the other markers, the levels are relatively consistent. These data illustrate at the level of individual markers the possibility that English results could be considerably improved in a more sophisticated approach using lexico-syntactic knowledge patterns in part-of-speech tagged corpora. However, in French the need for such developments appears to be considerably less.

In order to better evaluate the potential for using part-of-speech tagged texts, results obtained using Syntex (Bourigault et al. 2005) may be used. In a sample of contexts of the verbs *result* and *résulter*, 47 of 50 English occurrences were identified as expressing CREATION (while two involved tagging problems and a third was unclassified), and in French, 49 of 50 expressed this relation (the remaining context was also unclassified). Thus the two markers appear to provide much more similar performance in an approach using lexico-syntactic knowledge patterns than in simpler character-string-based techniques. Moreover, the markers are very efficient for identifying pertinent contexts in this kind of approach, identifying them as promising for KRC extraction tools.

The difference noted for the markers *induce* and *induire* (Figure 19) is also largely due to the presence of a significant amount of categorial ambiguity in both languages, but particularly in English. However, the source of this ambiguity is somewhat different. The challenge with this marker lies generally in the differentiation between forms that in this project were considered to be verbal and those that were considered to be participial adjectives. Many applications (e.g., part of speech taggers) do not differentiate between these forms, and thus the possibilities for using such tools to reduce the impact of this phenomenon are more limited. Another approach might lie

in simply considering these to constitute a unique marker that occurs in two (or more) forms.

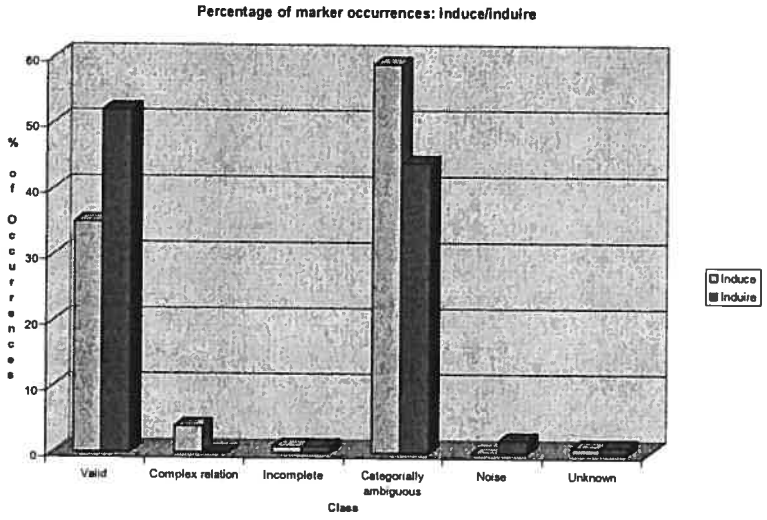


Figure 19. Marker precision: *induce* / *induire*

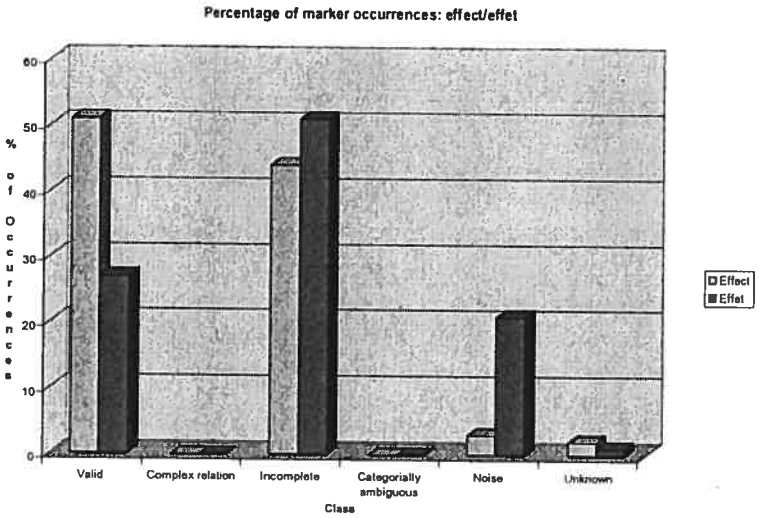


Figure 20. Marker precision: *effect* / *effet*

When the case of the markers *effect* and *effet* (Figure 20) is considered, the interlinguistic difference may be observed to result from the proportion of noise

observed in French. All 21 of these cases involve the occurrence of the expression *en effet*, which constitutes a noise level of over 20% in the results for this marker. Conversely, although a similar expression, *in effect*, does exist in English, it was not observed in the sample analyzed. Clearly, in French it is necessary to deal with this phenomenon, for example by explicitly excluding this form from the character strings used to identify occurrences, or by using a part-of-speech-tagged and parsed corpus that identifies this expression as a separate unit in itself. However, the need for these measures is not significant in English, as the noise level is quite low.

If data extracted using Syntex are analyzed, a more accurate picture of the possibilities of using this marker in a more sophisticated, lexico-syntactic pattern-based approach may be obtained, as Syntex distinguishes between occurrences of *en effet* and of *effet* alone. In a sample of 50 English contexts containing *effect* identified using Syntex, 27 cases of CAUSE–EFFECT relations were observed, while 22 contexts were incomplete and one was unclassifiable. In 50 French contexts, 22 cases of CAUSE–EFFECT relations were identified, with one case of noise resulting from a tagging problem concerning an occurrence of *en effet* and the remaining 27 cases identified as incomplete. This shows that although a more sophisticated approach to the identification and processing of contexts may be beneficial, particularly in French in the case of this marker, some differences in the productivity of the markers appear to remain.

The markers *effect* and *effet* also show a certain amount of polysemy as discussed in Section 4.7, corresponding to both cases of CREATION and of MODIFICATION. For the purposes of this research, both of these sub-relations were considered to be pertinent. The distribution of occurrences of *effect* and *effet* among the sub-types was proportionally somewhat different, with 16 cases of CREATION and 38 cases of MODIFICATION in English (a proportion of 30% CREATION and 70% MODIFICATION), and in French 16 cases of CREATION and 11 cases of MODIFICATION

(60% CREATION and 40% MODIFICATION).[126] This difference in distribution indicates a strong potential for variation in the productivity of even similar polysemous markers for identifying occurrences of specific sub-relations of CAUSE–EFFECT.[127]

Although the samples analyzed were limited, they nevertheless showed a strong potential for variation in precision between various groups of markers, as well as between individual pairs of markers. These results indicate the need to carefully evaluate a range of markers in light of more data extracted specifically for this purpose, to confirm and further analyze the trends identified in these observations. The confirmation of these observations in other corpora and using other approaches would also ensure that factors linked to the corpora or the methodology used are evaluated. Nevertheless, the results show that there is likely to be significant variation from marker to marker and as a function of other factors such as the markers' part of speech class or the relation indicated, and that coherent trends may be difficult to identify and isolate.

4.7 Polysemy of pattern markers

In addition to the fact that markers may produce noise (as in the case of *et*, as described in Section 4.6 above), the potential for candidate markers to indicate more than one type of relation or sub-relation considered to be pertinent in this project, or to indicate either a "core" CAUSE–EFFECT relation or a more complex relationship with a causal component, may be observed. These phenomena are discussed below.

---

[126] A similar distribution was found in the sample of Syntex data: 10 cases of CREATION and 17 cases of MODIFICATION in English, for a total of 27 CAUSE–EFFECT relations (i.e., 37% CREATION and 63% MODIFICATION). In French, 12 contexts expressing CREATION and 10 cases of MODIFICATION were noted, for a total of 22 CAUSE–EFFECT relations (i.e., 55% CREATION and 45% MODIFICATION).

[127] This potential for variation was noted for many of the verbal markers observed in this research, as described in Marshman and L'Homme (2006, 2006a).

4.7.1 Markers associated with more than one (sub-)relation

In each language, a few markers were associated with two or more types or sub-types of relations. The frequency of the phenomenon in the sample analyzed was relatively comparable, with five such markers observed in English and four in French, all indicating CAUSE–EFFECT relations in at least some cases.

The five cases observed in the English results (3% of the total number of markers, and 4% of the CAUSE–EFFECT markers) involved the association of markers with two distinct CAUSE–EFFECT sub-relations. These are shown in Examples 18 to 27:

18. Although atherosclerosis is a multifactorial disease, often occurring as a **complication of** hypertension, obesity, and diabetes… (Umehara et al. 2004)
19. Ultimately, these pathways synergize to construct a scaffold on which the **complications of** diabetes **in** the vasculature and heart may be built. (Yan et al. 2003)
20. Direct evidence for an important **role for** myeloperoxidase **in** lipid oxidation in vivo comes from recent studies… (Brennan and Hazen 2003)
21. … the **role of** radiation therapy **for** invasive breast cancer treated with BCS is now well accepted. (Meric-Bernstam 2004)
22. … they are important targets of the biological **effects of** fractalkine (ie, chemotaxis, adhesion, and activation)… (Umehara et al. 2004)
23. No study evaluated the associations between statins' **effects on** LDL oxidation and lipid levels. (Balk et al. 2003)
24. Witztum's group 40,41 has developed a range of antibodies directed against oxidation-**dependent** epitopes in LDL (anti-oxLDL)… (Griendling and FitzGerald 2003a)
25. First, the estrogen-**dependent** step in mammary gland development, the ductal elongation that takes place during puberty… (Sicinski and Weinberg 1997)
26. There was no consistency among these patients with respect to prior chemotherapy (1 **for** metastatic disease, 6 adjuvant, 3 chemonaive). (Housmaninger et al. 2004)
27. … 28% beginning new medications **for** cholesterol, blood pressure, or diabetes …. (Berra 2003)

In each of these cases, the first example was classified as an occurrence of CREATION; in almost all cases the second was classified as MODIFICATION (for the marker *dependent* the sub-relation identified was MAINTENANCE/PERMISSION). In addition, one case of a PREVENTION relation was observed for the marker *role*, although this occurred in a quite unusual context, as shown in Example 28:

> 28. Although the **role of** dietary and vitamin antioxidants **in** the development of breast cancer is not conclusive in human studies... (Kang 2002)

The four cases (2% of the total number of markers, and 3% of the CAUSE–EFFECT markers) observed in French are illustrated in Examples 29 to 36:

> 29. Les **complications de** l'ostéolyse maligne **dans** le cancer du sein engagent rarement le pronostic vital immédiat, mais sont source d'une morbidité importante. (Tubiana 2001)
>
> 30. Les interactions entre système rénine-angiotensine et **complications** vasculaires **du** diabète constituent un autre exemple de l'implication du TGF-ß. (Michel 2004)
>
> 31. L'athérosclérose est considérée actuellement comme une **réponse** inflammatoire **aux** lésions de la paroi artérielle. (Duriez 2004)
>
> 32. Cependant, la **réponse** osseuse **au** traitement reste toujours difficile à évaluer de par la faible spécificité de la scintigraphie osseuse.... (Leriche et Bonneterre 1997)
>
> 33. Les avancées de la chimiothérapie **anti**tumorale ont été obtenues grâce à des médicaments ayant une nouvelle structure chimique... (Lavelle and Jehanno 1998)
>
> 34. ... en situation métastatique associé à l'exemestane ou à une chimiothérapie **anti**tubuline, ou en néoadjuvant. (Guastalla et al. 2004)
>
> 35. Les résultats obtenus avec le paclitaxel en monothérapie **dans** le cancer du sein métastatique ont tout naturellement conduit à associer ce médicament aux anthracyclines... (Ferrero et al. 2003)
>
> 36. **Dans** les cellules AT exposées aux rayonnements ionisants, l'induction de p53 est réduite et très retardée... (Angèle et al. 2001)

However, unlike the regularities observed in English, the distribution of the occurrences of these ambiguous markers among the relations and sub-relations is quite

different. In the cases of *complication* and *réponse*, the divergence noted in English is present, with the first case considered to be an example of CREATION, and the second of MODIFICATION. However, in the case of *anti-*, Example 33 indicates not CREATION but DESTRUCTION, while Example 34 indicates MODIFICATION. Finally, the case of *dans* illustrates a more serious difficulty, as Example 35 indicates MODIFICATION, but Example 36 expresses an ASSOCIATION. Clearly, such cases of inter-relational ambiguities are even more critical to deal with than intra-relational variations between sub-types.

As this phenomenon was so rarely observed, statistical comparison between the English and French data cannot be considered reliable. Therefore, the interlinguistic comparison of the phenomenon observed will be restricted to qualitative comments.

In both languages, while approaches relying on the representation of the basic marker element alone would encounter difficulties resulting from this polysemy, promising avenues for disambiguation using formal criteria such as marker form and pattern structure (including the form of related elements participating in these structures) were observed in most cases. However, also in both languages, some markers (such as *dependent* and *anti-*, for example) seem likely to require other strategies for disambiguation.

It is difficult to draw conclusions on the basis of the small samples observed in this case, but it is interesting to note the relative regularity of the ambiguity observed in the English (between the CREATION and MODIFICATION sub-relations), and the wider variability (including inter-relational variation) in the French. However, the observation of similar ambiguity in the markers *complication* in English and French suggests that some parallels may exist in the two languages and could be considered in developing bilingual tools. Both of these observations suggest that more data should be gathered in order to further evaluate this phenomenon and the challenges it may pose for bilingual pattern-based tools, as well as potential strategies for resolving these kinds of ambiguities.

4.7.2 Complex relations denoted by markers

Another type of marker polysemy, noted in the results of the marker precision evaluation, involved markers denoting relationships between elements that — while including a CAUSE–EFFECT component — were more complex than those considered for the purposes of this research. This phenomenon was observed in the case of the markers *induce*, *inhibit* and *inhiber*. These three markers may not only denote a "core" relation of CREATION or DECREASE (depending on the marker), but also more complex relationships that involve the causing or decreasing of the functioning of an affected element (e.g., a molecule).

While these may be interesting to consider in at least some contexts, they should nevertheless be distinguished from the core relations considered in this research in order to prevent misinterpretation. While for the markers *inhibit* and *inhiber* similar, significant proportions of these more complex relationships were observed, a more complex relationship was observed in only a few cases in English for *induce*, and was not noted in French for *induire*.[128] Clearly, for some markers it will be necessary to determine whether these more complex relationships are to be considered for context extraction, and if so in what capacity (i.e., included with the core relations or as a separate category). More discussion and additional, similar cases may be found in Section 5.5.3.1, as well as in Marshman and L'Homme (2006, 2006a).

It is clear that the phenomenon of marker polysemy that has been widely observed in other projects is not fully explored in this study. This is in large part due to methodological choices: the comparative orientation of this work requires a relatively general analysis of the relation occurrences and markers observed, while the evaluation of polysemy requires a more specific analysis of individual patterns. Because of the approach used in this study, the relatively low numbers of occurrences of each marker

---

[128] The existence of a corresponding complex relationship indicated by *induire* was nevertheless noted in another study of contexts from the French corpus in Marshman and L'Homme (2006a).

identified did not provide as many opportunities to observe polysemy as projects that evaluated large numbers of occurrences of the same marker.

Moreover, the limitations imposed by other choices in the methodology, and primarily the exclusion of occurrences of complex relations from the initial analysis of relation occurrences, reduces the range of contexts that could be evaluated in the first phase of analysis of marker polysemy. In addition, in the study of marker precision that provided the data for the second type of analysis of marker polysemy, the focus was placed on the most frequently observed markers in the relation occurrences identified, which are also likely to be among the most useful for identifying the type of relation occurrences targeted in this research. Further evaluation of more occurrences of a wider range of these markers would doubtless reveal more about this phenomenon.

Finally, the term-based approach to the observation of relation occurrences may have contributed to the relatively restricted range of polysemous markers identified. As noted in a number of projects — including Marshman and L'Homme (2006, 2006a), which focused on English and French verbal markers identified in the course of this research — close associations may often be observed between specific senses of markers and the terms or classes of terms with which they are used.[129] The use of a set of 15 domain candidate terms in each language is thus likely to have restricted the possibilities for observing polysemy.

---

[129] Marshman and L'Homme (2006) identified an average of approximately 3 senses per marker for a set of 14 English verbal markers that were selected for study because they had been identified as being ambiguous in the corpora used for this research. Marshman and L'Homme (2006a) identified an average of 2.5 senses per marker for a set of 38 of French verbal markers observed in this research that were frequent in the French corpus. Both projects distinguished "core" causal, complex causal and non-causal senses on the basis of paraphrases identified for the markers in contexts extracted from the corpora, and included them in analysis.

4.8 Pattern variation

In this Section, the interlinguistic comparison of the variation observed in the form of markers and structure of patterns will be described. The principal types of pattern variation evaluated in this research — variations in marker form (including the specific case of variations in voice of pattern markers) and variations in pattern structures (including the specific case of variation involving the presence of relative pronoun constructions) — are described below.

This analysis was of course carried out in English only on the 70 markers (18 for ASSOCIATION and 52 for CAUSE–EFFECT) that were observed more than once in the contexts analyzed, and in French on the 65 markers (13 for ASSOCIATION and 52 for CAUSE–EFFECT) that were observed twice or more in the contexts analyzed. The complete data on which these discussions are based are available in Appendix H.

4.8.1 Variation in marker form

Variation in marker forms (e.g., the addition or change of a preposition or conjunction appearing with the principal (generally open-class) marker element or the change in the order of complex marker elements) was frequently observed in the relation occurrences analyzed. This phenomenon may be observed in Examples 37 to 45:

37. Abnormal endothelium-dependent vasomotor responses predict the long-term progression of atherosclerosis and **associated** coronary events.... (Davignon 2004)

38. ... the initiation and progression of cardiovascular dysfunction **associated with** diseases such as hyperlipidemia, diabetes mellitus, hypertension, ischemic heart disease, and chronic heart failure. (Taniyama and Griendling 2003)

39. Both glucotoxicity and lipotoxicity **play a** primary **role in** the development of diabetes. (Pantaleo and Zonszein 2003)

40. Could it be that BRCA1 and BRCA2 **play roles in** the development of hereditary cancers but not sporadic tumors? (Yang and Lippman 1999)

41. It has been recognized that atherosclerosis is an inflammatory disease **in** which various cytokines **play a** significant **role**... (Taniyama and Griendling 2003)

42. ... de nombreux traitements ont des **effets** rhéologiques... (Boisseau 2004)

43. Les risques de saignements seraient reliés à l'**effet de** l'ail **sur** la coagulation. (Trahan 2002)

44. La structure chromatinienne **joue un rôle** majeur **dans** des processus tels que la transcription, la réplication et la réparation de l'ADN. (Chailleux et al. 2000)

45. Le **rôle des** estrogènes **dans** la prolifération des tumeurs mammaires hormonodépendantes a été montré depuis de nombreuses années. (De Crémoux 2000)

The ratio of marker forms observed relative to the numbers of markers in each language indicates the level of variability of these marker forms. The results observed are illustrated below in Table 60, which shows that the English markers overall and for each relation showed more pronounced variation than the French. It is clear that at least in English there is a fair amount of variation in marker forms, and that more than a single pattern form will be required to represent many of the markers observed.

Table 60. Comparison of ratio of marker forms to markers in English and French

	English	French	Difference
Marker forms per marker, overall	1.5	1.3	0.2
Marker forms per marker, ASSOCIATION	1.8	1.3	0.5
Marker forms per marker, CAUSE–EFFECT	1.4	1.3	0.1

However, given the variation in the numbers of markers in each group and of occurrences observed for each marker, these figures can only indicate potential trends. The variation observed per marker is of course influenced by the number of occurrences observed. In order to evaluate the level of variation more accurately (although still not strictly comparably), the mean number of forms observed for markers observed a given number of times can be calculated (Table 61). These calculations show the results observed for the markers observed between 2 and 8 times in the sample

analyzed (i.e., the range of frequencies that were observed in both languages and thus provide a basis for comparison). Overall, the level of marker variation in English is higher in most groups, although the degrees vary.

Table 61. Comparison of marker variation (by number of marker occurrences) in English and French

	English		French		
Number of marker occurrences	**Number of markers**	**Mean number of marker forms**	**Number of markers**	**Mean number of marker forms**	**Difference**
2	30	1.1	25	1.2	-0.1
3	6	1.7	12	1.3	0.4
4	3	1.7	11	1.4	0.3
5	8	1.3	5	1.2	0.1
6	5	1.8	3	2.0	-0.2
7	4	1.8	7	1.4	0.4
8	2	1.0	1	1.0	0.0
Total	58		64		

The data for each relation separately are shown in Table 62. In the individual relations, the small numbers of markers for the ASSOCIATION relation make it difficult to confirm the apparent trend towards higher variation in English, but it generally remains in the CAUSE–EFFECT groups. These data suggest that in the process of pattern set design it could be necessary to include more pattern forms in English than in French to account for this kind of marker variation; the investment of time and effort in this process may be substantial. Certainly, the possibility is worth investigating in further, more appropriately designed projects; these projects should permit the evaluation of larger samples of marker occurrences of comparable size, for a wider range of markers and markers of similar distributions among relations, part of speech classes, etc., in order to neutralize other factors that may have contributed to the differences noted in this analysis.

Table 62. Comparison of marker variation for ASSOCIATION and CAUSE-EFFECT markers (by number of marker occurrences) in English and French

	English		French		
Number of marker occurrences	**Number of markers**	**Mean number of marker forms**	**Number of markers**	**Mean number of marker forms**	**Difference**
ASSOCIATION					
2	5	1.2	6	1.3	-0.1
3	2	1.5	0	n/a	n/a
4	1	1.0	2	1.0	0.0
5	3	1.3	2	1.0	0.3
6	0	n/a	1	3.0	n/a
7	1	2.0	1	1.0	1.0
8	1	1.0	0	n/a	n/a
Total	13		12		
CAUSE-EFFECT					
2	25	1.1	19	1.2	-0.1
3	4	1.8	12	1.3	0.5
4	2	2.0	9	1.4	0.6
5	5	1.2	3	1.3	-0.1
6	5	1.8	2	1.5	0.3
7	3	1.7	6	1.5	0.2
8	1	1.0	1	1.0	0.0
Total	45		52		

4.8.1.1 Variation in voice of verbal markers

One specific variation in marker form observed in the corpus was the occurrence of verbal pattern markers in the passive rather than active voice, as in Examples 46 to 49:

46. Plasma levels of PAI-1 are regulated on a genetic basis, and its expression can **be augmented by** insulin resistance and other factors such as abnormal adiposity, hypertriglyceridemia... (Pantaleo and Zonszein 2003)

47. Oxidative stress **has been linked to** the activation of both NF-[kappa]B and AP-1. (Granger et al. 2004)

48. Dans les cellules hormono-dépendantes, la transcription de CatD **est contrôlée par** les oestrogènes. (Chailleux et al. 2000)

49. Comme toute réponse immunitaire, la réponse anti-tumorale doit **être déclenchée par** des cellules présentatrices d'antigènes. (Catros-Quemener et al. 2003)

As easily observed in Examples 50 to 53, the voice in which verbal pattern markers occur in contexts can have a significant effect on pattern marker form:

50. Specifically, mitogenic effects of oxidized LDL on vascular smooth muscle cells, which contribute to the atherogenic process appear to **require** the activation of SK. (Saba and Hla 2004)

51. Therefore, it is currently suggested that ER[alpha] function may **be required for** maximum activation of IGF-signaling pathways. (McCance and Jones 2003)

52. ... these findings, together with those in chronic atherosclerosis, importantly **link** ligand-RAGE interaction **to** the pathogenesis of exaggerated neointimal expansion... (Yan et al. 2003)

53. Oxidative stress **has been linked to** the activation of both NF-[kappa]B and AP-1. (Granger et al. 2004)

Differences in voice may affect the order of the components of complex markers, as well as the placement of the related elements relative to the marker. One particular case of this kind of variation involves the inversion of the order of participants in asymmetric relations such as CAUSE–EFFECT, which would be pertinent for applications that attempt to identify related elements and assign a role in a relation to them. Moreover, the insertion of additional marker elements may also be observed in the case of passive transformations of verbs (e.g., *for* in Example 51), which in some cases may also entail the change of pattern markers from simple to complex, with the accompanying differences in the potentials for performance and for difficulties (cf. Section 2.6.1).

The proportion of the verbal marker occurrences in passive voice is illustrated in Table 63, which shows a statistically significant difference between the languages (p = 0.002), with the passive proportionally more common in English (observed in 14% of the English occurrences of verbal markers and only 4% of the French occurrences). This corresponds to 5% of the total relation occurrences in English, but only 1% in French.

Table 63. Comparison of the proportions of verbal marker occurrences in passive and active voice in English and French

	EN	FR	Total
Passive	24	5	29
Active	151	135	286
Total	175	140	315

Of the 16 different English markers observed in passive form, 8 also occurred in active voice in the sample analyzed, while 6 occurred only once in the sample (so that no conclusions can be drawn about their invariability), and the remaining 2 were observed only in passive form. Thus, only 20% of the 10 markers that were observed more than once occurred exclusively in passive form. As such, 68 (81%) of the total of 84 verbal markers were found in the active form only; 30 of these markers were observed more than twice in the sample analyzed, and 20 (67%) of this latter group were observed in active form only. In French, of the 140 verbal marker occurrences, only 5 (4%) were observed in the passive voice. Two of the four markers observed in passive form were also observed in the active voice in the sample, while one was observed only once and the other occurred exclusively in passive voice. Thus 4 of the total of 77 verbal markers were observed in the passive voice, and 3 (12%) of the 25 markers observed twice or more in the sample, while 22 (88%) of this latter group were observed exclusively in the active voice.

These data indicate that it is important to take this variation into account in designing English pattern forms; moreover, this need will often involve developing supplementary pattern forms for verbal markers. (Such forms would be required to account for the appropriate marker forms and — in the case of asymmetric relations such as CAUSE–EFFECT and/or applications that attempt to identify related elements and the roles they fill in relations — the placement of these elements). This is likely to be accompanied by a corresponding increase in the investment of time and effort required to for pattern set development. It is likely possible, however, to take advantage of regularities in passive structures, and to apply similar models for a variety of different

verbal markers. In French, however, the phenomenon is much less significant, and it is debatable whether the investment required to create supplementary pattern forms would be justified.

Some specific observations related to this phenomenon in the two data sets are discussed below.

4.8.1.1.1 Differences related to variation in voice of verbal markers

The use of passive voice in English was commonly associated with structures in which an additional actant — particularly a human actant such as a researcher — were present on a semantic level but not realized on a surface level. Many of the structures of the markers showed a potential for variation in the realization and/or number of the actants. As illustrated in Examples 54 and 55, underlying tri-actantial structures may present (i.e., X implicates Y in Z, X correlates Y with Z),[130] but in few of these cases is the first actant realized in the contexts observed.

> 54. ...several other factors (eg, oxidative stress) ... have also been **implicated in** the development of CVD. (Granger et al. 2004)
>
> 55. Cell adhesion molecules **have** also **been correlated with** CHD. (Rackley 2004)

In many cases, as with markers such as *find... in*, *report... in*, *detect... in* and *note... in*, this first actant is a researcher, who has perceived an ASSOCIATION between two or more variables. This phenomenon may be linked to the scientific method and the effort to maintain objectivity in research, as well as to the stylistic conventions that thus favour (the appearance of) this objectivity in reporting results, as discussed in Section 2.6.2.1.

As illustrated in Examples 56 to 57, cases in which the first actant when expressed indicates the results of research (again with all of the possibilities of

---

[130] This latter example can be contrasted with another possible structure of this marker, Y *correlates with* Z.

subjectivity that the interpretation of these results by a researcher or researchers involves) are also common; this can be observed in markers such as *implicate... in* and *link... to*:

> 56. There is a large body of evidence that **implicates** inflammation and adhesion molecules **in** the pathogenesis of CVD, including atherosclerosis, stroke, and myocardial infarction. (Granger et al. 2004)
>
> 57. ... these findings, together with those in chronic atherosclerosis, importantly **link** ligand-RAGE interaction **to** the pathogenesis of exaggerated neointimal expansion... (Yan et al. 2003)

The potential for the expression of an additional actant, (e.g., *evidence*, *findings*), with the active marker form recalls the observation by Condamines (2002) that patterns expressing relations may not always be binary, and may also complicate the analysis of the context at both a syntactic and semantic level (e.g., given the fact that the source of an observation is specified, which may have an effect on the certainty of a given context).

The failure to express the first actant for some of these markers did not affect the consideration of occurrences for the purposes of this project, since the two elements that were related by the relation in question were nevertheless present. However, if one of the actants of a marker in a bi-actantial structure was not expressed (e.g., X *was produced*, X *was initiated*), this necessarily excluded contexts from study, because the information about the relation present was incomplete. This phenomenon may pose considerable challenges for semi-automatic applications, as precision would be affected if incomplete contexts were retained in the results of extraction using these markers.[131]

In the French results, only one of the markers observed in the passive voice corresponded to the kind of tri-actantial markers observed in the case of the English.

---

[131] It is in part for this reason that pattern forms are often required to link two elements defined using formal criteria. However, this kind of requirement may have an undesirable effect on recall in many applications.

However, alternate constructions, such as those involving impersonal pronouns, were observed. These cases are illustrated in Examples 58 to 60:

> 58. … la perfusion d'angiotensine II induit la formation d'anévrismes, qui **a été reliée** à l'activation des leucocytes circulants. (Michel 2004)
>
> 59. … si l'on **inhibe** l'activation de la guanylate cyclase par NO, on **induit** l'apoptose… (Kolb_2)
>
> 60. On **associe** maintenant une faible capacité aérobie et une mauvaise composition corporelle **aux** maladies cardiovasculaires et au diabète, un manque de souplesse aux maux de dos, etc. (Béliveau and Léger 2004)

While more data are required in order to evaluate the extent of these differences, these observations suggest that the use of the passive voice to maintain objectivity could be more prevalent in English. While the effect on knowledge extraction using pattern-based tools is minor (since the impersonal pronouns are barely more informative than the ellipsis of the human participant), this difference constitutes an example of the types of subtle differences that may be observed between the languages.

4.8.2 Variation in pattern structures

Pattern structures may vary in several ways; among the most important of these are the insertion of additional but regular elements (e.g., copula verbs) within a pattern form, and the variation in the order in which pattern elements appear (i.e., the configuration of the related elements and the marker within the pattern structure). The former phenomenon is illustrated in Examples 61 to 64, and the latter in Examples 65 to 68:

> 61. … the initiation and progression of cardiovascular dysfunction **associated with** diseases such as hyperlipidemia, diabetes mellitus, hypertension, ischemic heart disease, and chronic heart failure. (Taniyama and Griendling 2003)
>
> 62. …the missense mutation in Lp-PLA2 is **associated with** development of atherosclerosis in the elderly. (Caslake and Packard 2003)

63. Ce dernier est en effet capable de stimuler le recrutement et l'assemblage des sous-unités p47phox et p67phox, étape **nécessaire à** l'activation de la NADPH oxydase. (Bonnefont-Rousselot et al. 2002)

64. Ainsi une activation de caspases, clivant sélectivement certains substrats, est **nécessaire à** l'induction de prolifération de lymphocytes T (Kolb 2001)

65. It has been recognized that atherosclerosis is an inflammatory disease **in** which various cytokines **play a** significant **role**... (Taniyama and Griendling 2003)

66. Elevated compartmentalized cortisol may **play a role in** the pathogenesis of insulin resistance in animals... (Pantaleo and Zonszein 2003)

67. La première **conséquence** fonctionnelle majeure **de** l'activation des plaquettes est le changement de conformation des glycoprotéines GP IIb/IIIa présentes à leur surface... (Collet et al. 2004)

68. ... ils peuvent aussi et simultanément, être la **conséquence de** l'oxydation des lipoprotéines de basse densité (LDL)... (Bonnefont-Rousselot et al. 2002)

It should also be noted that marker variation is also likely to be linked to the types of structures in which markers are observed, as in Examples 69 to 74:

69. The inflammatory **marker** C-reactive protein (CRP) can indicate low-grade chronic inflammation... (MacKenzie 2004)

70. As carotid IMT is a good early **marker of** atherosclerosis and risk of cerebrovascular ischemic events... (Zambon et al. 2003)

71. If so, how could prior assessments of the health **effects of** hormone replacement therapy (HRT) have been so different? (Grimes and Lobo 2002)

72. Recognition of the **effects of** influenza **on** CHD provides the medical community with a valuable opportunity to further reduce cardiovascular death and morbidity. (Madjid et al. 2004)

73. ... de nombreux traitements ont des **effets** rhéologiques... (Boisseau 2004)

74. Les risques de saignements seraient reliés à l'**effet de** l'ail **sur** la coagulation. (Trahan 2002)

Table 64 — based on data that appears in Appendix J — illustrates the ratio of pattern forms to markers in the two data sets, which is indicative of a relatively high level of variation in pattern structures observed in the analysis.

Table 64. Comparison of ratio of pattern structures to markers in English and French

	English	French	Difference
Pattern structures per marker, overall	1.9	1.7	0.2
Pattern structures per marker, ASSOCIATION	2.4	1.9	0.5
Pattern structures per marker, CAUSE–EFFECT	1.8	1.7	0.1

The English appears to show more variation in pattern form overall than the French relative to the number of markers observed, particularly in the ASSOCIATION relation. As the differences that were observed in this respect may be closely linked to the higher numbers of occurrences of markers in English, however, a more accurate — but still indicative rather than precisely comparable — picture can be obtained by comparing levels of variation for markers observed the same number of times. These are shown in Table 65.

Table 65. Comparison of pattern structure variation for markers of both relations (by number of marker occurrences) in English and French

	English		French		
Number of marker occurrences	Number of markers	Mean number of pattern forms	Number of markers	Mean number of pattern forms	Difference
2	30	1.3	25	1.3	0
3	6	2.0	12	1.8	0.2
4	3	1.7	11	2.2	-0.5
5	8	1.9	5	2.2	-0.3
6	5	3.0	3	2.3	0.7
7	4	1.8	7	1.9	-1
8	2	2.0	1	1.0	1.0
Total	58		64		

These data show relatively little variation between the markers in the two languages, with no coherent trend revealed overall: some of the groups indicate slightly higher variation in English, while others indicate the reverse.

The data for the two relations are shown in Table 66. As in the overall statistics, the level of variation indicates that many of the markers observed are likely to participate in two or more distinct structures.

Table 66. Comparison of pattern structure variation for ASSOCIATION and CAUSE-EFFECT markers (by number of marker occurrences) in English and French

English			French		
Number of marker occurrences	**Number of markers**	**Mean number of pattern structures**	**Number of markers**	**Mean number of pattern structures**	**Difference**
ASSOCIATION					
2	5	1.2	6	1.3	-0.1
3	2	2.5	0	n/a	n/a
4	1	2.0	2	2.5	-0.5
5	3	2.0	2	2.5	-0.5
6	0	n/a	1	3.0	n/a
7	1	2.0	1	2.0	0.0
8	1	2.0	0	n/a	n/a
Total	13		12		
CAUSE-EFFECT					
2	25	1.4	19	1.3	0.1
3	4	1.8	12	1.8	0.0
4	2	1.5	9	2.1	-0.6
5	5	1.8	3	2.0	-0.3
6	5	3.0	2	2.0	1.0
7	3	1.7	6	1.8	-0.1
8	1	2.0	1	1.0	1.0
Total	45		52		

Once again, these data do not show any strong trend towards higher variability either of the two data sets: there are too few markers considered in the ASSOCIATION relation to conclude as to whether the potential trend towards higher variability in the French is real, and the variability from group to group in the CAUSE–EFFECT relation makes conclusions difficult. As such, no evidence of a specific need to further evaluate the variability of pattern structures in either of the two languages was found. However,

these data are not ideally suited to this evaluation, and an opportunity to study variation in a more appropriate context (i.e., with sets of larger, consistent numbers of occurrences for a wider range of markers) could provide a better idea of the overall level of pattern structure variation and of the possibilities of observing trends in this respect in groups of markers established on the basis of parts of speech, relations, or language.

4.8.2.1 Variations in pattern structure involving relative pronouns

One specific type of pattern structure variation observed in the project, which also presents some commonalities with the phenomenon of anaphora (Section 4.9.2.1) involved the presence of relative clauses. This phenomenon is illustrated in Examples 75 to 87. Examples 75 to 78 involve a fairly straightforward structure, in which the antecedent of the relative pronoun, the related element, immediately precedes this pronoun. As such, while the pronoun does take the place of an antecedent, these occurrences pose few of the problems of anaphora, constituting rather a fairly stable variant of pattern form.

> 75. In this environment, HDL changes into a molecule that **promotes** LDL oxidation. (Cabe 2000)
>
> 76. This results in proinflammatory responses and autoimmune reactions, which **contribute** to the atherosclerosis. (Gupta et al. 2004)
>
> 77. ... la perfusion d'angiotensine II induit la formation d'anévrismes, qui **a été reliée à** l'activation des leucocytes circulants. (Michel 2004)
>
> 78. ... c'est une protéine qui **intervient dans** la régulation de la prolifération des cellules. (La Recherche 2002)

This is likely to be best dealt with in pattern design by adapting pattern forms (particularly of verbal patterns) or strategies for identifying related elements as necessary, to accommodate the structures observed. This might involve, for example, allowing variations on pattern forms such as *X that promotes Y* as well as *X promotes Y*, and so on. Nevertheless, this would multiply the number of pattern forms required to

cover the possible variations in a corpus, and in turn require time and effort to develop and computer resources to apply.

However, in addition to these relatively simple structures, more complex ones were also identified. Hierarchically related elements were also often observed in structures including relative pronouns, as in Examples 79 to 81:

79. Antioxidants are molecules that can prevent or reduce the extent of oxidation to the oxidizable substrate. (Kang 2002)

80. La chimiothérapie et l'hormonothérapie sont des traitements systémiques qui **ont pour but de diminuer** la récidive, surtout systémique. (Martin 2003)

81. Les PPAR sont des récepteurs de la superfamille des récepteurs stéroïdiens qui **modulent** la transcription de gènes contenant des éléments de réponse du proliférateur de peroxisome. (Guastalla et al. 2004)

In Example 82, the order of pattern marker elements has been modified. In Example 83, another noun appears between the antecedent noun phrase and the relative pronoun associated with it, and in Example 84, another type of cause (a causal event) has been inserted within the relative clause:

82. It has been recognized that atherosclerosis is an inflammatory disease **in** which various cytokines **play a** significant **role**... (Taniyama and Griendling 2003)

83. A pleiotropic effect reported for CCBs that might **affect** the development of atherosclerosis is the ability of these agents to reduce oxidative modification of LDLs and membrane lipids. (Mason et al. 2003)

84. ... cell recruitment into the developing plaque is enhanced by IL-18, which on ligation to its receptor on ECs, **induces** the expression of the adhesion molecules ICAM-1 and VCAM-1. (Szmitko et al. 2003)

The antecedents of relative pronouns are also not always simple noun or noun phrases, as in Examples 85 to 87, in which the pronouns replace propositions.

85. ...cyclin D1 is frequently overexpressed in human breast DCIS specimens (9, 13), which **confers** a high **risk for** the

development of infiltrating ductal carcinoma. (Wang et al. 2003)

86. ... the other goes through a different class of molecules known as Shc, which **leads to** the activation of the mitogen-activated protein kinase (MAPK) pathway.(Pantaleo and Zonszein 2003)

87. Les monocytes sont alors activés en macrophages (Ma) ce qui **contribue** probablement **à accroître** l'oxydation des LDL... (Arnal et al. 2003)

All of these variations may constitute challenges for applications that attempt to find contexts containing previously identified terms occurring contiguously with relation markers, that use patterns specifying the forms of related elements that can occur with markers, or that attempt to identify related elements automatically.

In regard to this criterion, only a small difference is observed between the data sets (Table 67): 41 cases of structures involving relative pronouns were found in English (constituting 9% of the relation occurrences), while in French 30 cases were found (8%), indicating that although this phenomenon was slightly more frequent in English, the difference in the proportions of relation occurrences affected is not significant (p = 0.740).

Table 67. Comparison of the proportions of relation occurrences containing structures involving relative pronouns (VRp) in English and French

	EN	FR	Total
VRp+[132]	41	30	71
VRp-	401	319	720
Total	442	349	791

As seen in Table 68 and Table 69, in the 41 English occurrences of relative pronouns, 3 pronouns were used: *that* (25 occurrences, 61%), *which* (13 or 32%) and *who* (3 or 7%). These figures produce a ratio of occurrences per marker of 13.7. In the 30 occurrences in French, 5 relative pronouns were found: *qui* (with 21 occurrences or

[132] As noted in the key to interpreting Chi-square tables presented at the beginning of this thesis, + in this table indicates the presence of the criterion being evaluated in the contexts indicated, and – its absence. This convention is retained, where applicable, in the Chi-square tables throughout this thesis.

70% of the total), *ce qui* (4) and *c'est... qui* (2), *dont* (2) and *que* (1). The occurrence to marker ratio of 6.0, reflecting the observation of a wider variety of pronouns in the French data as compared to the English, may indicate that the task of representing these structures may be somewhat more complex in this language.

Table 68. English relative pronouns observed

Pronoun	Occurrences
that	25
which	13
who	3
Total	41

Table 69. French relative pronouns observed

Pronoun	Occurrences
qui	21
ce qui	4
c'est (article)... qui	2
dont	2
que	1
Total	30

Representing such structures — and the pronouns involved in them — in pattern forms may provide access to a significant number of contexts that might be overlooked by more conventional patterns that impose restrictions on the context surrounding marker occurrences. Moreover, the relatively small number of different markers in both languages makes this a relatively achievable goal. However, the task in French could be just slightly more complex, as a wider variety of pronouns was observed in the data in this language.

However, the fact that the distribution of markers is somewhat different — given the concentration of occurrences in a single marker in French, *qui*, and a more widespread distribution of occurrences in English between three markers, *which, that* and *who* — may lead to subtle differences in the possibilities for designing and using these markers. At a formal level, in French, including just a single pronoun in pattern forms could locate a significant proportion of the occurrences of anaphora involving

relative pronouns, while in English, two or even all three of the pronouns observed would be needed to reach this level.

A finer-grained comparison reveals additional variations, as differences in pronoun distribution in the two data sets are tied to the information conveyed. In the French data, persons, things and phenomena may be represented by the relative pronoun *qui* (although the case of persons was not observed in the sample of contexts analyzed). However, in the English data *who* indicates a human antecedent, while *which* and *that* indicate things and phenomena, as in Examples 88 to 90:

> 88. The locoregional management of patients with stage IIIC disease who **respond to** chemotherapy should be individualized. (Shenkier et al. 2004)
>
> 89. ... free radical-scavenging abilities that may **contribute to inhibition of** lipoprotein oxidation. (Davignon 2004)
>
> 90. As compared to the treatment with novantrone, which demonstrated **anti**-tumor **efficacy** with an optimal T/C value of 56.8%... (Du et al. 2003)

This distinction could provide useful information, for example, for applications that perform some level of semantic analysis (e.g., that use semantic classes of actants to disambiguate markers), which would potentially be able to identify contexts involving anaphoric expressions referring to persons without needing to locate antecedents.[133]

Another aspect of the English variation that is potentially useful in interpreting contexts is the distinction between the use of *that* to introduce restrictive clauses (i.e., indicating characteristics of a given element that are particular to a given situation) in contrast to the use of *which* to introduce non-restrictive clauses (which rather provide an inherent characteristic of an element and thus a piece of information that applies in all

---

[133] The distinction in pronouns may also facilitate the work of applications that attempt to use information about semantic classes to aid in the process of resolving anaphora (i.e., locating the antecedents of anaphoric expressions), e.g., by matching terms found in the sentence to a resource that indicates their semantic classes. However, this kind of application is beyond the scope of the discussion in this project.

situations).[134] Access to this kind of information through the differences in relative pronouns is thus possible in English but not in French; unfortunately, this distinction is not universally respected and the possibilities of exploiting it in information retrieval are limited by the degree to which it has been implemented in the corpus texts themselves. A distinction that appears to be more widely — although not universally — respected, and one that exists in both languages, is the use of commas before and after a non-restrictive clause but not a restrictive one. This may also be taken into account when designing pattern forms and in attempting to identify the nature of information provided in such clauses.

The analyses of pattern structure variation and of variation involving relative pronouns do not provide conclusive evidence that it would be worth researching possible interlinguistic differences in pattern form variation further, although very slightly higher levels of variation may be observed in English. What variability was observed may be explained in part by the use of optional elements within a single pattern form (e.g., in the case of copula verbs that may or may not precede participial adjective markers), or distinct forms of patterns, both of which would require an investment of time and effort required to develop these patterns and to refine their form to ensure good performance. Further research, if undertaken, could focus on these aspects in order to evaluate the question more fully. In variations involving the use of relative pronoun structures, the regularities of some structures and parallelism in the forms observed in the two corpora show potential for the development of pattern forms that can deal with the phenomenon. However, more complex structures may be difficult to represent in either language. Moreover, in the analysis of the specific relative pronouns noted, the French occurrences showed a higher level of variation that would likely have to be reflected in the development of pattern forms.

---

[134] These may be contrasted in structures such as *drugs that reduce cholesterol levels* — a characteristic that does not apply to all drugs and that is thus analogous to a statement using a quantifier to indicate uncertainty, such as *some drugs reduce cholesterol levels* — and *statins, which reduce cholesterol levels*, which is analogous to *all statins reduce cholesterol levels*.

For the group of variations as a whole (of marker forms, in voice of verbal pattern markers, and in pattern structures), a tendency towards higher variation in English was observed, although this variation was observed in different degrees: most pronounced in the case of the active/passive variation of verbal marker forms, slightly less so in the case of marker variation, and only very subtly (if at all) in pattern structure variation. The data suggest that more research should be done to evaluate the cumulative effect of these types of variation on the process of pattern set design, and to determine if variation could necessitate the development of additional pattern forms (or optional variations on pattern forms), requiring a greater investment of time and effort in identifying and refining these structures in English.

Clearly, variation is an issue that must be taken into account in both languages. This raises a significant point in relation to the analysis of the numbers of markers required for pattern-based tools: as multiple pattern forms or pattern structures may be required for a single marker, the investment of time and effort required to add a marker to a pattern set is increased accordingly. Comparison of the languages in terms of the process of pattern set design thus requires that these factors be considered together.

The data analyzed above in Sections 4.2 and 4.4 suggest that in order to retrieve similar numbers of contexts in the two languages, more markers may be required in French; when the factor of pattern variation (i.e., the fact that each marker may require more than one pattern form for many applications) is taken into account, the difference between the two languages is likely to be even more pronounced. The compounding of these factors could counteract somewhat the potential tendency to higher variation in marker form in English. The end result may be a more comparable number of pattern forms required in the two languages than it might appear from an analysis of each factor individually. More structured research into these factors could clarify their interactions.

The different sources of variation are also important to take into account in the analysis of their impact on different types of pattern-based applications. In the case of the simplest strategies for locating relation occurrences, involving the representation of

markers only (e.g., as character strings or regular expressions possibly accompanied by information about POS class), pattern structure variation is relatively unimportant, but the number of different markers required is very significant. Variation in marker form is also pertinent to many of these kinds of applications, as applications that specify the most complete (and thus often most precise) forms are likely to be affected by this variation. More highly developed pattern forms that specify the structures in which markers occur and/or that attempt to identify related elements automatically will of course be affected by all of these factors, including that of pattern form variation.

4.9 Number and form of the elements linked by the markers

Next to be analyzed is the number and form of the elements linked by a relation marker. Like the markers, in many applications (and specifically those that impose restrictions on the form of related elements, either by specifying the POS class of elements surrounding relation markers, or by searching for markers in proximity to previously identified terms or candidate terms), these elements must be adequately described and represented in order for (complete) KRCs and the information they convey to be successfully identified in corpora.

4.9.1 Multiple elements sharing a role in a relation

In both corpora, the prototypical X + [MARKER] + Y pattern was often modified by the apparition of two or more elements in one of its slots; this occurred in 43% of the English relation occurrences and 49% of the French, as shown in Table 70.

Table 70. Comparison of proportions of occurrences of multiple elements (ME) sharing a role in a relation in English and French

	EN	FR	Total
ME+	189	169	358
ME-	253	180	433
Total	442	349	791

This clearly indicates that pattern forms must attempt to take this kind of variation in the canonical pattern form into account, or risk overlooking a significant proportion of the potentially useful contexts in such corpora. It also is apparent that while this phenomenon does not occur with significantly different frequency in the two data sets ($p = 0.112$), there is a chance that further data would reveal a trend towards significance in the higher proportion of French relation occurrences that involve multiple related elements.

Conceptually, contexts in which multiple elements share a slot in a pattern contain more (or more specific) information than a standard context linking only two elements. This fact underlines the importance of retrieving and analyzing such contexts.[135] However, accessing this information requires formal adaptation.

This phenomenon — and any difference in its prevalence — may affect both the development and performance of pattern sets at a formal level. A relatively minor effect may be noted for tools that use simpler pattern forms (e.g., character strings, lexico-syntactic forms of markers alone), and specifically those that extract contexts of fixed length, as the length of the contexts required for complete information to be located may increase.[136] A more significant effect may be observed in the substantially increased complexity of designing patterns that specify the forms (e.g., part of speech classes) in which the elements linked by a relation are expressed. Pattern forms must take these structures into account in order to ensure that the structures indicated for relation elements do not exclude these cases *a priori* because of their variation from "standard"

---

[135] By extension, the potential for observing a higher proportion of contexts containing multiple elements in French may indicate that French contexts could provide access to more information.

[136] If this phenomenon were observed to be more common in French, longer contexts might be required in this language for KRC extraction using tools that extract fixed-length contexts.

form, and that all of the elements linked by the relation marker are identified.[137]

These effects are even more significant for pattern-based tools that attempt to identify related elements automatically; these applications must not only recognize such contexts and ensure that all of the related items are extracted, but also must analyze the structures in which these items appear in order to distinguish between the separate elements and present these to the user.[138] Finally, pattern-based tools that attempt to identify contexts containing specific terms or candidate terms in connection with may also be affected by the presence of multiple related elements. As in the case of the other types of applications identified above, pattern forms must be adapted to accommodate the variation involved in the occurrence of multiple elements, and particularly to allow for other elements occurring between the marker and the term or candidate term in question. In addition, as the form in which terms and candidate terms appear may be affected by the presence of other elements sharing a slot in a knowledge pattern, further complicating the identification of pertinent contexts and potentially reducing the recall of these applications.

In this section, the results of the observations of multiple elements that share a role in relation occurrences will be discussed, according to a typology established from the results observed in the corpora. The presence of variant expressions of a single related element (including the specific case of abbreviations and symbols) will be described, followed by the phenomena of conjunction, disjunction and conjunction/disjunction of related elements and finally of elements linked by GENERIC–SPECIFIC relations. The frequency of the different forms of this phenomenon is summarized in Table 71.

---

[137] A higher proportion of occurrences of this phenomenon in French would indicate that it is particularly important to develop these strategies in this language, as not doing so would have a particularly great impact. Conversely, the return on investment for the development of adequate strategies for analyzing these cases would thus be higher in French.

[138] Once again, a higher prevalence of this phenomenon in French would increase the importance of taking these cases into account and the impact that difficulties in this process would have on results.

Table 71. Detailed comparison of proportions of relation occurrences involving multiple elements sharing a role, in English and French

	EN	FR	Total
Abbreviations and symbols	22	19	41
Other variants	4	8	12
Conjunction	127	116	243
Disjunction	17	18	35
Conjunction/ Disjunction	2	1	3
Generic(s) and specific(s)	68	61	129
All multiple elements	189	169	358
None	253	180	433
Total	442	349	791

The ways in which multiple elements that share a role in a relation — and a slot in a knowledge pattern — are related to one another will affect the type of additional information that is expressed and thus the applications for which it may be useful. This in turn affects the strategies that are likely to be useful for accessing and processing this information. The specific types of relationships observed between elements are described individually below. In addition, two related phenomena will be discussed: the ellipsis of part of multiple elements, and the repetition of a pattern marker or of part of a complex pattern marker in the presence of multiple elements.

4.9.1.1 Variant expressions of a single related element

The first category of multiple elements participating in a relation involves the presence of two or more expressions denoting a single concept participating in a relation, as in Examples 91 to 94:

91. The most common **cause of** brain infarction is hardening of the arteries (atherosclerosis). (DiGiovanna 1999)

92. Cette stratégie à pour effet de réduire le cholestérol total et le cholestérol LDL (ou mauvais cholestérol) de l'ordre de 10 à 20 % et de 12 à 16 % respectivement. (Blais 2001a)

93. … toute variation brutale **entraîne** la rupture de l'équilibre et l'oxydation des éléments sensibles de la cellule, une situation que l'on qualifie de "stress oxydant". (La Recherche 1997)

94. La radiothérapie adjuvante sera offerte à toute femme ayant subi une mastectomie partielle, que ce soit **pour** une tumeur infiltrante ou in situ (intracanalaire). (Martin 2003)

As may be observed in Examples 93 and 94, this phenomenon may be observed in fairly complex structures or with interruptions of additional elements, which can pose significant challenges for automating the retrieval of such elements by automatic applications.

One specific manifestation of this phenomenon, shown in Examples 95 to 97, is significantly more frequent than other types in the corpus texts. In these cases an abbreviation or symbol (e.g., denoting a molecule, substance, disease or treatment) appears with the full form of the term or other linguistic unit.

95. The inflammatory **marker** C-reactive protein (CRP) can **indicate** low-grade chronic inflammation… (MacKenzie 2004)

96. More recently, it has become apparent that reactive oxygen species (ROS) also **play a role in** the development of vasculopathies… (Griendling and FitzGerald 2003a)

97. F. Perret … a ensuite analysé le **lien** existant **entre** traitement hormonal substitutif de la ménopause (THSM) **et risque de** cancer du sein chez les sujets à risque [12]. (Cottu and Espié 1999)

Such contexts may provide pertinent information in addition to the primary relation involved, making these contexts quite information-rich. Moreover, the adequate representation of these kinds of structures at a formal level is important for designing pattern forms that correspond to the context in which pattern markers occur and in the automatic identification of related elements. Analysis of these occurrences can be challenging, especially in cases in which an abbreviation or other variant is given for one part of a more complex element involved in a relation, as in Examples 98 to 101:

98. Because LDL **upregulates** angiotensin II receptor type 1 (AT1) receptor expression, … (Griendling and FitzGerald 2003)

99. ...the other goes through a different class of molecules known as Shc, which **leads to** the activation of the mitogen-activated protein kinase (MAPK) pathway. (Pantaleo and Zonszein 2003)

100. Cette protection **passe** vraisemblablement **par** une activation de PKG, protéines kinases dépendantes du GMPc, mais les substrats de ces kinases sont encore mal defines... (Kolb 2001)

101. Les molécules qui **modulent** sélectivement l'activation des récepteurs hormonaux (SERM) par compétition avec les oestrogènes circulants... (Vinatier and Orazi 2003)

In these cases, reliably identifying the correct form of a participant in a relation may be difficult in an automatic approach, although the more knowledge can be gathered about the possible forms of such cases, the more successful such attempts are likely to be.

The co-occurrence of two or more variant expressions of an element was not observed in a very large number of occurrences in either language. The overall frequency was similar in the two data sets (26 occurrences in English and 27 in French, which constitute approximately 14 to 16% of the occurrences of multiple elements and between 6 and 8% of the total relation occurrences). The data from the two corpora are thus quite comparable in this respect (according to the Chi-square test, $p = 0.300$ for the comparison of these cases as a proportion of relation occurrences and $p = 0.555$ as a proportion of occurrences of multiple elements).

The proportions of abbreviations and symbols as compared with other types of variants were also relatively comparable, with the majority of cases involving abbreviations (85% in English and 70% in French) and far fewer (15% in English and 30% in French) involving other types of variants. (A Chi-square test to compare these proportions reveals a non-significant difference, $p = 0.215$.)

Each category and the contexts in which they appeared are described below.

4.9.1.1.1 Abbreviations and symbols

The proportions of relation occurrences in which abbreviations or symbols accompanied full forms of expressions indicating a concept participating in a relation are comparable in the two data sets; the phenomenon occurred in 12% of the occurrences involving multiple elements in English and 11% in French, and in 5% of the total relation occurrences, as shown in Table 72 ($p = 0.769$).

Table 72. Comparison of proportions of relation occurrences involving abbreviations or symbols (AB) in English and French

	EN	FR	Total
AB+	22	19	41
AB-	420	330	750
Total	442	349	791

In the majority of cases, the abbreviation for a term was presented in parentheses immediately following the full form (as in Example 102); however, in both corpora some cases were found with the full form appearing in parentheses (as in Example 103).

102. The presence of TNF- [alpha], IL-6, and other cytokines **cause** hepatic **production of** C-reactive protein (CRP) ... (Pantaleo and Zonszein 2003)

103. We studied the expression of DMBT1 (deleted in malignant brain tumor 1), a putative tumor **suppressor** gene, in normal, proliferative, and malignant breast epithelium... (Braidotti et al. 2004)

These observations indicate that the phenomenon will likely present similar opportunities — and challenges — for exploiting the additional information in these contexts, both in designing pattern forms that can process these occurrences, and in using the supplementary information contained in the contexts (e.g., for identifying relations of SYNONYMY). Given the proportion of relation occurrences affected in the two corpora, the potential for identifying pertinent information over and above the core relation expressed, and the pertinence of representing structures in which this phenomenon occurs in order to ensure that applications that attempt to do so can recognize and/or identify and extract related elements accurately, it may be beneficial to

take the phenomenon into account in designing pattern forms. The relatively stable structure in which these items often occur shows promise for representation. As in general structural variations observed were fairly minor — and the relation of SYNONYMY is symmetric in any case — these are not likely to critically affect the representation of the phenomenon in pattern forms or the interpretation of the information indicated. The major challenges for pattern design — the ambiguity of paralinguistic indicators such as parentheses and the occurrence of abbreviations within more complex elements — are also likely to be pertinent in both languages.

Examples 104 and 105 illustrate a phenomenon particular to French that may be very pertinent for the extraction of information from corpora such as this one: the use of English-language terms and/or abbreviations in French texts. As these Examples show, there may be a certain amount of variation between occurrences, as both full terms and abbreviations may be English forms, terms in full may be presented in French but accompanied by the English forms of symbols or abbreviations, and so on.[139]

> 104. L'hypertrophie des cellules musculaires lisses induite par l'angiotensine II résulte de l'activation **des** protéines kinases mitogéniques (MAPKs)... (Bonnefont-Rousselot et al. 2002)
>
> 105. Les ERO formées par la NADPH oxydase des cellules musculaires lisses sont également impliquées dans l'activation par la thrombine **du** facteur de transcription hypoxia-inducible factor-1 (HIF-1)... (Bonnefont-Rousselot et al. 2002)

Not surprisingly, no cases were found in which French terms were used in English texts. This apparent influence of English is a phenomenon that is worth monitoring in a larger context, in order to determine whether these phenomena are likely to pose difficulties for work on French texts. Certainly, the automatic extraction of such

---

[139] In addition, relatively atypical formulations may be found in French under the influence of the English, for example the pluralization of the abbreviation MAPK in Example 104 above. The processing of such variations might prove to be problematic: if an application attempts to identify the base form of related elements, this phenomenon could pose problems because it is atypical in French and thus might not be properly analyzed using standard rules; the impact of this phenomenon is nevertheless not likely to be very significant given the infrequency with which it was observed.

elements by applications used to enrich ontologies or terminological resources could pose difficulties in many contexts of use: for example, it is unlikely that these abbreviations would be considered good candidates for inclusion in terminological resources with full status as headwords for entries or term records. However, given that they are clearly used in the literature (as their presence in the corpus attests), their inclusion in resources as possible (if perhaps abusive) variants of French terms could be helpful to users. Their potential for helping to establish interlinguistic equivalence between terms is also not negligible. For these reasons, such occurrences might be beneficial (and thus important to take into account) in pattern-based tools intended for some specific purposes, but not others.

4.9.1.1.2 Other variants in expression of a related element

Other types of variants were found in 4 contexts in English (2% of the relation occurrences with multiple related elements and 1% of the total relation occurrences) and 8 contexts in French (5% of the occurrences of multiple elements and 2% of the total). The Chi-square test does not reveal any significant difference ($p = 0.113$), although a higher proportion of occurrences was observed in French.[140] Both apposition and parentheses were used to introduce variants in the two data sets, as illustrated below in Examples 106 to 109:

106. PD98059 and U0126 also **block** activation of MEK5, ... the kinase that activates ERK5, the sole member of the fourth MAPK family. (Force et al. 2004)

107. The most common **cause of** brain infarction is hardening of the arteries (atherosclerosis). (DiGiovanna and Adams 1999)

108. Cette protection passe vraisemblablement par une activation de PKG, protéines kinases dépendantes du GMPc, mais les substrats de ces kinases sont encore mal définis... (Kolb 2001)

---

[140] Although the numbers of occurrences observed are low, the expected values required for the application of the Chi-square test are above 5, indicating that the test can be accurately applied in this case.

109....des **facteurs de** transcription comme myc, jun, fos ou erbA (récepteur de l'hormone thyroïdienne T3).... (Blanchard 2003)

Given its rarity, in and of itself, the phenomenon would not likely be particularly productive to account for in pattern design and in the processing of contexts extracted to exploit the additional information present. However, both formal and conceptual similarities with other phenomena such as the introduction of abbreviations may mean that some variants — specifically those indicated by parentheses — could be processed in parallel with these more frequent cases. The ambiguity of the paralinguistic indicators used, however, introduces its own challenges. Moreover, variants introduced by apposition would be much harder to identify automatically. Example 94 above, in the presentation of the synonym *tumeur intracanalaire* accompanying *tumeur in situ*, also illustrates potential difficulties in representing the structures in which variants occur because of the presence of another term and the modification of the form of the complex term (cf. Section 4.9.1.2 on the disjunction of related elements and Section 4.9.1.4 on ellipsis of part of complex terms). The association of several phenomena observed here can clearly complicate the task of automatic analysis and identification of related elements.

One difference was observed, however, and should be evaluated for its significance for pattern design: this is the presence of lexical indicators in French (*ou* in Example 92, *que l'on qualifie de* in Example 93), but not in English. These kinds of indicators — which are similar to some that have been studied in lexical knowledge-pattern-based approaches to finding synonyms or metalinguistic information about terms (e.g., Rodriguez Penagos 2004, 2004a) — raise some interesting questions.

First and foremost is the task of differentiating these indicators from the elements that they accompany, which involves an analysis of the forms present and potentially requires a list of potential indicators of this type. Second is the nature of these indicators themselves, which may be quite ambiguous (e.g., in the case of *ou*, which also indicates disjunction of related elements, cf. Section 4.9.1.2), and may also

vary in their productivity (e.g., frequency in corpora). Preparing pattern forms that would be able to identify these kinds of variants in the expression of related elements — and differentiate them from other occurrences — could thus be a difficult task, and one that in light of these data does not appear likely to be rewarded by a significant return on this investment. Alternatives might include considering the use of pre-established patterns developed in research projects focusing on the extraction of SYNONYMY or metalinguistic information, although their integration in more complex pattern forms would not be without its difficulties.

4.9.1.2 Conjunction and disjunction of related elements

The conjunction and disjunction of elements that may share a role in a relation, as below in Examples 110 to 121, are among the semantically and formally simpler cases of multiple related elements. These occurrences may be relatively straightforward in form, as in Examples 110 to 113.

> 110.... PD98059 and U0126 also **block** activation of MEK5... (Force et al. 2004)
>
> 111. Increased glucose levels, FFAs, inflammatory cytokines, and oxidative stress **cause** activation of NF-[kappa]B with initiation and/or perpetuation of the inflammatory process as shown in Figure 4. (Pantaleo and Zonszein 2003)
>
> 112. L'obésité, le syndrome métabolique et le diabète **accroissent** notablement le **risque de** maladies cardiovasculaires. (Lambert 2002)
>
> 113. Dans ce travail [12], l'exercice physique n'a pas eu d'**effet sur** le cholestérol total ou le LDL cholestérol. (Ferrières 2004)

They may also involve more complex markers or structures, as in Examples 114 to 117.

> 114. Oxidative stress has been **linked to** the activation of both NF-[kappa]B and AP-1. (Granger et al. 2004)
>
> 115. This enhances retention of the lipoprotein and possibly **triggers**, along with oxidation, the formation of a recognizably foreign substance which requires macrophage recruitment to the locality... (Caslake and Packard 2003)

116.... lorsque l'association de metformine et d'une sulfonylurée ne **permet** pas une **maîtrise** optimale du diabète ou ne peut être utilisée **en raison d'**une contre-indication ou de l'intolérance à l'un de ces médicaments (Leblond 2001)

117.Les stratégies hormonales **de prévention** pourraient ainsi **concerner** à la fois les tumeurs sporadiques et les tumeurs génétiques. (Vinatier and Orazi 2003)

Quite complex structures may also be observed, as in Examples 118 to 121, in which the conjunction or disjunction links only parts of more complex elements.[141]

118.The acceptance that endothelin may **play an** important **role** in the pathogenesis and clinical manifestations of certain cardiovascular disorders has paved the way for the development... (Ram and Venkata 2003)

119....excessive activation of cytotoxic lymphocytes **causes** incidental vascular and tissue damage... (Umehara et al. 2004)

120.Dans l'adénose sclérosante, affection bénigne du sein **caractérisée par** une prolifération des cellules épithéliales et myoépithéliales, Clarke et al. [13] ont montré ... (Angèle et al. 2001)

121.Par contre, p53 **réprime** la transcription de gènes anti-apoptotiques comme bcl-2 et comme la NOSi elle-même. (Kolb 2001)

Both the overall frequency and the distribution of these phenomena were observed to be relatively comparable in the two data sets, with the figures in French just slightly higher than in English: for conjunction 69% of the occurrences of multiple elements as compared to 67%, and 36% of the total relation occurrences as compared to 29%, and for disjunction 11% as compared to 9% of the occurrences of multiple elements and 5% as compared to 4% of the total relation occurrences. The proportions of relation occurrences containing conjunction and disjunction of related elements are compared in Table 73 ($p = 0.172$) and Table 74 ($p = 0.373$).

---

[141] Some particularities of this latter phenomenon are described further in Section 4.9.1.4.

Table 73. Comparison of proportions of relation occurrences involving conjunction of related elements (CR) in English and French

	EN	FR	Total
CR+	127	116	243
CR-	315	233	548
Total	442	349	791

Table 74. Comparison of proportions of relation occurrences involving disjunction of related elements (DR) in English and French

	EN	FR	Total
DR+	17	18	35
DR-	425	331	756
Total	442	349	791

For the case of multiple elements linked by conjunction/disjunction, with two occurrences in the English sample and only one in the French sample, it is not possible to draw any conclusions.

At a conceptual level, these contexts indicate relationships between multiple pairs of concepts, and thus are particularly information-rich. Given this richness and its importance for knowledge extraction, the data indicate that these phenomena are very important to take into account in designing pattern forms in both languages, in order to ensure that potential KRCs are recognized and that complete information is extracted. While the interlinguistic difference observed was not significant, it is nevertheless possible that the higher proportion of occurrences observed in the French data could indicate a somewhat greater impact in this language, increasing the return on the investment of time and effort in adapting pattern forms. Some potential applications for this information, once identified, are discussed in Section 5.5.3.3.

In both languages, fairly straightforward structures involving two or more elements presented separately were observed. However, in many other cases phenomena that can pose more serious difficulties for representation — such as the presence of conjunction or disjunction within a more complex element, or the appearance of the

elements that share a role in a relation in widely separated parts of a context — are likely to be encountered in both languages. The formal analysis and representation of such cases would require a significant investment of time and effort. However, on a positive note, the fact that similarities in the two data sets were strong suggests that this investment in one language is likely to be adaptable to the other as well, and thus could be far more easily justified than it would be in one language alone.[142]

One of the approaches that may be considered in the development of pattern forms that can accommodate the conjunction and disjunction of elements sharing a role in a relation is the creation of formal representations of the structures in which these elements may occur, and their integration into a range of pattern forms. These may also serve as the basis for distinguishing the separate elements sharing the relation role for applications that attempt this task, and identifying the relationship between them. Some of the lexical items that indicate conjunction or disjunction and that thus may be useful for this kind of task — as well as their frequency in the relation occurrences analyzed — are presented below. It is nevertheless important to note that the more complex structures indicating conjunction and disjunction would pose significant challenges in such an approach, and would likely be very difficult to represent.

4.9.1.2.1 Indicators of conjunction and disjunction[143]

The numbers of lexical units indicating conjunction and disjunction were similar in the two data sets, with 12 different markers of conjunction observed in English (Table 75) and 16 in French (Table 76). *And* and *et* were by far the most frequent, accounting for

---

[142] It is likely that the needs of a specific project will determine the strategies applied to deal with such cases; for example, when human interpretation of contexts is possible, this may be the most efficient option.

[143] In this section and the sections below, the term *indicator* will be used to refer to the lexical units that mark relationships between multiple related elements in the contexts analyzed. This is done in order to avoid confusion with the term *marker*, used to denote the lexical units that indicated the CAUSE–EFFECT and ASSOCIATION relations analyzed in this research. However, many of the indicators discussed are also markers of semantic or conceptual relations (e.g., SYNONYMY, GENERIC–SPECIFIC) in their own right.

121 of the 150 English cases and 101 of the 123 French cases, respectively.[144] However, in English the distribution of occurrences among the less frequent markers was very slightly more even, with a wider variety of more frequent markers observed (*both... and*, *as well as*, *with*).

Table 75. English indicators of conjunction of related elements

Indicator	Occurrences
and	121
both... and	7
by	5
as well as	4
with	3
both	2
through	2
together (with)	2
along with	1
all	1
apart from... also	1
on	1
Total	150

Table 76. French indicators of conjunction of related elements

Indicator	Occurrences
et	101
en	5
avec	3
par	2
à la fois... et	1
accompagné de	1
ainsi que	1
association de... et de	1
cependant	1
d'une part... et d'autre part	1
également	1
en association avec	1
non seulement... mais aussi	1
outre	1
puis	1
via	1
Total	123

[144] Moreover, they also appeared as part of more complex markers, such as *both... and*, *à la fois... et*, *association de... et* and *d'une part... et d'autre part*.

The distribution of the indicators, with their occurrences expressed as a percentage of the total, is illustrated in Figure 21.

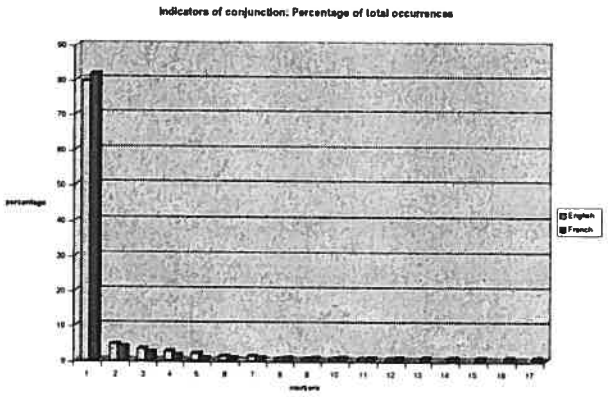


Figure 21. Indicators of conjunction: Percentage of total occurrences

In the context of pattern design, these data indicate that in both languages it seems advisable to ensure that patterns allow for conjunction of elements using *and* and *et*. In English it might also be marginally more productive to develop forms that include more of the indicators observed. The question nevertheless remains open, and would be worth investigating further: is the investment in time and effort to develop pattern forms to account for the many different indicators of conjunction that may be observed likely to be reflected in a return that justifies these measures? The necessary level of recall as well as the volume of data (i.e., the size of corpora) to be processed may help to provide an answer to this question: of course, the more data are available, the higher the return is likely to be. Another element that should be taken into account in answering this question is the form of the markers, and specifically the presence of more complex indicators in French (e.g., *association de… et*, *soit… soit*). The complexity involved in representing the structures in which these kinds of markers participate may be considerably higher, and these structures may introduce more noise in results than simpler indicators.

Moreover, each of the five complex indicators observed accounted for less than 1% of the occurrences analyzed, whereas in English one of the two complex markers, *both... and*, was observed in more than four times that proportion. This suggests that the English indicators may produce a more significant return on investment of time in developing pattern forms, while that of the French markers could be less satisfactory

In addition to lexical indicators of conjunction, occasionally in English — in quite irregular structures — paralinguistic indications of conjunction, including commas and forward slashes (/) were also observed. These structures may be more difficult to represent in pattern forms, due to their ambiguity. In some French occurrences, conjunction was also indicated by paralinguistic means, such as a comma (7 contexts):

122. Cette croissance annuelle de 14 à 16 % est vieillissement démographique, augmentation de l'espérance de vie, exigence d'une meilleure qualité de vie. (Chevallier et al. 2003)

123. Nous avons recherché chez tous les patients les **facteurs de risque** d'athérosclérose (diabète, hypertension artérielle, tabagisme, hormonothérapie, intoxication alcoolique, dyslipidémie, hérédité)... (Desauw et al. 2002)

While a formal representation of structures involving lexical indicators — at least of a prototypical nature — should be possible to develop, those involving paralinguistic markers may prove more difficult because of these markers' ambiguous nature. The prevalence of the phenomenon may be slightly higher in French, but the small numbers of occurrences make it difficult to draw conclusions.

The lexical indicators of disjunction in the two data sets were comparable in number and variety, as shown in Table 77 and Table 78. As in the case of conjunction, the prevalence of the prototypical indicators of disjunction, *or* and *ou*, indicates that a large proportion of occurrences of disjunction may be identified using a single marker, which would be relatively easy to include in pattern structures. It is notable, however, that disjunction occurred far less frequently in the results than conjunction; this indicates that investments of time and effort in creating pattern forms that allow for occurrences

of this phenomenon will have a significantly lower return than in the case of conjunction.

Table 77. English indicators of disjunction of related elements

Indicator	Occurrences
or	11
but not	3
unlike (with)	2
compared with	1
Total	17

Table 78. French indicators of disjunction of related elements

Indicator	Occurrences
ou	13
mais pas	1
parfois	1
plutôt que	1
soit... soit	1
Total	17

Observations indicate, however, that at least typical indicators of conjunction and disjunction could often be integrated into a single pattern form, with only the marker distinguishing between the two possibilities. This would offer considerable potential for formal representation of structures involving multiple related elements with relatively little investment of time and effort.

In both cases of conjunction/disjunction observed in English, the lexical indicator *and/or* was present, and in the single context in which conjunction/disjunction was observed in the French sample, the lexical marker *et/ou* was used. Thus the lexical indicators were also similar in nature. The inclusion of lexical markers of this type could also be envisioned in a single pattern form.

Another phenomenon was noted in both languages and may also be important to take into account when designing pattern forms: the use of specific lexical indicators, such as *by* and *through* in English and *en* in French, to link different types of causes (cf. the classification of the CAUSE–EFFECT relation by Nuopponen (1994; cf. Section

1.5.2.5), as well as the analysis of causation by Kahane and Mel'čuk (forthcoming; cf. Section 1.5.2.4)). Examples 124 to 128 illustrate this phenomenon. These contexts generally indicate both an entity (i.e., in Nuopponen's terminology, a causative agent) and an event (i.e., producing cause), that play a role in producing an effect. However, in some cases (as in Example 128) two or more distinct events may be indicated.

124. By inhibiting ACC, AMPK **elevates** fat oxidation. (Force et al. 2004)

125. The R +enantiomer of amlodipine... **prompts** the production of NO, ultimately through the activation of eNOS. (Mason et al. 2003)

126.... IL-18, which on ligation to its receptor on ECs, **induces** the expression of the adhesion molecules ICAM-1 and VCAM-1. (Szmitko et al. 2003)

127. Les agents anticancéreux interférant avec les microtubules cellulaires (taxol, colchicine, nocodazole, vinblastine, vincristine, 17-β-estradiol, 2-méthoxyestradiol) **stimulent** la transcription de Cox2 en favorisant la liaison du facteur de transcription AP1 à l'élément de réponse de l'AMP cyclique du promoteur de Cox2... (Guastalla et al. 2004)

128.... l'expression de Cox2 **favorise** la prolifération tumorale en inhibant l'apoptose, en stimulant la néo-angiogenèse et en favorisant le pouvoir invasif et métastasiant des cellules malignes. (Guastalla et al. 2004)

Both of these types may be identified as causes, although they are not connected by a typical lexical indicator of conjunction. These indicators may nevertheless provide valuable clues for applications that attempt to differentiate between different types of causes in order to sort contexts automatically or to identify cases involving a particular type of element to a user. However, these cases are also particularly difficult to represent formally, as they generally have relatively complex structures, often with the two elements separated by other elements in the pattern structure. This would pose challenges for applications that seek to recognize and analyze such cases. More data would be required to create adequate formal representations of these cases.

Moreover, while the forms of the elements linked by these indicators (e.g., the verbal form of the elements introduced by *by* or *en*) may facilitate the identification of

particular types of causes in both languages, such non-nominal forms may also pose problems for applications that specify that related elements must occur in particular (usually nominal) forms (see Section 4.9.2).

At a practical level, the representation of these kinds of combinations is unlikely to be implemented in the same way as those for other, simpler forms of conjunction or disjunction, as it would be very difficult to integrate a standard representation of the structures observed into more general pattern forms. This would reduce the possibilities for maximizing return on investment in developing pattern forms to deal with the phenomena.

4.9.1.3 GENERIC–SPECIFIC relations between elements

In the final category of multiple elements, the elements sharing a role are connected in a hierarchical relationship, as in Examples 129 to 134:

129. The endothelium contributes to the regulation of vascular tone, platelet aggregation, and other processes **relevant to** atherosclerosis. (Schwartz 2003)

130.... although hs-CRP and other inflammatory markers such as IL-6 may independently **predict** adverse cardiovascular events ... (Torres and Ridker 2003)

131....the human subjects had a modest but significant reduction in key **markers of** blood vessel inflammation: C-reactive protein, tumor necrosis factor, and the interleukins IL-1 and IL-6... (Cabe 2000)

132.... le cumul de certains gènes prédisposants et, bien sûr, les maladies métaboliques qui en **découlent**, c'est-à-dire le diabète, les dyslipidémies et l'hypertension artérielle, sont les **conséquences du** mode de vie adopté par les humains... (Essiambre 2003)

133. La chimiothérapie et l'hormonothérapie sont des traitements systémiques qui **ont pour but de diminuer** la récidive, surtout systémique. (Martin 2003)

134. L'obésité est un facteur susceptible d'**intervenir dans** de nombreuses maladies: maladies cardiovasculaires, diabète, hypertension artérielle, accidents vasculaires cérébraux, embolies pulmonaires, certains cancers, ostéoarthrite, affections de la vésicule biliaire, anomalies respiratoires, dont notamment l'apnée du sommeil. (Poirier 2003)

In these cases, although the elements that are connected can both or all be considered to participate in the same kind of CAUSE–EFFECT or ASSOCIATION relation, another potential relation, generally the GENERIC–SPECIFIC relation, is also present. (However, it should be noted that not all members of the class of generics are necessarily involved in the relationship indicated.)

As illustrated above, hierarchical relations may take either simple or rather complex forms (for example, with multiple levels of a given hierarchy represented), and are often accompanied by either conjunction or disjunction of two or more specific elements.[145] As a result, these contexts are relatively complex to analyze (at least automatically or semi-automatically), but also extremely valuable for information extraction. The value of these contexts increases exponentially, since they provide not only examples of two separate relations, with at least three separate relation occurrences, but also information about the potential for inheritance of relations from generics to their specifics (including, potentially, specifics not mentioned explicitly in the context itself). For example, in Example 131 above, it may be inferred that IL-1 and IL-6 are interleukins, that interleukins, along with C-reactive protein and tumour necrosis factor, are markers of inflammation, and that both IL-1 and IL-6 are also markers of this inflammation.

However, it is clear that the interpretation of the possible extension of the relation to other element pairs on the basis of inheritance from generic to specific relies

---

[145] In addition, the conjunction of an additional element, *cumul de... gènes prédisposants* may be observed in Example 132, and in Example 133, ellipsis of part of the more complex specific term, *récidive systémique*, is observed, no doubt due to the presence of the generic term and head of the specific term immediately preceding it.

on human input, and should be confirmed manually. This is certainly the case when there is a disjunction of specific elements, which may disqualify one of the specifics from participating in a given relation (e.g., in a structure such as *Xs such as X' and X'', but not X''', are associated with*...), but also may not (e.g., as in a structure such as *Xs such as X' or X'', are associated with*...). In addition, a number of other cases may also require careful analysis of the information present, and potentially other input, for judgments to be made.

A fairly large proportion of the relation occurrences analyzed involved elements linked by a GENERIC–SPECIFIC relation: 15% of the total relation occurrences in English and 17% in French, and 36% of the cases of multiple elements in both languages (Table 79).

Table 79. Comparison of the proportions of relation occurrences involving GENERIC–SPECIFIC relations between related elements (GS) in English and French

	EN	FR	Total
GS+	68	61	129
GS-	374	288	662
Total	442	349	791

This difference in the proportions of contexts containing this phenomenon is not statistically significant ($p = 0.428$ overall and $p = 0.981$ as a proportion of the occurrences of multiple elements), although the proportions are slightly higher in French.

As explained in Section 2.6.3, such occurrences are rich in valuable information and ideally, in the context of automatic relation extraction, should be retained and completely analyzed. In a pattern-based approach, this would generally involve developing pattern forms that represent structures such as those observed in the corpora, including the indicators of the relationship between these multiple elements.

In both languages, specifics were introduced using paralinguistic indicators and apposition. In English, apposition (16 cases) or paralinguistic indicators such as a colon or parentheses (in combination with another, lexical indicator) were used to introduce specifics, as in Examples 135 to 137. In 17 cases in French, apposition was used, as illustrated in Examples 138 and 139; in others, paralinguistic indicators such as a colon (e.g. Example 140) or parentheses (e.g., Example 141) were used.

135....this agent can be cleaved by cathepsin B, an enzyme present in biologically active macrophages 91 and **implicated in** the pathogenesis of atherosclerosis. (Jaffer and Weissleder 2004)

136.... IL-18, which on ligation to its receptor on ECs, **induces** the expression of the adhesion molecules ICAM-1 and VCAM-1 (Szmitko et al. 2003)

137. Emerging data reveals that a large number of additional proteins (i.e., growth factors) **influence** the transcriptional activation of ER[alpha] and possibly ER[beta]. (McCance and Jones 2003)

138. Ce dernier est en effet capable de stimuler le recrutement et l'assemblage des sous-unités p47phox et p67phox, étape **nécessaire à** l'activation de la NADPH oxydase. (Bonnefont-Rousselot et al. 2002)

139. Dans l'adénose sclérosante, affection bénigne du sein **caractérisée par** une prolifération des cellules épithéliales et myoépithéliales, Clarke et al. [13] ont montré... (Angèle et al. 2001)

140. L'obésité est un facteur susceptible d'**intervenir dans** de nombreuses maladies: maladies cardiovasculaires, diabète, hypertension artérielle, accidents vasculaires cérébraux, embolies pulmonaires, certains cancers, ostéoarthrite... (Poirier and Després 2003)

141.... déminéralisation **induite par** les traitements antitumoraux (hormonothérapie, chimiothérapie). (Tubiana-Hulin et al. 2001)

The use of parentheses to introduce specifics was more common in the French data (nine cases), while in the English data the two occurrences observed also included a lexical indicator. The use of these structures to express this relationship could introduce some ambiguity, as they may also be used to present abbreviations, symbols and

variants. The difficulties linked to this ambiguity could be more pronounced in French, although more data would be required to evaluate this possibility.

More striking differences were observed in terms of the lexical indicators used to identify the relationship and their relative frequencies. As shown in Table 80 and Table 81, in the 61 English occurrences associated with lexical indicators, 14 distinct indicators were observed (including *such as* (16 occurrences), *including* (14) and *is a* (13)). In the 38 French occurrences with such indicators, 15 were observed, the most frequent being *est un* (6), *comme* (5), *tel que* (5), and *autres* (5).

Table 80. English indicators of GENERIC–SPECIFIC relations between elements

Indicator	Occurrences
such as	16
including	14
is a (is an, is the, are [number], are)	13
(and) (several) other	7
as (an)	2
type of	1
among	1
another	1
become a	1
example of	1
eg	1
i.e.,	1
in the form of	1
include	1
Total	61

Table 81. French indicators of GENERIC–SPECIFIC relations between elements

Indicator	Occurrences
est un (est le, sont des, étaient des)	6
comme	5
tel que (tels que, telles que)	5
autre (autres)	5
en particulier	3
c'est-à-dire	2
notamment	2
y compris	2
principalement	2
comme tout	1
de	1

dont	1
regrouper (regroupent)	1
tel	1
type de	1
Total	38

Although more distinct indicators of this relation were found in English, given the numbers of occurrences the variety in French was proportionally higher, with a ratio of 2.5 occurrences per indicator as compared to 4.4 in English. Moreover, the distribution of the proportions of occurrences among the indicators show that the most frequent English items account for an overwhelming proportion of the occurrences, while in French the distribution is more even. The top 5 indicators in English account for 85% of the occurrences, while in French the top 5 represent only 63% of the occurrences, and more than twice as many indicators are required to attain the 85% level (Figure 22).

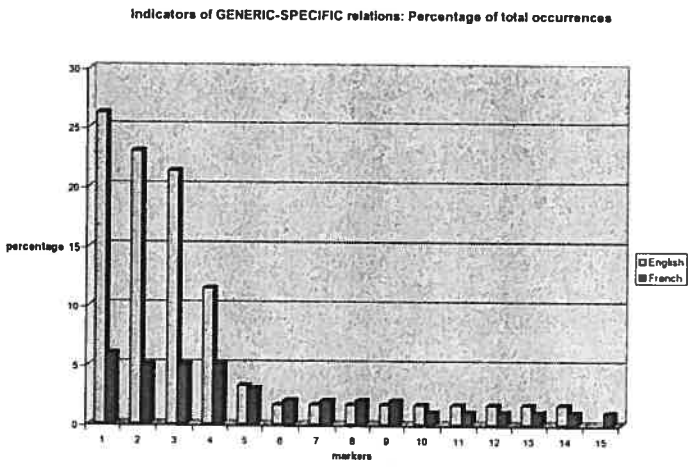


Figure 22. Indicators of GENERIC–SPECIFIC relations: Percentage of total

In this case, most results in the English data would be found using only a few of the more frequent indicators, while to retrieve the French occurrences pattern forms would need include a larger number of indicators. This could substantially increase the investment of time and effort required to create pattern forms that integrate these markers and properly represent the kinds of structures in which they may occur.

Many of the indicators of hierarchical relations identified in these contexts may also be used to identify GENERIC–SPECIFIC relations independently, and have been observed in previous studies of knowledge patterns (cf. Hearst 1992; Ahmad and Fulford 1992; Meyer et al. 1999; Meyer 2001; Marshman et al. 2002). It is possible that these existing data may be exploited for this application, facilitating the creation of pattern forms.

In evaluating and using the supplementary information conveyed by the presence of these elements, however, the need to take into account information conveyed by the presence of multiple elements in a single relation occurrence will nevertheless require additional research. For example, assisting users in identifying cases in which relations with a given concept may be inherited from generic to specific may be a complex task. It is clearly risky to assume that particular relations may be inherited from generic to specific; this kind of evaluation is necessarily best carried out by a human in light of more than a single relation occurrence in a text.

Nevertheless, some formal characteristics may be exploited in order to allow a user to access information useful in this decision-making process more quickly, easily and efficiently. Specific structures in which multiple relation participants linked by GENERIC-SPECIFIC relations occur may serve as guidelines for differentiating between cases in which a relation appears likely to be inherited and those in which this is not the case. Some of these are illustrated in Examples 142 to 148. In Examples 142 and 143, the relation indicated applies not only to the specifics, but also to the generics:

142. Oxidation **induces** neuronal cell death, <u>including</u> apoptosis of SNS neurons… (Harris and Matthews 2004)

143. La transfection du gène de la NOSi **inhibe** aussi l'activation des caspases, <u>y compris</u> la caspase 8… (Kolb 2001)

However, in Examples 144 to 148, the relation applies to the specifics, but is not applicable to all members of the generic class:

144. Interleukin-6 is an upstream proinflammatory cytokine that **induces** both CRP and fibrinogen expression. (Rackley 2004)

145. CRP can also **induce** monocytes to **express** tissue factor, a glycoprotein that **plays an** important **role in** coagulation. (Willerson and Ridker 2004)

146. CRP was recently shown to **reduce** synthesis of the vasodilator nitric oxide in cultured endothelial cells. (Rackley 2004)

147. L'obésité est un facteur susceptible d'**intervenir dans** de nombreuses maladies… (Poirier and Després 2003)

148. L'expression du gène cycline E est alors directement **sous la dépendance des** signaux extrinsèques, et ne **nécessite** plus une activation préalable de la cycline D1. (Blanchard 2003)

Generally speaking, cases in which generic-specific relations are indicated by the indicators *is a* and by apposition in structures such as those observed above appear rarely to indicate cases of the first type, and those indicated by other indicators (e.g., *including*, *y compris*) more likely to do so. More research into such occurrences would be required to determine to what extent — and how — formal indications may be used to guide users in making these decisions.

4.9.1.4 Ellipsis of part of complex related elements

An additional complication is present in some cases in which multiple elements are involved in a relation: the use of elliptical forms of terms or other linguistic units (i.e., forms in which a part of a complex unit has been omitted), as seen below in Examples 149 to 151. This phenomenon is prevalent in cases of conjunction or disjunction of complex elements (or even occasionally in hierarchically related elements).

149. These findings collectively indicate the significant **role of** oxidative stress **in** the development and progression of cancer. (Kang 2002)

150. There is evidence **implicating** oxidative stress **in** the pathogenesis of stroke, … myocardial infarction, … myocardial stunning, … atherosclerosis, … and congestive heart failure… (Granger et al. 2004)

151....the fractalkine/CX3CR1 system may nevertheless be **important in** the pathogenesis of atherosclerotic and coronary vascular diseases. (Umehara et al. 2004)

As shown above, a variety of forms may be observed. In Example 149 and its occurrence of a CAUSE–EFFECT relation between oxidative stress and the development and progression of cancer, there are two heads (*development* and *progression*) that share a single expansion (*of cancer*). However, the opposite is true in Example 150, in which a single head, *pathogenesis*, is followed by several expansions (i.e., oxidative stress is implicated in pathogenesis of stroke, pathogenesis of myocardial infarction, pathogenesis of myocardial stunning, etc.). In Example 151, the situation is even more complex, since there is a double separation: the fractalkine/CX3CR1 system is important in the pathogenesis of atherosclerotic disease and the pathogenesis of coronary vascular disease.

Similar forms and challenges were observed in French, as illustrated in Examples 152 to 159. In Example 152, the head of a complex unit is omitted; in Example 153, it is the expansion of a unit. Examples 154 to 159 illustrate some of the more complex structures that may be observed.

152. Les mécanismes responsables des différences d'**effet** entre les statines lipophiles et hydrophiles **sur** la prolifération des CML ne sont pas encore élucidés. (Nalbone et al. 2002)

153. Les graisses alimentaires peuvent aussi **nuire à** la coagulation et **à** la fibrinolyse plasmatiques, indépendamment de leurs effets sur la cholestérolémie. (Blais_2001a)

154. L'importance de ces facteurs a été reconnue par le fait qu'ils **stimulaient** la prolifération, la migration et la formation de tubes vasculaires in vitro par des cellules endothéliales... (Blot et al. 1999)

155.... les macrophages produisent et libèrent de l'IL-12, un puissant promoteur de la voie de différenciation Th1, et les cellules endothéliales des plaques **expriment** la P- et la E-sélectine qui recrutent préférentiellement les lymphocytes Th1. (Caligiuri 2004)

156. Cette protection **passe** vraisemblablement **par** une activation de PKG, protéines kinases dépendantes du GMPc… (Kolb 2001)

157. L'anion superoxyde produit pénètre dans le cytosol par des canaux anioniques et **conduit à** l'activation de facteurs de transcription comme NF-jB… (Bonnefont-Rousselot et al. 2002)

158. Les IL1, 2 et 6 **participent à** la prolifération et **à** la maturation des lymphocytes B et T. (Abrial et al. 2005)

159. L'**induction de** tumeurs bénignes ou malignes ovariennes **par** une stimulation continue des ovaires est une hypothèse qui a déjà été soulevée… (Sasco 2000)

Clearly, a pattern-based tool should extract the whole unit, given that a link, for example, between oxidative stress or the fractalkine/CX3CR1 system and *development, progression* or *pathogenesis* would not be very informative; that between oxidative stress and cancer, stroke, etc. is the most essential information for extraction.

Specific variations in the forms of the elements participating in such structures may also be observed (e.g., in Example 153, the plural form of *plasmatique* is present due to the conjunction of elements in the elliptical structure). This kind of variation could potentially be exploited to assist in identifying cases of ellipsis, but conversely could also pose difficulties for the recognition of related elements in some cases.

The challenges posed by this phenomenon for automatic identification of related elements, described above in Section 2.6.3, are significant. Moreover this kind of variation was observed relatively frequently, in approximately 16% of relation occurrences in English and 19% in French (Table 82).

Table 82. Comparison of the proportions of occurrences of ellipsis (E) of part of complex related element in English and French

	EN	FR	Total
E+	71	67	138
E-	371	282	653
Total	442	349	791

In both data sets, a substantial proportion of the total relation occurrences and of those containing multiple elements (38% in English and 39% in French) were affected by this phenomenon, which suggests that to obtain complete information in applications that attempt to identify related elements it would be necessary to develop strategies for breaking down these complex structures and/or identifying their individual elements. Moreover, patterns that specify the form related elements may take and applications that search for markers in connection with specific candidate terms should ideally allow for such structures to avoid excluding KRCs containing related elements that differ from the "standard" forms due to this phenomenon.

No significant difference was observed between the two data sets ($p = 0.248$), although the phenomenon was slightly more common in the French data. This suggests that the difficulties of dealing with this phenomenon may have a relatively similar impact in the two languages, but suggests that (especially if large amounts of data are to be processed), applications could confront slightly more difficulties in French. It is nevertheless necessary to evaluate this phenomenon in data gathered using a wider variety of terms, in order to determine whether methodological factors such as term choice could have an impact on the observations (e.g., through the tendency of a given term to participate in complex terms that may occur in elliptical form).

It is interesting to note that not only the frequency but also the structures found were similar in the two data sets, and that perhaps analysis of one of the languages could be used as a starting point for identifying structures in the other:

160. Recent data highlight mechanisms for how myeloperoxidase can **promote** lipid and lipoprotein oxidation in vivo. (Brennan and Hazen 2003)

161. L'activation du PPARγ par les thiazolidinediones **favorise** le flux d'acides gras et de triglycérides allant vers le tissu... (Gervois and Fruchart 2003)

162. These findings collectively indicate the significant **role of** oxidative stress **in** the development and progression of cancer. (Kang 2002)

163. Les lésions ou stimulations vasculaires ... **entraînent** une **augmentation de** libération et **de** synthèse de facteur Willebrand. (Drouet and Bal Dit Sollier 2002)

164. Activation of the HER-2/HER-3 pathway **induces** activation of the COX-2 promoter and expression of COX-2 mRNA and protein. (Witters et al. 2003)

165. Par contre, p53 **réprime** la transcription de gènes anti-apoptotiques comme bcl-2 et comme la NOSi elle-même.

Clearly, however, there would be some adjustments to be made; in the order of elements (e.g., Examples 160 and 161) or in some cases possible variations in use of noun and adjective forms as part of complex terms (e.g., Examples 162 and 163) would need to be taken into account.

In terms of the specific types of ellipsis observed (of the head or the expansion of a complex unit), the proportions of the relation occurrences and of the relation occurrences involving multiple elements showed some variation, but not to the point of statistical significance, as illustrated in Table 83 ($p = 0.066$) and Table 84 ($p = 0.533$). The higher proportion of ellipsis of the head of a complex element in French does nevertheless trend towards significance.

Table 83. Comparison of the proportions of occurrence of ellipsis of head (Eh) of relation occurrences in English and French

	EN	FR	Total
Eh+	51	56	107
Eh-	391	293	684
Total	442	349	791

Table 84. Comparison of the proportions of occurrences of ellipsis of expansion (Ee) of relation occurrences in English and French

	EN	FR	Total
Ee+	26	17	43
Ee-	416	332	748
Total	442	349	791

These data indicate that the proportions of relation occurrences in which the head of a complex unit was omitted showed a trend towards more frequent occurrences in the French data; more data could be analyzed to determine if this difference is in fact significant. If this kind of ellipsis did prove to be more frequent in this language, French could be more vulnerable to the difficulties linked to this phenomenon. One of these, for example, would be the fact that if problems occur in the identification of complete forms, the head of a more complex item may provide at least some pertinent information, while an expansion alone may be less useful. However, as discussed above, in some cases the essential information for information extraction is found not in the head of a related element (e.g., *progression, pathogenesis, development*) but rather in the expansion (e.g., *of cancer, of heart disease*).

More data would be required to develop strategies for automatically analyzing the different structures that may occur in order to target the most significant information. On a structural level, however, it is clear that the heads of related elements are more likely to occur in nominal form, creating fewer potential problems for KRC recognition by applications that specify the forms related elements may take.

It is interesting to note on this level that some differences in the prototypical forms of complex terms in the two languages may also affect performance. For example, complex terms composed of nouns and adjectives (i.e., NOUN + ADJECTIVE in English and ADJECTIVE + NOUN in French) would likely pose different challenges to pattern forms used to identify contexts and related elements using a more automatic approach. This may be observed when the following structures are considered:

ADJECTIVE *and* ADJECTIVE NOUN [MARKER] ADJECTIVE NOUN

ADJECTIVE NOUN [MARKER] ADJECTIVE *and* ADJECTIVE NOUN

NOUN ADJECTIVE *et* ADJECTIVE [MARKER] NOUN ADJECTIVE

NOUN ADJECTIVE [MARKER] NOUN ADJECTIVE *et* ADJECTIVE

The markers in the middle two structures are separated from the full forms of the related elements by the interruption of an elliptical form of a term, while those in the first and last structures are contiguous. Applications that reject contexts entirely if proximity conditions are not met will reject a similar number of contexts, but may exclude different kinds of contexts. If the examples illustrated had a similar, prototypical X(cause) + [MARKER] + Y(effect) structure, an application might reject proportionally more contexts in which an effect was present in elliptical form in English, and more in which an elliptical cause was present in French. The overall effect of this variation — given that ellipsis may occur in one or more of the roles in a given context — is hard to predict. However, variation in application performance in at least some contexts is very plausible. One can imagine, for example, problems posed when research focuses on specific terms or classes of terms that are more likely to play a specific role in a relation (e.g., in searching for information on the causes of a given disease). In these cases and under the conditions described above, the recall of an application in one language might be somewhat higher than in the other. (Nevertheless, given that the proportions of potentially affected contexts are reduced by each additional condition imposed (on the relation, pattern structure, class of related element, participation of that element in a given relation, and so on), the ultimate effect of such differences could well be insignificant. Much would depend on the volume of data being processed and the requirements in terms of recall for a given application.)

The phenomenon of NOUN NOUN terms in English (e.g., as in *breast cancer*) may also pose difficulties for applications that attempt to identify related elements automatically using POS classes, since their elliptical forms may be difficult to differentiate from complete terms in contexts such as Example 166:

> 166. Recent data highlight mechanisms for how myeloperoxidase can **promote** lipid and lipoprotein oxidation in vivo. (Brennan and Hazen 2003)

An automatic application could easily be misled into identifying either a single related element, *lipid*, or two distinct elements of which one is incomplete, *lipoprotein*

oxidation and *lipid* (rather than *lipid oxidation*). The structures more commonly used in French, including NOUN *de* NOUN and NOUN ADJECTIVE combinations, are not as likely to raise this kind of problem (although of course they may pose other difficulties).

As well, although only a few occurrences of this phenomenon were observed and thus statistical significance cannot be evaluated, more occurrences of multiple types of ellipsis were observed in the French results than in the English. A significant variation in this phenomenon would create additional complexities for any application attempting to identify related elements automatically within extracted contexts in French.

As noted above, more data and considerably more in-depth analysis would be necessary to develop formal representations of this phenomenon that could be implemented in pattern forms. It will be very important in gathering this data to include occurrences of as wide a variety as possible of complex elements (e.g., by using a substantially increased range of terms or even a methodology that is not term-based in gathering the data), in order to reduce the possibility of observing variations linked to the behaviour of specific terms.

4.9.1.5 Repetition of markers and marker elements

In connection with complex elements, the phenomenon of repetition of a marker or of part of a complex marker form in connection with multiple elements sharing a role a relation may be observed, and is illustrated in Examples 167 to 170.

167. Microalbuminuria (urinary ACR>2 mg/mmol) was **detected in** 32.2% of patients with diabetes and **in** 14.7% of patients without diabetes. (MacIsaac et al. 2004)

168. The mammographic density does not **increase with** tibolone, unlike **with** HRT. (Kocjan and Prelevic 2003)

169. Diverses recherches ont cependant montré que leur emploi au long cours, **conduit à** la prolifération de la glande mammaire et **au** risque accru de cancer de l'endomètre... (Kirkiacharian 2000)

170. L'activation des ostéoclastes est **responsable de** l'hyperrésorption osseuse et **de** la libération de facteurs de dégradation… (Tubiana-Hulin et al. 2001)

As may be observed in these contexts, commonly repeated items included prepositions such as *in*, *with*, *à* and *de*. Of course, cases such as these often involve the interruption of complex markers (discussed in Section 4.10.1.2). Clearly, this kind of separation makes the identification of all relations present in the contexts more challenging in more automated approaches, as the pattern form is complete in one occurrence and only partial in another. Complex pattern forms that would allow for repetition of a whole marker or a part of a complex marker would be more challenging to develop and would also be vulnerable to considerable noise, given the frequency with which these prepositions occur.

The phenomenon of repetition of a marker or part of a complex marker was relatively widespread in the French data, found in 7% of observed relation occurrences. In the English, however, it was quite rare, found in only 1%. Table 85 illustrates this distribution and shows a significant difference between the two data sets ($p < 0.001$).

Table 85. Comparison of the repetition of markers or marker elements (RM) in English and French

	EN	FR	Total
RM+	4	24	28
RM-	438	325	763
Total	442	349	791

The significance of this difference remains high ($p < 0.001$) when this figure is considered as a proportion of the occurrences of multiple related elements (3% of the 189 occurrences in English and 15% of the 169 occurrences in French).

Both this phenomenon and the difference between the languages have both positive and negative effects for pattern-based tools. The primary difficulty posed by the phenomenon is the risk that, for pattern forms that describe contexts in which markers

may occur and do not allow for this variation, this repetition may constitute an interruption of the form and interfere with the recognition of contexts or one or more of their constituents.

In contrast, however — and in French particularly — repetition of parts of complex markers may be a good indicator that multiple related elements are present, and may help in identifying these more easily (since applications could search, for example, for a preposition such as the frequently repeated *à* or *de* followed by a noun phrase). However, the frequency of these combinations means that this kind of strategy would likely be very vulnerable to noise or other difficulties, as in Examples to 171 to 174:

> 171.... une **diminution** importante **des** complications micro et macrovasculaires, de même qu'une diminution de la mortalité liée au diabète. (Gonzalez and Palardy 2004)
>
> 172. Dans des lignées de cancer du sein et de la prostate, BRCA1 inhibe la transcription **dépendante du** REα de gènes impliqués dans la prolifération cellulaire. (Pujol et al. 2004)
>
> 173. Même très modéré, il fait craindre la rechute, car les femmes pensent que cette séquelle est **due à** un traitement plus important à cause d'une tumeur agressive. (Bobin et al. 2002)
>
> 174. En conclusion, BRCA1 et BRCA2 pourraient **participer** activement **à** la prolifération et **à** la différenciation induite par les œstrogènes, en particulier au cours des périodes d'exposition hormonale... (Pujol et al. 2004)

In Examples 171 to 173, forms of *de* and *à* are found in the context of markers containing these elements, but do not introduce additional related elements. This is also the case in Example 174, but the structure is even more potentially problematic, since the second occurrence is preceded by a marker that can link multiple elements, *en particulier*, and thus closely resembles a structure in which repetition of part of a marker might occur.

Given this vulnerability, it is debatable whether there would be a significant return on investment for adapting pattern forms to use this phenomenon in identifying elements linked in relations; what is clear is that if this were done, the effort is likely to

be most profitable in French. It does seem evident, however, that French patterns should at least be adapted to allow the repetition of markers or marker elements before each one of multiple elements involved in a relation, to reduce problems in recognizing contexts. This is an additional, if minor, step to be taken in the design of pattern forms as compared to English, for which this phenomenon is likely not common enough to be worth taking into account.

4.9.2 Form of the elements linked by markers

While many research projects have focused on patterns linking nouns and noun phrases only, it is interesting to note that not all related elements are nominal in form. Non-nominal forms that may be found using patterns that allow for variation in the POS of related elements include some types of anaphora (i.e., those involving non-nominal anaphoric expressions), adjectives (Examples 175 to 178), verb phrases (Examples 179 and 180) and propositions (Examples 181 to 184).[146]

175. ONOO[middle dot]- is an important mediator of lipid peroxidation and protein nitration, including oxidation of LDL, which has dramatic **pro**atherogenic effects. (Griendling and FitzGerald 2003a)

176. CRP, sCD40L, and IL-18 are three inflammatory **markers** that result in endothelial activation. (Szmitko et al. 2003)

177. L'événement osseux, quant à lui, a été ainsi défini : ... nécessité d'une radiothérapie antalgique ou recours à un nouveau traitement **anti**tumoral en raison d'une progression osseuse. (Tubiana and Hulin 2001)

178. ...les **facteurs de risques** cardiovasculaires traditionnels tels que l'hypertension, les dyslipidémies, le diabète et l'obésité tronculaire sont en effet observés avec une fréquence croissante... (Duong et al. 2003)

179. Thus, the fractalkine/CX3CR1 system may **contribute to** the pathogenesis of vascular and tissue injury by enhancing cell adhesion... (Umehara et al. 2004)

---

[146] See Section 2.3.1.5.2.1 for a description of the importance of anaphora and Section 4.9.2.1 for an analysis of the observations in the corpora.

180.... p8, en s'associant aux protéines Smad, pourrait alors participer à un complexe transcriptionnel pour **régler** la transcription de gènes cibles impliqués dans le déroulement des phénomènes décrits ci-dessus... (Vasseur and Iovanna 2003)

181. CRP can also **induce** monocytes to express tissue factor... (Willerson and Ridker 2004)

182. The basic grafting procedure--bypass surgery--is performed 500,000 times a year in the U.S. to treat coronary arteries that are becoming blocked **as a result of** atherosclerosis. (Beardsley 2000)

183. Les cellules vasculaires sont avides de cholestérol, ce qui **aidera** le mélange gène-cholestérol à se coller assez longtemps sur la paroi pour **freiner** la prolifération des cellules... (Simard and Dussault 1997)

184. Lorsqu'il est activé, il induit une cascade de phosphorylations intracellulaires, **conduisant à** une transcription de protéines et à une croissance cellulaire accrues. (Cornez and Piccart 2002)

The development and refinement of pattern forms that can accommodate these types of elements poses certain challenges. An alternative is the use of an approach that does not place limits on the types of elements that may appear with markers (for example, a character-string-based approach using only the marker itself). The former option would require an analysis of significant numbers of occurrences of markers in order to identify the various forms that may occur with each one; the latter, due to its reduced specificity, would open the door to considerable amounts of noise in the results.

Another issue in this area is the way in which information extracted from contexts containing non-nominal related elements may be used. This of course will vary according to the user's goals and the approach envisioned. More highly automated applications, focused for example on automating ontology construction, are less likely to be able to find immediately useful information in such contexts, since it is very unlikely that propositions, adjectives or affixes will correspond to the nodes of an ontology. (However, these "non-standard" items may in fact be variants of terms corresponding to concepts in the sense of Daille (2005).) The same problem may be posed for semi-automatic applications used to establish links between entries in term

bases, although the difficulty is somewhat mitigated by the possibility of human interpretation of the results, allowing such structures to be recognized as denoting concepts that appear in the base associated with another (likely nominal) surface form. However, in semi-automatic approaches for purposes such as acquisition of domain knowledge and formulation of definitions — which are likely to be much more heavily dependent on user interpretation in any event — such cases should pose fewer serious problems, and may be as useful as those linking two nouns or noun phrases.

The vast majority of the occurrences of patterns observed linked elements in noun or noun form.[147] This indicates that pattern forms that specify the POS class to which related elements may belong — generally nouns or noun phrases — could identify a large proportion of the occurrences identified in this research, in either of the languages.

Some occurrences in both languages would nevertheless be excluded; the significance of this would depend on the situation in which a pattern-based tool is used. Cases in which high recall is desired would of course be more seriously affected. Moreover, if large amounts of data are to be processed, the numbers of occurrences excluded could become more significant. Conversely, applications in which little human interpretation is intended and/or that require a high level of correspondence between the items located in texts and terms or concepts in specific forms or types of forms that may appear in resources such as ontologies or term banks, may not be able to use occurrences involving non-nominal forms in any case. In such a situation the exclusion of these contexts could increase precision in the results.

---

[147] Of course, the approach used in the project is very likely to have contributed to this, as the terms chosen to generate the contexts for analysis were nouns or noun phrases, and in order for contexts to be retained and annotated these terms were required to be linked to the pattern marker (i.e., to denote one of the concepts involved in the relation). This means that at least one item participating in every relation annotated was required to be a noun or noun phrase. However, even taking this phenomenon into account, the proportions of related elements in noun form were very high.

In an analysis of the types of non-nominal forms observed, it becomes clear that the classes of these related elements are consistent in the two data sets, with adjectives, verb phrases and propositions identified in both English and French (in addition to cases of pronouns and some other types of anaphora). However, as illustrated in Table 86, the languages differ in the frequency with which this phenomenon was observed.

Table 86. Comparison of proportions of relation occurrences containing non-nominal (NN+) related elements in English and French

	EN	FR	Total
NN+	37	58	95
NN-	405	291	696
Total	442	349	791

In English, 3% of relation occurrences involved adjectival elements, 2% propositional and pronominal, and 1% verbal, for a total of 8%, while in French 6% contained adjectives, 5% pronouns, 4% propositions and 1% verb forms, for a total of 17%. Thus, while the different categories show perfect rank-order correlation, the proportions of occurrences involving these phenomena are quite different. The distribution among the different types may be represented as in Table 87.

Table 87. Comparison of the proportions of relation occurrences containing various types of non-nominal and exclusively nominal related elements in English and French

	EN	FR	Total
Adjectives	15	21	36
Propositions	8	15	23
Pronouns	10	17	27
Verbs	4	5	9
Nouns only	405	291	696
Total	442	349	791

If the proportions of contexts containing non-nominal related elements are compared using the Chi-square test, the difference is statistically significant ($p < 0.001$), with the phenomenon more common in the French results. The difference appears to

result primarily from the category of propositional and verbal elements (considered together to permit statistically valid evaluation) and of pronominal elements, which when considered individually are also significantly more frequent in French ($p = 0.033$ and $p = 0.045$, respectively). The proportion of adjectival elements also shows a trend towards higher frequency in French ($p = 0.079$).

This consistently higher proportion of related elements occurring in non-nominal form suggests that the exclusion of contexts containing such forms would have a greater impact in French, eliminating a disproportionately high number of potentially useful contexts in this language.

Another point may be raised in relation to applications that focus on the identification of relation occurrences involving specific terms or candidate terms. As the terms used in such cases are also likely to be nominal in form, again there is the possibility of a higher proportion of silences in the French results. Moreover, the excluded contexts may in fact contain potentially interesting variants of terms of interest for such applications (e.g., relational adjectives derived from terms, verb phrases or propositions that are equivalents of more conventional term forms including nouns derived from verbs, pronouns replacing terms).

In both corpora, regularities may be observed in the associations between specific markers and non-nominal elements. In English, for instance, the marker *marker* and in French the markers *facteur de risque* and *complication de* were observed with adjectival related elements, as in Examples 185 to 187:

185. The inflammatory **marker** C-reactive protein (CRP) can indicate low-grade chronic inflammation... (MacKenzie 2004)

186.... les **facteurs de risques** cardiovasculaires traditionnels tels que l'hypertension, les dyslipidémies, le diabète et l'obésité tronculaire sont en effet observés avec une fréquence croissante chez les patients VIH+ ... (Duong et al. 2003)

187. Les interactions entre système rénine-angiotensine et **complications** vasculaires **du** diabète constituent un autre exemple de l'implication du TGF-ß. (Michel 2004)

In addition, causal events as they appeared in conjunction with causal agents, as discussed in Section 4.9.1.2, were often indicated by verb phrases introduced by such markers as *by* and *en*.

Regular associations between specific markers (of relations or of the relationship between multiple related elements) and forms of related elements may be of use in developing pattern forms that can identify these kinds of contexts. If the importance of the phenomenon were considered to merit it, the analysis in some cases could even be extended to the identification of a nominal form from which these adjectives are derived and that constitutes a more promising element for further applications (e.g., *inflammation* for *inflammatory*, *vaisseau* for *vasculaire*). The parallels observed in the two data sets once again show promise for the adaptation of developments in one language for use in the other, although of course adjustments would be necessary.

4.9.2.1 Anaphora

Challenges in KRC identification and processing may be introduced when one of constituents of a KRC — and particularly one of the elements linked by a relation — is represented by an anaphoric expression. This phenomenon may involve the replacement of an element or part of an element by a pronoun (e.g., Examples 188 to 190), a possessive adjective (e.g., Example 191), a generic (e.g., Examples 190 and 192 to 195), or a quantifier (e.g., Example 196).

188. This **enhances** retention of the lipoprotein and possibly **triggers**, along with oxidation, the formation of a recognizably foreign substance... (Caslake and Packard 2003)

189.... les métastases à distance demeurant dormantes aussi longtemps que la tumeur primitive est en place, celle-ci exerçant un rétro**contrôle** négatif **sur** la croissance des micrométastases ... (Brain 2000)

190. L'importance de ces facteurs a été reconnue par le fait qu'ils **stimulaient** la prolifération, la migration et la formation de tubes vasculaires in vitro par des cellules endothéliales. (Blot et al. 1999)

191. Regulation of the apoptotic pathway by NF-[kappa]B may affect both the pathogenesis of breast cancer and its **response to** chemotherapy and radiation. (Garg et al. 2003)

192. While heredity can influence a person's **susceptibility to** development of the disease, a sedentary lifestyle and long-term obesity are key triggering events for most people. (Haskell 2003)

193. Clinical studies assessing the relationship between myeloperoxidase levels and acute coronary syndrome risks may help answer the overall **relationship between** this enzyme **and** development of the vulnerable plaque. (Brennan and Hazen 2003)

194.... Les résultats publiés apparaissent très encourageants, ne montrant pas d'**effet** apparemment **délétère de** ce traitement **sur** la maladie cancéreuse mammaire préexistante. (Gorins et al. 2003)

195. On se doit de dire ici qu'une réponse partielle était fournie par le fait que les seules tumeurs **induites** chez les souris D1-/- transgénisées **par** ces deux oncogènes étaient effectivement des tumeurs de la mamelle. (Larsen 2001)

196.... dyslipidaemia, malnutrition and inflammation [1*,2,3*], some of which **have** also **been linked to** the pathogenesis of anaemia itself. (Stevens and Levin 2003)

This phenomenon may have a significant impact on approaches used for identifying KRCs in corpora. The efficiency of tools that extract contexts of fixed length may be affected by anaphora (as more distant antecedents may not appear within contexts), and even those that work at a sentence level may also be vulnerable in some cases, as the antecedent of an anaphoric element may be extrasentential (as illustrated in Examples 188, 194 and 195). A user may thus require access to a larger context in order to identify pertinent elements in the relation. In more automatic applications, this problem is clearly more serious, as human involvement may be impossible.

Approaches based on identification of lexical markers linking two terms or candidate terms (or even, for example, nouns or noun phrases) may miss occurrences of relations because a given term form or part of speech is not present in context, or the

term form present in context may be significantly different from any form that is — or should be — included in a term base.

At a conceptual level, the impact of anaphora for any knowledge-extraction application is clear; the identification of concepts linked by a relation is more difficult for a human — and certainly extremely problematic for automatic applications — when concepts are represented entirely or in part by anaphoric expressions. While in some cases the anaphoric element may be useful to some degree (e.g., in instances where the generic that replaces a related element may provide some indication of the nature of the concepts involved in a relation), in others (e.g., in the case of pronouns and possessive adjectives), the anaphoric element itself is not useful for gaining knowledge. Moreover, even the replacement of a specific concept by a more generic one may be problematic, in that if only a generic is identified in the context, some information will be lost or — perhaps worse — unwarranted generalizations could be made about the existence of relationships involving all members of a class rather than only a specific one.

Even in cases in which anaphora occur in elements outside the structure of related elements, they can pose difficulties for the interpretation of contexts provided, as they indicate that a larger context is required for full comprehension of the information conveyed, including for example the specific nature of an interaction, the basis on which conclusions were drawn and the strength of this evidence, and so on. Certainly, if markers are modified or replaced by anaphoric expressions, the impact on KRC extraction is also significant; however, this was rarely observed in the data in this study.

Minimizing these difficulties requires careful design of pattern forms that take into account variations due to anaphora and the impact that these have on the usefulness of candidate KRCs for the intended application.

A comparable number — and considerable proportion — of relation occurrences containing anaphoric elements (in any part of the context) were observed (56 in English and 60 in French, constituting 13% of the total number of relation occurrences in

English and 17% in French). These figures can be represented as shown in Table 88, which reveals a difference that is just short of statistical significance ($p = 0.058$), with a strong trend towards a higher proportion of occurrences in French.

Table 88. Comparison of proportions of relation occurrences including anaphoric expressions (AE) in English and French

	EN	FR	Total
AE+	56	60	116
AE-	386	289	675
Total	442	349	791

If the proportions of occurrences of anaphora are compared (as some relation occurrences contained more than one case), this trend becomes more pronounced, with 61 cases in English and 66 in French.

From this comparison, it is clear that difficulties posed by anaphora for the identification of complete information about concepts and the relations linking them are likely to be fairly widespread in both languages. Careful handling of this phenomenon in applications for knowledge extraction (particularly those that aim to identify the elements linked by a relation automatically) is important, since a significant amount of potentially useful information could be lost if this phenomenon were not accounted for in application and/or pattern development (e.g., in ensuring that access to wider contexts is available and in designing pattern structures that admit anaphoric expressions). However, the higher prevalence observed in the French data suggests that this language may encounter somewhat more difficulties in this respect, both at the level of their form (in the challenges of adequately representing the phenomena in pattern forms) and their interpretation (in the identification of antecedents and the information they contribute to the context). Strategies designed to deal with this phenomenon may be more important to develop in this language, although more data would be necessary to confirm whether the apparent difference observed in this work is observed in other corpora and/or sets of relation occurrences.

These challenges — and the strategies required to deal with them — may vary in nature depending on the type of anaphoric element observed. The numbers of relation occurrences containing various types of anaphoric expressions are compared in Table 89, and the number of occurrences of each type in Table 90. The rank order of the different categories of anaphoric elements in the two data sets shows a weak positive correlation, with the second and third categories (pronouns and possessive adjectives) inverted. The comparison of the proportions of occurrences of various types of anaphoric elements in the two data sets using the Chi-square test (with the categories of pronouns and quantifiers combined in order to allow for statistically valid comparison) does not show any significant differences in the proportions of occurrences corresponding to these types (p = 0.200), although the proportion of possessive adjectives in French is slightly higher.

The difference in the prevalence of the phenomena observed in the two data sets results from an accumulation of smaller differences in most of the specific types of anaphora in this category. The proportions of relation occurrences that contained anaphoric elements in the form of pronouns and generics were not significantly different (p = 0.290 and p = 0.760 respectively), and the numbers of quantifiers playing this role were too low to be compared using the Chi-square test. However, a significantly higher proportion of the category of possessive adjectives was observed in French (p = 0.033). This indicates that this phenomenon would be particularly important to take into account in French; however, as even in this language it does not affect a large proportion of the contexts overall (6%), the decision to do so would depend on the situation in which an application is to be used.

Table 89. Comparison of the proportions of relation occurrences containing anaphoric elements of various types[148]

	EN	FR	Total
Quantifier	4	1	5
Pronoun	16	18	34
Possessive adjective	12	20	32
Generic	28	24	52

Table 90. Comparison of the proportions of occurrences of anaphoric elements of various types

	EN	FR	Total
Quantifier	4	1	5
Pronoun	16	19	35
Possessive adjective	12	22	34
Generic	29	24	53
Total	61	66	127

Some of the possibilities for analyzing contexts automatically and identifying antecedents of these anaphoric elements can be evaluated according to the nature of the anaphoric expressions observed. As shown in Table 91 and Table 92, 9 separate English pronoun forms were observed in 16 occurrences of this type, and 8 of these were marked for number and none for gender. In contrast, 5 pronouns were observed in French in 18 occurrences; four of these are marked for number and three for gender.

Moreover, while the proportion of the pronouns that are marked for number is high in both samples (50% and 80%), the lack of marking of pronouns for gender in English and the high proportion of the French pronouns that are so marked (resulting of course from the use of grammatical gender in French and not in English) may have some impact on possibilities for automatic processing of contexts as well as human interpretation in some cases. Pronouns that are marked for number and gender may offer

[148] As more than one type of anaphoric expression was noted in some contexts, these figures total more than the number of relation occurrences containing the phenomenon overall. For this reason, no total is given here.

possibilities for locating antecedents of pronouns automatically, as these criteria may help to identify the appropriate element in the context of occurrence of an anaphoric element, or at least to eliminate incompatible possibilities. The lack of marking for gender in English is thus a potential disadvantage for such applications in this language.

Table 91. English pronouns functioning as anaphoric elements

Pronoun	Occurrences
it, they	5
that, those	3
this, these	2
all	1
both	1
itself	1
ones	1
other	1
which	1
Total	16

Table 92. French pronouns functioning as anaphoric elements

Pronoun	Occurrences
il (ils, elle)	11
en	5
celles	1
lui	1
ce	1
Total	19

The variety of possessive adjectives in the two data sets, as shown in Table 93 and Table 94, was relatively parallel, with two forms found in English, and two in French.

Table 93. English possessive adjectives functioning as anaphoric elements

Pronoun	Occurrences
its	11
their	1
Total	12

Table 94. French possessive adjectives functioning as anaphoric elements

Pronoun	Occurrences
son (sa, ses)	10
leur (leurs)	10
Total	20

Once again, these (*its* and *their* in English and *son* and *leur* in French) are marked for the number of their antecedents in both languages; the marking for gender in French is however not pertinent in this context because it is a function not of the antecedent, but rather of the noun that it modifies. (The French possessive adjective forms of course also reflect the number of the elements they modify — e.g., in the distinction between *son* and *ses* — but this is also not pertinent for this analysis.) Thus in these cases, neither language shows particular advantages over the other.

The strong resemblances in the data sets in the form of structures involving generics also suggest that there are similar challenges and possibilities in the two languages. Anaphoric elements containing a generic generally took the form of a noun preceded by a definite article or demonstrative adjective — in English generally *this* and *these* and in French, *ce*, *cette*, and *ces* — as in Examples 197 to 201, although in some cases quantifiers were also present, as in Examples 202 to 204.

197. Although promising, this kinase is a critical **regulator of** many basic cellular processes, including development, cardiac growth and hypertrophy, and tumorigenesis. (Force et al. 2004)

198.... this phenomenon **contributed**, at least in part, **to diminished** atherosclerosis years later. (Yan et al. 2003)

199. A family history of breast cancer is recognized as one of the most important **risk factors for** the disease. (Yang and Lippman 1999)

200.... entraîne l'activation de la fonction tyrosine kinase du domaine intracellulaire du récepteur. Cette activation **entraîne** de nombreuses réponses cellulaires avec stimulation de la croissance et de la division cellulaire... (Penault-Llorca et al. 2002)

201. Ce processus peut **être modifié au cours de** certains phénomènes pathologiques comme le cancer, l'athérosclérose ou le diabète. (Blot et al. 1999)

202.... both ovarian steroids are known to **play key role** [sic] in mammary gland development during pregnancy. (Sicinski and Weinberg 1997)

203. Toxicity remains a major concern, because many of these kinases not only **play roles in** the pathogenesis of diseases but also function in pathways that regulate the most basic of normal cellular processes. (Force et al. 2004)

204. On se doit de dire ici qu'une réponse partielle était fournie par le fait que les seules tumeurs **induites** chez les souris D1-/- transgénisées **par** ces deux oncogènes étaient effectivement des tumeurs de la mamelle. (Larsen 2001)

In two other cases in French, a generic was introduced by *un tel* (Example 205).

205. Un tel programme d'exercice physique **permet** l'**augmentation du** HDL cholestérol ... (Ferrières 2004)

This phenomenon may affect an entire related element, as above, or only a part of a more complex element, as in Examples 206 and 207.

206. Long-term activation of these appropriate responses **leads to** left ventricular remodelling... (Stevens and Levin 2003)

207. L'activation de ces récepteurs **commande** la transcription des gènes insulinosensibles... (Leblond 2001)

Another notable point is that of the possibilities offered by the GENERIC–SPECIFIC link that holds between (at least some of) the anaphoric elements involving generic terms to replace a more specific one. Depending on the approach used and the information available, identification of antecedents might be assisted by using established GENERIC–SPECIFIC links between known terms. Alternatively, links between a generic anaphoric element and its antecedent, once identified by the user, could be added to the stock of information about relations between terms.

In a more specific analysis of only the proportions of relation occurrences in which an entire related element or the head of an element was an anaphoric expression (6% in English and 11% in French), this phenomenon was observed to be significantly more prevalent in French overall ($p = 0.021$). The data are shown in Table 95.

Table 95. Comparison of proportions of relation occurrences with related elements in the form of anaphoric elements (REae) in English and French

	EN	FR	Total
REae+	27	37	64
REae-	415	312	727
Total	442	349	791

This overall difference results primarily from smaller discrepancies observed in two specific types of expressions. When the types of anaphoric elements are considered individually as a proportion of the relation occurrences, higher — but not significantly higher — proportions of anaphoric expressions in pronoun and quantifier form (p = 0.156), as well as those involving generics (p = 0.109) are observed in French.

However, no differences are noted among the individual types as a proportion of the occurrences including anaphoric expressions replacing related elements, indicating that the proportions of occurrences belonging to each type are relatively similar in the two data sets. The figures are shown in Table 96.

Table 96. Comparison of proportions of relation occurrences with related elements in the form of anaphoric expressions, by type, in English and French

	EN	FR
Pronouns	10	14
Possessive adjectives	4	5
Quantifiers	1	1
Generics	12	17
Total	27	37

As the phenomenon of related elements in the form of anaphoric elements is significantly more frequent in the French data, more difficulties associated with it, including the challenges for formal representation in pattern forms (particularly in the case of pronouns) and the difficulty of obtaining complete information about the concepts linked by a relation, may be encountered in this language. Fortunately, much of this difference comes from the use of generics, which — although less precise than a

specific — at least provide some useful information, However, even without this category, the prevalence of the phenomenon in French is higher.

From these results it may be concluded that overall the anaphora observed show strong parallels in the data in the two languages, and that techniques for dealing with these phenomena in one language may prove to be good starting points for the development of strategies for use in the other, despite some relatively minor variations in the distribution of the various phenomena overall. The exceptions to this rule may be the use of possessive adjectives, which was significantly more frequent in the French data, and some of the characteristics of the pronouns observed that may affect the development of strategies for the resolution of anaphora for applications — or even humans — that attempt this task.

The connection between the prevalence of related elements in non-nominal form and that of certain types of anaphora (i.e., related elements that take the form of anaphoric elements such as pronouns and possessive adjectives) should not be disregarded. When the data are analyzed, it becomes clear that a relatively small proportion of adjectival related elements are cases of anaphoric elements in the form of possessive adjectives (4 of 15 in the English data and 5 of 21 in the French data). This proportion of the cases observed makes only a minor contribution to the higher prevalence of related elements in adjective form observed in the French data. However, in the case of pronouns, it becomes clear that those that are anaphoric expressions constitute the majority of the occurrences overall (all of the 10 cases in English and 14 of 17 cases in French). These contribute significantly to the higher prevalence of this phenomenon observed in the French data, although according to these limited data pronouns appear to be more widely used in contexts in this language both as anaphoric elements and as independent entities.

More research into the prevalence and forms of anaphoric expressions could help to determine whether the slight differences observed in many of these factors become significant in light of more data. It may also be interesting to study the types of elements

that are replaced by anaphoric expressions in contexts, to evaluate whether regularities may be observed in their nature or the nature of their antecedents, as well as whether methodological decisions in this research (e.g., the choice of corpus texts and text types or of candidate terms, or the requirement that relation occurrences involve the concepts denoted by these terms) have contributed significantly to the observations.

4.10 Challenges in using knowledge patterns and extracted contexts

In this Section, a number of difficulties for the identification, analysis, and ultimate re-use of knowledge-rich contexts will be discussed, as outlined in Section 2.4. These include interruptions of patterns, the presence of expressions of uncertainty in candidate KRCs, and text-related issues. In addition, proportions of relation occurrences in which a range of challenges were observed will be compared, to provide an overview of difficulties of the approach.

4.10.1 Pattern interruptions

In developing pattern forms for use in KRC extraction, it is clearly necessary to take into account the natural variability of language — and thus the frequency with which basic pattern forms (e.g., X + [MARKER] + Y) may be modified in texts, as well as the ways in which this may occur.

Various types of external elements may interrupt pattern forms at different locations; the part of the pattern form that is interrupted helps to determine the effect this phenomenon may have on the recognition, extraction and processing of KRCs, and the resulting effect on pattern-based tool performance. The different types of interruptions observed will be described below, with specific attention to those that are particularly pertinent for semi-automatic knowledge extraction: the interruption of patterns by other patterns, of complex markers, and of related elements.

A significant proportion of the contexts analyzed — 66% in English and 58% in French — contained pattern structures that were interrupted by external elements. The proportions of relation occurrences including this phenomenon are compared in Table 97.

Table 97. Comparison of the proportions of interrupted relation occurrences (INT) in English and French

	EN	FR	Total
INT+	293	202	495
INT-	149	147	296
Total	442	349	791

A significant difference in the frequency of pattern interruption was found in the two data sets ($p = 0.015$), with the phenomenon proportionately more frequent in English than in French. This indicates that although the phenomenon affects many relation occurrences in both languages, its impact in English may be greater.

This strongly suggests that this phenomenon will be essential to take into account in designing pattern forms for use in semi-automatic applications, particularly if these patterns restrict the structures in which markers may occur, and if applications attempt to identify the related elements in KRCs automatically. Unless a high level of silences is considered acceptable for a given application, pattern forms must allow for the insertion of these external elements within one or more of the elements of the pattern.

Investigating the specific source of this difference between the two corpora may help to determine exactly how this difference may affect the process of pattern design and application performance. Below, the proportions of relation occurrences in which complex markers and related elements were interrupted will be analyzed. First, however, the proportions of occurrences of other types of interruptions will be presented.

4.10.1.1 Interruptions of patterns

In a significant proportion of the relation occurrences observed, the relation marker was not contiguous with one or more of the linguistic expressions representing concepts involved in the relation. This is often the case, for example, when a sentence includes a relative clause, as in Examples 208 and 209.

> 208. Activation that endures beyond the resistance stage is hypothesized to **cause** disease. (Schwartz 2003)[149]
>
> 209. Dans les cellules, ce sont les facteurs d'échange qui collectent les signaux qui **déclenchent** l'activation des protéines G. (Cherfils and Pacaud 2004)

In addition, non-contiguity of pattern components may also be observed when more than one relation and pattern (i.e., pattern marker) is present in a given context, as in Examples 210 and 211.

> 210. Antioxidants are molecules that can **prevent** or **reduce** the extent of oxidation to the oxidizable substrate. (Kang 2002)
>
> 211. Les rétinoïdes **règlent** la croissance cellulaire [19], **modifient** la prolifération [20], **inhibent** l'ornithine décarboxylase [21], **facilitent** la différentiation et l'apoptose [22,23].

It may also occur when two or more elements share a role in a relation (cf. Section 4.9.1), as in Examples 212 and 213.

> 212. This **enhances** retention of the lipoprotein and possibly **triggers**, along with oxidation, the formation of a recognizably foreign substance... (Caslake and Packard 2003)
>
> 213. Les lésions ou stimulations vasculaires en particulier endothéliales **entraînent** une **augmentation de** libération et **de** synthèse de facteur Willebrand. (Drouet and Bal Dit Sollier 2002)

Interruption may also result from the modification of a relation marker or one of the elements linked by the relation, as illustrated in Examples 214 to 216.

---

[149] The presence of an expression of hedging, *is hypothesized to*, in this context of course also constitutes an interruption of the pattern structure. Hedges will be discussed separately in Section 4.10.2.2.

214. The response to injury hypothesis developed by Russell Ross in the late 1970s suggested that atherosclerosis, at least, **resulted from** an initial injury to endothelial cells, leading to impaired endothelial function... (Griendling and FitzGerald 2003a)

215. Valantine [26] a évalué l'**impact du** ganciclovir administré en prophylaxie immédiate après transplantation cardiaque **sur** l'athérosclérose du transplant au cours d'une étude randomisée contrôlée versus placebo chez 149 patients consécutifs. (Chidiac and Braun 2002)

216.... les LDL oxydées **induisent** à leur tour une activation de l'endothélium.... (Arnal et al. 2003)

Finally, this phenomenon may occur when other elements such as discourse markers and references are present, as in Examples 217 and 218.

217. Toxicity remains a major concern, because many of these kinases not only **play roles in** the pathogenesis of diseases but also function in pathways that regulate the most basic of normal cellular processes. (Force et al. 2004)

218. L'accumulation de la cycline D1 **résulte**, d'une part, **de** l'induction transcriptionnelle de son gène, et, d'autre part, **de** l'activation de la traduction de son ARN messager. (Blanchard 2003)

Moreover, two or more of these factors commonly co-occur within a single context, further complicating the structures of the patterns and the task of representing them.

Any element occurring between a relation marker and the elements it links can pose problems for recognition of contexts and identification of related elements if pattern forms specify structures in which pattern markers may appear. Pattern forms should allow for a certain amount of variation in structures in order to reduce silences that may result from interruptions. However, the extreme variability in the form of interruptions poses significant challenges for developing such pattern forms. The adaptation of patterns to allow for this phenomenon thus complicates the process of pattern design considerably, and moreover may introduce possibilities for noise in the results of extraction. A delicate balance between recall and precision is required to

obtain the best results for a given application; this balance may shift depending on the goals of a given project and users' needs.[150]

One additional difficulty at a formal level occurs in the analysis of the form of related elements by applications that automate this task: delimiting the elements that are linked by a marker may be particularly challenging. For example, modifiers of related elements may be difficult to differentiate from the elements themselves automatically; this may be particularly true as the typical forms of complex terms — some of the most interesting candidates for knowledge extraction — may be very similar to those of simpler items coupled with modifiers (e.g., in ADJECTIVE + NOUN or NOUN + ADJECTIVE form).

Approaches that target relations between previously identified candidate terms avoid this particular challenge in many cases.[151] The impact of this problem is also considerably reduced when specific terms or candidate terms are sought in combination with markers. These types of applications are more affected by non-contiguity of markers and related elements (cf. Sections 4.10.1.2 and 4.10.1.3).

In addition to its effect at a formal level, this phenomenon may raise concerns about the value of the KRCs for knowledge extraction. Modifiers of pattern markers or related elements that characterize some aspect of the relation being expressed may call into question the subsequent usability of the context for knowledge extraction. This is most particularly — but not exclusively — the case with modifiers that express some kind of uncertainty, as described in Section 4.10.2. Conversely, some modifiers may also provide additional, specific information about relations, which may increase the value of contexts for knowledge extraction, as in Examples 219 to 221.

---

[150] Morphological variation in marker forms may also result from the presence of external elements within pattern structures. While morphological variation of markers was not specifically considered in this project, it could certainly be an issue in some applications.

[151] Of course, if automatic candidate-term extraction tools are used, similar difficulties may be encountered.

219. Receptor-mediated leukocyte activation **leads to** conformational **changes in** LFA-1 structure... (Granger et al. 2004)

220. Le profil lipidique le plus fréquemment **retrouvé dans** le diabète de type 2 associe une élévation du taux plasmatique des triglycerides... (Fredenrich et al. 2004)

221. L'obésité, le syndrome métabolique et le diabète **accroissent** notablement le risque de maladies cardiovasculaires. (Lambert 2002)

Unfortunately, it may be difficult to differentiate between these two cases on a formal level, and human interpretation of each case may be required.

A very significant number of the relation occurrences observed were interrupted by an external element: 45% of the relation occurrences in English and 40% in French, constituting 71% and 75% of the interrupted relation occurrences respectively. These results are compared in Table 98, which indicates that the proportions of interrupted pattern occurrences in the two data sets are not significantly different ($p = 0.153$).

Table 98. Comparison of the proportions of relation occurrences with interruption of a pattern (INTp) in English and French

	EN	FR	Total
INTp+	201	141	342
INTp-	241	208	449
Total	442	349	791

However, the proportion of occurrences overall that were affected by pattern interruptions were somewhat higher in English, indicating that pattern design and/or application performance in this language could be somewhat more affected by this phenomenon than in French. This kind of difference would involve a greater investment of time and effort in developing pattern forms that could deal with this phenomenon, and conversely a higher risk of silences in the results due to the inability to account for all possible types of interruptions. The accurate identification of related elements would also be likely to pose additional challenges and require additional investment in English

due to this phenomenon. However, more data would be required to determine whether this apparent tendency could become significant.

One difference that can be identified — although it was not counted towards statistics of pattern interruption in this research — is the relative consistency of the article appearing with nouns in French (whether these are nouns that constitute parts of pattern markers, elements linked by these markers, or external elements that appear within patterns), while in English articles are less consistently present. In applications that work with contexts of fixed length (in characters or words), the presence of these articles may indicate a need to use a slightly longer context in French. However, this consistency does provide some advantages over the less predictable use of articles in English, as in this latter language, pattern forms that allow for the presence or absence of articles may be required. However, ideally, this problem would be dealt with relatively systematically (e.g., by always allowing for an optional article in many pattern forms) and should not pose serious problems for the development of pattern forms.

4.10.1.1.1 Multiple markers and pattern interruptions by other patterns

One special type of pattern interruption involves the occurrence of two separate patterns with their own markers in a single context. These patterns may link different elements in separate relations, or may denote a relation that holds between the same two elements. The presence of multiple patterns in a given context can raise some interesting questions for semi-automatic KRC extraction, as these contexts are often both conceptually information-rich and formally variable and difficult to represent in pattern forms.

One case in which the presence of multiple patterns and pattern markers can be fairly straightforward is in the presence of "chains" of relations, as in Examples 222 to 224.

> 222.The inflammatory marker C-reactive protein (CRP) can **indicate** low-grade chronic inflammation, which can **identify**

patients at risk for atherosclerotic complications. (MacKenzie 2004)

223. In a situation of stress, activation of counterregulatory hormones and release of cytokines **increase** insulin requirement **leading to** hyperglycemia. (Pantaleo and Zonszein 2003)

224. Les gènes BRCA1 et BRCA2 **sont impliqués dans** deux tiers des prédispositions génétiques **à l'origine d'**un **risque** majeur **de** cancer du sein. (Coupier and Stoppa-Lyonnet 2002)

In these contexts, different kinds of relations between separate pairs of elements may be observed, indicated by interconnected patterns that can be represented as follows:

225. X *indicates* Y (ASSOCIATION), *which identifies* Z (ASSOCIATION)

226. X *and* Y *increase* Z (INCREASE), *leading to* W (CREATION)

227. X *et* Y *sont impliqués dans* Z (CREATION) [*qui est*] *à l'origine de* W (CREATION)

At a formal level, these may be relatively easily recognized as corresponding to pattern forms, and thus should not pose serious problems for most knowledge extraction applications (except for the possibility of multiple occurrences in lists of results, with the same context presented once for each relation occurrence observed).

In processing contexts (e.g., sorting extracted KRCs according to the relations present), co-occurrences of relation markers associated with different (sub-)relations indicate that the sorting process must involve the specific occurrences of each marker. Clearly contexts cannot be sorted exclusively according to the presence of a marker anywhere in the context.

In addition, at a conceptual level, the question of the transitivity of relations — particularly of CAUSE–EFFECT relations — does remain to be resolved. Decisions must be made in interpreting contexts such as Example 223 whether to consider, for example, that activation of hormones and release of cytokines not only modify insulin requirements, but are also causally linked to hyperglycemia. (Cf. the analysis of causation by Kahane and Mel'čuk (forthcoming) described in Section 1.5.2.4.)

Evaluating the possibilities for this kind of analysis would involve the study of a significant and varied body of data.

However, these interconnected chains of relations are not the only combinations of patterns observed in the contexts. An additional type of pattern interruption by another pattern occurs in complex sentences. In the first case, the principal clause in a sentence is interrupted by a parenthetical clause that contains another relation, as in Example 228. This may pose difficulties for the identification of both relations present, because of interruptions in the case of the relation in the principal clause, and of unusual pattern form in the case of the parenthetical. In the second case, two clauses containing relations both involving a common element are juxtaposed, as in Examples 229 and 230, and may again pose problems because of the interruption of a pattern, in this case by insertions between the related element and the second pattern.

228. A genetic background that significantly **modulates** hepatic lipase activity in vivo may potentially **impact on** the risk of coronary heart disease (CHD) and possibly **affect** individual CHD response to lipid-lowering therapy. (Zambon et al. 2003)

229. Impaired ANS regulation is **associated with** greater platelet activation, **contributing to enhanced** aggregation and adhesion to vessel walls. (Harris and Matthews 2004)

230. .... l'interaction avec l'ERE concerné peut **conduire à** l'activation de la transcription d'un sous-groupe déterminé de gènes **permettant** une formation plus ou moins complète de leurs ARN messagers ... (Kirkiacharian 2000)

The necessity of taking these variations into account in designing pattern forms may complicate this task, and difficulties encountered may interfere with recognition of relations and reduce recall.

In other — and even more complex — cases, two or more markers may link the same two elements, as in Examples 231 to 235:

231. Overall, results of our investigation indicate that the **association between risk of** breast cancer **and** HRT varies by regimen. (Weiss et al. 2002)

232. Aldosterone has been **implicated** for many years **as** an **important** substance **in** the pathogenesis of heart disease. (Moore et al. 2003)

233. Endothelial cells **help create** this antithrombogenic surface. (Granger et al. 2004)

234. This effect was reversed by mevalonate and was attributed to the **inhibitory effect of** statins **on** promoter IV of MHC-II transactivating factor, **leading to suppression of** T-lymphocyte activation. (Davignon 2004)

235. Il en va de même après l'administration d'estrogènes lesquels réduisent la production d'Il-6 et inhibent la résorption induite par les ostéoclastes **contribuant** ainsi **à maintenir** une bonne minéralisation et à protéger de la fragilisation des os. (Kirkiacharian 2000)

In some cases, as in Examples 231 to 233, the two markers are associated with the same relation or sub-relation, and thus the relation expressed in the context is fairly easily identified. In these cases, the major difficulty posed by this phenomenon is the difficulty for pattern recognition posed by the interruption of pattern structures. However, in most cases observed in the corpora (e.g., Examples 234 and 235), the markers denoted different CAUSE–EFFECT sub-relations. As a result of this phenomenon, applications that attempt to classify knowledge-rich contexts according to the relation expressed may encounter problems. In these Examples, the combination of a marker of CREATION such as *effect of... on*, *leading to*, and *contribuer à* with markers of another type of CAUSE–EFFECT relation such as *inhibitory, suppression* or *maintenir*, indicating DECREASE, PREVENTION and MAINTENANCE, require that the context be sorted into one category or another, or appear in both (thus creating repetition in the results of the application).

Similar, but more complex, is the situation observed in Examples 236 and 237, in which markers of ASSOCIATION, *associated with*, *risk* and *risque de*, are combined with *reduced* and *influencer*, markers of CAUSE–EFFECT sub-relations (specifically DECREASE and MODIFICATION).

236. Strenuous PA was generally **associated with** a **reduced** breast cancer **risk**. (Dorn et al. 2003)

237. Gènes modificateurs. Facteurs génétiques modulant l'expression d'une maladie héréditaire (exemple : gènes **influençant** le **risque de** cancer conféré par une mutation germinale de BRCA1 ou BRCA2). (Bonadona and Lasset 2003)

Classifying such contexts requires a choice between maintaining the precision of the latter relation, which identifies the type of change likely to occur, and respecting the level of uncertainty remaining about the potential CAUSE–EFFECT link between the two elements, as indicated by the markers of ASSOCIATION. Such difficulties and some possibilities for dealing with them are discussed further in Section 5.5.3.4 and 5.5.3.5.

One phenomenon that must be discussed in this context is the fact that in many cases in which multiple markers linking the same element pair were present, one of the markers present was a strong indicator of the CAUSE–EFFECT relation, while the other was less clearly causal, but nevertheless determined the specific type of sub-relation that was present. This may be observed in Examples 238 to 242.

238. Loss of ER[alpha] in MCF-7 cells **causes reduced** expression of IGF-signaling molecules, **diminished** IGF signaling, and **failure** to proliferate in response to estrogen or IGF-1. (McCance and Jones 2003)

239. These results indicate that SNS activation may **contribute to impaired** endothelial function, possibly because of activation of [beta]-adrenergic receptors. (Harris and Matthews 2004)

240. As is the case for chemotherapy, radiation-induced NF-[kappa]B activation has been reported in a variety of cancer cell types, including breast cancer, **leading to decreased** apoptosis... (Garg et al. 2003)

241.... une athérosclérose prématurée **responsable d'**une mortalité coronarienne et neurovasculaire **augmentée**... (Meyer 2001a)

242. Lorsqu'il est activé, il induit une cascade de phosphorylations intracellulaires, **conduisant à** une transcription de protéines et à une croissance cellulaire **accrues**. (Cornez and Piccart 2002)

The evaluation and classification of such relation occurrences may be challenging, because the presence of a relation is most strongly indicated by markers such as *cause*, *contribute to*, *lead to*, *responsable de* and *conduire à*, but the sub-relation

present is determined by the additional element (i.e., *reduced*, *diminished*, *impaired*, *decreased* (DECREASE), *augmenté* and *accru* (INCREASE)).

In addition, some markers of ASSOCIATION were observed to be very commonly used with others defining the type of relation present. In fact, some of these markers were observed to occur exclusively or almost exclusively with another marker of a relation; such markers included *associated with*, *risk of* and *risk for* in English and *risque de* in French, as illustrated in Examples 243 to 252.

243. Moreover, calcification itself might be **associated with** an **increased risk for** subsequent breast cancer development. (Shaaban et al. 2002)

244. One registry of 727 consecutive patients found that an elevated baseline C-reactive protein before PCI was **associated with** progressive **increase in** death or myocardial infarction at 30 days. (Shah and Newby 2003)

245. Impaired ANS regulation is **associated with greater** platelet activation, contributing to enhanced aggregation and adhesion to vessel walls. (Harris and Matthews 2004)

246. .... there may be subsets of at-risk populations in which high plant-sterol levels significantly **increase** the **risk of** CHD. (Davidson and Toth 2004)

247. Pike argues that oral contraceptives may slightly **increase** the **risk of** breast cancer, a contention disputed by a number of other researchers. (Fackelmann 1992)

248. There is good evidence that HRT **increases** the **risk for** VTE. (Kocjan and Prelevic 2003)

249. .... cyclin D1 is frequently overexpressed in human breast DCIS specimens (9, 13), which **confers** a high **risk for** the development of infiltrating ductal carcinoma. (Wang et al. 2003)

250. L'**augmentation du risque de** cancer du sein **liée à** la prise de THS ... (Fournier et al. 2003)

251. L'obésité, le syndrome métabolique et le diabète **accroissent** notablement le **risque de** maladies cardiovasculaires. (Lambert 2002)

252.Les gènes BRCA1 et BRCA2 sont impliqués dans deux tiers des prédispositions génétiques **à l'origine d'un risque** majeur **de** cancer du sein. (Coupier and Stoppa-Lyonnet 2002)

As noted in Section 3.3.1.5.1.1, for the purposes of this project, such occurrences were associated with the marker that was most decisive in classifying the relation and/or sub-relation present (i.e., DECREASE or INCREASE in the case of the first set of examples, and ASSOCIATION in the second). This does, however, result in the identification of some candidate markers that are poorer indicators of relations if they occur independently, and the failure to annotate additional occurrences of strong markers. (Fortunately, these markers are also generally common, and were thus observed in other contexts.)[152]

A fairly high proportion of the contexts in both languages contained multiple markers (22% in English and 21% in French), and in many of these cases the principal pattern form identified was interrupted by this other marker, with this phenomenon observed in 15% of the relation occurrences in both English and French and 22.5% and 27% of the interrupted pattern occurrences, respectively. These figures are illustrated in Table 99 and Table 100, which reflect a strong similarity in the prevalence of the phenomena in the two data sets ($p = 0.709$ and $p = 0.833$ respectively).

Table 99. Comparison of the proportions of relation occurrences containing multiple markers (MM) in English and French

	EN	FR	Total
MM+	96	72	168
MM-	346	277	623
Total	442	349	791

[152] One way of taking this phenomenon into account in some pattern-based applications is the development of pattern forms for these markers that also require the presence of an additional, strong marker. This of course may be challenging, but may also provide improved results in many cases.

Table 100. Comparison of the proportions of relation occurrences with interruption of patterns by other patterns (INTpp) in English and French

	EN	FR	Total
INTpp+	66	54	120
INTpp-	376	295	671
Total	442	349	791

The frequency of the phenomena indicates that they will pose significant problems for automatic applications (in the processing and sorting of contexts and/or in application performance) unless pattern forms and candidate KRC processing strategies can be developed to deal with such issues — with a reasonable investment of time and effort.

The types of contexts in which multiple markers were observed show strong parallels between the two data sets. The combinations of types of markers are often similar, often with a strong ASSOCIATION or CAUSE–EFFECT relation marker coupled with a weaker but more specific marker. In addition, some individual markers in each language are commonly observed in combination with others in such structures.

At a formal level, these regularities and similarities may present opportunities for adapting pattern forms, since a selection of the markers or types of markers that are most frequently seen in combination could be used in pattern forms that can process such contexts, without requiring the development of variations on pattern form for all markers. In addition, similarities in structures between the two languages could be useful, since pattern forms could possibly be adapted from one language to another.

Similarly, at a conceptual level, the development of strategies for dealing with the occurrences of different types of markers may also be facilitated by such regularities. Some suggestions to this effect are discussed in Section 5.5.3.5.

It would also be interesting to gather more data in order to more fully analyze the potential interlinguistic variations in the types of markers that often occur in combination with others. At a conceptual level, the similarities seem clear. However,

differences at a formal level may potentially be observed (e.g., in the proportions of the co-occurring markers that belong to various part of speech classes), and these could affect the possibilities for development of pattern forms. For example, a fairly significant proportion of the "relation-determining" markers observed in combination with other markers in both languages were nouns; however, a higher proportion of these markers were adjectives in French, and participial adjectives or verbs in English. Nevertheless, more data would be required to properly evaluate this kind of variation. Moreover, in comparing these figures it will also be necessary to take into account the overall distribution of the markers observed in the two languages, making this kind of comparison even more complex.[153]

4.10.1.2 Interruptions of complex markers

As noted in Section 2.6.1, complex pattern markers pose the unique challenge of being potentially interrupted by external elements, as in Examples 253 to 259. These interruptions may take the form of elements related to the wider discursive structure of the text (Examples 253 and 254) and modifiers of the marker or the relationship indicated (Examples 255 to 259), among other possibilities.

253.... endothelial CAM expression and several other factors (eg, oxidative stress) that **have** <u>also</u> **been implicated in** the development of CVD. (Granger et al. 2004)

254. La NADPH oxydase **jouerait** <u>donc</u> **un rôle** majeur lors des premières étapes du processus athéromateux (oxydation des LDL, adhésion monocytaire, accumulation de cellules spumeuses). (Bonnefont-Rousselot et al. 2002)

255. **Involved** <u>early on</u> **in** the inflammatory process, VCAM-1 recruits white blood cells, including monocytes and lymphocytes, to the surface of the endothelial cell... (Stix 2003)

---

[153] Conversely, the effect of this phenomenon on the proportions of markers in each class that were retained in this research would also be an interesting subject to evaluate in further work.

256.Reactive oxygen species **are produced** continuously **by** all cells in normal and pathological aerobic metabolism, from xenobiotics to ionizing radiation. (Kang 2002)

257.MMPs **have been** broadly **implicated** in a number of cardiovascular diseases, including atherosclerosis, ... aortic aneurysms, ... and heart failure... (Jaffer and Weissleder 2004)

258.It has been recognized that atherosclerosis is an inflammatory disease **in** which various cytokines **play a** significant **role**... (Taniyama and Griendling 2003)

259.Il est intéressant de noter que cet effet des oestrogènes **a été associé** chez le rat **à** une augmentation de l'expression de la connexine 43, qui est exprimée par les cellules endothéliales et musculaires lisses... (Feletou et al. 2003)

260.Les résultats publiés apparaissent très encourageants, ne montrant pas d'**effet** apparemment **délétère de** ce traitement **sur** la maladie cancéreuse mammaire préexistante. (Gorins et al. 2003)

Some markers are more susceptible to interruption than others; this is particularly true of the pattern X *plays a role in* Y. The marker may be interrupted by a number of different modifiers of the element *role*, among them the intensifiers *important*, *central*, *key*, *prominent, major*, *critical*, *significant*, and *crucial*, as illustrated in Examples 261 to 264. The French counterpart of this pattern, X *joue un rôle dans/lors de* Y, may also be interrupted by intensifiers such as *important*, *majeur*, *essentiel*, *clé*, *critique*, *fondamental*, *capital*, *principal*, *prépondérant*, *primordial*, *central*, and *crucial.* Another example is the pattern *effet de* X *sur* Y, which may be interrupted by a various modifiers. These cases are illustrated in Examples 265 to 268.

261.... endothelin may **play an** important **role** in the pathogenesis and clinical manifestations of certain cardiovascular disorders... (Ram and Venkata 2003)

262.Endothelial dysfunction **plays a** central **role** in the pathogenesis of CVD... (Pantaleo and Zonszein 2003)

263.Clearly, cyclin Dl **plays a** key **role** in mammary gland development... (Sicinski and Weinberg 1997)

264.The inflammatory process **plays a** prominent **role** in the pathogenesis of CVD... (Rackley 2004)

265. La NADPH oxydase **jouerait** donc **un rôle** majeur **lors des** premières étapes du processus athéromateux (oxydation des LDL, adhésion monocytaire, accumulation de cellules spumeuses). (Bonnefont-Rousselot et al. 2002)

266. La structure chromatinienne **joue un rôle** majeur **dans** des processus tels que la transcription, la réplication et la réparation de l'ADN. (Chailleux et al. 2000)

267. Les résultats publiés apparaissent très encourageants, ne montrant pas d'**effet** apparemment délétère **de** ce traitement **sur** la maladie cancéreuse mammaire préexistante. (Gorins et al. 2003)

268. Il a été décrit un **effet** synergique **des** oestrogènes et **de** l'IGFI **sur** la transcription de pS2. (Chailleux et al. 2000)

In addition to the potential conceptual impact of these modifications (for example, when a modification indicates some level of uncertainty about a relation and thus casts doubt on the value of a potential KRC for knowledge extraction), interruptions of markers have clear formal implications. These may be reflected not only in a need to allow for interruption of markers as they are represented in pattern forms, but also in possible modifications to the central elements of the marker that may be associated with the phenomenon, as in Example 261, in which *a* becomes *an* before a modifier beginning with a vowel (*important*). This adds yet another layer of complexity to pattern design in order to allow for the identification of such variations.

As in any pattern refinement process, the goal in creating pattern forms to accommodate these interruptions is to identify forms that locate a maximum of pertinent contexts without an inordinate amount of noise. In order to do this, researchers may need to evaluate the frequency with which individual markers tend to be interrupted, and the types of elements that may interrupt them. For example, the frequent interruption of markers such as *play a role in* by adjectival modifiers indicates a need to design a pattern form for this marker that can identify occurrences interrupted in this way.

Moreover, if the interrupting elements are pertinent in other respects — for example, if they are intensifiers, hedges or other modifiers of the relation in question (Cf. Section 4.10.2) — they may be of value in the process of sorting contexts and/or

evaluating the information contained in potential KRCs in and of themselves. Sets of modifiers that frequently collocate with markers may be identified and could help in this task, if they can be consistently linked to a given function in intensifying or attenuating a relationship, or in otherwise characterizing it.

In another type of phenomenon, complex markers may also be interrupted by one of the elements linked by the relation, as in the case of patterns such as Z *implicates* X *in* Y, *importance of* X *in* Y, *effect of* X *on* Y and *role of* X *in* Y in English and *effet de* X *sur* Y and *déclenchement de* X *par* Y in French, as shown in Examples 269 to 274.

269. There is evidence **implicating** oxidative stress **in** the pathogenesis of stroke, myocardial infarction, myocardial stunning, atherosclerosis, and congestive heart failure. (Granger et al. 2004)

270. The **importance of** glycaemia **in** the development of microalbuminuria has also been demonstrated in the Framingham Offspring Study (MacIsaac et al. 2004)

271. Recognition of the **effects of** influenza **on** CHD provides the medical community with a valuable opportunity to further reduce cardiovascular death and morbidity. (Madjid et al. 2004)

272. As with heart failure, the **role of** aldosterone **in** the pathogenesis of hypertension has also been studied for decades. (Moore et al. 2003)

273. Les risques de saignements seraient reliés à l'**effet de** l'ail **sur** la coagulation. (Trahan 2002)

274. Lorsque la plaque est rompue, le **déclenchement de** la coagulation **par** les cellules inflammatoires aboutit à la thrombose… (Collet et al. 2004)

In these cases, the interruption tends to be both typical of a given marker form and relatively regular in its own form, and would therefore be fairly easy to take into account in designing pattern forms. However, in other cases, these markers may be interrupted in a less regular way; for instance, in Example 275, the marker *implicated in* is interrupted by not only the modifier *for many years*, but also by *as an important substance*, in which *substance* is a generic of *aldosterone*:

275. Aldosterone has been **implicated** for many years as an important substance **in** the pathogenesis of heart disease. (Moore et al. 2003)

Interruptions of complex markers were observed in 73 contexts in English and 39 in French, constituting 16.5% of English relation occurrences and 11% of relation occurrences in French. These constitute 25% of the interrupted pattern occurrences in English and 19% in French, and 28% of complex marker occurrences in English and 19% in French). These data are shown below in Table 101, Table 102 and Table 103.

Table 101. Comparison of proportions of relation occurrences containing interruptions of complex markers (INTcm) in English and French

	EN	FR	Total
INTcm+	73	39	112
INTcm-	369	310	679
Total	442	349	791

Table 102. Comparison of proportions of interrupted relation occurrences containing interruptions of complex markers (INTcm) in English and French

	EN	FR	Total
INTcm+	73	39	112
INTcm-	220	163	383
Total	293	202	495

Table 103. Comparison of proportions of occurrences of complex markers containing interruptions of complex markers (INTcm) in English and French

	EN	FR	Total
INTcm+	73	39	112
INTcm-	191	167	358
Total	264	206	470

While the proportion of interrupted contexts involving interruptions of complex markers was not significantly different ($p = 0.143$), the English data do show a higher proportion of interruptions of this type. Moreover, the proportions of relation occurrences containing this phenomenon and the proportion of interrupted complex

marker occurrences were significantly higher in English ($p = 0.032$ and $p = 0.028$ respectively).

These figures suggest that while the difficulties involved in such interruptions affect a significant proportion of relation occurrences in both languages, their impact may be greater in English.[154] Pattern design in this language may thus be considerably more complex, because representing variation in marker forms will pose particular challenges.

However, a more detailed evaluation reveals that a fairly large proportion of these interruptions (54 in English and 16 in French) belong to the more "regular" category of interruptions, i.e., complex markers that are interrupted only by one of the elements that they link. These interruptions are among the more straightforward to account for in designing pattern forms, posing far fewer problems because of the consistency of the occurrence and form of the interruptions.

When the occurrences of interrupted markers that belong to this category are set aside, no significant difference between the two data sets is observed, as illustrated in Table 104 ($p = 0.154$). Moreover, the proportion of occurrences is somewhat higher than expected in French, rather than in English.

Table 104. Comparison of proportions of relation occurrences with interruptions of complex markers other than by related elements (INTcmo), in English and French

	EN	FR	Total
INTcmo+	19	23	42
INTcmo-	423	326	749
Total	442	349	791

[154] This is further supported by the observations that the proportions of complex marker occurrences (which are therefore vulnerable to this kind of interruption) are comparable in English and French (cf. Section 4.5.2).

This suggests that the phenomenon of interruption of complex markers by related elements may be more prevalent in English than in French, both as a proportion of the total number of relation occurrences as illustrated in Table 105 ($p < 0.001$), and as a proportion of complex marker interruptions ($p = 0.001$), as shown in Table 106.

Table 105. Comparison of proportions of relation occurrences with interruptions of complex markers by related elements (INTcmre) in English and French

	EN	FR	Total
INTcmre+	54	16	70
INTcmre-	388	333	721
Total	442	349	791

Table 106. Comparison of proportions of complex marker occurrences with interruptions by related elements (INTcmre) in English and French

	EN	FR	Total
INTcmre+	54	16	70
INTcmre-	210	190	400
Total	264	206	470

These data indicate that in the two languages, complex marker interruptions are more likely to come from different sources. In the English data, the overall proportion is much higher, but much of this difference comes from the category of "regular" interruptions that are most likely to be taken into account in the design of basic pattern forms, likely increasing the complexity of this task. In contrast, interruptions in the French occurrences tended to be of a more unpredictable type that would likely involve more challenges in adapting pattern forms, and that could cause problems for KRC recognition due to their unpredictable nature. The variation in the prevalence of the two specific phenomena between the two data sets provides a striking example of the subtle differences that can affect the development and performance of various types of pattern-based tools, as well as the need to fully understand the phenomena observed in order to predict the effect that these differences may have in a specific use situation.

It is interesting to note that if the interruption of complex markers by related elements alone is eliminated from the overall figures (as it is generally more a regular occurrence than a specific difficulty), the proportions of interrupted relation occurrences in the two data sets are extremely similar, with 244 occurrences in English and 191 in French ($p = 0.894$). This type of interruption is thus identified as the major source of the difference between the English and French data in respect to interruptions.

This in turn clarifies the impact the interlinguistic difference observed for the phenomenon of interruption in general is likely to have. As complex marker interruption of this type primarily affects the complexity of developing pattern forms — specifically of representing marker forms and the structures in which they appear — it is thus in the investment of time and effort required to develop these forms that English appears in these data to be likely to present more difficulties than French. This phenomenon affects a wide range of pattern-based applications, as it must be taken into account in pattern forms from character strings representing relation markers to highly specific representations of KRC structures. However, as this is among the more regular forms of interruption at a formal level, the impact of the difference between the two languages is likely to be considerably lower than if another type of interruption had been involved. Moreover, the likelihood of effects at the level of application performance, due to the difficulties of comprehensively representing all forms of interruptions, are not likely to be as high as they would be expected to be with other types of interruptions.

It is interesting to contemplate the possibility of a link between the higher prevalence of the class of nominal markers (cf. Section 4.5.1.2) and of marker interruption by a related element in the English data, as a large number of the interrupted markers were noun-based. Further research could clarify how these factors may be inter-related.

Another interesting difference was observed in some recurrent structural differences in the two data sets. These may be illustrated, as in Examples 276 to 283, by variations in commonly modified markers such as those found in the English patterns X

plays a role in Y and *role of* X *in* Y, as well as their French counterparts X *joue un rôle dans* Y and *rôle de* X *dans* Y. In the case of X *plays a* ADJECTIVE *role in* Y and X *joue un rôle* ADJECTIVE *dans* Y, modifiers interrupt the markers themselves. However, in the case of structures such as ADJECTIVE *role of/for* X *in* Y and *rôle* ADJECTIVE *de* X *dans* Y, in English the intensifier occurs outside the pattern entirely and thus does not interrupt the marker, while in French the marker is interrupted:

276. Clearly, cyclin Dl **plays a** key **role in** mammary gland development... (Sicinski and Weinberg 1997)

277. La NADPH oxydase **jouerait** donc **un rôle** majeur **lors des** premières étapes du processus athéromateux (oxydation des LDL, adhésion monocytaire, accumulation de cellules spumeuses).

278. William Osler ... was one of the first to propose a major **role for** acute infection **in** the pathogenesis of atherosclerosis. (Madjid et al. 2004)

279. Direct evidence for an important **role for** myeloperoxidase **in** lipid oxidation in vivo comes from recent studies by Zhang et al. (Brennan and Hazen 2003)

280. These findings collectively indicate the significant **role of** oxidative stress **in** the development and progression of cancer. (Kang 2002)

281. Plusieurs auteurs ont évoqué un **rôle** potentiel **des** Herpesviridae **dans** leur physiopathologie. (Chidiac and Braun 2002)

282. Le **rôle** potentiel **du** tamoxifène **dans** la prévention du cancer du sein est basé sur ... (Serin and Escoute 1998)

283. Ils suggèrent par ailleurs un **rôle** important **de** l'apoptose des cellules endothéliales **dans** le mécanisme d'érosion ... (Mallat and Tedgui 2004)[155]

Whether these differences in structures affecting pattern interruption are systematic enough to make a significant difference in tool design and performance overall remains to be determined, of course, as significantly more data and analysis

---

[155] Examples 281 to 283 are taken from the corpus used for this project, but were not part of the set of relation occurrences analyzed in this project. They are provided here simply to illustrate the potential for variation.

would be required to investigate the phenomenon. These observations do nevertheless suggest that regular structural differences may be important factors in the interruption of marker forms, and that identification of pattern forms for use in new languages should include a phase of evaluation and adaptation to deal with such variations.

4.10.1.3 Interruptions of related elements

In addition to relation markers, the elements that they link may also be interrupted in various ways, for example by additional related elements (e.g., abbreviations, generics or specifics of related elements, as in Examples 284 and 291), anaphora (Example 285), quantifiers (Examples 286 and 292), modifiers (including intensifiers and hedges) (Examples 287, 288 and 293), negation (Example 289), references (Example 290), and other discourse-related elements (Example 294).

284. TNF-[alpha]-regulated SK activation is likely to be **important in** nuclear factor-[kappa]B (NF-[kappa]B) activation and inhibition of apoptosis. (Saba and Hla 2004)

285. Long-term activation of these appropriate responses **leads to** left ventricular remodelling... (Stevens and Levin 2003)

286. The acceptance that endothelin may **play an** important **role in** the pathogenesis and clinical manifestations of certain cardiovascular disorders... (Ram and Venkata 2003)

287. Emerging data reveals that a large number of additional proteins (i.e., growth factors) **influence** the transcriptional activation **of** ER[alpha] and possibly ER[beta]. (McCance and Jones 2003)

288. The 26S proteosome, **responsible for** the degradation of the inhibitory I[kappa]B[alpha] protein and subsequent activation of NF-[kappa]B... (Garg et al. 2003)

289. Oral but not transdermal HRT **induced** APC resistance... (Seed and Knopp 2004)

290.... chemotherapy,41 tamoxifen,42 and RT 43 all **act to reduce** LR independently of surgery. (Naik et al. 2004)

291. La chimiothérapie et l'hormonothérapie sont des traitements systémiques qui **ont pour but de diminuer** la récidive, surtout systémique. (Martin 2003)

292.... la p53 qui **régule** la transcription de diverses molécules impliquées dans l'apoptose (bax, inhibiteurs de kinases dépendantes de cyclines) (Kolb 2001)

293. L'expression du gène cycline E est alors directement sous la dépendance des signaux extrinsèques, et ne **nécessite** plus une activation préalable de la cycline D1.

294. Les monocytes sont alors activés en macrophages (Ma) ce qui **contribue** probablement **à accroître** l'oxydation des LDL (flèches pointillées). (Arnal et al. 2003)

In some of these cases (e.g., Examples 288 and 290), the interruption occurs between distinct elements that share a role in a relation, applying to one or more of these elements. However, this is not always the case. Interruption often occurred within the form of a more complex related element (e.g., a proposition).

The automatic identification of related elements can thus be complicated by the insertion of external items within these elements, particularly in the case of the more complex structures observed above. Identifying the base forms of these elements (those that are suitable for inclusion in term banks or for labelling nodes in ontologies, for example) may be difficult or even impossible for automatic applications. The variability in the form and nature of these interruptions makes this task all the more challenging. Moreover, the phenomenon may interfere with the recognition of KRCs by tools that use pattern forms specifying the structures in which relation markers may occur, if these impose restrictions on the form of the elements linked by markers. Finally — and perhaps most strikingly — applications that use previously identified terms or candidate terms as starting points for extraction may confront severe difficulties in identifying such contexts because of variations in form due to these interruptions.

This phenomenon was observed in 7% of the relation occurrences and 11% of the interrupted occurrences in English and 12% of the relation occurrences and 21% of the interrupted occurrences in French, as shown in Table 107 and Table 108. The statistical evaluation of the differences observed reveals that this phenomenon is

significantly more frequent in French both as a proportion of the total relation occurrences ($p = 0.021$) and of those that were interrupted ($p = 0.002$).

Table 107. Comparison of the proportions of relation occurrences containing interrupted related elements (INTre) in English and French

	EN	FR	Total
INTre+	32	42	74
INTre-	410	307	717
Total	442	349	791

Table 108. Comparison of the proportions of interrupted relation occurrences involving interrupted related elements (INTre) in English and French

	EN	FR	Total
INTre+	32	42	74
INTre-	261	160	421
Total	293	202	495

In both languages, therefore, the proportions of contexts that pose this kind of problem for automatic identification of related elements (either as terms or candidate terms or represented by POS classes or other elements as part of pattern forms), should be fairly substantial. However, the fact that a greater proportion of contexts was affected by this phenomenon in French suggests that greater problems with the identification of (appropriate forms of) related elements in that language may be encountered by applications that attempt to automate this task, or in the recognition of KRCs involving previously identified terms or candidate terms.

Moreover, the specific forms of interruptions of related elements show significant variation, and thus the development of forms that can account for this kind of variation would be extremely challenging. Further research with more data would be required in order to identify any regularities that could be exploited for such applications.

4.10.2 Expressions of uncertainty

In addition to challenges in identifying and extracting KRCs and the information they convey, pattern-based tools may also confront difficulties in evaluating the value of the information in these contexts (e.g., by excluding candidate KRCs from results if there are indications that the information may not be reliable), or assisting users in doing so (e.g., by ranking contexts to present those that appear most pertinent first in order to save a user time and effort in interpreting the information retrieved, or even attributing a level of certainty to information as suggested by indicators in the context).[156] One pertinent phenomenon involves indications of reservation, doubt or uncertainty about the relation expressed in a candidate KRC, which for some applications may make the context unusable, and for others may require special treatment.

As expressions of uncertainty affect the value of contexts for knowledge extraction — the goal of the applications studied in this research — this phenomenon can affect any type of pattern-based tool. Those that depend heavily on human interpretation of extracted contexts may encounter fewer severe problems due to this phenomenon, but even these tools may take the phenomenon into account. An application that can classify relation occurrences finely and accurately according to their levels of certainty can help users to locate reliable information for a given application quickly and easily.

The preference for a particular approach depends largely on user needs in a situation; strategies for implementing it depend on the form expressions of uncertainty take and the possibilities they offer for representation and automatic processing.

From a formal perspective, in addition to the challenges at a conceptual level that affect all pattern-based tools, there are also in many cases difficulties for tools that use

---

[156] Moreover, even if a tool does not attempt to perform this task automatically, users of the contexts retrieved must be sensitive to indications of the potential value of information extracted in a particular situation.

pattern forms representing structures in which markers occur, as expressions of uncertainty may often appear within these structures (a phenomenon explored in Section 4.10.1, affecting the complexity of pattern design as well as application performance).

Expressions of uncertainty were common in the KRCs identified in both languages, appearing in 40% of the English occurrences and 26% of the French (Table 109). According to a Chi-square test of these results, this type of phenomenon was significantly more frequent in English than in French ($p < 0.001$).

Table 109. Comparison of relation occurrences containing expressions of uncertainty (EC) in English and French

	EN	FR	Total
EC+	175	90	265
EC-	267	259	526
Total	442	349	791

When various types of expressions of uncertainty are analyzed (Table 110), the rank order of the various types of expressions shows a weak positive correlation. This indicates that while there are similarities in the types observed, these are somewhat mitigated by the differences in their proportions, particularly of hedges and modal verbs observed. The results of a Chi-square test comparing the proportions of expressions belonging to each category do not identify any significant difference ($p = 0.234$).

Table 110. Detailed comparison of proportions of relation occurrences containing various types of expressions of uncertainty in English and French

	EN	FR
Quantifiers	32	21
Hedges	105	52
Modal verbs	56	20
Negation	21	15
All types[157]	175	90

[157] This figure reflects the number of individual contexts affected and not the sum of the occurrences observed, as two or more (types of) expressions of uncertainty were observed in some contexts.

Below, each type of expression is analyzed in more detail, and the observations in the two corpora compared using the Chi-square test.

4.10.2.1 Quantification of related elements

In a number of the knowledge-rich contexts observed in the research, related elements were accompanied by a quantifier. Examples 295 to 299 illustrate cases in which quantifiers used with elements must be taken into account in order to achieve a complete and accurate understanding of the relation indicated in the context.

295.... these mutations are **responsible for** 30 to 80 percent of all hereditary forms of the diseases. (Pistoi 2001)

296. Most strokes **result from** atherosclerosis in arteries either within the brain or leading from the heart to the brain. (DiGiovanna and Adams 1999)

297. **Associations between** lymph node metastases, various clinicopathological features, **and** development of distant metastasis were assessed with the Pearson [chi]2 test. (Susnik et al. 2004)

298. Les chercheurs estiment aujourd'hui que 5% des cancers du sein **sont dus à** une mutation du gène BRCA1. (Dussault 1997)

299. Certaines mutations **engendrent** aussi des protéines oncogéniques... (Chène 1999)

As above, quantifiers of related elements may indicate proportions of a total precisely (e.g., *30 to 80 percent of, 5% de*) or approximately (e.g., *most*, *various*, *certains*).

Quantification of related elements is pertinent on two levels. First, it may be considered in the context of pattern interruptions, since quantifiers often occur between a marker and one of the elements it links. This phenomenon is discussed in Section 4.10.1, and will not be further considered here. More significant in this discussion, in terms of the further usefulness of contexts, is uncertainty as to the validity of a relation identified in the context. Quantification of related elements often indicates some kind of condition on the involvement of elements in a relation (e.g., indicating that a relation holds between only some members of the class of elements indicated).

When quantifiers are used to indicate that a given relation is not universally present between all members of a class, as in the case of markers such as *X percent of* or *most*, the contexts — while still potentially useful for the applications mentioned above — must be considered with a certain amount of reserve.[158] The case of *various* requires somewhat more interpretation, but nevertheless indicates uncertainty by signalling a distinction between the relation as it would be interpreted from an unqualified statement (in which the default interpretation would be that the relation holds between all members of the class) and the statement as presented above, indicating that some of the members of the class participate in the relation.

The use that can be made of the information extracted may depend on the goals of the terminologist and the possibilities for indicating this kind of uncertainty about a relation between elements in the application of the information obtained. For example, in applications intended to assist in automatically linking term records in a resource such as a terminological knowledge base (e.g., by indicating a CAUSE–EFFECT relation in the form of a link between term records for the concepts denoted by the terms *BRCA1 mutation* and *breast cancer*), contexts such as Examples 295 and 298, which indicate the presence of a relation in varying proportions of cases ranging from very low to very high, are not likely to be sufficient to justify a connection. However, terminologists working on manually enriching term records in light of these contexts might instead consider adding a note about the potential relationship to one or both term records, or including such a context in one of the term records to make the information available to users without representing it formally.

---

[158] The (authors' opinion of the) certainty of a relation may be also expressed by the use of quantifiers such as *all*, *tous* and so on, which correspond to the criterial level of certainty discussed by Barrière (1996) (cf. Section 2.4.1). These kinds of expressions thus may increase the value of the context for further applications (e.g., acquiring domain knowledge, formulating definitions or establishing links between entries in terminological resources). They are not, however, discussed here, as this analysis focuses on difficulties for use of contexts. The use of markers corresponding to the excluded range of possibilities (e.g., *no*, *none*, *not... any*) of course also pose problems in using contexts. Such cases will be discussed in Section 4.10.2.4, which focuses on negation.

Some tools may be able to take advantage of regularities in the use of some of these quantifiers in order to identify a given level of certainty and potentially exclude or sort affected contexts using this criterion. This might involve, for example, implementing a scale of certainty levels on which each quantifier can be placed, such as that developed by Barrière and her colleagues (Barrière 1996; Barrière and Hermet 2002), and using this scale to sort contexts containing expressions of possibility (e.g., *some*), probability (e.g., *most*) and so on. However, dealing with more specific expressions of quantification, such as those involving numbers or percentages, poses more challenges for formal evaluation of certainty levels. Moreover, the complexities involved in differentiating between quantifiers that express certainty (e.g., *all*) and those expressing uncertainty, and in processing combined forms such as *30 to 80 percent of all* or *virtually all*, could complicate this task.

In addition, while in most cases observed in the data quantification applied to a related element in its entirety, in a few it applied to only part of a more complex element or to one of multiple elements sharing a role in a relation, posing additional challenges for formal representation and analysis of the phenomenon.

Quantification of a related element was observed in 7% of relation occurrences and 18% of cases involving expressions of uncertainty in English, and 6% and 23% in French, respectively. These figures are represented in Table 111 and Table 112.

Table 111. Comparison of proportions of relation occurrences involving quantification of related elements (QR) in English and French

	EN	FR	Total
QR+	32	21	53
QR-	410	328	738
Total	442	349	791

Table 112. Comparison of proportions of relation occurrences containing expressions of uncertainty involving quantification of related elements (QR) in English and French

	EN	FR	Total
QR+	32	21	53
QR-	143	69	212
Total	175	90	265

These results reveal that quantification of related elements occurred in similar proportions of the relation occurrences observed in the two data sets ($p = 0.495$) — with the proportion in the English data only very slightly higher — and also of those containing expressions of uncertainty ($p = 0.331$), in this case with the proportion in the French somewhat higher.

Overall, this suggests that the return on the investment of time and effort in accounting for the possibility of quantification within pattern forms, as well as that of developing strategies for automatically identifying such cases in extracted contexts in order to identify them for a user or sort them according to this phenomenon, is likely to be fairly comparable in the two languages.

However, the complexity of the task may not be as similar. More forms of quantifiers were observed in the English data than in the French. A total of 14 distinct lexical indicators of quantification was observed in 29 occurrences in English (i.e., a ratio of 2.1 occurrences per marker), with 7 indicators in 19 occurrences in French (i.e., a ratio of or 2.7 occurrences per marker). These markers are illustrated in Table 113 and Table 114.

Table 113. English quantifiers of related elements observed

Quantifier	**Occurrences**
a (large) number of	4
several	4
various	4
a number of	2
a variety of	2

certain	2
many (of)	2
range of	2
some (of)	2
another	1
more than	1
most	1
multiple	1
virtually all	1
Total	29

Table 114. French quantifiers of related elements observed

Quantifier	Occurrences
certains (certaines)	7
de nombreux (de nombreuses)	4
un sous-groupe déterminé de	3
divers (diverses)	2
la plupart de	1
plusieurs	1
un de	1
Total	19[159]

In addition, in English three relation occurrences involved quantification by percentages and three by a number, and in French three contexts involving quantification by a number were also found. Examples 300 to 302 illustrate these cases.

300. In high-risk populations (i.e., Ashkenazi Jewish), the threshold is lower; for example, 12% of the cases of breast cancer and 48% of ovarian cancer in Ashkenazi women **were related to** a BRCA mutation. (Khoury-Collado and Bombard 2004)

301. In 2000 there were over an estimated over 1 million new cases and approximately 373,000 deaths **from** breast cancer worldwide, an age standardised death rate (ASR) of 12.51 per 100,000. (Carrick et al. 2004)

302. Pour 17 patientes, il y avait un haut risque de récidive pariétale **du fait de** la présentation clinique : récidive inflammatoire (4 cas), récidive multifocale (5 cas), nodules de perméation (5 cas)… (Racadot et al. 2003)

[159] One context in this set contained two occurrences of quantifiers. Multiple occurrences of other types of quantifiers discussed below were also observed.

These results, although they are limited and require confirmation in light of more and more varied data, suggest that a larger variety of expressions may be used for quantification in English. This would require additional time and effort if quantifiers are to be accounted for in pattern forms and/or used as cues for sorting contexts according to the certainty of the information they convey. This underlines an interesting point, that the balance between the possibilities for recall offered by a given strategy and the time and effort it may take to implement such a strategy may often vary inversely, requiring that a choice be made according to the priorities set for a given project.

In both corpora, a range of both relatively standard expressions and of expressions that were more varied in form and usage were observed. Of course, the latter are likely to pose more challenges for automatic processing.

4.10.2.2 Hedging

As discussed in Section 2.4.2.2, various authors have discussed expressions used to express some uncertainty regarding statements made in texts. Some working with knowledge patterns (e.g., Pearson 1998) have dealt with relatively restricted sets of these expressions in analyzing the value of contexts for extracting knowledge and the possibilities of evaluating this value using formal cues, but it is possible to examine a wide range of possible means of expressing uncertainty about or restrictions on statements made in texts.

In this research, hedging was observed to focus on several different aspects of a relation or the basis on which it is asserted. One group of hedges refers to the necessity of some kind of interpretation of results, showing a tendency that opposes that of scientific style's usual concern with maintaining (an appearance of) objectivity (e.g., in English *appear to*, *seem*, *view*, *suggest*, *hypothesis*, *theory*, *controversy*, *dispute* and *presumption*, and in French *sembler*, *suggérer*, *considérer*, *apparemment*, *vraisemblablement*, *hypothèse*, and *débats*). A second group indicates restrictions on the consistency of a given relation's occurrence (e.g., in English *likely*, *generally*, *normally*,

often, *in some cases*, *tend to*, *potential* or *possible*, and in French *susceptible de*, *tendance à*, *possible*, *éventuel*, *potentiellement* and *probablement*).[160] Another type of hedge qualifies the importance of the role played by a given element (e.g., *in part* in English). Still another qualifies the degree of the relationship (e.g., *little* and *slightly* in English and *essentiellement*, *plus ou moins*, *moins*, *moindre* and *peu ou pas* in French).

Another type of hedging related to the discourse structure in which a statement is made involves the use of items such as *although* or *nevertheless* in English and *bien que* in French; these indicate some kind of reservation about the statement made, and often occur in combination with other expressions of uncertainty, as illustrated in Examples 303 to 306.

> 303. Although it seems at present that there is no **effect of** HRT **on** breast cancer mortality, more studies are needed to clarify this issue. (Kocjan and Prelevic 2003)
>
> 304. Although their study has shown the **importance of** ARHI inactivation **in** breast tumor pathogenesis, the technique they used is real-time PCR… (Wang et al. 2003)
>
> 305.… these findings suggest that it is likely that the fractalkine/CX3CR1 system may nevertheless **be important in** the pathogenesis of atherosclerotic and coronary vascular diseases. (Umehara et al. 2004)
>
> 306. Bien qu'une RCH ne **garantit** [sic] pas définitivement **contre** une récidive, sa valeur puissante pronostique est confirmée dans de nombreuses analyses multifactorielles. (Brain 2000)

Finally, hedging may be accomplished using descriptions of the availability, sufficiency or reliability of data to justify conclusions. These tend to be formally both more complex and more variable than those observed above, often taking the form of phrases or propositions, as in Examples 307 to 310:

> 307. … the risk of mortality from breast cancer **related to** HRT could not be determined. (Watkins 2003)

---

[160] Quantification of the types of elements that may participate in a given relation, using markers such as *certain*, *some*, or *virtually all* (cf. Section 4.10.2.1), may of course also play a similar role.

308. To date, only limited data **correlate** Bcl-XL expression **and** breast cancer treatment response in humans. (Garg et al. 2003)

309. No study evaluated the **associations between** statins' effects on LDL oxidation **and** lipid levels… (Balk et al. 2003)

310. De plus, certaines données indiquent que l'extrait d'ail vieilli **réduirait** l'oxydation des LDL. (Trahan 2002)

The most obvious challenge hedges pose at a formal level, for patterns that specify the structures in which relation markers occur, results from their interruption of pattern forms, complex relation markers, or related elements. This may require adaptation in pattern design and/or interfere with the recognition of KRCs or the elements involved in them. As interruptions were discussed in Section 4.10.1, this will not be discussed further here. The insertion of hedges (e.g., verbs or verb phrases) within pattern forms may also change the morphological forms of markers observed in relation occurrences (e.g., from X *causes* Y to X *seems to cause* Y).[161]

At a conceptual level, the interpretation of hedging is of course far simpler for human users than for automated applications. Nevertheless, the level of certainty or uncertainty indicated by a given expression can often be extremely challenging to evaluate in both cases. As studies such as those carried out by Barrière (1996, 2002; cf. also Barrière and Hermet 2002) have illustrated, some associations between specific expressions and levels of certainty can be established. These may assist in evaluation of this phenomenon and ultimately in applications such as automatic sorting of extracted candidate KRCs or elimination of those that are considered to be too uncertain for use. The possibilities for implementing such strategies, however, hinge on the regularity of the expressions observed. While the more predictable (and generally simpler) means of expressing uncertainty could often be listed, automatically identified and used in a sorting process, less frequent or more variable means will still be difficult to evaluate automatically. Moreover, the question of whether contexts associated with a given level

---

[161] While this issue will not be discussed further here, as the morphological variation of markers was not considered in this research, some examples may be found in the sample contexts provided for markers in Appendix H.

of certainty are useful for knowledge extraction is one that can only be answered by users in light of their specific goals and intended applications of this information.

Hedging occurred in 24% of the relation occurrences and 60% of those including expressions of uncertainty in English, and 15% and 58% respectively in French. These data are presented in Table 115 and Table 116.

Table 115. Comparison of the proportions of relation occurrences containing hedging (HG) in English and French

	EN	FR	Total
HG+	105	52	157
HG-	337	297	634
Total	442	349	791

Table 116. Comparison of the proportions of relation occurrences containing expressions of uncertainty that involved hedging (HG) in English and French

	EN	FR	Total
HG+	105	52	157
HG-	70	38	108
Total	175	90	265

In the two data sets, hedging occurred in a very similar proportion of the relation occurrences involving expressions of uncertainty ($p = 0.727$). This reflects the prevalence of the phenomenon in both corpora, and its role as one of the primary means of expressing uncertainty.

However, hedging was present in a significantly lower proportion of the total relation occurrences in the French data than in the English ($p = 0.002$), suggesting that the value of strategies developed for dealing with the phenomenon as observed in this research could be particularly high in this language. The difference in prevalence in the relation occurrences indicates that this phenomenon contributes significantly to the overall difference observed in the category of expressions of uncertainty as a whole.

The higher prevalence of the phenomenon is a difference that is worth investigating further, in order to determine potential sources of this variation. It is possible, for example, that the concepts denoted by the terms chosen for use in each of the corpora participate in relations characterized by varying levels of certainty; more data gathered using a wider range of terms or another methodology designed to neutralize this potential contributor to the difference observed could clarify this issue.

Another focus for further investigation could be the prevalence of hedging in relation occurrences involving various classes of terms, to determine whether this factor affects the interpretation and/or description of the relations that may be observed (e.g., if authors present observations of relations involving entity, activity, process or pathology concepts with more or less certainty, potentially reflecting the possibilities for observing the real-world objects they represent and/or connections between these concepts and others).[162] The evaluation of more relation occurrences involving each type of concept would provide an opportunity to study this factor as well.

4.10.2.2.1.1 Types of expressions used for hedging

Hedging can be accomplished using several different lexical means, including verbs, adjectives and adverbs (and verb, adjective and adverb phrases) as well as more complex units such as propositions. These may be observed in Examples 311 to 318:

> 311.... effects of oxidized LDL on vascular smooth muscle cells, which contribute to the atherogenic process appear to **require** the activation of SK. (Saba and Hla 2004)
>
> 312. Activation that endures beyond the resistance stage is hypothesized to **cause** disease. (Schwartz 2003)
>
> 313. Strenuous PA was generally **associated with** a reduced breast cancer risk. (Dorn et al. 2003)

---

[162] It might be particularly interesting to investigate the prevalence of hedging in contexts indicating relations involving artefact and activity concepts. These are likely to be associated with specific and observable goals and thus could be less likely to be described in statements necessitating hedging.

314. Although interactions between cardiovascular ANS regulation and endothelial function are likely **involved in** CVD development, further research is needed to determine whether ANS and endothelium interactions are a plausible pathway... (Harris and Matthews 2004)

315. Il semble exister un **lien** très étroit **entre** le syndrome de lipodystrophie, l'hyperlipidémie, l'intolérance au glucose **et** le diabète, bien que chacun de ces troubles puisse survenir isolément. (Baril and Junod 2004)

316. Cette prolifération musculaire lisse **participe à** la constitution de la plaque athéroscléreuse et à l'éventuelle réduction de la lumière artérielle... (Teiger 2001)

317. Les espèces lipidiques oxydées **responsables de** ces effets sont essentiellement des dérivés d'oxydation des phospholipides tels que le POVPC. (Bonnefont-Rousselot et al. 2002)

318. L'**induction** de tumeurs bénignes ou malignes ovariennes **par** une stimulation continue des ovaires est une hypothèse qui a déjà été soulevée... (Sasco et al. 1997)

These markers may occur in simple forms, or as part of far more complex structures and combinations. In some of these more complex cases, one or more hedges may appear in conjunction with other types of expressions of uncertainty, such as negation, modal verbs, or quantification, as in Examples 319 to 324.

319. Although it seems at present that there is no effect of HRT on breast cancer mortality, more studies are needed to clarify this issue. (Kocjan and Prelevic 2003)

320. It is possible that basal IGF activation of the ER may **be necessary for** maximal estrogen-mediated activation and may, in part, **explain** the synergy observed between the two mitogens. (McCance and Jones 2003)

321.... the findings suggest that dietary intake of fat and fiber do not **play a** major **role in** the development of breast cancer. (Fackelmann 1992)

322. Il pourrait paraître illogique qu'une augmentation de la concentration extracellulaire de potassium puisse **provoquer** l'hyperpolarisation des cellules musculaires lisses... (Feletou et al. 2003)

323. Les résultats publiés apparaissent très encourageants, ne montrant pas d'**effet** apparemment délétère **de** ce traitement sur

la maladie cancéreuse mammaire préexistante. (Gorins et al. 2003)

324. En revanche, les deux études randomisées les plus récentes ont désormais démontré qu'une chimiothérapie d'induction peut augmenter les possibilités de chirurgie sans **diminuer** significativement les taux de survie... (Lerouge et al. 2004)

These complex forms are of course challenging to take into account in designing pattern forms that attempt to describe the context in which markers may occur, and in any application that attempts to evaluate levels of certainty on the basis of these formal markers. This kind of interpretation may be particularly challenging in cases involving multiple expressions of uncertainty, particularly as the interactions of these various elements with one another or with other elements may either reduce or increase the level of uncertainty present in a given context.

In addition to these lexical indicators, hedging may also be indicated by non-lexical means, for example by question forms (3 occurrences in the English data) as in Examples 325 and 326 or even potentially verbs in the future tense (1 occurrence in the English and French data) as in Examples 327 and 328:

325.... the following critical question remains unanswered: 'is oxidation **important in** human atherosclerosis?'. (Brennan and Hazen 2003)

326. Could it be that BRCA1 and BRCA2 **play roles in** the development of hereditary cancers but not sporadic tumors? (Yang and Lippman 1999)

327.... among ER-positive tumors, nearly 70% of those that are also progesterone receptor (PR)-positive and 25-30% of PR-negative tumors will **respond to** hormonal therapy. (Vogel 2003)

328.... cette augmentation **va favoriser** l'adhésion, l'agrégation plaquettaire et la coagulation par l'interrelation avec le facteur VIII de la coagulation. (Drouet and Bal Dit Sollier 2002)

Hedging was indicated in some other occurrences in English by variation in the form of modal verbs (cf. Section 4.10.2.3), e.g., from *may* to *might* or *can* to *could*. The use of the latter forms was observed in 7 contexts, including Examples 329 and 330:

329. Conversely, angiotensin II can upregulate the expression of receptors for oxidized LDL 25 and could in fact **contribute to** oxidation of LDL. (Griendling and FitzGerald 2003)[163]

330. Moreover, calcification itself might be **associated with** an increased risk for subsequent breast cancer development. (Shaaban et al. 2002)

A similar phenomenon observed in French was the use of conditional verb forms to indicate hedging, as in Examples 331 to 334. In particular, pattern markers or parts of pattern markers themselves fairly often occurred in conditional form in the French data, as in Examples 333 and 334. As in Example 331, this method of hedging may occur in conjunction with other indicators of uncertainty (e.g., *susceptible de*).

331. Cette oxydation serait susceptible d'**entraîner** l'altération de diverses structures nerveuses. (La Recherche 1997)

332. En conclusion, BRCA1 et BRCA2 pourraient **participer** activement **à** la prolifération et à la différenciation induite par les œstrogènes... (Pujol et al. 2004)

333. Dans le cas des tumeurs, l'expression de p8 **faciliterait** la transcription de gènes indispensables à la progression tumorale. (Vasseur and Iovanna 2003)

334. La NADPH oxydase **jouerait** donc **un rôle** majeur lors des premières étapes du processus athéromateux (oxydation des LDL, adhésion monocytaire, accumulation de cellules spumeuses). (Bonnefont-Rousselot et al. 2002)

The variability in the means used to express uncertainty in these examples clearly illustrates the difficulties that can confront automatic applications attempting to deal with hedging at a formal level in context identification and/or evaluation.

Some of the simpler lexical means of hedging (e.g., adjectives, adverbs) may be fairly easily represented in pattern forms, particularly as they often appear in relatively regular structures associated with specific types of markers (e.g., adjectival hedges preceding nominal markers, as in *possible association between* X *and* Y, or adverbial

---

[163] The contrast between the levels of uncertainty conveyed by the forms *can* and *could* is particularly evident in this example, as the two forms co-occur.

hedges preceding verbal or participial adjective markers, as in X *is generally associated with* Y or X *slightly increases* Y). However, the variability of the form and location of the more complex items (particularly propositions) makes representing these types of hedges particularly challenging. Combinations of multiple hedges or hedges and other expressions of uncertainty are of course also extremely challenging to account for in pattern forms.

When the proportions of occurrences of various categories of hedges are compared (Table 117), it is clear that only a weak positive correlation is observed: the elements used to indicate hedging — particularly verbs and non-lexical means (e.g., verb tenses) — vary in their prevalence in the two data sets.

Table 117. Comparison of occurrences of different types of expressions used for hedging in English and French

	EN	FR
Adjectives	12	7
Adverbs	25	9
Conjunctions	7	3
Nouns	4	4
Verbs	41	8
Propositions	26	12
Non-lexical means	11	16
Total[164]	126	59

An evaluation of the individual categories using the Chi-square test (for those that presented a sufficient number of occurrences for evaluation), indicated very similar distributions in most categories as a proportion of the total occurrences of hedging, although in English the proportions of adjectives and nouns were very slightly lower than in French, and of adverbs was somewhat higher. English did show a significantly higher proportion of verbs ($p = 0.003$). However, the most striking difference comes

[164] Although most comparisons in this thesis are based on the proportions of contexts containing each phenomenon, this total reflects the total number of occurrences in order to more accurately reflect the proportions of each type of hedge.

from the proportions of contexts containing hedging using non-lexical means, which was significantly higher in French ($p = 0.002$). Of course, the primary difference is the use of conditional verb forms in French, observed in 15 contexts (13% of the contexts containing hedging). In English the closest approximation was the use of modal verbs in past tense (observed in 7 contexts, or 7% of the contexts containing hedging).

This variation indicates that strategies for dealing with the phenomenon of hedging in the two languages would likely benefit from targeting different types of expressions. Moreover, approaches used to deal with the different types of hedges (particularly those involving lexical or non-lexical means) may be quite different in themselves. The use of conditional verb forms, for example, offers some interesting possibilities for sorting contexts, particularly when it is the marker itself that occurs in conditional form. This could involve implementing an analysis that links the conditional form of a marker observed in KRC detection with a given level of uncertainty. This kind of analysis would be considerably different from those likely to be useful in the English, which should focus more on the identification and evaluation of elements external to the pattern forms observed, such as verbs, adjectives or adverbs linked to the markers.

While the proportions of occurrences of various types of hedges in the two data sets differed, there were nevertheless some general resemblances in many of the simpler forms of hedging, and on a conceptual level, similarities in the types of hedges observed. Thus, some potential for adapting strategies for use in both languages may be observed. Some common principles guiding approaches to processing this phenomenon, and certainly the clarification and analysis of the underlying phenomena, could be of significant benefit to users in both languages.

Further research, in addition to analyzing more occurrences of each phenomenon in order to better evaluate the structures that may be observed and how they may be implemented in pattern-based tools, could investigate the potential for observing variations in the types of hedges used in connection with specific markers, marker types (e.g., POS classes), relations, or classes of terms denoting the concepts participating in

these relations. Such an analysis could help to target the sources of the differences observed more exactly and to evaluate whether — and if so, how — methodological choices in this research or the corpora analyzed may have contributed to them.

4.10.2.3 Modal verbs

As noted above in Section 2.4.2.3, another type of expression of uncertainty that frequently interrupts pattern forms involves the use of modal verbs, as in Examples 335 to 338:

> 335. Accumulating data indicate that dysregulation of NF-[kappa]B may **contribute to** the pathogenesis of some breast cancers. (Garg et al. 2003)
>
> 336. The inflammatory marker C-reactive protein (CRP) can **indicate** low-grade chronic inflammation, which can **identify** patients at risk for atherosclerotic complications. (MacKenzie 2004)
>
> 337. .... l'interaction avec l'ERE concerné peut **conduire à** l'activation de la transcription d'un sous-groupe déterminé de gènes... (Kirkiacharian 2000)
>
> 338. Les graisses alimentaires peuvent aussi **nuire à** la coagulation et à la fibrinolyse plasmatiques, indépendamment de leurs effets sur la cholestérolémie. (Blais 2001a)

By explicitly characterizing the relations expressed in these contexts as possible, these verbs restrict the certainty of the information that can be extracted from them. As mentioned above in the discussion on hedging in Section 4.10.2.2.1.1, the uncertainty expressed by modal verbs in English is further increased when they occur in the past tense, as in Examples 339 and 340, and a similar effect is observed when they occur in conditional form in French, as in Examples 341 and 342.

> 339. New findings published online by the journal Science reveal the crystal structure of the BRCA2 protein and demonstrate how mutations in the gene could **contribute to** tumor growth. (Graham 2002)
>
> 340. Moreover, calcification itself might be **associated with** an increased risk for subsequent breast cancer development. (Shaaban et al. 2002)

341. Ainsi la mutation du gène BRCA1 pourrait **empêcher** la réparation de gènes et **induire** une prolifération anarchique des cellules… (La Recherche 1997)

342. BRCA1 et BRCA2 pourraient **participer** activement **à** la prolifération et à la différenciation induite par les œstrogènes… (Pujol et al. 2004)

For at least some applications, valid and important information can be obtained from contexts including indicators of uncertainty such as modal verbs. However, locating this information faces certain challenges at a formal level, as pattern forms would need to be adapted to account for variations introduced by the presence of these verbs in pattern structures. The restricted list of verbs and forms observed and the relative regularity of the structures in which they occurred (i.e., always preceding a verb, often either a verbal relation marker or a copula verb used in combination with participial adjective, adjective or nominal markers) offer some possibilities for formal representation of the phenomenon.[165] Identifying specific structures in which modal verbs apply to the relation expressed in a context is particularly important, not only because of the effect that the presence of these elements may have on the form of the verb they precede, but also because the mere presence of these elements within an extracted context does not necessarily affect the validity of the relation expressed in this context (e.g., if modal verbs apply to other elements appearing within a pattern structure).

At a conceptual level, for example in applications that attempt to identify levels of certainty indicated by these items and sort or otherwise process contexts accordingly, problems may result from the polysemy of these verbs, and the subtle shades of meaning they can express. The modal verb *can*, for example, may indicate ability, possibility or permission (Swan 1995: 104–109); *may* may indicate possibility,

---

[165] As in the case of hedges, the presence of modal verbs within pattern structures introduces morphological variation of markers from expected forms (e.g., as in X *causes* Y and X *may cause* Y). However, this variation as well is quite regular and could likely be represented in a relatively straightforward and standard way in pattern forms.

permission, requests, suggestions and criticisms (Swan 1995: 322–328). Moreover, a given modal verb may indicate varying degrees of certainty in different types of contexts (Swan 1995: 335) (a phenomenon observed in the corpora in the past and conditional forms identified). The sorting of contexts according to precise senses or levels of certainty associated with the individual verbs would thus be a complex task, especially given that even individual forms of verbs may be associated with different levels of certainty.[166] Nevertheless, if the situation allows for such human intervention, it should be possible to distinguish occurrences containing these items from others automatically, and to present them to a user for precise evaluation.

Modal verbs were found in 12% of the relation occurrences and 31% of those containing expressions of uncertainty in English, and 6% of relation occurrences and 22% of the occurrences with expressions of uncertainty in French. These data are shown in Table 118 and Table 119.

Table 118. Comparison of the proportions of relation occurrences involving modal verbs (MV) in English and French

	EN	FR	Total
MV+	55	20	75
MV-	387	329	716
Total	442	349	791

Table 119. Comparison of the proportions of relation occurrences with expressions of uncertainty involving modal verbs (MV) in English and French

	EN	FR	Total
MV+	55	20	75
MV-	120	70	190
Total	175	90	265

[166] However, as mentioned in Section 2.4.2.3, the fact that restricted numbers of forms and/or senses of modal verbs are likely to be observed in specialized languages has been noted, e.g., by Sager et al. (1980).

As observed in the case of hedging, the difference between the data sets in the proportions of relation occurrences containing expressions of uncertainty in which these expressions took the form of modal verbs was not statistically significant (p = 0.201), although the proportion in English was slightly higher. However, the proportion of the total relation occurrences observed including modal verbs was very significantly higher in the English data (p = 0.001), suggesting that this means of expressing uncertainty is more common overall in English. These results indicate that investing time and effort in developing pattern forms that take this potential variation into account, and/or strategies for identifying and processing these contexts automatically is likely to provide a significantly higher return in English.

When a more detailed analysis of the modal verbs and forms identified is carried out, in order to further explore the possibilities for developing automatic strategies for dealing with such cases, one challenge may be found in the fact that a wider variety of distinct modal verbs was found in the English data (Table 120), while in the French data all of the occurrences were of the verb *pouvoir* (Table 121).

Table 120. English modal verbs observed

Modal verb	Occurrences
may	35
can	11
might	6
could	3
will	1
Total	56

Table 121. French modal verbs observed

Modal verb	Occurrences
peuvent	8
peut	7
pourraient	3
pourrait	2
Total	20

This suggests that in English, reaping the benefits of developing pattern forms and processing strategies for dealing with occurrences of modal verbs would nevertheless require a somewhat more substantial investment of time than in French. If levels of certainty are to be assigned to each of the verbs (and potentially to specific forms) to assist in automatically classifying contexts, this complexity will be even more significantly increased. Once again, with the prevalence of a phenomenon in a given language, the complexity of representing it formally also increases.

Given the difference in prevalence of various phenomena analyzed above, it would be possible to envision the development and use of different approaches to identifying uncertainty in candidate KRCs in the two languages. While in English the identification of certain types of lexical hedges and of the use of modal verbs could be productive, in French, pattern forms could be adapted to take into account other indications of uncertainty, such as the use of conditional verb forms (considered in this analysis as a type of hedging using non-lexical means). However, these phenomena are likely to differ substantially in both their impact on KRC identification (e.g., in interfering with the recognition of candidate KRCs) and pattern design (e.g., in the need to adapt pattern forms to modified structures and/or to allow for interruptions), as well as the strategies required to resolve these issues. Thus these factors are likely to be best considered separately in application and pattern set development, although advances made in addressing the challenges they pose may ultimately complement one another in the two languages.

4.10.2.4 Negation

Perhaps the strongest indicator of "unreliability" of information present in candidate KRCs is the presence of negation. In fact, the use of the term *uncertainty* in this case is perhaps strictly inaccurate, as negation often does not express any doubt at all about a statement. Rather, it may explicitly and categorically deny that statement. Nevertheless, following authors such as Barrière (2002), this phenomenon can be considered to be

closely related to those described in this category and thus to be best addressed in this framework.

Negation within a context describing a relation (either of the relation itself or of some element of it) can certainly call into question the validity of the information contained in that context for further use, and thus is an important factor to consider in many pattern-based applications. As observed above in Section 2.4.2.4, some authors have chosen to disregard all contexts containing negation in semi-automatic applications, in order to set aside those that might be misleading if used. However, as will be discussed below, the rejection of all contexts containing negation would result in the reduction of recall. Moreover, for applications such as definition formulation and domain knowledge acquisition, contexts containing negated statements about relations may still provide useful information, and thus may be of interest to users. However, the status of these contexts is necessarily different than that of contexts without negation.

Negation was observed in a number of the contexts containing potentially pertinent relation occurrences in the two corpora, as in Examples 343 to 348. Considerable parallels in the types of phenomena were observed in the two corpora.

343. Lower levels of plasma folate and vitamin B6, however, were not **associated with increased risk of** breast cancer in an early prospective nested case-control study with 195 case-control pairs. (Zhang 2004)

344. ... the findings suggest that dietary intake of fat and fiber do not **play a** major **role** in the development of breast cancer. (Fackelmann 1992)

345. Oral but not transdermal HRT **induced** APC resistance measured by the alteration of the effect of APC on thrombin generation. (Seed and Knopp 2004)

346. Dans ce travail [12], l'exercice physique n'**a** pas **eu d'effet sur** le cholestérol total ou le LDL cholestérol. (Ferrières 2004)

347. La grossesse n'**a** que peu ou pas d'**effet sur** le risque de récidive de cancer du sein. (Debourdeau et al. 2004)

348. Les AINS sont **capables d'activer** la transcription de leur propre enzyme cible, notamment Cox2 (mais pas Cox1) via l'activation de PPARγ. (Guastalla et al. 2004)

The semantic implications of negation can be complex. As observed in Examples 349 to 354, at a general level the value of a given context for extracting information may or may not be affected by negation:

349. The cytotoxic **effects of** SC236 and docetaxel were not **affected by** HER-2/neu expression. (Witters et al. 2003)

350. This is in stark contrast with the properties of fibroblasts where the ectopic expression of cyclin D1 shortens the G1 phase but is not **sufficient to trigger** S-phase entry. (Sicinski and Weinberg 1997)

351. ... the findings suggest that dietary intake of fat and fiber do not **play a** major **role in** the development of breast cancer. (Fackelmann 1992)

352. Could it be that BRCA1 and BRCA2 **play roles in** the development of hereditary cancers but not sporadic tumors? (Yang and Lippman 1999)

353. The mammographic density does not **increase with** tibolone, unlike **with** HRT. (Kocjan and Prelevic 2003)

354. Augmentation de la survie globale et sans récidive par la suppression ovarienne (**induite** ou non **par** chimiothérapie) et la prescription de tamoxifène (Debourdeau et al. 2004)

In the case of Example 349, the expression of the MODIFICATION relation itself is negated, but the addition of a condition in Example 350 may indicate that the relation is potentially — but not certainly — present (i.e., expression of this cyclin may help to trigger entry into this phase although it is not sufficient to do so). In Example 351, the negation applies not to the marker of the relation or to the relation itself, but rather to an intensifier of the relation, *major*. While the relation — and thus the context — remains potentially valid, its expression is hedged by this combination of an intensifier and negation. (Cf. Section 4.10.2.2.) In Example 352, the negation applies to only one of the pairs of elements indicated; while it is not possible to draw a conclusion about the validity of the relation, due to the question structure, it is suggested that a relation may hold between the genes and hereditary cancers, but that no such relation links sporadic

tumours with these genes. A similar phenomenon is observed in Examples 345 and 348 above, as well as Example 353, in which more than one pair of elements is present, and the relation of one pair is affected by the negation in *but not*, *unlike* or *mais pas* while the other remains unaffected. In Example 354, a similar situation is also observed, as the relation may or may not hold (as indicated by *ou non*). (Cf. also Section 4.9.1.2 on disjunction of related elements.)

Negation may also be combined with other expressions of uncertainty (e.g., *X may not Y*, *X does not always Y*) or may occur in contexts placing conditions on a given statement, and as a result may vary in its impact on the validity of a relation expressed. Examples 355 to 360 below illustrate some problematic contexts including negation:

355. Although Ras is not often mutated **in** breast cancer... (Wang et al. 2003)

356.... found no significant **associations between** sequential HRT **and** breast cancer **risk**... (Weiss et al. 2002)

357. Unlike combination HRT, therapy with estrogen alone did not appear to have any **effect** (either favorable or adverse) **on** heart disease... (Aschenbrenner 2004)

358. Chemotherapy containing platinum ... might not **increase** survival... (Carrick et al. 2004)

359. Toutefois, certaines de ces mutations n'**affectent** pas la **capacité d'**APC d'**induire** la dégradation de la caténine ß ... (Blanchard 2003)

360. Notons toutefois que les mutations ne sont pas le seul phénomène **empêchant** la protéine de jouer son rôle. (Chène 1999)

It is thus clear that in automated applications, the use of negation to either sort or eliminate contexts according to the certainty or uncertainty of the information they contain will face numerous difficulties. The value of information expressed in contexts containing negation is likely to vary according to the application envisaged for the ultimate use of this information, and at least some applications may need to adapt pattern forms to accommodate this kind of variation.

This raises some interesting questions about the usefulness of individual contexts and the advantages and disadvantages of automatic filtering (i.e., exclusion) or sorting of contexts containing negation. If all relation occurrences containing negation are rejected, valid occurrences would almost inevitably be lost.[167] A more conservative strategy, and one that would be particularly pertinent in applications involving human interpretation of candidate KRCs, would involve the sorting of contexts to present users with non-negated occurrences first and to indicate that negated occurrences require interpretation. However, given the varying scope and types of negation observed, properly processing such contexts to retain only those relations that are not negated or developing strategies to sort contexts containing negation automatically would require a detailed analysis of the many forms in which negation can occur, based on considerably more data. (Fortunately, the regularity of some of these structures observed could provide a starting point for strategies for taking on the task in certain cases.)

Negation was observed in 5% of relation occurrences and 12% of those containing some kind of expression of uncertainty in English, and 4% and 17% respectively in French. These data are shown in Table 122 and Table 123. Neither of the differences observed between the two data sets is statistically significant ($p = 0.761$ and $p = 0.294$ respectively), although negation accounts for a slightly higher proportion of the expressions of uncertainty observed in the French data.

Table 122. Comparison of the proportions of relation occurrences containing negation (NG) in English and French

	EN	FR	Total
NG+	21	15	36
NG-	421	334	755
Total	442	349	791

[167] Cf. Bowden et al.'s (1996) description of negative triggers.

Table 123. Comparison of the proportions of relation occurrences containing expressions of uncertainty that involved negation in English and French

	EN	FR	Total
NG+	21	15	36
NG-	154	75	229
Total	175	90	265

The effects of negation on requirements for pattern forms and/or developing strategies for evaluating occurrences involving negation thus appear likely to be of comparable importance in English and French. Moreover, given the quantitative and qualitative parallels observed in the corpora, the representation of the more complex forms of negation may also pose significant challenges in the two languages.

Given the relatively infrequent occurrences of the phenomenon and its variable but potentially significant impact on the value of the information contained in the relation occurrences in which it is found, it is important to consider the investment of both time and effort required to process occurrences of negation at a formal level. One factor in this evaluation is the form in which negation is observed.

In both languages, negation was generally indicated by an independent marker, although some cases of negation using affixes were observed, as in Examples 361 and 362. When this is the case, this affixation — like any other modification of a marker or related element — may make it difficult to identify for applications in which negated relations are considered pertinent.

361. (In)activation of aromatic amine carcinogens **is catalysed by** metabolic enzymes including N-acetyltransferase 1 (NAT1)... (Van der Hel et al. 2003)

362. Les protéines mutées sont incapables **de provoquer** l'**élimination des** cellules ayant, par exemple, un ADN endommagé par les UV. (Chène 1999)

A significant difference for approaches attempting to identify standard forms of negation that may be allowed for in patterns is the wider variety of indicators of negation observed in French. A total of 8 different indicators (Table 125) were observed

in 15 occurrences of negation in French (for a ratio of 1.9 occurrences per indicator), compared to 5 indicators (Table 124) in 22 occurrences in English (4.4 occurrences per indicator). As would be expected, the indicators of negation in the French data were also more often complex in form (e.g., *ne... pas*, *ne... jamais*, *mais pas*) than in the English, with 5 of 8 forms observed to be complex in French and 1 of 5 in English (*but not*).

Table 124. English indicators of negation observed

Indicator of negation	Occurrences
not	13
but not	3
no	2
in-	2
un-	2
Total	22

Table 125. French markers of negation observed

Indicator of negation	Occurrences
ne... pas	6
sans	2
non	2
ne... aucun	1
ne... que	1
in-	1
mais pas	1
ne... plus	1
Total	15

These data suggest that more distinct indicators of negation may need to be taken into account in French when developing applications that are capable of identifying and/or classifying contexts containing this phenomenon. Moreover, while the representation of structures involving negation is of course a challenge in the two languages, it may be even more complicated in French, given that the indicators of negation tend to be complex.

More complex or variable structures were also noted, as in Examples 363 to 364:

363.... n'a mis en évidence aucun **bénéfice** attribuable à l'HTR pour **contrer** l'athérosclérose, ni aucune **réduction** de la mortalité entre les groupes traités par rapport au groupe témoin. (Bouchard 2001)

364.... lorsque l'association de metformine et d'une sulfonylurée ne **permet** pas une **maîtrise** optimale **du** diabète ou ne peut être utilisée **en raison d'**une contre-indication ou de l'intolérance à l'un de ces médicaments... (Leblond 2001)

The complexity of the structures in Example 363 as well as the absence of one part of a complex indicator of negation (*pas*, *plus*, etc.) in Example 364 could pose challenges for pattern design and for applications in automatic identification and analysis of negated relation occurrences. The additional semantic complexities of the variations between the different markers of negation (e.g., *ne... pas*, *ne...plus*, *ne... que*) in French may also be pertinent for some more semantically rich processing tasks.

The identification of complex forms poses slightly more significant challenges for automatic applications attempting to identify negation for the purposes of either eliminating or classifying contexts; this is particularly true in less regular structures involving a change in order of the elements of an indicator of negation, the separation of these elements by large numbers of words, or cases in which potentially complex markers are incomplete (e.g., Example 364 above). The complexity of the task of recognizing and/or representing negation formally may therefore be somewhat higher in French.

Future research on more data could help to clarify the ways that negation can affect the value of candidate KRCs at a conceptual level, as well as the possibilities for representing this phenomenon in pattern forms in order to process contexts containing negation appropriately for a given application. It nevertheless appears from this research that in both languages the complexity of the task of properly and precisely processing such occurrences automatically could be all but prohibitive, given the semantic and formal variability observed. While some simpler structures that could potentially be exploited were observed, a significant investment of time and effort would be required

to meet many of the challenges observed. From the evaluation of the indicators of negation observed, it appears that even the more basic strategies for identifying and processing negation could be more complex to develop in French.

4.10.3 Text-related issues

As described in Section 3.3.1.5.3, particularities of certain texts may interfere with the recognition of KRCs. These may include minor problems such as spelling or punctuation errors that may interfere with the analysis of a context (and therefore the recognition of a pattern form) or with the recognition of a marker. In addition, some problems may interfere with the interpretation of the information contained in a context (e.g., the evaluation of whether a relation is present or what kind of relation is present, what elements are involved in this relation, and how). Examples 365 to 369 illustrate some of the phenomena observed.

365. In animal models of diabetes, antioxidant defense capacity is **diminished is** [sic] certain tissues. (Griendling and FitzGerald 2003)

366. The presence of TNF- [alpha], IL-6, and other cytokines **cause** hepatic **production of** C-reactive protein (CRP)... (Pantaleo and Zonszein 2003)

367. .... examination of coronary arteries showed an **interaction between** social environment and social status **on** the development of atherosclerosis. (Schwartz 2003)

368. Ainsi, les effets procoagulants des cellules endothéliales **est augmenté** lorsque celles-ci sont infectées par du virus herpes simplex [39] ou par du cytomégalovirus [40]. (Lizard et Gambert 2001)

369. ...le haut taux élevé de croissance cellulaire **observé au niveau des** tumeurs surexprimant HER2... (Cornez and Piccart 2002)

Among the relation occurrences observed, a very comparable proportion — 6% in English and 7% in French — were observed to contain some kind of text-related issue ($p = 0.454$) (Table 126). A very slightly higher proportion of such cases was found in French.

Table 126. Comparison of the proportions of relation occurrences containing text-related issues (TR) in English and French

	EN	FR	Total
TR+	27	26	53
TR-	415	323	738
Total	442	349	791

The variability of this phenomenon is such that the classification and quantification of the problems is very complex, and as such is beyond the scope of this project. However, clearly tools may not locate some potential KRCs in both languages due to these problems.

4.10.4 Difficulties overall

As discussed above in Section 3.3.1.5.4, a number of pattern characteristics and external difficulties may affect pattern-based tool performance and the use of extracted candidate KRCs. It is thus useful to consider the total proportions of relation occurrences in which one or more of these phenomena (related elements in non-nominal form, anaphora, pattern interruptions, expressions of uncertainty, and text-related issues) were observed, to gauge the proportions of contexts that diverge from prototypical, easily interpreted pattern forms.

If the sum of all of the contexts containing at least one of these phenomena is considered, the results are very striking: in English, 333 of the 442 relation occurrences identified — 75% — fall into this category, and in French, this figure is 252 or 72%.[168] This indicates that the proportion of relation occurrences not admissible in the most conservative approaches that exclude variants of restrictive, prototypical forms of relation occurrences is very high, and that — particularly in cases in which available

[168] This figure excludes the interruption of complex markers by related elements alone, which — as noted in Section 3.3.1.5.4 — is not generally considered as a difficulty as such, although it does add to the complexity of developing pattern forms. If the cases including this phenomenon are included, the figure rises to 364 occurrences (81%) in English and 260 (74%) in French, showing a significantly higher prevalence of the phenomenon in English ($p = 0.006$).

data are limited and/or high recall is desired — such approaches may lead to an unacceptable level of silences in the results of KRC extraction. In these cases, it would often be advisable to use a semi-automatic technique that relies more on human interpretation and/or human or automatic sorting of contexts rather than their exclusion or elimination from results based on the presence of these kinds of phenomena.

Moreover, the absence of a significant difference between the two data sets ($p = 0.319$), with only a slightly higher proportion of occurrences in English, does not indicate that either language will be significantly more or less vulnerable to such difficulties.

This similarity also underlines an important observation that can be made in the results of the analysis carried out in this research, relating to the interaction of the multiple factors analyzed in this project. Although significant interlinguistic differences were observed in a number of the factors included in this measurement (e.g., the higher prevalence of non-nominal related element forms, of interruptions of related elements, and of certain types of anaphora in French; the higher prevalence of certain types of expressions of uncertainty in English), overall similarities may camouflage underlying differences. Comparable results may in fact be obtained in the two languages by applications operating with similar restrictions, but these data suggest that in the pursuit of improvements in performance (e.g., the development of strategies for improving recall) in each language, tool developers would do well to focus on different difficulties, in order to address the most significant challenges in each language.

Of course, both the impact of addressing these difficulties and the requirements for doing so vary substantially. Certainly the potential for and necessity of addressing each one in a given context will vary depending on the needs of users and the application envisaged for a pattern-based tool. In addition, the results of this analysis reveal that both languages can certainly benefit from developments in any of these aspects: only very rare phenomena were observed frequently in one language but not in the other.

The development of effective strategies in one language — in which a given phenomenon is more frequent and in which therefore more data are available and the strategies are more likely to be profitable — may therefore open doors for the subsequent adaptation of strategies to a new language. While certainly not all approaches will be directly transposable from one language to another, enough similarities were noted in this research to suggest that comparing and contrasting the languages in light of strategies developed to deal with particular phenomena may be excellent starting points for improving performance overall.

The general significance of the observed similarities and differences will be discussed further in Chapter 5.

Université de Montréal

Lexical Knowledge Patterns for Semi-automatic Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Study of English and French

par

Elizabeth Marshman

Département de linguistique et de traduction

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)

en traduction

option terminologie

avril 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Lexical Knowledge Patterns for Semi-automatic Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Study of English and French

présentée par :

Elizabeth Marshman

a été évaluée par un jury compose des personnes suivantes :

Patrick Drouin, Président-rapporteur

Marie-Claude L'Homme, Directrice de recherche

Sylvie Vandaele, Co-directrice

Nathan Ménard, Membre du jury

Lynne Bowker, Examinatrice externe

André Ferron, Représentant du doyen de la Faculté des études supérieures

5 Discussion

5.1 Introduction

In this Chapter, the results of the comparisons described above in Chapter 4, and in particular the similarities and differences highlighted, will be discussed in light of various aspects of pattern-based tool development, performance and use.

The discussion will begin with the description of the effect these similarities and differences will have on the design of pattern-based tools, pattern forms and pattern lists for semi-automatic tools (Section 5.2), which will be followed by a discussion of the factors that may affect tool performance (Section 5.3), and finally a discussion of the effect on the ultimate usefulness of extracted KRCs (or other information) for terminological research and other applications (Section 5.4). Some additional observations made and challenges encountered in the course of this research will be discussed in Section 5.5, and a brief discussion of semi-automatic and automatic approaches to knowledge extraction in terminology work will be presented in Section 5.6. The chapter will conclude with an analysis of the limits of this research in Section 5.7.

In this project, as stated in the Introduction, the basic approach envisioned was that of the semi-automatic extraction of KRCs for terminological research, including domain knowledge acquisition. This choice was made because this application can benefit from information of a wide variety of natures and forms, and thus the information pertinent for such an application includes that useful for other, more specifically designed applications. Such an inclusive description can then be analyzed from the perspective of information that is pertinent in more restrictive applications, while also allowing the cost of such restrictions (e.g., in lost contexts or information) to be evaluated. Less restricted patterns observed in such studies may later be refined in order to adapt them to other applications, while work that begins with restricted patterns does not provide information about types of relation occurrences that are not retained by

these pattern forms because of occurrences of variations on these forms or other difficulties, and thus are harder to adapt for use in other, less restricted applications. Therefore, the discussion of the observations of this research includes an overview of aspects of the contexts observed that are pertinent for basic approaches such as semi-automatic KRC extraction using simple marker forms, accompanied by a discussion of how many of the phenomena may affect possibilities for developing more sophisticated tools that impose restrictions on pattern forms and contexts retained for various purposes.

Each phenomenon will be discussed here in terms of its individual effect on performance in the contexts in which it is pertinent, and some observations of general trends likely to affect the various aspects of pattern-based tool development and use will be made. While clearly the convergence and interaction of factors in tool development and performance will be critical in determining the ultimate effectiveness of a pattern-based approach in the two languages, a discussion of the individual factors can help to highlight some of the similarities and differences that should be taken into account in designing tools for specific goals, and the ways that these differences may come into play in the performance of pattern-based tools.

These may inspire further research on particular factors or combinations of factors as they apply to specific projects It is with this perspective that the observations in the research will be discussed below: as indications of a need to further evaluate and examine some of the factors evaluated in this research, with specific applications in mind, in light of hypotheses that may be drawn from the results of this study.

The phenomena evaluated, and their pertinence for the three aspects of pattern-based knowledge extraction identified above, are summarized in Table 127.

Table 127. Summary of factors analyzed and interlinguistic comparisons

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Number of relation occurrences observed	4.1	Tool performance	All	Yes $p < 0.001$	Fewer relation occurrences in French data, particularly of the ASSOCIATION relation Phenomenon could be linked at least in part to the terms used to generate concordances
Number of different markers observed	4.2	Pattern design	All	Indications	A wider variety of markers apparent in French data
Number of occurrences of markers	4.4	Pattern design Tool performance	All	Indications	Indications that the English markers observed are in most cases more frequent than the French and that relation occurrences are more concentrated among the more frequent markers in English
Types of pattern markers observed	4.5		All		
Part of speech classes of markers	4.5.1		Lexico-syntactic		

[169] Potential types of pattern forms specified are: *Character strings/Regular expressions*, in which the marker is represented using either of these means; *Lexico-syntactic*, in which the part of speech class of the marker and/or of surrounding elements is specified, with the option of also specifying potential distances between separate elements in the pattern form; *Related element structures*, in which the part of speech class of the related elements is specified or automatic identification of related elements attempted using this structure; *Specific related elements*, in which a term, other lexical unit, or class thereof is coupled with a pattern form. *All* denotes that all of these types of forms may be affected.

[170] A p value of 0.05 or less was considered to be statistically significant. The possibility of a trend towards significance is considered to exist when the p value for a given difference was higher than 0.05 but less than 0.1. *Indications* identifies cases in which statistical tests are not considered to be strictly reliable but in which a potential for variation was suggested.

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Individual markers	4.5.1.1	Pattern design	Lexico-syntactic	Trend $p = 0.066$	Although difference is non-significant, slightly more verbs observed in English and adjectives and adverbs in French for the two relations In the ASSOCIATION relation, a trend towards higher prevalence of function words observed in French as well as somewhat more adjectives and adverbs, and more verbs observed in English
Marker occur-rences	4.5.1.2	Pattern design Tool performance	Lexico-syntactic	Yes $p = 0.001$ Yes $p = 0.001$ Yes $p = 0.011$	Overall, proportions of markers in POS classes significantly different Adjectives more frequent in French data than in English Nouns more frequent in English data than in French Similar trends observed in the two relations separately
Simple and complex markers	4.5.2	Pattern design Tool performance	All	No	Very similar proportions of complex and simple markers observed in the two data sets

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Marker precision	4.6	Pattern design Tool performance	All	Indications	More valid occurrences retrieved with French markers in the sample evaluated More categorial ambiguity observed with English markers in the sample evaluated If categorial ambiguity is excluded from consideration, more valid occurrences retrieved with English markers in the sample evaluated Differences also observed between markers of different POS classes and for different relations, potentially indicating effects on performance from interaction with other factors
Polysemy of pattern markers	4.7	Pattern design Tool performance Information evaluation and use	All	Not statistically evaluated	Similar cases of ambiguity noted in the two data sets
Pattern variation	4.8		All		
Variations in marker form	4.8.1	Pattern design	All	Indications	Some suggestions that variability of the English markers is higher than of the French markers, particularly for the CAUSE–EFFECT relation

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Variation in voice of verbal markers	4.8.1.1	Pattern design Information evaluation and use	All	Yes $p = 0.002$	Passive voice significantly more commonly observed for English markers
Variation in pattern structures	4.8.2	Pattern design Tool performance	Lexico-syntactic Character strings/ Regular expressions	No	Inconsistent results observed; no conclusions can be drawn
Variations in pattern structure involving relative pronouns	4.8.2.1	Pattern design Information evaluation and use	All	No	Very similar proportions of relation occurrences involving these structures observed in the two data sets More variety in relative pronouns observed in the French data
Number and form of the elements linked by the markers	4.9		All		
Multiple elements sharing a role in a relation	4.9.1	Pattern design Tool performance Information evaluation and use	All	No	No significant difference in proportion of relation occurrences involving multiple elements sharing a role in a relation, although prevalence somewhat higher in the French data
Variant expressions of a single related element	4.9.1.1		All		

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Abbreviations and symbols	4.9.1.1.1	Pattern design Tool performance Information evaluation and use	All	No	Very similar proportions of relation occurrences containing abbreviations pr symbols in the two data sets.
Other variants in expression of a related element	4.9.1.1.2	Pattern design Tool performance Information evaluation and use	All	No	Difference observed is not significant Although small numbers of occurrences make comparisons difficult, slightly more occurrences were observed in the French data
Conjunction and disjunction	4.9.1.2	Pattern design Information evaluation and use	All	No	Proportions of relation occurrences containing conjunction and disjunction of related elements somewhat but non-significantly higher in French In both data sets and types of relations between the elements one indicator of the relationship accounted for a large proportion of occurrences
GENERIC–SPECIFIC relations between elements	4.9.1.3	Pattern design Tool performance Information evaluation and use	All	No	Proportion of relation occurrences involving the phenomenon only slightly higher in French More variety noted in the indicators of the relation in the French data, while fewer indicators in English accounted for a higher proportion of occurrences observed

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed? [170]	Details of difference
Ellipsis of part of complex related elements	4.9.1.4	Pattern design Tool performance Information evaluation and use	All	Trend $p = 0.066$	Ellipsis of the head of a complex related element somewhat more prevalent in the French data, trending towards significance Non-significant differences observed for ellipsis overall and for ellipsis of expansions
Repetition of marker or part of marker	4.9.1.5	Pattern design Tool performance	Lexico-syntactic Related element structures	Yes $p < 0.001$	Phenomenon significantly more prevalent in the French data
Form of elements linked by markers	4.9.2	Pattern design Tool performance Information evaluation and use	All	Yes $p < 0.001$ Yes $p = 0.033$ Yes $p = 0.049$ Trend $p = 0.079$	Non-nominal elements significantly more frequent in French data Propositional and verbal elements significantly more frequent in French data Pronominal elements significantly more frequent in French data Higher prevalence of adjectival elements in French data trends towards significance

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Anaphora	4.9.2.1	Pattern design Information evaluation and use	All	Trend $p = 0.058$ Yes $p = 0.033$ Yes $p = 0.021$	Higher prevalence of anaphoric expressions in French data trends strongly towards significance Significantly higher proportion of anaphoric elements in the form of possessive adjectives in the French data Significantly higher proportion of anaphoric expressions replacing all or the head of a related element Some variation in the potential of pronouns occurring as anaphoric reference both in terms of variety (higher in English) and the possibilities for locating antecedents (more promising in French)
Challenges in using knowledge patterns and extracted contexts	4.10		All		
Pattern interruptions	4.10.1	Pattern design Tool performance	All	Yes $p = 0.015$	Significantly higher prevalence observed in English data
Interruptions of patterns	4.10.1.1	Pattern design Tool performance	Lexico-syntactic Related element structures	No	Non-significant difference with somewhat higher prevalence observed in the English data.

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Multiple markers and interruptions of patterns by other patterns	4.10.1.1.1	Pattern design Tool performance Information evaluation and use	All	No	Very similar proportions of relation occurrences involving this phenomenon observed in the two data sets Some regularities also observed in structures in the corpora
Interruptions of complex markers	4.10.1.2	Pattern design Tool performance	All	Yes $p = 0.032$ Yes $p < 0.001$	Significantly higher proportion of English relation occurrences involve interrupted complex markers Difference observed largely due to significantly higher proportion of marker interruptions by related elements in English
Interruptions of related elements	4.10.1.3	Pattern design Tool performance	Related element structures, Specific related elements	Yes $p = 0.021$	Significantly higher proportion of related elements interrupted in the French data
Expressions of uncertainty	4.10.2	Pattern design Information evaluation and use	All	Yes $p < 0.001$	Phenomena significantly more prevalent in English data
Quantification of related elements	4.10.2.1	Pattern design Information evaluation and use	All	No	Similar proportions of relation occurrences involving this phenomenon observed in the two data sets, with prevalence in the English data only slightly higher More variety in quantifiers observed in the English data

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Hedging	4.10.2.2	Pattern design Information evaluation and use	All	Yes $p = 0.002$ Yes $p = 0.003$ Yes $p = 0.002$	Phenomenon observed significantly more frequently in English data Verbal hedges significantly more prevalent in the English data Non-lexical means of hedging significantly more prevalent in the French data
Modal verbs	4.10.2.3	Pattern design Tool performance Information evaluation and use	All	Yes $p = 0.001$	Phenomenon observed significantly more frequently in the English data More variety observed in the modal verbs and forms in the English data
Negation	4.10.2.4	Pattern design Information evaluation and use	All	No	Very similar proportions of relation occurrences involving this phenomenon observed in the two data sets More variation noted in indicators of negation in the French data
Text-related issues	4.10.3	Tool performance Information evaluation and use	All	No	Similar proportions of relation occurrences involving this phenomenon observed in the two data sets, with proportion only slightly higher in the French data

Factor analyzed	Section	Stage of pattern-based extraction affected	Primary types of patterns affected[169]	Significant difference observed?[170]	Details of difference
Difficulties overall	4.10.4	Pattern design Tool performance Information evaluation and use	All	No	Relatively similar proportions of relation occurrences involving this phenomenon observed in the two data sets, with a slightly higher proportion of occurrences in the English data

5.2 Tool and pattern design[171]

In the process of tool and pattern design, a number of decisions must be made that will affect the kinds of potential KRCs that are identified and retained for use, and the means used to identify these contexts.

Tool design should take into account the purpose for which a given tool will be used, the needs of users, and the ways these users will participate in the evaluation of the information identified. These factors will influence the balance of precision and recall that is desired, as well as the approaches needed to achieve this balance.

However, the possibilities of pattern-based approaches depend on the ways relations are expressed in each language, and the decisions that are made may affect the languages to different degrees depending on language-specific factors. Achieving comparable performance may depend on the ability to recognize and manipulate these factors. Adjusting expectations of the complexity of the task of designing and developing bilingual pattern-based tools and of expected performance in the process is also essential, and will rely on observations of pertinent phenomena such as those discussed in this research.

---

[171] In this discussion, the various factors evaluated will be indicated in bold, in order to facilitate consultation.

5.2.1 Factors affecting approaches to pattern discovery

This study demonstrated that a term-based approach to pattern discovery was effective in corpora both languages, and that this kind of approach allowed for the identification of a relatively wide (and more or less comparable) range of markers in both English and French.

The distribution of relation occurrences among **semantic classes** in both languages echoed Bodson's observations (2005) of associations between specific classes of terms and relations. Specifically, the terms denoting processes were found to be particularly productive for locating CAUSE–EFFECT relations, and those denoting pathologies were particularly effective for identifying ASSOCIATION relations. Including a fairly high proportion of terms representing these classes in term-based pattern discovery approaches is likely to allow for the identification of a range of markers in both English and French. However, specific associations between terms belonging to these classes and the markers that occur with them could limit the range of markers observed if such a choice were made.

Observations of the term pairs that were **equivalents** and those that were not suggested that the equivalents could be most effective for identifying comparable numbers of relations in the two languages, although the variation observed between individual term pairs was higher than that between the term classes as a whole. The data do not allow for this possibility to be fully evaluated, however, and more research would be essential to determine the contribution that this and other factors may have made to the observations. If further evaluation supports this possibility, bilingual approaches that use relatively large sets of equivalent terms belonging to particular classes may be the most promising avenues for term-based pattern identification. An alternative technique, less used in the field to date (the exception being Barrière 2001, 2002), would involve the manual analysis of a small corpus in its entirety, or of a sample of randomly selected contexts from a corpus, in order to observe potentially useful pattern markers.

Additional refinement of pattern discovery approaches may take into account the types of markers that are likely to be observed in each language; pattern discovery strategies may involve targeting specific **part of speech classes** — such as verbs, as in the case of the research carried out by Garcia (1997) and Feliu (2004), cf. also the observations of Barrière (2001) — in marker identification. The observations in this project identified verbs and participial adjectives as very prevalent classes of markers — particularly of the CAUSE–EFFECT relation — in the two corpora, but also revealed that other classes such as nouns are both numerous and productive.

Some interlinguistic differences, however, may affect the choices made in specific cases. **Adjectival markers** were observed to be both more numerous and more productive in the French data, and therefore may be stronger candidates for pattern discovery in this language. From another perspective, their exclusion from pattern discovery approaches would affect this language more than English, reducing the numbers of markers that may be observed and thus the potential for recall in pattern-based tools in which the markers located are used. Conversely, the choice to limit pattern discovery to the observation of **verbal** markers would provide access to a slightly wider proportion of the markers observed in the English data, suggesting a greater potential for identifying KRCs than in French. In addition, the proportions of **nominal** marker occurrences indicating the ASSOCIATION relation was substantially higher in the English data, suggesting a somewhat greater need to consider these types of markers for the relation in that language.

The observations in this study also revealed variation between the relations in the numbers and types of markers observed. In addition to the need to evaluate the numbers of markers necessary or advisable in pattern sets for each relation, the differences also highlight the importance of adapting pattern discovery approaches in the two languages to retrieve the kinds of markers most commonly used to indicate specific relations.

5.2.2 Factors affecting the number and choice of markers

A number of the factors evaluated in this research may affect the number of markers that are used in pattern sets for any type of pattern-based tool. These factors include the **numbers of markers** that are observed to indicate a given relation, as well as the **distribution** of relation occurrences among these markers and the **numbers of occurrences** of markers, which determines the number of potentially useful contexts that can be retrieved and thus the overall productivity of a pattern set.

The ratio of the **number of markers** observed relative to the numbers of relation occurrences was almost universally higher in the French data (and particularly so for the ASSOCIATION relation), although statistical confirmation of the significance of this difference was not possible. This consistency nevertheless suggests that further research should be undertaken to investigate this potential difference with more data. The presence of more distinct markers for the relations suggests that a wider variety of markers may be necessary in order to retrieve the same number of candidate KRCs in French.

This observation was also supported by the **distribution** of the relation occurrences among the markers observed. The occurrences of the two relations together were more concentrated among the most frequently observed markers (and therefore those that are of particular interest for use in pattern sets) in the English data than in the French, suggesting that a pattern set consisting of a limited number of these promising markers would be more productive (i.e., locate a higher proportion of relation occurrences) in English, and that in order to achieve the same results in French, more markers would be required. This difference was nevertheless much smaller for the most frequent markers of the ASSOCIATION relation independently; the largest contribution to the difference was observed in the markers of CAUSE–EFFECT relations.

Once again, although the difference is subtle, in the English data the marker sets observed tended overall to be somewhat more frequent per 1,000 corpus tokens than the

French markers, in both relations together and individually. The frequency of the markers of the ASSOCIATION relation was significantly higher in both corpora, but the trend towards higher frequency was also larger in this case.

Overall, although no formal statistical confirmation of this trend was possible in the data gathered in the study, the factors that affect the numbers of markers likely to be required in pattern sets and the potential for productivity of pattern sets showed quite consistent tendencies suggesting that more French markers may required to retrieve the same numbers of candidate KRCs in the two languages. The effect of these tendencies is moreover likely to be larger than it at first appears, as the identification of markers is only the first step in a labour-intensive process of evaluating marker performance and designing and refining pattern forms.

Once the numbers of markers required are evaluated, the choice of the types of markers for inclusion in pattern sets is the next step in pattern set development.

As was the case in pattern discovery approaches, decisions involving the choice of markers for inclusion in pattern sets may be affected by the **types** (e.g., **part of speech class) of markers** that are commonly used to express a given relation in a language. Differences on this level primarily affect the representation of markers in lexico-syntactic pattern forms, but may also be pertinent in applications that use character strings. The prevalence of adjectival markers for the CAUSE–EFFECT relation in the French data and of nominal markers in English for the ASSOCIATION relation suggests that these types of markers are good candidates for inclusion in pattern sets.

Such differences may also be associated with variations in **marker precision**, which may also affect the choice of marker types for inclusion in pattern sets (cf. Section 5.3.2). For example, in the analysis of the precision of a small sample of nominal and verbal markers in both languages, it appeared that in both corpora contexts containing nominal markers were considerably more likely to be **incomplete** (i.e., not to contain an explicit indication of one or more of the elements linked in a potentially

pertinent relation) than those containing verbal markers, therefore indicating that nominal markers may produce a larger amount of noise in the results of extraction (although such noise may nevertheless be useful at some level). This observation parallels that of Barrière (2001) in her analysis of English CAUSE–EFFECT markers.

These observations suggest that the process of marker selection should take into account interlinguistic differences in marker part of speech, and that language-specific processes of marker discovery may be necessary in order to discover the most productive markers in a given language. Marker sets of comparable distribution among POS classes may not retrieve comparable proportions of relation occurrences in corpora in English and French. However, the potential for variations in performance linked to the characteristics of markers retained may also affect the precision of marker sets, requiring an analysis of the interaction of these often conflicting factors in pattern set design.

5.2.3 Factors affecting the design of pattern forms

Once promising markers have been selected, strategies must be developed for representing these markers in a form that tools can apply for KRC extraction. This step involves choosing how precise the pattern forms should be in their description of markers and the contexts in which they occur. Simpler forms such as character-string representations of markers principally confront difficulties in the representation of marker forms, but may allow more noise in results of extraction and provide little basis for further automatic processing of contexts. Conversely, more complex pattern forms that specify the context in which markers may appear in addition to the form and characteristics of that marker (e.g., lexico-syntactic knowledge patterns) face challenges related not only to the representation of marker forms but also to the identification and analysis of pattern structures — and in particular markers' relationships with elements participating in a relation — and to the representation of the related elements. These three factors will be discussed below.

5.2.3.1 Factors affecting the representation of markers

Among the factors that must be taken into account in representing pattern forms are whether markers are **simple or complex**, how **variable** they are likely to be, and what forms this variation may take (e.g., **interruptions of markers, variation in marker form**). These affect the representation of markers in any pattern-based tool.

No significant differences were observed in the proportions of **simple and complex marker occurrences** in the two data sets, with the distribution between the two categories approximately equal in the CAUSE–EFFECT relation and with more complex markers the case of the ASSOCIATION relation. The two languages thus seems likely to present similar challenges, while the representation of markers of ASSOCIATION is likely to be more complex than that of CAUSE–EFFECT relation markers due to the greater potential for variation and the need to account for this in representing these markers.

The representation of complex markers is complicated by the potential for their interruption by external elements, requiring careful representation of these markers to ensure that pertinent occurrences are identified (i.e., that an excessive number of potentially useful contexts is not excluded by unduly restrictive forms, but that the levels of noise are also not too high due to excessively permissive forms). **Interruptions of complex markers** were significantly higher in the English data, indicating a higher level of complexity in representing markers in this language. However, the difficulty associated with this task and the strategies that are most appropriate for dealing with the phenomenon are affected by the specific types of interruptions observed.

Interlinguistic differences were observed primarily in the **interruption of markers by one of the elements that they link** (e.g., *association of* X *with* Y, *correlation of* X *with* Y, *effect of* X *on* Y, *role of* X *in* Y). This constitutes the most regular and predictable form of marker interruption, and one that is most likely to be dealt with systematically in the design of pattern forms, either as character strings or in

more restricted pattern forms. These markers are largely nominal, suggesting that the effect of **part of speech class variation** in markers (e.g., a higher prevalence of nouns in the English data) may be linked to this phenomenon.

In **other types of interruption of complex markers**, the situation is considerably different, as the phenomenon is slightly — but not significantly — more prevalent in the French data. These interruptions are considerably more irregular in both their occurrence and their form (i.e., part of speech class and even length) than those described above, and as such pose more challenges in designing pattern forms if these occurrences are to be found. However, the frequency with which specific markers are interrupted (e.g., the commonly observed modification of markers such as *role of... in* or *rôle de... dans*) and the relatively regular form of these interruptions may provide a basis for dealing with a certain proportion of this phenomenon.

Another concern in the representation of markers is that of **variation in marker form** (e.g., the potential for the presence or absence of elements in addition to a "base" marker or the change in the order in which elements of a complex marker appear in a text, as well as variations associated with a change in the voice of verbal markers). The levels of variation were fairly significant in the two languages, indicating that multiple marker forms could be required for a number of markers in pattern-based tools (or at least those that attempt to use as complete a marker form as possible, as in the strategy adopted in this research). Although the differences observed were not very large and precise statistical confirmation was not possible, the level of this kind of variation was observed to be slightly higher in the English data, particularly for the CAUSE–EFFECT relation. This indicates a potential need for more pattern forms or more flexible pattern forms in this language.

The combination of these factors suggests that the representation of markers may be somewhat more complex in English due to higher variability in marker forms overall, and particularly in the CAUSE–EFFECT relation.

5.2.3.2 Factors affecting the representation of pattern structures

The choice of types of pattern structures for use in a pattern-based tool should take into account the possibilities and challenges of representing commonly observed KRC structures and their elements, as well as the ways in which the structures observed may reflect the kind of information that is pertinent and meet user needs in the situation in which a tool is to be used. More specific forms such as lexico-syntactic patterns that represent the structures in which markers occur can allow a tool to target contexts that correspond to specific forms identified as the best candidates for identifying relation occurrences, and can therefore reduce noise. This kind of description can also provide a first step towards analysis of these contexts to extract and/or evaluate the information they contain. However, these restrictive forms are also far more labour-intensive to develop and are vulnerable to problems resulting from variation in such structures in texts. Simpler forms are of course less likely to be affected by these phenomena and are less problematic to develop, but conversely are generally less precise.

The choice to use more or less specific and restrictive pattern forms should involve the consideration of the number of potentially useful contexts that such pattern forms will allow a tool to identify, the complexity of the task of designing these pattern forms, and the impact that higher or lower recall and/or precision will have on the effectiveness of a tool for a particular use. Evaluating these factors involves estimating the frequency of phenomena affecting them in each language, including **variations in pattern structure** and **interruptions** of these structures.

In both languages, the fairly high level of **pattern structure variation** observed suggests that for tools using specific pattern forms, multiple pattern forms per marker are likely to be required in order to find all pertinent relation occurrences. The observed inter-corpus variability in the structure of patterns indicated by a given marker was nevertheless very uneven, with no consistent trend towards a higher level in either data set observable. More data collected using a more appropriate methodology would be

necessary to properly evaluate this phenomenon, but these data do not provide any evidence that could support a hypothesis regarding potential interlinguistic differences.

A specific contributor to pattern variation in both languages is the presence of **structures involving relative pronouns**, observed in similar proportions of the two data sets. This phenomenon would require adaptation of pattern forms for contexts to be located, although regularities in the types of structures identified for different markers show promise for developing strategies involving labour-saving, standardized adaptations of pattern forms. However, the different numbers of relative pronouns observed in the two corpora (with more variety observed in the French data), as well as some differences in the nature of the pronouns and the information they convey, indicate that representation of this phenomenon may be slightly more complex in French.

A significant variation in the form of verbal markers was observed in the appearance of these markers in the **passive** as well as active voice; this phenomenon affects not only the form of the marker itself but also the structure in which it participates (generally including inversion of the order of relation participants in the case of markers of asymmetric relations such as CAUSE–EFFECT), often requiring adjustment in or addition of pattern forms. The much higher prevalence of this phenomenon in the English data indicates a probable need for more pattern forms in this language to deal with the phenomenon, particularly for tools that attempt to identify the participants in a relation and to assign specific roles to these.

The observations of interlinguistic variations in the number of markers required for retrieving a given number of contexts and the number of marker and pattern forms required for each marker suggests an interaction of these factors that is important to take into account when evaluating possibilities for developing pattern-based tools. The fact that more markers may be required in French and that a relatively high proportion of pattern structures to markers was observed in both corpora indicates a greater impact of the difference than might otherwise be expected. However, as the challenges involved in the languages are associated with different factors that may affect distinct tool types, the

impact on individual projects will likely vary with the choices made in designing the tool and choosing the approach. It is essential to understand the differences that may be observed in order to determine what factors are likely to be pertinent and should be evaluated in planning a specific project.

Even if restrictive pattern forms are developed to represent potential variations in pattern structures, these forms encounter another very problematic phenomenon in use: **interruption of pattern structures by external elements**. The interruption of pattern forms was observed in 40 to 45% of the relation occurrences observed in this research, indicating that restrictive pattern forms that do not allow for this phenomenon will not permit the identification of a large proportion of potentially useful contexts. This is particularly important to take into consideration when developing a pattern-based approach, both in the choice of the kinds of patterns for use and in the development of pattern structures if more restricted forms are used.

While the difference observed was not significant, the proportion of interrupted occurrences was somewhat (but not significantly) higher in the English data, suggesting that in this language it may be even more important to adapt pattern-based tools to deal with the phenomenon, or a higher proportion of potentially useful contexts may be lost. Further study could confirm whether this difference becomes significant in light of more data.

A specific type of interruption observed involved the co-occurrence of **multiple relation markers in a single context** and particularly cases in which these markers linked the same pair of elements. This often involves a significant variation in pattern form from cases in which a single marker is found, because of the interruption of a pattern structure by this additional marker. This phenomenon was observed in a fairly substantial — and similar — proportion of relation occurrences in the two data sets, indicating that this phenomenon is important to take into account if specific pattern forms are used. Some recurring structures observed in the results in both corpora indicate that the formal representation of at least some cases of this phenomenon for use

in pattern forms is a possibility, although more data would be necessary to evaluate both this possibility and the potential differences between the languages.

Given these data, it appears that some slight trends towards increased challenges in English may be present, but that the major contributions are likely to come from specific sources such as the presence of passive forms. Firm conclusions about other contributions to increased difficulties in English cannot really be drawn.

5.2.3.3 Factors affecting the representation of related elements

Related elements of course form part of the structure of knowledge patterns, and in more restrictive pattern forms, the form of related elements may be specified not only in order to target contexts that are likely to express relations of interest, but also as a precursor to the automatic identification of these elements.

Pattern forms that include a representation of related elements should ideally be adapted to reflect ways in which related elements may be expressed in texts. This involves taking into account the forms (e.g., **part of speech classes**) individual related elements may take, as well as variations in the structures in which they occur (e.g., **interruptions**, **multiple elements sharing a role in a relation**).

Tools that target relations between specific types of items (e.g., terms) that are assumed to take a particular form (e.g., nouns or noun phrases) often use pattern forms that specify these part of speech classes. However, it became clear in this research that this kind of approach would exclude a small proportion of potentially useful relation occurrences, and moreover that the proportions of **non-nominal related elements** observed were somewhat different in the two data sets.[172] The proportion of non-nominal elements observed was significantly higher in the French data, suggesting that

---

[172] As discussed in Section 4.9.2, footnote 147, in this research at least one candidate term in nominal form was required to be linked to the relation marker for a relation occurrence to be retained for analysis, potentially reducing the frequency with which this phenomenon was observed as compared to its actual prevalence.

this phenomenon could lead to the exclusion of more potentially useful contexts in this language. The proportions of a number of individual types of non-nominal elements (i.e., adjectives, pronouns, propositions and verbs) were also higher in the French data. This indicates that the choice to specify the form of related elements in patterns, and the potential effect on performance in French, should be carefully considered in light of the interlinguistic differences observed. One possibility for dealing with this phenomenon involves adapting specific pattern forms containing markers frequently observed with non-nominal related elements (e.g., *risk factor*, *marker*, *facteur de risque*, *complication de*) to allow for this phenomenon. Regularities observed in the two data sets suggest that such an approach could permit some of these contexts to be retained.

A specific case of variation in the forms of related elements may involve the presence of **anaphoric expressions** replacing all or part of a related element. The higher prevalence of this phenomenon in the French data indicates a particularly strong potential for observing challenges related to anaphora in this language. Adaptations required may include developing pattern forms that admit the occurrence of non-nominal elements such as pronouns, or fully representing structures such as those involving possessive adjectives or combinations of demonstrative adjectives and generic nouns in order to clearly identify cases in which anaphora take these forms and to allow for further processing of these cases if desired for a given application.

The potential for observing **multiple elements sharing a role in a relation** is also important for such patterns to accommodate in order to completely and accurately extract (and possibly subsequently identify) of all related elements. The somewhat higher prevalence of the phenomenon (as well as of a number of the sub-types evaluated) observed in this language suggests that such measures may be slightly more important in French.

One means of dealing with this phenomenon is to create formal representations of the relationships and structures in which multiple elements may participate. Regularities and interlinguistic similarities in the structures observed for some of the

sub-types of the phenomenon suggest possibilities for using parallel approaches in the two languages, although some of these may be more straightforward to develop than others. However, some interlinguistic differences were noted in the cases of the indicators of some relationships between related elements. While in the case of conjunction and disjunction the occurrences observed showed a relatively comparable prevalence of the prototypical indicators *and, et, or* and *ou*, in the case of GENERIC-SPECIFIC relations between multiple related elements, the most frequent indicators of the relation in English accounted for a very high proportion of the occurrences, while in the French data the distribution was much more even. This suggests a need to include more French indicators linking multiple related elements if contexts containing these, or the elements themselves, are to be identified and possibly further analyzed to identify the specific information present.[173]

The challenges of dealing with this kind of phenomenon are increased by a related one, the **ellipsis of part of one or more complex related elements** that share a role in a relation. Pattern forms must take this phenomenon into account to identify contexts containing relations and to offer possibilities for further processing of these contexts, for example in identifying related elements automatically. The variability of the structures in which this phenomenon is observed (e.g., the **ellipsis of a head of a complex item** in some cases and **of an expansion** in others, as well as ellipsis occurring within even more complex structures that correspond to related elements) poses many challenges for formal representation and pattern design. Although similar phenomena were observed in relatively comparable proportions in the two corpora, the French results indicated a somewhat higher prevalence of the omission of a part of complex related elements, and a significantly higher prevalence of the ellipsis of the head of such

---

[173] It is very interesting to consider here the unequal distribution of occurrences between the markers of GENERIC–SPECIFIC relations between related elements in the two languages, as it parallels that observed in the CAUSE–EFFECT relation in this research. While this is a small sample, it nevertheless identifies this phenomenon as one that is potentially observable in other relations, highlighting its potential impact and suggesting a need for further evaluation.

elements, indicating the importance of considering the phenomenon in this language. Interlinguistic differences in the typical structures that may be observed (e.g., the part of speech classes of elliptical related element forms observed in proximity to markers) also indicate that considerable interlinguistic adjustments may be necessary in order to ensure that pattern forms reflect usage in the two languages. Evaluation of a larger sample of data would be necessary to accurately describe the structures observed and their prevalence.

Finally, tools that search for specific terms or candidate terms in connection with markers may encounter difficulties linked to most of the factors discussed above. Certainly the prevalence of **non-nominal** variants of the more usual nominal term forms, the presence of variants involving **anaphoric expressions** or of **elliptical forms of complex terms** may interfere with KRC recognition using standard representations of these terms, and alternate means may be necessary to locate all occurrences of relations involving a concept denoted by a term.

When these — admittedly often closely related — factors are considered together, it becomes clear that a higher proportion of the relation occurrences in the French data involved challenges in this phase of pattern development, suggesting that the task may be more time-consuming and difficult in this language.

Thus, when the factors of marker types and representation, pattern structure development and the representation of related elements are considered together, some significant differences between the two data sets suggest that the development of pattern-based tools is likely to involve different needs and different challenges in the two languages. Some of these differences (e.g., in the case of marker part of speech categories) do not necessarily pose more challenges in one language or another, but would need to be taken into account in developing tools and considered in their potential implications when linked to other factors. The major challenges in French appear to be

linked more to the choice and number of markers required and the representation of the elements that they link. Those identified in English, while more subtle, may involve variations in the form of markers and/or of pattern structures from "standard" forms. The choices of approaches for use in different languages and the planning of tool development projects may take these factors into account in order to organize work and address critical issues that may interfere with tool performance in each language.

5.3 Pattern-based tool performance

Pattern-based tool performance may be evaluated on several levels, including the potential for recall offered by markers, the precision of these markers, and the potential for recognition of KRCs, identification of related elements, and sorting and further processing of contexts retrieved.

5.3.1 Factors affecting potential for recall

The potential of any lexical pattern-based tool is first determined by the density of relation occurrences associated with lexical markers in the corpora it is used to process, and then by the proportion of these that are associated with markers included in its pattern sets.

Although the first results observed in this research suggested that a significant difference in the **density** of such relation occurrences was lower in the French corpus analyzed, further analysis suggested that this difference could be attributable to the choice of terms used. In the analysis of relations involving pairs of equivalent terms, the proportions of occurrences observed in the two data sets were very comparable, with the English showing only a slightly higher return for the two relations together and the CAUSE–EFFECT relation alone, and the French showing a slightly higher proportion of occurrences of the ASSOCIATION relation. This supports the assertion that in general, despite some minor variations, a pattern-based approach can be equally productive and useful in the two languages for these relations, and encourages the continuation of

further research into the development of bi- and multilingual tools. More specifically, it also indicates a need to evaluate the various factors that may have contributed to the difference observed in the term pairs, in order to determine with more precision the source of the variation and to identify the best strategies for selecting terms (or even alternate methodologies) for use in similar projects in the future.

The potential for identifying relation occurrences then depends on the markers observed and their distribution in the corpus. As discussed above in Section 5.2.2, some indications of slightly higher **numbers of occurrences** and lower **variety** of the markers observed in English were observed, suggesting that pattern sets in the two languages containing comparable numbers of markers (e.g., a set of the most frequent markers observed in a pattern discovery project) could provide access to fewer potentially useful contexts in French. A difference can then be predicted in the baseline potential for performance (i.e., recall) of similar pattern sets in the two languages (although of course this is only the first of many factors influencing the ultimate effectiveness of tools).

5.3.2 Factors affecting precision

The analysis of the **precision** of a small sample of markers in both languages allowed for the evaluation of a certain number of criteria that may affect the precision of pattern-based tools. (As discussed above in Section 4.6, however, the size of the sample precludes wide generalizations and rather suggests some possibilities for further research.)

In the analysis of a set of 10 markers that represented a comparable distribution between relations and part of speech classes in the two data sets, the French markers tested retrieved a significantly higher proportion of **valid contexts** than their English counterparts. The French markers evaluated also were observed to express **complex relationships** more frequently than the English markers tested. Conversely, the English markers tested were observed to present very significantly higher levels of **categorial ambiguity**, which may very plausibly contribute to the difference in precision observed.

ambiguity, which may very plausibly contribute to the difference in precision observed. This likely also explains the fact that significantly more **noise** was also observed in the contexts extracted using the French markers than in those found using the English markers.

Significant differences were thus observed in the kinds of challenges confronted by the various markers and their counterparts in the other language. Individual markers showed differing levels of vulnerability to challenges such as the prevalence of noise or categorial ambiguities. These results underline the potential for significant variability from marker to marker and the need to consider the impact these variations may have on the overall performance of pattern-based tools, as well as to base further evaluation on a wider range of markers in order to provide a more comprehensive view of possible trends.

One possibility for dealing with categorial ambiguity in the results of pattern-based extraction involves the use of more sophisticated approaches using lexico-syntactic pattern forms in part-of-speech tagged corpora. The potential impact of this kind of technique on the precision of markers as evaluated in this study can be illustrated by a comparison of the results of the precision evaluation of ten markers in each language as described above, but with the cases of categorial ambiguity (as identified in human evaluation) eliminated.[174] In this case, it was the English markers evaluated that were observed to be somewhat more precise than their French counterparts (although only a trend towards significance was noted). Moreover, an evaluation of two marker pairs in which one member was particularly vulnerable to one of these difficulties, in contexts extracted using the tool Syntex from part-of-speech tagged and lemmatized versions of the corpora, indicated that this approach allowed for a large proportion of these challenges to be overcome and for more comparable results

---

[174] Clearly, this measure can only be seen as indicative, as human evaluation (as in this work) and automatic part-of-speech tagging cannot be expected to produce identical results, and automatic tagging is far more likely to be used in the context of semi-automatic knowledge extraction.

to be obtained for the pair of markers. (However, it is of course also essential to consider the additional challenges confronted in these more sophisticated approaches, for example in terms of the complexity of representing marker and pattern forms. Moreover, some categorial ambiguities, while distinguished from valid hits in this context, of course may also indicate occurrences of the desired relation.)

When the **precision** data were evaluated in light of the **part of speech class** of the markers analyzed (i.e., nouns and verbs), a striking difference in the proportion of incomplete contexts was observed. In both languages, the proportion of incomplete contexts containing nominal markers was much higher than that of contexts containing verbal markers, suggesting that nominal markers are likely to produce more unusable contexts in the results of KRC extraction. When the proportions of individual markers in each part of speech category for both relations are compared in the two data sets, the proportions of nouns were similar in the two languages, indicating the likelihood of relatively comparable effects of this factor. However, a somewhat higher proportion of verbs in the English data and function words in the French may indicate that if pattern sets reflect the observed distribution, relatively precise verbal markers could contribute more to the results in English, while function words (which may be presumed to be less precise, particularly in light of Barrière's (2001) results and in the observations for the marker *et*) may contribute more in French. Given the more pronounced differences observed in the case of the individual relations, the likelihood of differences seems greater. The complex interactions of individual marker part of speech class, the proportions of occurrences associated with the classes of markers and the precision of markers and marker classes merit further large-scale and in-depth analysis in light of the differences observed.

In terms of the **polysemy of markers**, similar (and small) numbers of markers were observed to present ambiguities, in both the analysis of the numbers of markers that were identified in the sample as being associated with more than one of the relations or sub-relations retained in this research, and in the analysis of the number of markers

studied in the analysis of the data on marker precision that presented both a core causal relation and another more complex relation with a causal component.[175]

From these observations, it appears that the choice of markers (in terms of their part of speech classes and individual vulnerabilities to difficulties) and their representation in pattern forms is likely to have a significant effect on the overall precision of a tool. Differences were certainly observed that merit further investigation in light of more data, particularly to determine how various factors interact.

5.3.3 Factors affecting KRC recognition

The performance of any pattern-based tool will of course be affected by variations in marker form (e.g., interruptions of complex markers, many types of marker variation) that are not accounted for in the process of pattern design. As discussed in Section 5.2.3.1, the higher prevalence of **variation in marker form** (suggested by evaluations of variation in marker form overall and clearly indicated in the specific case of variation in the voice of verbal markers) and of **interruption of complex markers** in the English data suggests a possibility of encountering more such difficulties in this language. (Although as discussed above, the relatively regular interruptions of many English markers by one of the elements linked to them may be possible — if not always easy — to account for in the design of pattern forms in many cases, while the less regular interruptions of French markers in fact may in fact pose more difficulties at this level.)

In tools that use restrictive pattern forms that represent the structures in which markers occur, **interruption of these structures** that is not accounted for in pattern design will of course also interfere with KRC identification and extraction. As discussed in Section 5.2.3.2, this kind of interruption was observed to be somewhat more common in the English data, although this difference was not statistically significant and would

---

[175] Because of the limited numbers of polysemous markers identified, it is not possible to draw general conclusions about the prevalence of the phenomenon.

require more evaluation for certainty. No consistent trend towards a higher level of other kinds of **variation in pattern structures** in either corpus was observed.

Tools that use patterns that specify the form of the elements linked by a pattern marker are affected by the various factors discussed in Section 5.2.3.3, and the failure of pattern forms to account for these phenomena will interfere with the recognition of potentially useful contexts. The higher prevalence of phenomena such as **non-nominal related elements**, **multiple elements sharing a role in a relation** and **ellipsis of part of a related element**, as well as greater **variety of indicators** of some types of relations between these multiple elements in the French data signals a potential for observing more difficulties in this language in the performance of such tools if these phenomena are not acceptably accounted for. Pattern forms that do not take this phenomenon into account may be unable to (accurately) identify contexts containing relations, particularly if a non-standard form of a related element (e.g., the expansion in a complex item of which the head has been omitted) occurs closest to the marker identified.

For pattern-based tools that focus on the retrieval of relation occurrences in which previously identified (candidate) terms are found, variation in the form in which these elements are observed can interfere with the identification of potentially pertinent KRCs. In some cases, **non-nominal forms** may be used in connection with a marker (e.g., *inflammatory marker* rather than *marker of inflammation*, or *coronary arteries that are becoming blocked due to...* rather than *coronary artery occlusion due to...*), reducing recall for these tools. The higher prevalence of non-nominal related elements in the French data suggests that this language may be more vulnerable to the phenomenon.

Cases of **anaphora** in which a related element is replaced can also interfere with the identification of KRCs by these tools; once again this phenomenon was observed to be significantly more prevalent — and thus likely to be more problematic — in French. Of course, many occurrences of non-nominal related elements involved anaphoric elements such as pronouns and possessive adjectives; however, even the proportions of

anaphoric elements in nominal form (involving the use of a generic term) were higher in the French data.

The **interruption of related elements** can also contribute to reducing the recall of such tools by changing the form in which these elements are observed; the higher prevalence of the phenomenon in the French data is likely to result in a greater impact in this language.

Finally, the occurrence of **multiple elements sharing a role in a relation** may also cause significant difficulties for these types of tools. Pattern forms that require contiguity between the marker and specific (candidate) term to be identified may not permit many potentially pertinent occurrences to be identified if another related element (and often a marker of the relation between the two elements) occurs between the two. This phenomenon was observed to be slightly more prevalent in the French data than in the English, indicating a greater potential impact in the former language. Particular difficulties in these cases can result from the **ellipsis of the head or expansion of a complex element** that is frequently observed. This phenomenon was observed in a fairly similar proportion of cases in the two data sets, but was somewhat more frequent in the French data. Overall, then, these results suggest that tools that search for relation occurrences involving a specific (candidate) term will be more likely to confront difficulties linked to a number of separate factors in French. Moreover, different strategies would be required to deal with these diverse phenomena, making it difficult to reduce the difference between the languages by adjusting a specific parameter. It appears that it will be necessary to accept that such tools are likely to show differences in performance (specifically in recall) in the two languages.

Interlinguistic differences in the likelihood of encountering difficulties with KRC identification were thus observed for various types of pattern-based tools. However, it is interesting to observe how the expected impact of these differences varies from one type of tool and/or aspect of the KRC-identification task to another. This illustrates the degree to which the approach adopted — which is likely to be a function of user needs

and the situation of use — determines how that tool is likely to perform in the two languages.

In these observations, it appears that English tools may face more challenges in the design of pattern structures in general, while in French the representation of related elements is likely to be particularly problematic. The overall effect of these factors may equalize interlinguistic differences to some level in some contexts. However, the source of difficulties determine the strategies that will be effective for minimizing them and the ways in which they may be taken into account in the process of tool development.

5.3.4 Factors affecting the identification of related elements

Tools that attempt to automate the identification of related elements depend largely on the representation of the forms of these items and on their position in contexts relative to the markers observed. Variations at either level may interfere with this process.

Non-nominal related elements, observed to be more prevalent in the French data, may be particularly difficult for these kinds of tools to identify. The interlinguistic differences observed may contribute to making this type of task more difficult in French.

Patterns that specify the forms in which related elements occur may also confront difficulties linked to the **interruption of related elements** (e.g., by abbreviations, anaphoric expressions, other related elements and the indicators of the relationships between elements). The irregularity of the occurrence of this phenomenon makes it difficult to take into account in the design of these forms, and thus is likely to result in difficulties. While the phenomenon was not excessively frequent, its significantly higher prevalence in the French data suggests that it will have a greater impact on the identification of related elements in this language.

The occurrence of **multiple related elements sharing a role in a relation** also poses challenges for the identification of these elements, particularly in cases in which

elliptical forms are present. The slightly higher prevalence of the phenomenon in general and of occurrences of elliptical forms in particular in the French data suggests a possibility that tools in this language may encounter more difficulties related to this type of variation from standard pattern forms. However, the evaluation of the forms in which this phenomenon was observed, complemented by additional data, may suggest some strategies for dealing with these cases automatically.

One factor in French that may facilitate the analysis of such cases is the **repetition of markers or part of complex markers** (or even of simple markers in a few cases) before each of two or more related elements. These repetitions (e.g., of prepositions such as *à* or *de*) before each element may serve as cues for the identification both of the presence of multiple related elements and of these elements themselves if pattern forms are adjusted to take this phenomenon into account. (Of course, if they are not, the phenomenon is rather likely to pose additional problems for this task (e.g., to constitute interruptions of forms and therefore interfere with (accurate) KRC or related element recognition). The additional dangers of structural ambiguity associated with the phenomenon may also complicate analysis.) The significantly higher prevalence of this repetition in the French data suggests that it is worth investigating to assist with the analysis of contexts containing multiple related elements in this language; its near-absence in the English data suggests that this kind of development would not be productive in this language.

When related elements are replaced (entirely or in part) by **anaphoric expressions**, the identification of related elements is problematic at both a formal and conceptual level. **Non-nominal anaphoric elements** may not be recognized by pattern forms or located by standard strategies for identifying related elements, while even nominal forms are not as likely as their non-anaphoric counterparts to provide complete and precise information and thus are of dubious value for identification. Forms that are interrupted by anaphoric elements are also of uncertain — or at least, lesser — value for extraction. Any anaphora will require some kind of process of resolution (either human

or automated) in order to identify the precise information conveyed in a text, often necessitating access to a context larger than a single sentence. The French results once again indicate a higher prevalence of the phenomenon (significantly higher in the case of anaphoric expressions that replace the head of a complex related element or an entire related element), suggesting that this language will be more vulnerable to problems linked to the phenomenon than English.

The **variability of pattern structures** also of course plays an important role in the difficulties of identifying related elements automatically. The unpredictable nature of the placement of these elements — particularly for example when different types of related elements (e.g., a causal agent and a causal event) co-occur within a single context — makes the identification of these elements problematic, because pattern forms that can consistently represent such cases would be very difficult to develop. Additional variations in pattern form, such as the **interruption of pattern structures**, also make the identification of related elements more complex, particularly if these interruptions introduce structural ambiguities (e.g., if they involve the insertion of an external element similar in form to the related element between a marker and the element to which it is truly linked). The somewhat higher prevalence of the latter phenomenon in the English data, while not statistically significant, suggests that this language could be more vulnerable to such difficulties. Conclusions about the contribution of **variability of pattern structures** are not possible in light of the data gathered in this research.

One factor observed in the English data, which may somewhat reduce difficulties in this task, is linked to the higher proportion of the English relation occurrences observed that involve the **appearance of one of the elements participating in a relation within a complex marker form** (e.g., *association of* X *with* Y). While this complicates the representation of the marker in the phase of pattern development, in tools that attempt to identify related elements this kind of structure may facilitate both the identification and the delimitation of at least one of the related elements. (Challenges

of course remain in cases in which the element occurring within the structure is modified or when the structure is otherwise interrupted.)

These observations suggest that tools that attempt to identify related elements automatically in the two languages are likely to be affected by problems at different levels. In French, additional challenges appear more likely to involve factors internal to the elements themselves, while the English data may suggest a slight tendency towards the observation of difficulties linked to pattern structures in this language. The types of differences involved suggest equally different strategies for dealing with this issue. The implementation of these strategies may be common to both languages (e.g., in the case of the resolution of anaphora or the design of pattern forms that can analyze structures in which multiple related elements occur) or specific to one (e.g., in the use of repeated markers to help identify multiple related elements). In addition, some strategies may be used in both languages, but are likely to be more productive or more straightforward to implement in one than the other (e.g., strategies involving the appearance of a related element within a complex marker form in English, or integration of indicators of relationships between multiple related elements into pattern forms to facilitate the identification and analysis of these cases).

5.3.5 Factors affecting processing and sorting of KRCs

Once candidate KRCs are identified, the possibilities for further processing these contexts (e.g., to refine their classification, to eliminate those that are not considered to be useful for a particular application, or to sort them in order to save a user time and effort by presenting those that are most likely to be useful first in a list of results) depend in large part on formal elements of these contexts. These may involve the classification of sub-types of relations according to markers observed (including cases in which **multiple markers** are present, or those involving **polysemous markers**), the identification of cases in which **multiple related elements** share a role in a relation (and the **relations** that hold between these elements) and the presence of **anaphora** and of

expressions of uncertainty. Of course, automatic **identification of related elements** may also be particularly useful in this context (for example, permitting the grouping of KRCs according to the elements involved to give users a coherent idea of the type and variety of information retrieved); the possibilities and challenges of this task were discussed above in Section 5.3.4.

The classification of contexts according to the sub-relation present (particularly of the CAUSE–EFFECT relation), as indicated by the marker observed, offers significant possibilities for assisting users in targeting specific types of information and in obtaining an overview of the types of relationships that may be observed in a given text collection. Barrière's (2002) classification of CAUSE–EFFECT relations used in this project was largely satisfactory for sub-categorizing the occurrences of relations observed in both the English and French corpora, although some challenges were observed (cf. Section 5.5.3.2).

Marker **polysemy** is of course a challenge for the automatic classification of contexts according to the relation and sub-relation present. In both languages, a few cases in which markers were associated with more than one of the relations or sub-relations retained for this analysis were observed, indicating that such difficulties are likely to be encountered in both languages. Moreover, as some parallels were observed in the prevalence, the types of phenomena and potential strategies for dealing with the ambiguities observed, the implications for bilingual approaches may be similar. However, more data on the phenomenon would be required in order to draw any conclusions. More detailed evaluations such as those carried out in Marshman and L'Homme (2006) and Marshman and L'Homme (2006a) could — and did — reveal additional nuances of markers' semantic content that can provide valuable data for context classification. The difficulties observed are of course closely dependent on the markers observed; however, some possible techniques for resolving these (e.g., the use of actantial structures and of the semantic classes of the actants involved in them), have been suggested in English (e.g., Marshman and L'Homme 2006) and French (e.g.,

Bodson 2005), as well as in other languages (e.g., Feliu 2004; Weilgaard 2004), and may offer strategies for this kind of processing.

The presence of **multiple markers** in a given context, particularly when these markers link the same element pair but are generally associated with different (sub-)relations, may pose challenges for the sorting of contexts. As this phenomenon was observed in similar proportions of occurrences in the English and French data — and moreover often in similar structures — some possibilities for developing automatic processing techniques to deal with these cases in both languages appear promising. Some of these possibilities are discussed in more detail below in Section 5.5.3.5.

Contexts in which **multiple elements share a role in a relation** are often particularly information-rich not only because they indicate that a relation holds between more than one pair of concepts, but also because they often indicate more than one type of relation. The identification of these contexts as particularly worthy of evaluation, or even of the different types of relationships present in these contexts, can constitute an interesting addition to KRC extraction. The slightly higher prevalence of this phenomenon in the French results observed would suggest that this kind of approach could be more productive in this language; however, the French data also showed slightly more complexity in the task of representing some types of this phenomenon, one of the precursors to developing strategies for context sorting. Some of the potential applications for this kind of information in sorting contexts are discussed in more detail in Section 5.5.3.3.

The resolution of **anaphora** is a complex task that far exceeds the scope of any analysis that can be made from the data in this project, and as such will not be discussed in detail here. However, as the presence of the phenomenon may constitute a criterion for sorting contexts extracted using pattern-based tools, it is interesting to evaluate the prevalence of the phenomenon in the two data sets. The greater prevalence in the French data, particularly in the case of related elements, suggests that a larger number of

contexts may be identified as problematic (and therefore either eliminated from results or separated from others not presenting this problem) in this language.

The possibilities for identifying contexts containing anaphora are most likely to involve the representation of the various forms these may take. Regularities in the occurrences analyzed in the two corpora show promise for developing similar strategies in both English and French. However, some differences may be observed in the nature of the items observed that can affect possibilities for further processing. Various anaphoric expressions (e.g., pronouns, possessive adjectives) may provide some information about their antecedents that can help in the process of interpreting contexts and resolving anaphora. However, the pronouns in the French data offer additional details that may aid in identifying antecedents by indicating their grammatical gender (or even real gender, in the rare case of human antecedents).

Indications of the certainty — or more importantly, uncertainty — of the information contained in a given context may also be extremely valuable to a user of a pattern-based tool. The possibilities for identifying the level of certainty present in a given context automatically depend largely on the potential for identifying **expressions of uncertainty**. The higher prevalence of the types of markers of uncertainty evaluated in this project in the English data suggests that this kind of approach could be particularly productive if implemented successfully. However, the complexity of such a task is significant, and will depend largely on the strategies available for dealing with different types of these expressions.

The use of **quantifiers of related elements** was observed to be slightly higher in the English data than in the French, but the variety of these quantifiers indicates a potential for significantly more challenging implementation in the former language, as both the recognition of quantifiers and the association of a level of certainty with each one would involve a greater investment of time and effort. The proportion of relation occurrences including **modal verbs** was significantly higher in the English data, and once again the variety of distinct items observed was higher in this language, reflecting

a situation similar to that involving quantifiers, with more possibilities for sorting contexts in this language but a greater investment required to implement sorting. The use of **negation** is very similar in both prevalence and overall form in the two data sets, indicating that the possibilities for sorting offered by this are likely to be comparable. However, the task in French could be somewhat more challenging because of the complex forms of the markers of negation observed, which could complicate the process of representing and thus recognizing and sorting contexts containing negation in this language. The most difficult expressions of uncertainty to process automatically, principally because of the variation in their form and placement relative to the components of knowledge patterns, are **hedges**. For this reason, while these items were much more prevalent in the English data, given the challenges of automatic processing the possibilities for sorting or otherwise evaluating candidate KRCs according to the presence of the phenomenon may in the two languages are quite difficult to evaluate and compare.

Some exceptions can nevertheless be noted, as the means used for hedging can have a significant effect on the possibilities for automatic processing. Some simple and recurrent markers (e.g., adjectives, adverbs, some verbs) may be identified and exploited for context sorting, although more unpredictable and complex forms (e.g., propositions, some verb phrases) may be too difficult to use for this task. The principal interlinguistic differences observed involved the prevalence of verbs in the English data, which in many cases (e.g., *suggest*, *tend*, *appear*) appear in regular structures that could be used in automatic processing, an approach that would not be likely to be as productive in French. Conversely, the prevalence of **non-lexical means of hedging** (e.g., the use of conditional verb forms) in French would offer possibilities for automatic processing using very different strategies involving the evaluation of the inflected forms of markers that may be observed. This difference suggests the possibility that human evaluation of hedging may be particularly important in English, whereas the means of expressing uncertainty in French may be more easily dealt with automatically, at least for preliminary sorting.

Thus, differences both in performance in tasks involving the identification of levels of uncertainty and in the means necessary for accomplishing these tasks are likely in the two languages, requiring adjustments in approaches implemented in English and French in order for tools to take advantage of the possibilities available — and to minimize problems — as much as possible. Nevertheless, it is clear that many of the tasks discussed above are extremely complex, and that much work remains to be done in these areas.

5.4 Use of extracted KRCs and other information

Once KRCs have been extracted and processed to the degree considered appropriate — and practical — for a given application, the usefulness of the information contained in these contexts may still vary depending on criteria including the availability of this information, its validity and the possibilities for using it in a given situation.

Anaphora in relation occurrences, discussed in various contexts above principally in terms of their form, of course are primarily important because of the effect they have on the availability of complete information in a given relation occurrence. The significantly higher prevalence of this phenomenon observed in the French data — and particularly in cases in which an entire related element or the head of a complex item was replaced — may be reflected in a higher proportion of problematic contexts in this language. The need for additional strategies for obtaining information (e.g., human evaluation, access to a larger context) is likely to be particularly important in these cases.

The immediate usefulness of the information in candidate KRCs extracted by tools may be influenced substantially by the **form of related elements**, although this impact varies significantly depending on the situation in which the information is to be used. For purposes such as domain knowledge acquisition and the writing of definitions, the form in which related items occur is often not particularly critical. However, for

applications that focus on establishing relationships automatically (e.g., between term records, or between nodes in an ontology), forms that do not correspond to those that are (or should be) used as terms or as labels for concepts may not be usable, or may require additional processing before they can be used. The challenges of processing these cases automatically are many, and in a large proportion of cases human interpretation may be necessary. The prevalence of non-nominal related elements can be a good indicator of the proportions of relation occurrences likely to fall into this category, as these are less likely than nouns to be terms (or concept labels). The higher proportion of such cases in the French data indicates a possibility for greater challenges in this language; it may be more important in French to develop and implement strategies such as those focusing on the identification of nominal bases from which adjectival forms are derived or the nominal derivatives that may be associated with verbs in order to help resolve some of these difficulties.

Challenges may also be observed in the case of contexts in which **different types of causes** (e.g., causal agents and causal events) are observed: for some applications only one of these may be pertinent, while in others they both may be considered (although the causal events are of course more likely to pose difficulties similar to those described above). The parallels in the prevalence and forms of this phenomenon in the two data sets suggest there is a similar need to deal with these cases, and that there are possibilities for developing similar strategies in the two languages.

Expressions of uncertainty of course are also primarily important because of the restrictions they indicate on the validity or reliability — and therefore usefulness — of the information in KRCs. Of course, the level of certainty required in a given situation depends heavily on the application to which this information is to be put. The prevalence of these expressions — and particularly of hedges and modal verbs — in the English data indicate that in situations in which contexts involving uncertainty are not considered usable (or are considered separately from contexts containing more certain information), a greater proportion of the occurrences in this language may be affected.

Thus, as is the case with many of the factors evaluated here, the two languages each present some particular difficulties, which may affect overall performance. The sources of these challenges often differ, however, which indicates corresponding differences in the possibilities and strategies available for dealing with them. In the relation occurrences observed in both languages, however, the subtleties involved in the evaluation of the usefulness of information and the challenges of representing the factors that can contribute to this evaluation provide strong indications of a need for considerable human intervention in the analysis of candidate KRCs retrieved by pattern-based tools.[176]

5.4.1 Synthesis

Although the initial hypothesis in this research focused on the probability of observing differences in evaluations of candidate KRCs in English and French, the most important conclusion that can be drawn from these observations comes not from differences in the data analyzed in the two languages, but from similarities. The general possibilities of using pattern-based approaches of various types, and of the challenges associated with them, show a strong general resemblance in English and French. Moreover, many strategies that could be developed to exploit these possibilities and overcome the challenges may be useful in both languages. This indicates a promising future for the development of tools that can function adequately in the two languages.

Important observations also arise from differences observed between the occurrences of ASSOCIATION and CAUSE–EFFECT relations. Despite the close conceptual links between these relations, the number of occurrences, the markers associated with them, and some of the characteristics of the patterns in which the markers participate

---

[176] These observations are of course coloured by the methodological choices made in this research and the consideration of a wide range of contexts retrieved. More restrictive approaches may minimize many of the factors that contributed to challenges observed, although of course these advantages are almost inevitably accompanied by a decrease in recall.

showed significant differences. Moreover, these differences may affect those observed between the languages. Thus it is clear that the possibilities for developing tools should be evaluated not only as a function of language but also of the relations in question. More data — particularly for the ASSOCIATION relation — should be gathered to assist in this task.

When observations of specific factors in the data in English and French are compared and contrasted, it must be kept in mind that specific characteristics may have differing implications for pattern-based tool development and use, and for the ultimate use of information retrieved. Some differences may clearly be identified as presenting particular challenges in the data analyzed in English or French, suggesting that more difficulties may be encountered in a given language. In other cases, differences are simply that: differences. While these should — in fact, we argue, must — be taken into account in the various phases of development and use of pattern-based tools in order to make informed decisions and to guide the choice of strategies, particular advantages or disadvantages in a given corpus (and by extension, potentially in one of the languages) cannot always be identified in light of the data in this research.

The trends identified in the English and French relation occurrences analyzed are summarized in Table 128. Observations in the corpora used in this research suggest that a number of aspects of the processes of pattern-based tool development and use may be more affected by challenges in one language than the other.

Table 128. Summary of interlinguistic variations observed by phases of tool development and use

Phase of tool development or use	Contributing factors	Particular challenges observed in data in:[177]
Tool and pattern design		
Tool design		
Pattern discovery	Semantic classes, Relation density, Marker POS	Both languages
Number of markers	Number of markers, Number of occurrences of markers, Marker variety	French (indications)
Choice of markers	Marker POS, Simple and complex marker forms, Marker precision	Both languages
Pattern design		
Representation of markers	Simple and complex marker forms, Variation in marker form, Interruptions of complex markers	English
Pattern structure design	Variation in pattern structures, Pattern interruptions, Multiple markers, Voice of verbal pattern markers	English
Representation of related elements	Non-nominal related elements, Interruption of related elements, Anaphora, Multiple elements sharing a role in a relation, Ellipsis of part of complex related element	French
Tool performance		
Potential for recall	Number of markers, Number of occurrences of markers, Marker variety	French (indications)
Precision	Marker precision, Marker polysemy, Marker POS	Both languages
KRC recognition: character strings and lexico-syntactic patterns	Variation in marker form, Variation in pattern structures, Interruptions of complex markers, Pattern interruptions	English
KRC recognition: representation of or search for specific related elements	Non-nominal related elements, Anaphora, Interruption of related elements, Multiple elements sharing a role in a relation, Ellipsis of part of complex related element, Variation in pattern structures, Pattern interruption, Interruption of complex markers	French
Identification of related elements	Non-nominal related elements, Interruption of related elements, Anaphora, Multiple elements sharing a role in a relation, Repetition of (part of) marker	French
Processing and sorting of contexts	Multiple markers, Marker polysemy, Multiple elements sharing a role in a relation, Anaphora, Expressions of uncertainty	Both languages

[177] When factors were not clearly identified as posing greater challenges for applications in one or the other of the corpora, their impact on both languages was noted.

Phase of tool development or use	Contributing factors	Particular challenges observed in data in:[177]
Use of extracted contexts		
Availability of information	Anaphora, Non-nominal related elements	French
Usability of information	Non-nominal related elements, Conjunction of different types of causes	French
Reliability of information	Expressions of uncertainty	English

The potential for recall observed in the French markers appears to be slightly lower, resulting in a potential need for more markers in pattern sets, and consequently a likely need for even more pattern forms. However, this difference in the numbers of pattern forms may be reduced slightly by the increased variability of some aspects of these forms in the English data. Applications that involve the identification, processing and analysis of related elements and the structures in which they appear seem likely to confront more of the challenges evaluated here in French as compared to English, and some of these factors may also influence the ultimate usability of the information extracted. Conversely, difficulties such as certain expressions of uncertainty appear more likely to be observed in the results of KRC extraction in English, although their implications for the usability of the information identified require further analysis in light of specific applications.

In considering the differences observed in the data in English and French, it is nevertheless essential to keep in mind that in the design and performance of pattern-based tools, many distinct factors that interact in complex ways depending on the needs of the specific situation in which tools will be used can affect the ultimate usefulness of these tools. The effects of interlinguistic differences may be cumulative in some cases, but may also in some ways balance one another. It is thus difficult to identify an overall tendency that indicates that more difficulties will be encountered in one language or another. Rather, differences suggest strategies for improving performance as required

and guiding choices in the development and use of tools in both languages in specific situations.

Most importantly, these differences indicate a number of promising avenues for future research, to evaluate the observed differences in larger samples of data, to better describe the phenomena observed, to more precisely identify sources of variation and to develop effective strategies for dealing with these phenomena in the two languages.

In addition to the observations of these factors and their similarities and differences in the results in the two data sets, this research provided an opportunity to observe issues that may suggest additional directions for future research and development of various approaches in the field, as well as possibilities for adjusting and improving the type of methodology used in this project. These will be discussed below in Section 5.5.

5.5 Additional observations and challenges

This section will present some challenges related to the methodology used in this research, as well as some possibilities for further developing and adjusting it. The discussion addresses aspects of the corpus-building process, the choice and classification of terms for the initial concordances, the identification, annotation and classification of these relation occurrences, and some issues in the interpretation of different types of expressions of the concepts linked by these relations.

5.5.1 Corpus building

The process of corpus building is always influenced by both the criteria judged to be pertinent in selecting texts for a given purpose and the availability of resources that can satisfy these. In this section, two aspects of the process will be discussed.

The first of these involves the availability of texts classified according to subject headings identified in medicine. This kind of resource was readily available — and used — for the English texts, while in French alternative strategies were used in the corpus-building process. Subsequently, however, an alternative approach was identified that could allow a more parallel approach to be used. The CISMeF gateway (www.cismef.org), a directory that provides access to "quality-controlled" French-language Internet resources indexed using the U.S. National Library of Medicine's Medical Subject Headings (MeSH) could offer an excellent complement to the approach used here to build the English corpus, and is well worth exploring in future work.

The second factor for discussion involves additional text-classification criteria that could be very useful in the specific context of this kind of research. A classification based on what is referred to in the domain as *quality of evidence* or *grades of evidence* (GRADE Working Group 2004, 2004a, 2004b; Higgins and Green 2005; Liberati et al. 2001; Schünemann et al. 2003; University of California at San Francisco–Stanford University Evidence-based Practice Center 2001; Upshur 2003) could provide important information about the types of texts in which contexts and markers were identified.

In evidence-based medicine, judgments on the strength of evidence (i.e., the reliability of evidence for drawing conclusions, particularly on the presence of causal relationships on the basis of associations, as reflected in the criteria proposed by Hill (cf. Section 1.5.2.1)) may be based on criteria that in large part reflect the context in which associations were observed. Streiner and Norman (1998: 29–71) describe a number of criteria that assist in the ranking of study designs from those providing the strongest evidence, generally double-blind randomized controlled trials, to the weakest, clinical case studies. Judgments on the quality of evidence may also be assigned at a textual level, generally based on the kind(s) of study or studies that are discussed in a particular text, as well as the number of studies covered. The texts that are considered to be most reliable are those that report meta-analyses (i.e., that provide a synthesis of results of the major studies that have focused on a given subject); these may roughly correspond to at

least some of the review articles included in the corpora used for this research. The research articles included in the corpora cover a variety of other study types that provide a somewhat lower range of evidence strengths. Thus, the quality of evidence provided by each article in the corpora is reflected to some extent in the criteria used in corpus building in this project, although the quality of evidence is not explicitly stated. More exact indications could be made according to specific grades of evidence assigned to articles in databases and other resources.

A more precise division of the corpus texts used in this research into sub-corpora according to the criterion of quality of evidence and an analysis of observations in the texts according to these categories could provide extremely interesting material for identifying the types of markers that may be associated with each type of article. Moreover, the analysis of the use of ASSOCIATION and CAUSE–EFFECT markers could provide interesting material for discussion of the precision with which these markers are used by authors. However, the complexity of such a task is undeniable (given, for example, that in different sections of articles, authors may refer to or even cite a particular study and its conclusions, allude to generally accepted information, or make direct observations and interpretations of these observations). A multicriterial analysis of texts and text segments in which markers were observed would likely be necessary to produce a fine-grained representation of the interactions between these factors and relation markers observed.

This kind of analysis could nevertheless be worthwhile, particularly since it could also be useful in another context: the presentation of results by a semi-automatic knowledge extraction tool. If such a tool could assign to each candidate KRC an indication of the grade of evidence associated with a text (or text segment) in which that context appeared, a user would have a better basis for evaluating the results of KRC extraction. This kind of distinction thus provides some extremely interesting possibilities for future research.

5.5.2 Choice of terms for initial concordances

Several issues linked to the choice and classification of terms used to generate the concordances that were analyzed in this work, the methodology used to deal with these questions, and possible adjustments in the methodology in future work can be discussed in light of observations in the course of this research. The issues addressed below include the choice to group terms denoting pathologies in a class of their own, some specific associations observed between markers and classes of terms or specific terms, and difficulties associated with the classification and/or use of particular terms, as well as the fact that some of the terms chosen could be markers in themselves. Finally, the effect of term choice — specifically the use of equivalent and non-equivalent terms to generate the initial concordances — on the results observed is also worthy of discussion.

The choice of terms for generating the concordances for analysis involved identifying both terms of interest in the domains of study and terms that represented a balance of various classes in order to reflect a broad range of possibilities for observing relations and markers. According the first criterion, it is clear that terms denoting pathologies are likely to be of central interest; in terms of the second, however, these terms posed challenges for the identification of appropriate classes. After consultation of lexical and conceptual resources, including WordNet and the UMLS, as well as analysis of occurrences in the corpora, it appeared very difficult to definitively and consistently classify even a single term as a state, process or an entity.[178] The choice to establish a separate class for these terms was made in order to allow for these important concepts to be observed while maintaining as good a balance as possible between the two languages. However, it is clear that this class is not at the same level of abstraction as the others used, and that ideally comparable levels of abstraction would be used in every case. One

---

[178] As indicators of very different points of view on just one term, for example, collocations such as *large/small tumour* (which is appropriate to the point of view of tumours as concrete entities) and *advanced tumour* (which involves viewing tumours as processes or possibly states) can be cited. The former view of the term is also reflected in the first definition given in the UMLS Metathesaurus, although the UMLS classification of the concept linked to by the term places it in the category of pathologic functions.

solution to this type of challenge could be a comprehensive evaluation of each term and potentially the contexts in which they occurred, in order to attempt to identify the best applicable class either overall or for each occurrence. However, this was considered to be too labour-intensive an approach in this kind of project. Conversely, the cost of excluding these terms was seen as too costly in terms of domain coverage. Therefore, classifying these terms separately was chosen as the best solution.

Moreover, the evaluation of this class allowed for the identification of some specific associations between these terms and the relations in which they often participated, as well as their use with some specific markers of these relations. These suggest that exploring more specific classes and their participation in various relations is a promising avenue for further research. Thus the choice made presents both challenges and opportunities for insight into the relations in the domain and the concepts that participate in them.

In another type of specific association, in the results of the analysis there appeared to be specific associations between markers (or groups of markers) and terms (or groups of terms). One good example may be seen in the connection between the French markers of CREATION *activateur de* (NOUN + PREPOSITION or ADJECTIVE + PREPOSITION) and *activer* (VERB) and the term *transcription*, with which this marker was exclusively associated.[179] Another example is found in the frequent occurrences of combinations such as *tumour suppressor gene* and *gène suppresseur de tumeur*. Once again, these kinds of associations are interesting subjects for further evaluation, as they may affect not only the potential productivity of these markers but also possibilities for marker disambiguation, for example using semantic classes of elements occurring in conjunction with markers.

---

[179] It is notable that *activation*, related to the markers mentioned here, was one of the initial terms used to generate concordances in the two languages and identified as a potential relation marker. However, it was not identified in this analysis other than in cases in which it occurred as the term used to generate the concordances, and thus was excluded from the analysis, as described in Section 3.3.1.1.

Another issue in term choice involves the resources used to assign a class to candidate terms. The use of an established and widely used resource allows for the classification of candidate terms in a systematic and coherent way, and can offer a method of minimizing bias in the results obtained. Nevertheless, as discussed in Section 3.2.1.2.1, no resource perfectly adapted to this kind of task is available, and the assignment of a single class — and moreover a class that reflects a place assigned to that term or the concept it denotes in a much larger structure defined according to numerous criteria — to a given term often does not reflect the specificities or variability of the usage of that term in practice. One example is that of the terms *tumour* and *tumeur*, mentioned above. While these issues of course present challenges in application, particularly in natural language processing, the use of the UMLS nevertheless allowed for the interlinguistic and inter-class balancing required in this project.[180]

A more specific issue related to classification and term choice is linked to the selection of the term *récidive* for the French research. As described in Section 3.2.1.2.1, footnote 74, *récidive* was classified neither as a disease or syndrome, nor as a pathologic function; however, on the strength of its less specific classification as a phenomenon or process and its definition, which clearly linked the term to a pathology, its inclusion in this category was considered to be acceptable. However, the results observed in relation identification — which was considerably lower than that of the other members of this category in both English and French — suggest that this term is significantly different in its performance for this kind of research. In light of these observations, and of those described above, it seems that the supplementing of high-level semantic classes of terms used in term-based approaches by more specific criteria when possible, in order to provide a better portrait of terms' characteristics and likely performance, is advisable.

---

[180] Moreover, the use of the terms *tumour* in English and *tumeur* in French, and the observation of similar phenomena in both cases provides a certain parallelism for the purposes of comparison in this research.

As noted in Section 3.3.1.1, a few of the terms used to generate the initial term-based concordances analyzed were observed in the course of the evaluation to constitute markers of relations in and of themselves. In the annotation of these concordances, a choice was made not to annotate occurrences of relations marked by these terms in the concordances generated with the term, in order to minimize bias in the results in terms of the frequency of relations observed and of individual markers of these relations.

The inclusion of these candidate terms was considered to be justified in the context of this project, as these terms may in fact be particularly good candidates for description using a pattern-based approach to identify KRCs: the concepts they denote are not likely to be satisfactorily described or defined using traditional approaches such as a generic and differentiating characteristics. Evaluation in KRCs allows for the observation of usage and collocates of these terms and thus can provide information that is particularly helpful for terminological description, including differentiation between multiple senses, if necessary.

However, it is clear that this phenomenon nevertheless did have an effect on consistency in annotating relation occurrences. It may be advisable in future work in pattern discovery to consider this factor, and to develop strategies to maintain consistency as far as possible. In comparative studies, it would be possible either to exclude these terms, or to consider them as markers but to ensure that — insofar as possible — similar cases are studied in the two languages. Both of these strategies, however, also clearly present disadvantages and challenges of their own. Since these terms — as potential domain-specific markers of relations (given their specificity as suggested by the results of the TermoStat analysis and in their identification as potential relation markers in the analysis of the concordances) — occupy a special place in the language of the field, they could be interesting subjects for study in and of themselves. It thus seems advisable to pursue research including these candidate terms, while keeping in mind the potential challenges in the process of evaluation due to their dual natures.

In addition to these points, a major issue in the evaluation of the methodology involving the selection of terms for generating the concordances analyzed involves the use of some terms that were equivalents in the two languages, and of some that were not. The impact of this choice on the results, and what this can suggest about this general approach for pattern discovery, is discussed below.

5.5.2.1 Use of equivalent and non-equivalent candidate terms

As described in Section 3.2.1, the term-based approach used in this research was intended to parallel that likely to be used by a terminologist working in the field, who would be likely to be searching for information about terms to be described, and might often, in today's context, use computer tools to assist in this process.

Moreover, given the results observed in work such as that of Bodson (2005), steps were taken to ensure that the choice of terms did not unduly bias the results due to associations of given classes of terms (e.g., denoting entities, activities, and so on) with specific relations in which they are likely to participate (and indirectly — or even directly — with the markers used to express them). These criteria produced lists composed in large part, but not exclusively, of equivalent pairs in the two languages.

However, as discussed briefly in Section 4.1, differences were observed in some analyses of data obtained using terms that were identified as equivalents in the two languages and those that were not, suggesting that the number of relations observed may be more influenced by the choice of specific terms than expected, and that this performance may have implications for observations of other aspects of the approach's performance.

The discrepancy between the numbers of relations identified in the two data sets in the overall results evaluated in this project originally inspired the comparative analysis of the numbers of relations observed for the pairs of equivalent candidate terms as compared to non-equivalents used to generate the initial concordances. This study

revealed that the proportions of relations observed in the group of equivalents, and particularly of ASSOCIATION relations, were much more comparable than those involving the non-equivalent terms, and suggested that the difference in the numbers of relations observed in the two data sets could be linked to inter-term differences.

It is possible that in a term-based approach to bilingual pattern discovery, equivalent terms may be particularly good candidates for research into developing bilingual tools, as they may be more likely to produce comparable results in terms of the number and distribution of relation occurrences that are observed and therefore that can be analyzed to create pattern sets and evaluate possibilities for performance.

Differences in a number of other factors were also observed in the analysis of data gathered using equivalent terms in the two languages as compared to that found using non-equivalent terms. While some of the variations may be fairly easily explained by direct relationships between the terms observed and the criterion in which variation was observed, others appear to result from more indirect and/or complex interactions of factors that are worthy of further evaluation in larger amounts of more comparable data.

Further study may shed light on the ways in which various factors inter-relate and may help to expand knowledge about the performance of knowledge patterns and the strategies that may be used to develop pattern-based tools and to maximize their efficiency. Some areas in which this kind of evaluation may be of particular interest involve the occurrences of complex and simple markers, the form of elements linked by pattern markers, the prevalence of anaphoric references, and the occurrence of some types of expressions of uncertainty, such as modal verbs, all of which showed some variation between the two groups.

However, the data available do not allow for the role of equivalence between candidate terms used to extract contexts for analysis to be evaluated systematically and differentiated from other factors that may contribute to the differences observed.

First, while the data for the groups of terms appear to indicate differences, the results for individual terms showed considerable variation in the numbers of relation occurrences in which they participated (as illustrated in Section 4.1), making it more difficult to observe coherent trends.

In addition to the differences in individual terms, the two groups of term pairs show different distributions among the various semantic classes considered (e.g., with a large proportion of the terms denoting pathologies constituting equivalents in the two languages, but most of the terms denoting processes not equivalent). Because many of these criteria may be linked directly or indirectly to differences in terms or in term classes, it is difficult to trace the exact origins of many differences observed. Neutralizing differences in classes would involve a term-by-term comparison, which is not only an extremely detailed and thus labour-intensive process, but also one that would involve far smaller numbers of occurrences of each of the phenomena analyzed, often making comparisons of prevalence in the two data sets statistically impossible or unreliable.

Moreover, because pattern characteristics and difficulties can be inter-related in complex chains (e.g., a marker may be associated with a specific term class or term; this marker may participate in specific pattern structures; these structures may be particularly vulnerable to certain kinds of difficulties such as interruptions; and so on), it may not be possible to explain all of the variations observed without an extremely fine-grained, multicriterial evaluation of all of the occurrences.

Therefore, the differences in the various factors could not be reliably evaluated in detail on the strength of these data. However, this variation does appear to merit further study using a methodology that allows for the effects of equivalence to be properly evaluated in the context of pattern-based tools (i.e., using an appropriate methodology and a sufficient amount of data).

If further evidence is found indicating that the choice to use equivalent and non-equivalent terms has an impact on results of relation and marker discovery, this may suggest modifications to the methodology that could be considered in future work — provided that the challenges and potential for bias in other areas inevitably associated with these choices are also taken into account.

Rather than relying exclusively on the use of semantic classes to reduce potential bias, one approach to pattern discovery in future work might be to couple a technique based on the results of an evaluation of terms' specificity in a given corpus, such as that produced by TermoStat, with consultation of existing terminological resources in order to identify term pairs that are considered to be equivalents. This approach certainly is not perfect (given, for example, that no such resource is exhaustive and that the coverage of terms in the corpora evaluated could well be uneven), and would be likely to reduce the specificity of the terms retained somewhat (as some of the more specific items in one corpus might not have equivalents that are equally specific in the other), but it could increase the chances that the sets of markers observed in the two corpora will be comparable.

Another potential approach would be similar to that used for example by Barrière (2001), which involved a comprehensive, manual analysis of a smaller corpus of texts (or, alternatively, of random samples of texts). This approach would eliminate bias linked to term choice entirely, and moreover — depending on the scope of the analysis — could provide a more complete overview of the relations and markers present in a given corpus. It would be an interesting approach, for example, if corpus size were limited, or an exhaustive coverage of patterns in a corpus were the goal.

However, such analyses involve their own challenges and raise some different questions about whether the data obtained are representative. First, a comprehensive manual approach entails the limits of human processing and the selection of a corpus or sample of a corpus of manageable size for this kind of analysis. The challenges posed by the selection of this kind of sample could involve difficulties related to whether a sample

is representative (e.g., a reduction of the variety of authors and texts that can reasonably be represented, or the sections of various texts that are evaluated). Moreover, an approach intended to locate data for use in terminology, but that does not begin with terms, may not provide as accurate a reflection of the use to which the patterns located will eventually be put (particularly in applications that use sets of candidate terms or specific terms for searching). The contemplation of an approach involving the random selection of contexts may also raise questions about criteria for identifying pertinent contexts. A (candidate) term-based approach provides some measure of assurance that the relations identified involve at least one concept that could potentially be researched by a terminologist in the course of conceptual analysis and terminological description. However, a manual approach would involve the definition of criteria for the admissibility of relations observed according to their usefulness for an intended application, the nature of the elements that are linked, and so on. The task of establishing these kinds of criteria could prove to be quite complex. In addition, the use of a term-based methodology for pattern discovery, which is commonly used in the domain in unilingual projects, offers possibilities for observing phenomena that are likely to confront others using the methodology, for highlighting — as these results clearly do — some of the difficulties and challenges likely to be encountered this kind of approach, and for suggesting ways of improving it.

In considering the two types of approaches available for such projects, it may be interesting to consider the differences observed in the markers observed in projects that focused on the same relation but differed in the approach taken. One such basis for comparison may be found in the data gathered in Barrière (2001, 2002) and Marshman (2002) (cf. Sections 2.1.8 and 2.1.9). A brief overview of the markers observed in these projects is provided below.

Barrière's work (2001, 2002) involved the complete, manual analysis of an 80,000-token, 5,500-sentence corpus on composting, in order to identify lexical patterns for the CAUSE–EFFECT relation. Marshman (2002, 2002a, 2004) targeted lexical markers

of the same relation classified according to the same typology, but in a 225,000-token, 7,600-sentence corpus on biopharmaceuticals, using a term-based approach. This second research project involved the analysis of occurrences of twelve terms, accounting for approximately 4,000 contexts in total (although several contexts contained more than one of the terms in question, and thus would have been viewed more than once).

In comparing the results, a striking amount of overlap between the patterns identified was found, but with some interesting differences related to the nature of the patterns located. In Barrière's articles, a list of 42 verbal markers (e.g., *cause* [VERB], *destroy*), 19 with associated nominal derivatives (e.g., *cause* [NOUN], *destruction*), as well as 5 non-derived nouns is given. Of the verbs, 26 or approximately 62% were also located in Marshman.[181,182] Approximately the same proportion (63%) of the nominal derivatives listed was also found. However, of 28 conjunctional patterns identified by Barrière (e.g., X *so that* Y, *since* X, Y), only 8 or approximately 29% were found in Marshman, and only 20% of the nouns not derived from verbs. While these figures are indicative only, they nevertheless suggest that specific types of markers might be favoured in each approach.

Interestingly, Barrière's further analysis (2001: 145; cf. Section 2.1.8), indicates that she found verbal patterns to be much less noisy than conjunctional patterns (with a noise ratio of 0.31 as compared to the conjunctional patterns' 0.82), explaining her subsequent decision to concentrate on the former class. Thus, while it must be recognized that the patterns identified may be influenced by the approach used, it appears possible that a term-based approach will in fact favour the identification and

---

[181] It is important to note nevertheless that the criteria for identifying marker occurrences used in this thesis (cf. Chapter 3) are somewhat stricter than those used in Marshman (2002). In this research (following models used by many researchers, including Bodson (2005)), for an occurrence of a relation to be retained a connection between the term used to generate the concordance and the marker was required, while in Marshman (2002) occurrence in the context alone was considered to be sufficient.

[182] As Marshman (2002) used a character-string-based approach, verb and noun forms often corresponded to a single marker.

study of markers that are likely to be particularly useful for subsequent applications in terminology. Nevertheless, formal evaluation would be necessary in order to explore this possibility and to permit the drawing of conclusions, as these data constitute only a small, *ad hoc* sample.

In conclusion, the use of a term-based approach involving both equivalent term pairs and non-equivalents in this study has revealed some interesting results that — if confirmed in more targeted studies — will be pertinent for bilingual work and could be taken into account in future projects, and has provided concrete data on which future research may be based.

5.5.3 Challenges in identifying and classifying relations

Among the issues related to the identification and classification of relation occurrences that may be discussed in light of the results of the research are the criteria for retaining occurrences of CAUSE–EFFECT relations, the challenges of using the relation classification for the CAUSE–EFFECT relation, and possibilities for enhancing the classifications of both CAUSE–EFFECT and ASSOCIATION relations, as well as for classifying candidate KRCs containing multiple markers.

5.5.3.1 Criteria for retaining CAUSE–EFFECT relation occurrences

As noted above in Section 1.5.2.7, a decision was made in this research to set aside occurrences of relationships that involved an element of causation but were considered to be too complex and specific to be used for the purposes of terminology work as envisaged here (e.g., relations that may be indicated by markers such as IRRITER or NETTOYER). This distinction may be compared to that made by Kahane and Mel'čuk (forthcoming), between verbs of causation and causative verbs (cf. Section 1.5.2.4).

This is not to state, however, that more complex relationships and the markers that denote them may not be of interest for knowledge extraction in certain specific contexts. Reference may be made here to the observations by a number of researchers, including Sager (1990) and Nuopponen (2005) that the types of relationships that are important for a given project are likely to depend on the specific context and goals of that project, and to the work of researchers such as Faber et al. (2006), who noted — in the context of a study focusing on the creation of frame-based representations of relations in domain-specific applications (in their case, the field of coastal engineering) — that in particular, domain-specific situations, more specific types of relationships may be extremely useful for structuring knowledge.

Two specific observations of difficulties related to the distinction between core and complex causal relationships may be made, in light of the analysis carried out in this work. The first pertains to the forms in which a given CAUSE–EFFECT relation is manifested at a textual level, while the second relates to ambiguities of markers retained in this analysis.

A potential inconsistency in relation identification related to the distinction between core and complex relationships may be observed in examples based on those of Kahane and Mel'čuk (forthcoming). Occurrences such as *Le va-et-vient des voitures cause l'irritation de Zoé* or *Zoé tue la grenouille (en l'écrasant)* could be retained as occurrences of pertinent conceptual relations in this research and *causer* and *tuer* identified as markers of the sub-relations of CREATION and DESTRUCTION, respectively, since these correspond to the CAUSE–EFFECT relations of passing cars causing irritation to occur or Zoé causing the frog to cease to exist (at least as a frog, i.e., a living thing). However, alternate expressions of the same realities such as *Le va-et-vient des voitures irrite Zoé* or *Zoé écrase la grenouille* would not, because the complexity of the relationships expressed between Zoé and the passing of the cars or between Zoé and the frog, as observable at a semantic level in the meanings of IRRITER ("provoquer chez [quelqu'un] un certain énervement pouvant aller jusqu'à la colère" (*Lexis* 1992)) and

ÉCRASER ("déformer ou… aplatir [quelque chose] par pression ou par choc" (*Lexis* 1992)). This is considered here to be an inevitable drawback of pattern-based approaches, one that must be recognized, but is also, in our opinion, acceptable in the context of research projects such as this one.

The second issue involves markers observed in this research that are ambiguous in the sense that in some cases a marker indicates a "core" relation as retained for the purposes of this research, while in others it corresponds to a relationship that includes an element of causation but is complex enough that it was excluded from this study. The phenomenon is reflected in a brief discussion of such marker polysemy as observed in the data evaluated in the study of marker precision (Section 4.7.2). It was also discussed in Marshman and L'Homme (2006) and Marshman and L'Homme (2006a), in which analyses of the different senses of the markers identified in this project — both those senses that correspond to the core relations and those that are more complex — were carried out.

A number of examples may be identified: *induce* and *induire* (which can indicate the causing, i.e., CREATION, or INCREASE of some kind of process, the CREATION of a molecule, or another more complex causal sense involving the modification of a molecule — generally an enzyme — so that it becomes functional); *activate* and *activer* (which can indicate either CREATION or the transformation of a molecule to make it functional); and *block* and *bloquer*, which can indicate "core" PREVENTION in addition to more complex relationships, e.g., of preventing access to and/or the functioning of something (generally a protein or receptor), or preventing passage through something (generally a blood vessel or a channel in a cell membrane). The case of *block* is illustrated in Examples 370 to 372:

> 370. AT1 receptor antagonists **block** the oxidative stress… (Granger et al. 2004)
>
> 371. Other methods of attenuating the effects of aldosterone involve inhibiting the aldosterone synthase enzyme or blocking the aldosterone receptor. (Moore et al. 2003)

372.... blood clots (thrombi), which partially or completely **block** the vessel... (DiGiovanna and Adams 1999)

Interestingly, the more specific senses are often linked with specific terms or classes of terms in specialized domains (e.g., blocking of receptors, blocking of blood vessels, stimulation of cells or of pathways, activation of cells or molecules, induction of enzymes).

The usefulness of such specific relations may vary according to the application envisaged in terminology work, and thus these cases may be interpreted from several perspectives. One of these — which parallels that explored in Marshman and L'Homme (2006, 2006a) — may allow the various senses involving causation to be retained in an analysis, as long as those that involve the primary, "core" causal sense can be differentiated from those that involve a more complex sense with a causal component.

Alternatively, strategies may be adopted to limit results to the core senses that were originally targeted. This is an approach that may be chosen in a number of approaches to KRC extraction, and the one used in this thesis (largely in an effort to obtain a reasonably consistent, comparable and manageable range of data that reflects the most appropriate contexts for pattern-based extraction). Projects focusing on the establishment of direct links between concepts in concept structures included in or underlying the structure of terminological resources, for example, will not likely be able to make widespread use of the more specific types of relationships without entering into an extremely detailed (and thus time-consuming) analysis of all of the kinds of CAUSE–EFFECT-based relations present in a domain and how they connect to one another.

For example, in the case of the verbal marker *block* that consistently indicates PREVENTION of some kind, it would nevertheless be misleading to link concepts in a concept system using identical PREVENTION links in the case of both the blocking of a process's occurrence and the blocking of a blood vessel or of a receptor (since it is not, of course, the existence of the blood vessel or receptor that is prevented). Additional information would have to be added to the link to specify what exactly is prevented,

thereby multiplying the types of links required in a resource — and the difficulties of managing these — exponentially. This becomes very difficult or even impossible as the scale and scope of a terminological resource — and thus the range of potential sub-types of relations — increases (e.g., particularly if a large domain or several different domains are covered). It might be equally misleading, in applications that attempt to sort candidate KRCs by the relation present, to group together all of the contexts in which any PREVENTION-based relationship is observed, as the above relationships are quite distinct.

The choices made in this respect should reflect the application intended for the results of knowledge extraction: tools focusing on more automatic applications or the extraction of knowledge for the purposes of linking concepts in a system or terms in a term base are likely to focus more on the core relations, while domain knowledge acquisition and the creation of definitions may make use of information about the more complex relationships these markers may indicate, provided that the nature of this information can be clearly differentiated from core senses to prevent confusion.

At an interlinguistic level, the regularities observed, both in the polysemy of markers in the two data sets and in associations with specific terms or classes of terms suggest some possible methods for automatic sorting and/or elimination of these more specialized contexts, depending on the needs of a given project. However, this kind of technique would also confront a number of challenges (cf. Marshman and L'Homme 2006).

5.5.3.2 Challenges in Barrière's CAUSE–EFFECT relation classification

As described above in Section 3.3.1.2.2, Barrière's classification of the CAUSE–EFFECT relation was chosen for use in this project; it was considered to be the most appropriate for this purpose because it targets important differences between types of CAUSE–EFFECT relations in the medical domain, reflects a number of common characteristics of these relations that were identified as important to take into account in analyses from a variety

of points of view (e.g., cognitive to linguistic), and is particularly suited to this kind of project because it allows for marker-based sorting of contexts according to these criteria. Paralleling observations in Barrière (2002) and Marshman (2002), the system was found to be adequate for the task of classifying relation occurrences in the majority of cases observed in this project. Nevertheless, like any such system it does have limitations. In this section, some of the most widely observed difficulties will be outlined.

One difficulty observed in the use of this classification was related to differentiating between potential classifications of ambiguous pattern markers. Such markers, including *stimulate/stimuler*, *inhibit/inhiber*, and *suppressor/suppresseur*, may potentially convey two or more sub-types of CAUSE-EFFECT relations (e.g., CREATION or INCREASE; DECREASE, DESTRUCTION or PREVENTION), and it may be very difficult even for a human to determine which is pertinent in a given case (certainly without reading a much larger part of the original document, which is of course what a pattern-based KRC-extraction tool seeks to spare the terminologist wherever possible).[183] Certainly, this phenomenon poses even more difficulties for automatic context sorting.

For the purposes of this project and its descriptive orientation, occurrences of these markers were classified according to the best possible interpretation of the individual occurrences in light of further research (both in the text itself and of supplementary resources where necessary).

However, to manage such cases in practice (i.e., in attempting to sort KRCs retrieved using knowledge patterns) requires another approach. Some possibilities for disambiguation may be observed in specific cases. In Example 373, for example, a rare case in which the sense of the marker is clarified is found: by opposing *stimuler* with *provoquer*, which is less ambiguous, the author has clearly indicated that the growth of

---

[183] Moreover, in some cases it may be challenging to determine the exact relation present even in light of a reading of a larger segment of or even a whole text.

these lesions is increased — not caused — by hormones. This kind of input, however, would be very difficult to represent formally in a consistent way, and would likely not be widely useful for distinguishing senses in an automatic application.

373. Si les lésions, **provoquées ou stimulées**, peu importe, par les hormones n'étaient que des lésions de faible malignité, totalement curables ? (Bouchard 2001)

In other cases, modifiers of these markers may clarify at least to some extent the type of relation that is present, as in Examples 374 and 375:[184]

374. La toxicité de l'acide flavone-acétique (FAA ou flavone-8-acetic acid) sur des cellules cancéreuses mammaires in vitro **est** totalement **inhibée par** un inhibiteur de NOS… (Gauthier et al. 2004)

375. La production des autres hormones surrénaliennes (testostérone, déhydrotestostérone androstènedione, progestérone et 17-hydroxyprogestérone) **est** très partiellement **inhibée**. (De Crémoux 2000)

In Example 374, *totalement* helps to eliminate DECREASE from the possible relations; by the same token, in Example 375, *très partiellement* eliminates PREVENTION and DESTRUCTION, leaving DECREASE as the best interpretation. Once again, however, formal approaches to disambiguation based on these indications would be difficult to develop.

Alternatives for the disambiguation of at least some of the occurrences and senses of these markers may include the use of the actantial structures in which they appear and the semantic classes of their actants. As mentioned above in Section 2.2, this possibility has been explored in research projects such as those of Feliu (2004), Weilgaard (2004) and Bodson (2005), as well as in Marshman and L'Homme (2006), which focused specifically on some of the ambiguous markers observed in this project.

In an alternative approach, Barrière (2002: 12–104) noted some cases of similar ambiguities, and either assigned two possible sub-relations to a single marker, or

---

[184] These two examples are not part of the set of contexts identified and analyzed in this research, but are provided here to illustrate phenomena that may be observed in pattern-based candidate KRC extraction.

classified more ambiguous markers in a more general category, such as EXISTENCE, INFLUENCE or CAUSAL.

From the observations made in the course of the research, an addition or modification may be proposed. Most often in the ambiguous cases evaluated in this research, ambiguity occurred between sub-relations that can be characterized as "positive" or "negative" effects, between CREATION, MAINTENANCE and INCREASE on the positive side, and DESTRUCTION, PREVENTION and DECREASE on the negative side.[185] As mentioned in the description of the relation classification (Section 1.5.2.8.3), the sub-relation of MODIFICATION may be used in cases where it is impossible or very difficult to identify with certainty whether a change results in an INCREASE or DECREASE of the effect. In parallel, it could be possible to integrate into the classification — or otherwise implement — a similar generic classification that could be used in cases of doubt, where it is not feasible to assign a single sub-relation for a relation expressed in a given context, and in which the ambiguity occurs between sub-relations in both the EXISTENCE and INFLUENCE categories.

As in the case of MODIFICATION, a more specific classification is of course desirable, but not at the cost of mis-classification of occurrences due to guesswork or excessive generalization. The establishment of such relation categories would allow misclassifications to be avoided, but also to retain the option of classifying occurrences and patterns more specifically than the generic CAUSAL that was Barrière's ultimate solution to such problems. This intermediate level of granularity may be sufficient for user needs in some applications, for instance if all a user needs to know is if a given factor exerts a positive or negative influence on an event.

It is interesting to note that in these cases very similar markers presenting parallel senses and ambiguities in English and in French were found. This would be an

---

[185] This rejoins observations made by Nazarenko (2000) in her analysis of CAUSALITY in French lexical semantics (cf. Section 1.5.2.3).

advantage for developing bilingual tools, as strategies for disambiguating these markers may also be similar, and thus the work required to resolve these ambiguities (if possible) or to develop strategies for dealing with them may be useful in both languages.

5.5.3.3 Possible complements to the CAUSE–EFFECT relation classification

In addition to some potential adjustments of Barrière's classification in itself, it is possible to consider adding another layer to this classification in order to further refine the granularity of context sorting. One method of doing so would be to consider integrating criteria taken from such classifications as Nuopponen (1994; Section 1.5.2.5). While some of these were considered to be too specific for the orientation and goals of this research (cf. Section 1.5.2.8.5), observations in the data in the two languages did suggest some ways in which certain aspects could be implemented in future research.

Nuopponen's classification specified, among other questions, whether causes produce an effect independently (closely linked to the idea of sufficient causes) or are one of several that contribute either alternately or together to produce it, whether a cause has one or more effects and whether these effects are co-occurring or alternative.

At a conceptual level, this aspect of CAUSE–EFFECT relations is clearly significant. It is nevertheless important to keep in mind that, when analyzing CAUSE–EFFECT relations, one is necessarily drawing conclusions about connections between two elements, identifying a cause (or possibly more than one cause) to the exclusion of all other factors in the environment that may encourage or allow this connection to exist or may fail to prevent it (e.g., the absence of counteracting causes as identified by Nuopponen (1994), predispositions, the fundamental laws of nature and of chemistry, biology, physics, etc.) In medicine, researchers and other specialists frequently have only a partial picture of the processes that are taking place in the complex system that is the human body, and any number of environmental factors may play a role in a given event. Thus, the identification of any relationship between two elements, such as a cause

and an effect, could be said to be conditional on environmental factors. Moreover, as new environmental factors that influence these relationships are identified, researchers' comprehension of causal links may evolve to include additional pertinent conditions. Moreover, at a textual level (and particularly when conditions have already been specified), authors may not specify at every mention of a relation whether the cause is independent or not, contributing or alternative, or as yet uncharacterized. However, the contributing nature of causes in some cases is reflected in the markers used to express these relations, e.g., *contribute to/contribuer à*, *participate in/participer à*, *involved in/impliqué dans*, *play a role in/jouer un rôle dans*.

Thus, the results indicate that while it is almost impossible to determine with certainty from a given knowledge-rich context whether a given cause is sufficient to produce an effect (according to knowledge at a given moment or in the long term), in some cases, it is possible to conclude when it is not, on the basis of overt description of a cause as contributing. Moreover, the interlinguistic similarities observed in these markers suggest that this phenomenon is similar in the two languages, and that it may thus be exploited in both English and French.

The next question for consideration is then how this type of information distinguishing (definitively) contributing causes from others can be accounted for in the classification of occurrences, i.e., whether it should be associated with and applied in the processing of occurrences containing individual patterns, or whether this kind of information should be integrated into the relation classification itself, as reflected in Garcia (1997; Section 1.5.2.8.2). Moreover, in considering the potential for inclusion in a relation classification, determining how this distinction may be integrated into a hierarchical structure such as the one created by Barrière (2002) (e.g., as a separate high-level class, or as a subordinate class of each of the sub-types already identified) requires evaluation.

In the observations in the corpora, the markers observed to convey this kind of information were included in the category of CREATION markers. For marker-based

classification, it could thus be possible to consider integrating a sub-classification of the CREATION sub-relation into the relation hierarchy, to deal with these cases.

At a fundamental level of the definition of relation types, however, another point can be raised in the case of the INCREASE sub-relation. From one perspective, the element responsible for increasing another (e.g., causing it to occur more, causing more of it to exist) may be identified as a contributing cause of the latter's existence or occurrence (as, clearly, the former element cannot be the only cause of this existence or occurrence, as the latter already existed or occurred before its intervention). This would then suggest a need to reflect this reality in the classification system used if the contributing nature of causes is to be reflected in a systematic way. Reflecting this shared aspect of some occurrences of CREATION and all of the cases of INCREASE in a relation classification, however, is considerably more complex.[186]

Another potentially interesting complement to the classification used here is also based on the classification developed by Nuopponen (1994) and addresses the number of causes or effects observed in a given relation, but suggests different strategies for implementation as a complement to a classification such as that used in this project. This approach would involve the evaluation of the form of the elements linked by markers, and particularly the presence of multiple related elements sharing a role in a relation and joined by conjunction or disjunction.

The presence of multiple elements filling a slot in a knowledge pattern can clearly indicate cases in which (an author states that) two or more causes may contribute to an effect, or two or more effects may result from a given cause. Automatic processing of these kinds of structures may thus allow for semi-automatic sorting of contexts expressing relations with multiple as opposed to single causes or effects.

---

[186] Moreover, the issue of whether the mirror image of this reasoning can be applied in the case of the DECREASE sub-relation raises some equally — if not more — complex questions.

However, the further classification of causes as co-operating or alternative and effects as co-occurring or alternative may pose more challenges: at a contextual and even a textual level, it may be difficult to determine exactly which situation is present when conjunction of multiple elements occurs. One possibility involves an analysis based on certain markers of conjunction. While many (including the most frequent *and*) do not clarify the kind of relationship that is likely to be present, others, such as *together with* or *along with* may suggest that causes are co-operating or that effects are co-occurring, rather than alternative. The case of disjunction, however, is more straightforward, in that it more reliably indicates alternative causes and effects. Once again, it seems that while some aspects of Nuopponen's classification may be relatively straightforward to implement on a formal level to complement other classifications used for context sorting, further refinement is likely to require human interpretation and intervention. This is particularly true when the possibilities of nuances in the ways in which conjunction and disjunction may be interpreted are considered (e.g., differences between exclusive and inclusive disjunctions).

It is interesting to note that the kinds of approaches that could be used to reflect these different aspects of a single basic distinction are considerably different. One could involve a refinement of a relation classification and is largely reflected in differences in markers, while the other would be most usefully implemented in analyzing pattern structures.[187] As such, it appears that if this task is pursued, evaluation of the presence of contributing causes should take place at multiple levels. This task would, however, be relatively complex to automate in a coherent way, and it seems likely that a fairly significant element of human evaluation would be necessary to obtain a complete, precise and reliable picture of the relationships described in various contexts.

---

[187] Additionally, if the structures in which multiple elements sharing a role in a relation can occur are evaluated, this opens possibilities for the extraction of additional information from contexts containing multiple elements sharing roles in a relation, including those that indicate SYNONYMY or GENERIC–SPECIFIC relations between these elements. While these relationships do not directly affect the principal CAUSE–EFFECT relation identified, they nevertheless are likely to be useful for conceptual analysis and may be usefully extracted.

Another potential application of some of the observations made by Nuopponen (1994, 2005), in this case concerning the interpretation of alternative effects in CAUSE–EFFECT relationships, involves a parallel with expressions of uncertainty such as quantification, hedging or the use of modal verbs. The existence of alternating effects indicates that a given effect will not always occur as a result of a cause, establishing a potential parallel with quantifiers such as *some* or *many*, indicators of hedging as *generally* or *sometimes* or modal verbs such as *can* or *may*. As this can be observed at a surface level in the disjunction of related elements (discussed in Section 4.9.1.2), it offers an additional strategy for the automatic evaluation of certainty levels. However, as this phenomenon is not always explicitly marked, once again it can offer only a partial contribution towards the evaluation of these factors. It may nevertheless be worth exploring, especially since this phenomenon may be pertinent for some applications in other relations as well (e.g., GENERIC–SPECIFIC, as in *a bicycle is a type of vehicle or of sports equipment*, or PART–WHOLE, as in *a wheel is a part of a bicycle or of a car*).

A final possibility for a refinement of the classification of contexts according to additional criteria indicated in Nuopponen (1994) is the possibility of distinguishing at a formal level between different types of causes (e.g., causal agents, causal events). In some cases in which multiple elements share a role in a relation, an element corresponding to each type of cause may be identified; this kind of conjunction of elements may be indicated by such structures as *By* X-*ing,* Y *causes* Z (cf. the discussion in Section 4.9.1.2). Formal analysis of these phenomena poses challenges due to the complexity of such structures. Nevertheless, if strategies are developed to deal with these cases, they may allow for different types of elements to be distinguished automatically, and for the different types of causes to be presented to the user, providing a more fine-grained classification of the information present and/or allowing for the most useful cases for a given purpose to be retained or sorted within a list of results.

All of these possibilities, although complex to implement, could constitute useful additions to the basic classification if more precise relation descriptions were desired.

5.5.3.4 Possible refinements of the classification of ASSOCIATION relations

As discussed above in Section 1.5.1, in this project, while two specific types of ASSOCIATIONS (RISK and CORRELATION) were considered for the purposes of this research, they were not specifically identified. Some observations of these specific ASSOCIATIONS will be presented below, and possibilities for refining the relation classification discussed.

5.5.3.4.1 Risk

As stated in Section 1.5.1.4, the concept of "risk" has been included in this classification of ASSOCIATION — given that (in the specialized domain of medicine) risk is calculated based on observations of co-occurrence of two variables — but that it differs from other types of associations, for example, in its directionality (i.e., the clear identification of an outcome and a factor presumed to affect it).

In many cases, relatively coherent classes of elements are associated with these markers; the prototypical combinations would involve the association between one of a number of entities (e.g., a treatment, a characteristic, or a test result — indicating the presence of a particular molecule in the blood, for example) and a disease or disorder, with the former identified as indicating a risk of the latter. Some surface variants in this form may be found in cases in which a particular aspect of the disease is mentioned, as in Example 376, or when an anaphoric expression is present, as in Example 377:

> 376. Hyperhomocysteinaemia is a **risk factor for** the development of CHD... (Mackness et al. 2004)
>
> 377.... an elevated serum creatinine level ... was associated with an approximately twofold higher risk of overall and CVD mortality. After adjustment for cardiovascular risk factors ..., elevated serum creatinine remained an independent **risk factor for** these outcomes as well as total CVD, congestive heart failure (CHF) and claudication. (Coresh et al. 2004)[188]

[188] This example is not part of the set of relation occurrences analyzed in this project, but is provided here to illustrate the phenomenon in question.

Another type of variation is observed when one of the variables involved is expressed in adjectival form (e.g., *cardiovascular* or *cardiovasculaire*), as in Examples 378 and 379:[189]

> 378. However, in response to the traditional cardiovascular **risk factors**, such as hypertension, diabetes, and hypercholesterolemia, the endogenous defenses of the vascular endothelium begin to break down. (Szmitko et al. 2003)
>
> 379. En plus de ces deux critères majeurs de sélection, on suggère de prescrire un bilan lipidique aux enfants démontrant d'autres **facteurs de risque** cardiovasculaire (obésité, tabagisme, hypertension, diabète, consommation d'aliments riches en matières grasses, prise de médicaments augmentant les lipides plasmatiques et sédentarité)… (Lambert 2002)

The use of specific markers to denote this special kind of ASSOCIATION, as well as the important difference in the nature of the relation and the needs for processing the results of extraction (e.g., in preserving the order of the elements observed, in order to identify which factor is the assumed cause and which the assumed effect) indicates that in future work — in which sufficient amounts of data are available — it would be interesting to distinguish these occurrences from others of ASSOCIATION. Fortunately, these relative formal regularities provide a promising starting point for this kind of differentiation. Some challenges will of course be encountered, and further evaluation would be helpful for developing strategies.

In another phenomenon observed in the relations, the *risk* and *risque* families of markers were often observed in combination with other types of markers. With one group of additional markers (e.g., *associated with*, *effect of*, *role of... as*, *effet sur*, *effet de, du fait de*, *à l'origine de*), a reinforcement of the ASSOCIATION (or potential CAUSE–EFFECT) relation expressed in the context may be observed. In the second, very common group (e.g., *increase*, *increased*, *augmentation*), the additional markers further describe

---

[189] It should be noted that this form is elliptical; the risk is of cardiovascular disease or other similar disorders of the heart and blood vessels. This may pose challenges for interpreting such contexts, particularly in more highly automated applications.

the type of relationship that exists between the two elements (often a MODIFICATION (INCREASE or DECREASE) in the likelihood of the event in question). While the *risk*-based marker is considered to determine the relation present and thus to take precedence over any CAUSE–EFFECT markers present (cf. Section 3.3.1.5.1.1), it could be interesting to take into account the more specific information conveyed by some of the additional markers, particularly of MODIFICATION. Given the relatively stable structures observed in many of these combinations, the development of strategies for further refining the classification of such contexts appears promising.

5.5.3.4.2 Correlation

As noted above in Section 1.5.1.3, in contrast to other types of ASSOCIATION in which the relation holds between static variables, CORRELATION involves a relationship between dynamic, continuous variables at a series of points, in which the value of one changes with the value of the other. However, general usage does not necessarily reflect this distinction, posing challenges for automatic identification of this kind of relation in corpora.

CORRELATION, as one would expect, may be indicated by markers such as *correlate*, *correlate with*, *corrélation entre* and *corrélé avec*. In other cases, this kind of ASSOCIATION may be possible, but unconfirmed, as in Example 380:

> 380.... on observe une nette **augmentation du** risque **avec** la durée d'utilisation... (Clavel-Chapelon and Hill 2000)

This may often occur in cases in which combinations of elements indicating MODIFICATION, INCREASE or DECREASE of a variable are observed, along with an indication that this MODIFICATION occurs in ASSOCIATION with another variable (e.g., the markers *in*, *with*, *dans* and *avec*). It may be very difficult to determine whether the relation involves two dynamic variables (i.e., is a true CORRELATION), or whether it is a simple case of an ASSOCIATION involving one dynamic and one static variable, as in Examples 381 and 382:

381. Two trials showed **improved** DFS **with** anthracycline-based chemotherapy… (Shenkier et al. 2004)

382. A further aspect of the **change** of atherogenicity of lipoproteins **with** HRT was tackled by Wakatsuki et al.…. (Seed and Knopp 2004)

One possible criterion for the decision may depend on the nature of the related elements. In Example 380, both elements (*risque* and *durée*) are continuous variables (i.e., it is possible to characterize risk as high or low, duration as short or long, and so on); thus they can potentially be compared at different values to determine if a dynamic ASSOCIATION (i.e., CORRELATION) is present. However, in Examples 381 and 382, the elements *chemotherapy* and *HRT* are less easily discussed in this way, and therefore are less likely to be involved in a true CORRELATION (at least, without some additional modification identifying what aspect of these treatments was being compared at different values). However, the evaluation of the various concepts involved in a potential correlation would involve highly developed analysis of corpus texts.

The variability in the expression of these kinds of relations suggests that differentiating automatically between ASSOCIATION and CORRELATION would involve a very significant investment of time and effort, and could not necessarily be successfully achieved in many cases. The relative infrequency of this specific ASSOCIATION in this research — or at least, of the cases in which it can be confirmed — does not suggest that this investment would provide a significant return.

Thus, if the development of a more specific classification of types of ASSOCIATION relations at an automatic level is to be considered, the sub-type of RISK appears to be the more promising avenue for development. CORRELATION, in contrast, appears to be best evaluated by a human user, perhaps using a tool that provides a primary sorting of contexts in order to identify some that are more likely to involve this sub-type of ASSOCIATION on the basis of the marker that is present, and perhaps even providing more specific indications of specific types of these associations as indicated by the presence of additional markers.

5.5.3.4.3 Challenges in interpreting ASSOCIATION relations

In Section 4.9.1, some fairly clearly defined cases in which multiple elements may share a role in relations were described, and some possibilities for processing them discussed. However, in some cases of ASSOCIATION relations, it is difficult to interpret exactly how multiple elements are related. This is the case, for example, in Examples 383 to 386:

> 383. Part 1 will provide a brief overview of the **link between** inflammation, endothelial dysfunction, **and** atherosclerosis... (Szmitko et al. 2003)
>
> 384. **Associations between** lymph node metastases, various clinicopathological features, **and** development of distant metastasis were assessed with the Pearson [chi]2 test. (Susnik et al. 2004)
>
> 385. Il semble exister un **lien** très étroit **entre** le syndrome de lipodystrophie, l'hyperlipidémie, l'intolérance au glucose **et** le diabète, bien que chacun de ces troubles puisse survenir isolément. (Baril and Junod 2004)
>
> 386.... un suivi attentif permettant d'établir les **liens entre** les anomalies lipidiques, le tabagisme, l'hypertension artérielle, le diabète **et** la maladie coronaire. (Bauduceau et al. 2004)

Here more than two elements are implicated in an ASSOCIATION relation, but there is no clear proof of whether a series of binary ASSOCIATION relations involving separate element pairs is present, or whether all of the elements are somehow linked in the same relation (i.e., whether there is a question of the co-occurrence of all factors, or some kind of constellation of factors that occur more or less consistently together). This makes the use of the information conveyed by the context challenging: analysis of a larger context (e.g., the paragraph or even full text) or further research may be needed to confirm the relation(s) present. The occurrence is nevertheless useful — at least for applications in which human interpretation of a larger context or other resources is possible — as it does indicates a relation of interest.

In addition to the challenges this phenomenon poses at a conceptual level, this kind of variant certainly has implications for the recognition of contexts using restrictive pattern forms (for example, a form that allows for only a single element to be present

with — or in the cases shown above, within — a marker would likely exclude such occurrences). Ideally, pattern forms intended for applications in which human interpretation is possible should allow for such variations. Some regularities in the occurrences of this phenomenon may be exploited for this purpose. This phenomenon was observed in the case of ASSOCIATION relations indicated by the markers *association between... and* and *link between... and* in English, and by *lien entre... et* in French. Thus, it appears to be linked primarily to specific markers (for which appropriate forms could be developed), and also shows similarities between the two languages. This raises the possibility that strategies for dealing with such contexts in one language might be useful in the other as well.

5.5.3.5 Occurrences of multiple patterns and/or markers

Many of the relation occurrences observed in this research contained multiple candidate patterns or pattern markers indicating a relationship between the same two elements. As described in Section 3.3.1.5.1.1, each context was associated with the relation identified for the context as a whole, and the marker that corresponded to this relationship was annotated as the principal one located in the context, although the presence of (an) additional marker(s) was noted.

However, this solution at the pattern identification stage does not resolve the issues that would be encountered at the stage of pattern-based tool use, and particularly of relation identification. In order to present a user with a sorted list of contexts expressing various relations or sub-relations, it is necessary to assign each context to one or more groups, and this task is most likely to be carried out on the basis of the markers present. The presence of more than one marker thus poses difficulties.

There is a danger, of course, in presenting misleading evidence in this classification; for example, classifying contexts containing both a marker of MODIFICATION and of ASSOCIATION — a common combination — as occurrences of MODIFICATION relations would imply that a CAUSE–EFFECT relation has been determined,

whereas the presence of the patterns indicating ASSOCIATION indicate that this is not (or at least, not necessarily) the case. However, pertinent information would be lost if the contexts were classified as indicating ASSOCIATION alone; certainly users would like to know, if possible, what specific types of changes are observed in ASSOCIATIONS and CAUSE–EFFECT relations.

One possible method for dealing with multiple relations and patterns in contexts would be to establish a hierarchical system that determines how contexts containing structures involving markers that generally indicate different relations or sub-relations should be classified. However, here again the dilemma of preserving maximum information with minimum chance of misleading the user must be confronted. For example, in the case of contexts containing markers of CREATION and of MODIFICATION, the most pertinent information appears to be that of the MODIFICATION and its nature, with the marker of CREATION bolstering the causal element in the relation. In this case, the contexts could thus be classified relatively safely according to the MODIFICATION sub-relation present. The precedence of markers of the MODIFICATION sub-relation is therefore different when such markers are combined with a marker of ASSOCIATION and with another CAUSE–EFFECT marker. Thus, if a relation hierarchy were created to manage these types of occurrences, first priority in classification would likely be assigned to ASSOCIATION markers, followed by markers of MODIFICATION, and finally to other types of markers. However, this is only a partial portrait of the occurrences found in the corpus, and such a hierarchy would need to be thoroughly tested and evaluated before being implemented. Furthermore, the introduction of additional relations or sub-relations into such a system — should the need arise — would likely increase the complexity of the task considerably.

Another option would simply be to classify contexts containing multiple relation markers into separate categories, according to the combinations of markers observed. However, this would clearly increase the complexity not only of sorting the contexts,

but also of consulting them in the final results. As the goal of tools is to simplify the consultation of information for the user, this does not seem to be advisable.

5.5.4 Variation in expression of related elements: Some implications for knowledge extraction

The use of a term-based methodology such as that used in this project for identifying and extracting relation occurrences may also be affected by the phenomenon of variation in the forms in which concepts are represented at a textual level. The specific phenomenon of terminological variation can occur in many forms that are important to take into account in applications such as information extraction (cf. Daille 2005 for an overview of these kinds of variations, and Condamines and Rebeyrolle (2001) for a brief discussion of this phenomenon in pattern-based applications). In addition, concepts may be represented by non-terminological units; some examples of these were given in the discussion of non-nominal related elements such as propositions. Of course, any tool that searches for specific terms will encounter problems when such variations occur.

In the results of the analysis in the English and French data, some recurring variations were observed, and could be taken into account in planning knowledge-extraction approaches to maximize recall, group together similar occurrences of relations and/or to target specific types of information. One such widespread variation often occurs in the description of the effects of a given treatment. For instance, in Examples 387 to 393, a molecule or drug is said to have a given effect:

387. Interleukin-6 is an upstream proinflammatory cytokine that **induces** both CRP and fibrinogen expression. (Rackley 2004)

388. These studies found no consistent associations between statins' **effects on** CRP and lipid levels. (Balk et al. 2003)

389. Statins do not **affect** fibrinogen levels, and limited data suggest little effect on lipid oxidation, tissue plasminogen activator, or plasminogen activator inhibitor. (Balk et al. 2003)

390. Consistent with previous studies, Virulizin had a high level of **anti**-tumor **activity** against human breast, ovarian and prostate tumor xenografts. (Du et al. 2003)

391. [I]n future tamoxifen may even **help to prevent** cancer development. (Health News 1991)

392. Les mécanismes responsables des différences d'**effet** entre les statines lipophiles et hydrophiles **sur** la prolifération des CML ne sont pas encore élucidés... (Nalbone et al. 2002)

393.... ce qui **implique** Cox2 **dans** la prolifération tumorale... (Guastalla et al. 2004)

In Examples 394 to 400, however, alternative expressions may be observed, in which it is not a molecule or drug (i.e., an entity) but rather its presence or administration (i.e., an event) that is identified as the cause of another phenomenon.

394. The presence of TNF- [alpha], IL-6, and other cytokines **cause** [sic] hepatic **production of** C-reactive protein (CRP)... (Pantaleo and Zonszein 2003)

395. Administration of Virulizin showed **anti**-tumor **efficacy** in the treatment of human pancreatic cancers and melanoma in previous preclinical studies... (Du et al. 2003)

396. The results of this study suggest that tamoxifen use may **play a** greater **role in** the development of the endometrioid histologic subtype of endometrial cancer... (Slomovitz et al. 2004)

397. As compared to the treatment with novantrone, which demonstrated **anti**-tumor **efficacy** with an optimal T/C value of 56.8%... (Du et al. 2003)

398.... there is no support for using these markers to identify patients likely to **benefit from** statin treatment. (Balk et al. 2003)

399. Un traitement de 3 mois par la pravastatine **entraîne** une **baisse du** contenu en lipides (et de leur oxydation) des plaques carotidiennes humaines... (Nalbone et al. 2002)

400. Même si ses mécanismes pathogéniques restent incomplètement compris, l'expression de Cox2 **favorise** la prolifération tumorale en inhibant l'apoptose, en stimulant la néo-angiogenèse et en favorisant le pouvoir invasif... (Guastalla et al. 2004)

Cases such as Examples 394 to 400 may be interpreted in several ways: the complex unit may be identified as representing a concept that participates in the relation;

the entity may be identified as the pertinent item (i.e., as corresponding to the concept that should be identified as participating in the relation); the entity may be seen as an instrument and the event as the real cause of the effect; or the two elements (the entity and event) may be seen as cooperating causes. The choice made in this research was the first, to retain the complete form identified in the context as the textual representation of the concept involved in a relationship, without differentiating between events and entities.[190] However, while the annotation of different types of causal elements was not carried out in this research, certainly this possibility exists for situations in which a greater level of detail in classification is required. The choice of interpretation in these cases will affect the choice of strategy for dealing with this variation. The relative regularity of some of these structures may offer possibilities for analysis to identify a component that is considered to be most pertinent for a given application, if desired.

These variations in the expression of related elements may also be pertinent in subsequent applications of the information extracted. Given that more complex structures are less likely to correspond to entries in term databases or nodes in ontologies (i.e., *molecule X* is more likely to be the focus of a term record than *treatment with molecule X* or *administration of molecule X*), the information extracted from contexts such as Examples 394 to 400 may be — while admittedly less elliptical on a conceptual level — less immediately usable.

Another form of variation with a similar impact may be observed when a participant in a relation may be identified at a surface level either as a human patient or as the disease from which this person suffers (or the manifestation of that disease, as in the case of a tumour). This is illustrated in Examples 401 and 402:

---

[190] Some cases in which both entity and event causes of a given effect were present were nevertheless discussed briefly in Section 4.9.1.2.

401. The locoregional management of patients with stage IIIC disease who **respond to** chemotherapy should be individualized. (Shenkier et al. 2004)

402. The treatment of patients with LABC whose tumours do not **respond to** anthracycline-containing chemotherapy is unclear. (Shenkier et al. 2004)

A similar variation may be observed in the expressions of ASSOCIATIONS between a given variable and a disease or other disorder or characteristic; while in some cases the disorder itself is indicated, in other cases, such as Examples 403 to 405, reference is made to the patient group that has this disorder or characteristic:

403. Microalbuminuria (urinary ACR>2 mg/mmol) was **detected in** 32.2% of patients with diabetes and in 14.7% of patients without diabetes. (MacIsaac et al. 2004)

404. Endothelial cell function is **impaired in** patients with atherosclerosis and could antecede the development of overt evidence of the disease. (Griendling and FitzGerald 2003)

405. The presence of TNF- [alpha], IL-6, and other cytokines cause hepatic production of C-reactive protein (CRP), which has been shown to be **elevated in** patients with insulin resistance 66,67 as well as in T1DM patients... (Pantaleo and Zonszein 2003)

A similar phenomenon was also observed in the French data, in the expression of ASSOCIATIONS involving the explicit description of a variable (e.g., treatment with a given drug) as a characteristic of a patient group, as in Example 406:

406. Par ailleurs, le diabète est apparu moins fréquemment **chez** les patients qui étaient traités avec le losartan plutôt qu'avec l'aténolol. (Garnier 2002b)

For the purposes of this project, and according to the definition given for the ASSOCIATION relation in this research, both types of contexts can nevertheless be considered: medical research is of course often carried out in patient groups in order to observe links between given variables, and thus the same ASSOCIATION might be said to exist between the characteristics of patient groups (e.g., microalbuminuria and diabetes) or in the patient groups in which these characteristics were observed (e.g., patients with

microalbuminuria, patients with diabetes). Cases such as those above simply illustrate a variation in the choice of expression for the two variables observed.

While the validity of the relation occurrence was considered to be equal in the two types of expressions for the purposes of this project, however, the facility of identifying the related elements in automatic applications is not. The more complex forms such as those involving the mentions of patient groups are unlikely to correspond to entries in term bases or other terminological resources, and thus would need to be processed either manually or through relatively highly developed automatic analysis before the information extracted from the corpus could be linked to an appropriate term or term record.

This issue may also affect tools that attempt to disambiguate pattern markers using the semantic classes of the elements that they link in contexts. Moreover, tools that use semantic classes of related elements for pattern refinement or marker disambiguation — a possibility discussed, for example, in Marshman and L'Homme (2006) — would need to account for the possible variations in surface realization of conceptually equivalent relations. If this is not done, disambiguation procedures may result in the elimination of pertinent contexts.

While these challenges are certainly significant, their presence in both languages in fairly similar forms indicates possibilities for adapting approaches developed in one language for use in the other. This may significantly increase return on the investment of time and effort in developing these strategies.

5.6 Discussion of semi-automatic and automatic approaches

This research has not focused explicitly on the comparison of automatic and semi-automatic knowledge extraction approaches (i.e., those that attempt to maximize the automation of knowledge extraction and those that assume a certain — and potentially substantial — degree of human interpretation of the results produced by a computer

tool). However, the methodology used in this approach was chosen in large part in order to provide access to data that can inform decisions about the kinds of approaches that may be appropriate for a specific application or situation, including determining appropriate levels of automation. As such, the results provide a basis for discussing the challenges that are involved in increasing the level of automation in pattern-based tools and their likely effect on the proportions of relation occurrences that may be retrieved from a corpus.

The most striking measure of the challenges of automating the extraction and processing of KRCs is that presented in 4.10.4, describing the proportions of the relation occurrences observed that involved at least one of a set of challenges that constitute departures from the most restrictive forms of patterns and relation occurrences, widely accepted as candidates for use in highly automated applications.[191] The proportions of relation occurrences in this category are strikingly high: 75% of the occurrences identified in the English data and 72% in the French data.

Moreover, this figure is only a proportion of occurrences that could potentially be retrieved using the lexical knowledge patterns identified in this evaluation, and the proportion of the total number of relations present in the corpora that would not be located by such restrictive patterns would be even higher. In addition, this level of recall would be reduced by additional difficulties related to phenomena such as variations in pattern or marker forms that are not accounted for in pattern design.

The impact these phenomena have on the potential for identifying occurrences of relations automatically is thus both undeniable and extremely significant. In situations in which a high level of recall is desired or required — for example, when a tool is used in an effort to obtain a complete picture of the information conveyed in a text or text collection, or when a limited amount of data or data with limited redundancy is available

---

[191] The set of challenges included non-nominal related elements, anaphoric expressions, unpredictable interruptions, expressions of uncertainty and text-related issues.

for analysis — a level of silences of over 70% would clearly be unacceptable. Additional strategies would need to be implemented to overcome some of these challenges and provide access to more occurrences, while taking into account the impact these phenomena can have on the usefulness of this information. Conversely, if a large amount of fairly redundant data is available and the goals of using a tool involve obtaining only the most straightforward and/or certain relation occurrences observed, even a fraction of the relation occurrences found in a corpus may be sufficient, and these restrictive pattern forms may be adequate.

It is our belief that terminologists and terminographers are most likely to benefit from access to a wide range of potentially useful information that they can evaluate themselves in order to determine its applicability for a given application. Computer tools are certainly not adequate to take over the kinds of evaluation that this work requires. Rather, they should facilitate rapid and efficient access to textual information — but not at the expense of its completeness. While setting a goal of 100% complete information retrieval with a pattern-based tool is clearly not realistic, a level of recall of less than 25 to 30% of the occurrences that could potentially be located using such an approach is just as clearly undesirable in this kind of work.

Increasing the potential for recall then involves developing strategies for dealing with some of the difficulties identified. These might include measures as simple as ensuring that a tool facilitates access to original texts to help in tasks such as the manual resolution of anaphora, or as complex as the implementation of formalisms that assist in identifying and evaluating pattern and marker interruptions or expressions of uncertainty, or in identifying links between non-nominal forms of related elements and the more conventional nominal forms to which they may correspond. The observations in this research may provide a starting point for developing some of these strategies.

However, every effort to increase recall can be expected to lead to a decrease in precision, which of course reduces the savings in time and effort that pattern-based tools are intended to provide, requiring that users evaluate a larger number of contexts that are

ultimately not pertinent. Restricting the permissible degree of variation from prototypical forms in certain areas is one way of controlling this increase in noise.

Another option could be offered by a hybrid approach that could identify the most prototypical occurrences of relation forms and either process them automatically or present them to users as the most promising of the contexts retrieved, while still retrieving additional, less prototypical contexts as a complement to this information. The types of variations from prototypical form present in these additional contexts could also be used to provide an indication of their potential value and challenges. While a comprehensive approach that can identify and deal with the various types of difficulties would involve a significant investment of time in tool and pattern design, the possibility of gradually expanding the coverage offered by a basic and restrictive tool — perhaps by first targeting the issues that are identified as most pertinent for a given application, relation or language — could provide a strategy for expanding coverage and improving tool performance.

5.7 Limits of this work

This work has, in our opinion, shed light on some interesting and pertinent aspects of the nature and behaviour of knowledge patterns and pattern markers in English and French that are worthy of consideration in the development and use of pattern-based tools, and of further research. However, it is also important to recognize the limits of this study, considering both the perspective of the work and the methodology used.

The broad perspective on potential pattern-based approaches and thus on what may constitute useful information envisioned in this project necessarily limits the specificity with which each individual type of application may be considered. While we felt that this kind of inclusive perspective was necessary in light of the novelty of interlinguistic evaluation in the field, it is clear that the results of this study should be supplemented by additional data selected to reflect the primary concerns of specific

approaches and situations. The results of this work therefore constitute an indication of avenues for further research in many areas, and a starting point for making observations and formulating hypotheses.

The relations chosen as the focus of the research are also clearly pertinent to the scope of the conclusions that can be drawn on the strength of this study. Even in these two relations, which present a certain number of similarities, significant inter-relational differences (in numbers of occurrences, in the nature and variety of markers, and even in differences between the English and French data) were identified. The likelihood of encountering equally — or even more — significant differences in knowledge patterns, pattern markers and their behaviour between these and other relations (for example, the commonly evaluated relations of GENERIC–SPECIFIC and PART–WHOLE) is high. As such, it is impossible to conclude that the interlinguistic evaluations in this research will apply to all relations. Rather, they indicate potential foci for research that should be expanded to address these other relations. The choice to evaluate ASSOCIATION, a relation that is closely linked to the fields of medicine and epidemiology, also imposes certain limitations on the scope of the conclusions that may be drawn from this project, as the relation may not be present or pertinent in other fields.

The choice of methodology for any work necessarily imposes certain limits on the scope of the observations that can be drawn from it. At this level, the nature of the corpora and terms used to extract the samples of contexts for analysis, the volume of data analyzed, and the specific types of analyses carried out are pertinent for evaluation.

As is the case with any corpus-based project, the results of this research are largely dependent on the corpora from which the observations were drawn. Due to practical limitations (including those on the availability of sources and texts) the size and scope of the corpora (i.e., domains, text types) are necessarily restricted. In addition, while the corpora were designed to be as comparable as possible, such limiting factors necessarily introduce the potential for variation resulting from the nature of the texts included and the distribution of different types of texts, as well as the overall size of the

two corpora. The conclusions drawn on the basis of this research thus apply to the types of texts chosen for inclusion in the corpora. Of course, the results of this research should be complemented by additional research in a variety of corpora representing different domains, sub-domains and text types, in order to better characterize the scope of the variations observed.

Moreover, while the corpora were designed to represent comparable sub-domains and aspects of those sub-domains (i.e., etiology, development, effects, diagnosis, prevention and treatment of the diseases in question) as well as text genres, at a more specific level there is inevitably variation in the content of the individual texts, which may directly or indirectly influence the observations in the research. Few options are available to eliminate such sources of variation. The most obvious is of course the use of parallel texts (i.e., an original and a translation). However, this approach would introduce the equally problematic possibility of language interference or calquing of structures from the source text in the target text.

The choice of term selection methodology also clearly affects the results obtained and the interpretations that can be made of these. The contexts analyzed were considered to be representative of those likely to be pertinent for terminologists in their work in concept analysis and terminological description, and the methodology to reflect commonly used approaches in pattern discovery (and therefore to provide a valid basis for comparison of the productivity that can be expected in such an approach). However, this methodology requires the selection among a number of candidate terms for use in extracting initial contexts, which may in turn be expected to influence the kinds of contexts that are retrieved and retained for analysis, and the types of relationships that may be observed. The investigation of other types of approaches could complement this term-based work and provide additional information.

As the comparison of the numbers of relations identified using the selected terms — in addition to the discussion in Section 5.5.2.1 — illustrates, the results may show differences related to the specific terms used. These may be linked to the status of these

terms as equivalents or non-equivalents. The evaluation of both equivalent and non-equivalent terms in this work has provided concrete data that may guide the development of future research and suggest strategies that may help to provide a comparable basis for future analyses. However, the range of other factors that may contribute to observed differences (for example, the volume of data, the varying distribution of term sets among different semantic classes and of relation occurrences among the relations and sub-relations, and of course the possibility of the interaction among these and other factors) does not allow for the effects of this specific aspect of term choice to be identified with certainty. As such, this question remains a subject for future evaluation that may shed additional light on the kinds of interlinguistic differences that may be observed.

The choice of a single classification for the types of CAUSE–EFFECT relations observed in the corpora — in this case that of Barrière (2002), which was considered to be appropriate both for the domain and for the application evaluated in this study — necessarily involved the setting aside of a number of other potential bases for classification, and as a result some potentially interesting criteria for relation classification were not exploited in this research. Different classifications of the relation that highlight additional distinctions among the specific types of relations could reveal more pertinent data about the relations in the domain and the markers that denote them in the two languages.

The volume of data analyzed of course also influences the strength of the conclusions drawn from this study. Given the fact that a range of phenomena — some more common than others — were analyzed, some of the more specific and rarer factors were evaluated in a smaller pool of data than those that were more common. The amount of data available for the analysis of these factors must be considered when evaluating the strength of the evidence. The use of statistical measures such as the Chi-square test was intended to assist in the consideration of this factor, but this test clearly cannot completely neutralize the influence of this variation.

Some aspects of the methodology and limitations on the data gathered imposed restrictions on possibilities for evaluating marker polysemy and precision. The evaluation of a sufficient variety of markers and number of contexts for each one to obtain an accurate portrait of the variations that may be observed was unfortunately beyond the scope of this project. Further research should be pursued to fill this gap. The structure of the study and the nature of the data observed also precluded the use of tests to confirm the statistical significance of certain differences observed (for example, of the frequency, variety and variation of the markers observed in the two data sets). As such, these evaluations do not offer statistically conclusive evidence but rather are indicative of the potential for interlinguistic variation and of the need for further study in a structure that allows for more precise evaluation from a statistical standpoint.

Moreover, from a statistical perspective, it should be kept in mind that a threshold of 0.05 for significance of results of Chi-square tests essentially admits the possibility that one in twenty tests may indicate a significant difference where such a difference is not in fact present. Given the numbers of Chi-square tests carried out in this study, the possibility of observing a small number of apparently significant differences as a result of chance should not be disregarded. (Clearly, however, the smaller the *p* value identified for a given difference, the less likely it is to be a result of chance alone.)

Finally, as noted in a more specific context above, the potential for complex direct and indirect interactions of many of the factors observed in this study poses immense challenges for the evaluation of the ultimate effect of the differences observed in the French and English data. It is far beyond the scope of this project to provide a specific "recipe" for the construction of a bilingual tool for KRC extraction that can produce perfectly comparable results. However, in highlighting a number of potential differences that are likely to be pertinent in specific cases, this work will allow researchers to identify potential sources of difficulties and subjects for future evaluation and research.

Conclusion

This research was carried out with two objectives: to observe lexical knowledge patterns for the conceptual ASSOCIATION and CAUSE–EFFECT relations in English and French specialized medical texts and to explore several aspects of their nature and behaviour, as well as their implications for the development and performance of pattern-based tools for extracting candidate knowledge-rich contexts (KRCs) from these texts and the ultimate use of these candidate KRCs for the purposes of terminological analysis and description; and to compare these observations in the English and French corpora to identify similarities and differences in the results that may affect these applications. This comparison did reveal both similarities and significant differences that should be considered in projects that aim to implement knowledge patterns for locating KRCs.

With reference to the work of a number of researchers (e.g., Pearson 1998; Meyer et al. 1999; Séguéla 1999; Barrière 2001, 2002; Condamines and Rebeyrolle 2001; Meyer 2001; Bowker 2003; Feliu 2004), various types of projects in which knowledge patterns may be used were identified, and some of the choices that must be made in the design and use of pattern-based tools revealed. Also identified were a number of characteristics of knowledge patterns and some additional challenges that may influence these tasks.

The evaluation of these characteristics and challenges began with the analysis of occurrences of candidate terms identified as specific to the corpora, in order to extract contexts indicating the relations of interest and the candidate knowledge patterns associated with them. These occurrences were then annotated to highlight the pertinent factors observed, and the prevalence and characteristics of the various factors evaluated.

The data gathered from the two corpora were then compared and contrasted, in order to reveal similarities and differences between the data in English and French and to identify aspects of pattern-based applications that may be affected by these factors.

The first concrete result of this study is the identification of a number of markers of CAUSE–EFFECT and ASSOCIATION relations, as well as data on several characteristics of these markers and the structures in which they participated that have been identified as pertinent by various researchers. Many of these markers are promising candidates for refinement and inclusion in pattern sets for semi-automatic KRC-extraction tools in specialized English and French medical texts similar to those used in this study, and the data gathered about their behaviour may help developers to determine the most appropriate situations in which to implement these markers.

The use of Barrière's classification of CAUSE–EFFECT relations provided evidence of the usefulness and appropriateness of this system for pattern-based tools operating in both languages, and also offered an opportunity to consider potential strategies for refining the classification if required in a given situation. Moreover, the study provided data that can assist in refining the analysis of the types of ASSOCIATION relations found in the corpora.

In adopting a broad perspective on KRC extraction that included a wide range of potentially useful information and pattern forms, the research provided access to data that may be useful not only for basic applications (e.g., using character strings representing markers), but also in a number of potential adaptations and refinements of the basic approach (e.g., the use of lexico-syntactic knowledge patterns or further processing of candidate KRCs).

This perspective also allowed many of the difficulties that may affect pattern-based tool performance (including several observed but not studied systematically in other projects) to be evaluated and quantified in the contexts analyzed in the two languages. These results revealed what may be termed the "opportunity cost" of the choice to limit analysis to contexts that correspond to a certain set of criteria (e.g., the part of speech of relation markers, form of related elements, pattern structures, lack of intervening elements). This cost, and in particular any differences observed in the data

in the two languages, are essential to consider in the process of developing and using tools for a given application.

At a general level, the data in the two languages showed some striking similarities. The knowledge-pattern-based approach was revealed to be productive in the two languages, and the presence of a number of recurrent and relatively frequent and precise markers of both relations in the two languages shows that pattern-based tools can certainly be effective for extracting information about these relations, which can assist in the task of conceptual analysis and terminological description. In the data analyzed in English and French, similar distributions of relation occurrences between the ASSOCIATION and CAUSE–EFFECT relations and between the groups of contexts identified using terms belonging to various classes (i.e., entities, activities, pathologies, processes) were observed. The forms of these markers (e.g., simple or complex) were largely similar. The prevalence of contexts containing two or more elements that shared a role in a relation, as well as the types of relationships that were observed between these elements, also showed strong similarities in the two data sets. Moreover, overall similarities were observed in the nature of many of the challenges affecting tool development and performance and the ultimate use of the KRCs extracted, as well as the proportions of relation occurrences identified as presenting these difficulties.

This research thus provides clear and concrete evidence indicating that the creation of pattern-based tools for use in a bilingual context is a promising avenue for development, as the overall possibilities for extracting information using this approach and many of the areas on which further research may be concentrated show similarities. Thus, many general strategies may be equally or at least similarly viable in both languages, facilitating bilingual tool development. Moreover, progress made in one language in some areas of the field is also likely to be profitable at least to a certain extent in the other language.

However, in other aspects of knowledge patterns' characteristics and the challenges in their use — and often underlying these overall similarities at a more specific level — some significant differences were observed. These indicate a need for further evaluation in light of the observations in this study, as well as for careful consideration in developing tools.

In designing pattern-based tools, it will be important to consider the fact that in the two languages, the part of speech classes of the markers — and even more so of marker occurrences — may vary, and thus different types of markers may be productive. Moreover, the potential for locating candidate KRCs using the marker sets observed — a function of their variety and frequency in corpora — was observed to differ in the two data sets. This indicates that developers must take into account the possibility that more markers may be required in pattern sets in French in order to retrieve a number of contexts that is comparable to English results. Moreover, as markers were observed to appear in a variety of forms and structures in the two languages, the need for more markers in French could lead to a significant increase in the number of pattern forms required for some applications.

The process of designing pattern forms for use in KRC-extraction tools may also be better adapted to the two languages in light of the information gathered in this research. Especially pertinent is the observation of challenges particular to one or other of the English and French data sets in regard to different factors that influence aspects of pattern development. This suggests that while there are certainly challenges in this area in both languages, the sources in each — and therefore the choices to be made and the strategies for dealing with issues — may well be quite different. The frequency with which complex markers in the English data were interrupted indicates a particular need for representing these markers in a way that allows contexts in which this phenomenon occurs to be identified. The potential for increased variability in some aspects of marker forms in this language also contributes to the need for more or more flexible pattern forms in this language. The prevalence of variation in the form of related elements in the

French data (e.g., non-nominal elements, anaphora, interruptions) may influence the choice of approaches in pattern design, particularly for tools that attempt to target contexts with specific structures or that attempt to sort contexts according to such criteria. Both of these differences may affect the choices made in terms of the type and specificity of pattern representation and/or in the investment of time and effort required to develop pattern sets. The fact that different components of the pattern are involved indicates that different types of tools may be affected (a wider range in English, and a more restricted range in French that involve the representation or processing of related elements at some level).

Once tools are developed, the comparability of their performance may also be affected in the two languages by issues that have not been (completely) resolved in planning stages. The potential recall of pattern-based tools will clearly be affected by the variety and frequency of markers, and there are indications in the data observed that the French markers identified may be less productive on this level. Conversely, observations in a small sample of data on marker precision indicated that the French markers evaluated (at least in the form of character strings) were somewhat more precise than the English. It is possible, if this trend continues in evaluations of more data, that some tools may provide fewer contexts, but a higher percentage of valid occurrences in this language. Clearly, the challenges mentioned above relating to the form and nature of the elements linked by markers in the French data could affect the performance of tools that represent these elements as part of knowledge patterns, or try to identify them automatically.

Possibilities for further processing and use of extracted candidate KRCs may also be influenced by interlinguistic differences. Expressions of uncertainty appearing in candidate KRCs may affect the information value of these contexts, and the form of some of these expressions may offer cues for sorting the contexts according to their value for relation identification. In other cases, the unpredictability of these expressions, rather than offering an opportunity, may interfere with identification and processing of

contexts retrieved. Interestingly, a higher prevalence of expressions likely to introduce both possibilities and challenges was noted in the English data. In the French occurrences, the presence of uncertainty was often indicated by the variation in the form of verbal markers — another potentially useful formal cue. These differences in the two data sets indicate that not only the possibilities of exploiting this kind of information, but also the strategies for doing so, will once again vary between the two languages.

When a set of statistically evaluated challenges that may affect these various aspects of pattern-based tool development and use are considered as a group (including anaphora, non-nominal related elements, expressions of uncertainty, and "unpredictable" interruptions of markers), the proportions of the relation occurrences affected in the two data sets is quite comparable. This reveals an interesting point in the interlinguistic comparison: various individual factors are likely to interact, often in quite complex ways, and differences in one aspect of pattern characteristics or challenges may in some sense compensate for another variation. (For example, it is possible to imagine a case in which more potentially useful French contexts are excluded by tools that require that markers occur contiguously with related elements expressed in nominal forms, but in which these same types of pattern forms exclude more contexts in English because of factors such as variations in marker forms and expressions of uncertainty that interrupt the pattern structure). However, as the sources of these issues differ, so do the possibilities and strategies for dealing with them. Improving pattern-based tool performance will likely involve focusing on areas that are specifically relevant to processing texts in a given language.

Another interesting observation made in this research is that considerably more obvious differences were identified between sets of contexts corresponding to the two relations than between those corresponding to the two languages. The number of relation occurrences, the number of markers, the types of markers and a number of other factors varied substantially from relation to relation. This raises an important point for future research: it is essential to study these aspects of pattern form and behaviour in

more detail for each relation and a wider range of individual relations in order to better evaluate the adjustments that may be necessary in tools that attempt to identify occurrences of different types of relations.

Moreover, some reflections on the methodology used to observe the knowledge patterns may also be drawn from this experience. The approach, one that is widely used in the field and focuses on the identification of relation occurrences and knowledge patterns in contexts identified using (candidate) domain terms, attempted to neutralize certain sources of bias (e.g., related to potential associations between semantic classes of terms and the relations in which they may participate). However, the bilingual orientation of this project and the use of terms that were equivalents and non-equivalents in the two languages allowed for the observation of potential variations between these two groups of terms (i.e., equivalent pairs and non-equivalents) in respect to many of the evaluated characteristics. The precise sources of the differences observed and the mechanisms that may produce them are interesting and important subjects for future work. Moreover, these results suggest that future research (particularly with a bilingual orientation) should at least consider the possibility of term-linked variation and evaluate methodologies accordingly.

This work has thus revealed both the considerable possibilities for developing tools that can support terminologists working in a bilingual environment, and the real need to be aware of specific and potentially language-linked issues in pattern-based tool development and use. It has offered concrete evidence of the impact that certain choices may have on the effectiveness of tools for identifying KRCs in texts, and has indicated some of the areas that are likely to be particularly pertinent in each language for further research. It has targeted some potential strategies for further work to help improve the results of pattern-based tools and to avoid specific pitfalls in these areas. It has also provided quantitative and qualitative descriptions of various factors that may affect tool development and performance, to help those carrying out further work in the field adjust

their expectations and better plan research and development of tools that can meet the specific needs of a situation.

In conclusion, this research has succeeded in its aim of providing a preliminary evaluation and interlinguistic comparison of a number of factors that have been identified as pertinent in extracting KRCs using knowledge patterns, to analyze how potential interlinguistic differences may influence the development and use of pattern-based tools. This is nevertheless one of only very few interlinguistic comparisons if its kind, and as such will need to be complemented by a range of other studies. The observations in this work have identified some areas in which further evaluation is essential, and raised a number of questions about developing and using pattern-based tools in English and French that may be researched in light of more data. However, the study has also revealed a promising future for developing bilingual applications, as indicated by some significant similarities in the two corpora.

Future work

Given the nature of this research, focusing on exploring a new perspective on pattern-based tool development and performance — and the resulting goal of providing a general overview of phenomena that may be observed in connection with lexical knowledge patterns, and how these may be affected by language — success in the project lies not identifying certain answers but rather in raising questions for further evaluation. In this, the research has succeeded: a number of the factors evaluated show a strong potential for interlinguistic difference, and many challenges for developing both unilingual and bilingual tools have been identified, quantified and characterized in the corpora evaluated here.

The next steps involve further study of the factors evaluated in light of the needs of specific applications and in additional data, in order to develop new strategies for dealing with challenges and creating the most effective bilingual tools possible. Each project may target the factors considered most pertinent for its specific goals, assisted by

the evaluation carried out here of the impact of the various phenomena in the English and French data.

Due to the orientation of this research and some resulting methodological choices, the data gathered in this project were not sufficient or appropriate for the comprehensive evaluation of some aspects of the use of pattern-based tools. Analysis of large numbers of occurrences of markers that are particularly appropriate for a given application will provide a more reliable basis for evaluating factors such as marker precision, marker polysemy, and variation in marker forms and pattern structures. Such an evaluation is essential for filling the gaps in knowledge about how pattern-based tools function in the two languages.

Moreover, the evaluation of how these various factors may interact and influence one another and the overall results of pattern-based tool design and use is a complex task that will require additional study. The development and implementation of pattern-based tools for specific applications, and the analysis of their performance in both languages, will provide concrete evidence that may help to examine these questions more closely.

One approach to observing more specific aspects of the phenomena studied here could focus on specific sub-corpora (in the spirit of evaluations such as those carried out by Condamines 2002 and Bowker 2003), to allow for a more comprehensive and detailed evaluation. The corpora used here could be sub-divided and analyzed according to language variety (geographical origin, level of specialization, etc.), text type or segment of text (e.g., research or review articles; texts representing different types or grades of evidence; abstracts, introductions, results, and conclusions), sub-domain (heart disease or cancer; etiology, development, diagnosis, treatment), and so on. Alternatively, more specifically oriented corpora, or corpora that focus on texts that differ from those used here in one or more of the criteria indicated above, could be built and analyzed. Each of these distinctions may reveal interesting phenomena that would allow for a

more fine-grained analysis of the phenomena likely to be observed both unilingually and bilingually.

Another more specialized study that is clearly worth pursuing is that of inter-relational differences that may be observed in many of the aspects of markers and the patterns in which they participate. This research provided an opportunity to explore differences between relations in respect to a number of factors; given the significant divergences observed in these evaluations, it would be interesting to evaluate other issues to examine the ways the identified differences — and others not yet revealed — may directly or indirectly affect tool development and performance. The evaluation of other relations both individually and in combination with those evaluated here could also reveal important phenomena affecting pattern-based tool development and use in the two languages.

In addition, a number of other observations made in the course of this project indicate a need for further study. Some specific links between relations and/or relation markers and particular items or classes of items with which they may be used were identified (similar to observations in Feliu 2004; Weilgaard 2004; Bodson 2005; and Marshman and L'Homme 2006). Further research into this phenomenon may be useful for applications such as pattern marker disambiguation, and thus in developing strategies for refining pattern forms and increasing precision of pattern-based tools. The evaluation of these specificities may also be pertinent in the choice of patterns used in a given application, and even in the description of relation markers as part of the phraseology of a given domain.

The comparison of sub-groups of the contexts located using equivalent and non-equivalent terms raised a number of questions. Some of these differences (e.g., in the numbers of relation occurrences observed) are more clearly and directly linked to the terms used, and thus more easily explained. Others do not appear to have such simple explanations, and seem likely to result from more complex interactions of various factors with one another. Further research with more data is advisable in order to

develop a more comprehensive understanding of the phenomena that contribute to these differences.

One approach to further research that would also provide a new perspective on the issues evaluated in this study at a more general level would be the evaluation of knowledge patterns as they may be observed in parallel texts (i.e., original texts and their translations). Such a study would provide another perspective on interlinguistic variations that may be considered as a complement to and as a basis for comparison with this work's observations of how conceptual relations are expressed in the two languages.

The final — but certainly not least important — of the suggestions for future work to be discussed here involves evaluating the possibilities and challenges of developing and using pattern-based knowledge extraction techniques in more languages. The Canadian context in which this work was carried out identified English and French as primary candidates for this kind of evaluation. However, the multilingual nature of terminology work makes the discovery of knowledge patterns and the evaluation of their effectiveness for identifying KRCs in other languages a promising avenue for further work. In a North and South American context, Spanish and Portuguese would be logical next languages for evaluation; at a more international level, it would be interesting to study languages that are likely to present very different opportunities and challenges in order to evaluate the potential for developing tools with as wide an applicability as possible.

Works Cited and Consulted

Ahmad, K. and A.E. Davies. (1994). "Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar." *Internationales Institut für Terminologieforschung Journal* 5(2): 22–52.

Ahmad, K. and H. Fulford. (1992). "Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology." (Computing Sciences Report CS-92-07). Guildford: University of Surrey.

Ahmad, K. and M. Rogers. (2001). "Corpus Linguistics and Terminology Extraction." In S.E. Wright and G. Budin, eds. *Handbook of Terminology Management*, vol. 2. Amsterdam/Philadelphia: John Benjamins. pp. 725-760.

Aijmer, K. (1986). "Discourse variation and hedging." In Aarts, J. and W. Meijs, eds. *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*. Amsterdam: Rodopi.

Aristotle. (1970). *Aristotle's Physics, Books I and II*. Trans., introduction and notes W. Charlton. Oxford: Clarendon Press.

Aristotle. (1998). *Metaphysics*. Trans. and introduction H. Lawson-Tancred. London: Penguin Books.

Aristotle. *Physics Book II*. trans. R.P. Hardie and R.K. Gaye. In *The Pocket Aristotle*. J. Kaplan, ed. trans. under the direction of W.D. Ross. (1958). New York: Pocket Books. 26–46.

Arkin, H. and R.R. Colton. (1963). *Tables for Statisticians*, 2nd edition. New York/London: Barnes and Noble.

Auger, A. (1997). "Repérage d'énoncés d'intérêt définitoire dans les bases de données textuelles." Doctoral thesis, University of Neuchâtel.

Barrière, C. (1996). "Including certainty terms into a knowledge base of conceptual graphs." In *Actes du Colloque Linguistique et Informatique de Montréal*, CLIM'96. 184–191.

Barrière, C. (2001). "Investigating the Causal Relation in Informative Texts." *Terminology*, 7(2): 135–154.

Barrière, C. (2001a). "Causal Links in Semi-Technical Texts: Discovery, Classification and Representation." Unpublished manuscript.

Barrière, C. (2002). "Hierarchical Refinement and Representation of the Causal Relation." *Terminology*, 8(1): 91–111.

Barrière, C. and M. Hermet. (2002). "Causality Taking Root in Terminology." In *Proceedings of Terminology and Knowledge Engineering, TKE 2002.* 15–20. Nancy, France, 28–30 August 2002.

Bodson, C. (2005). "Termes et relations sémantiques en corpus spécialisés : rapport entre patrons de relations sémantiques (PRS) et types sémantiques (TS)." Doctoral thesis, Département de linguistique et de traduction, Université de Montréal.

Borillo, A. (1996). "Diversités des sources : La relation partie-tout et la structure [N1 à N2] en français." *Faits de langues* 7: 111–120.

Bouaud, J., B. Bachimont, J. Charlet and P. Zweigenbaum. (1995). "Methodological Principles for Structuring an 'Ontology'." DIAM Rapport Interne RI-95-148. *IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing.* Montreal, Canada, 19–25 August 1995.

Boudreau, S. (2005). "Résolution d'anaphores et identification de chaînes de coréférence selon le type de texte." Master's thesis, Département de linguistique et de traduction, Université de Montréal.

Boudreau, S. and R. Kittredge. "Résolution des anaphores et détermination des chaînes de coréférences. Différences entre variétés de textes." *Traitement automatique des langues (TAL)* 46(1): 41–70.

Bourigault D., C. Fabre, C. Frérot, M.-P. Jacques and S. Ozdowska. (2005). "Syntex, analyseur syntaxique de corpus." In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, vol. 2. 17–20. Dourdan, France, 6–10 June 2005.

Bowden, P.R., P. Halstead, and T.G. Rose. (1996). "Extracting Conceptual Knowledge from Text Using Explicit Relation Markers." In N. Shadbolt, K. O'Hara and G.

Schreiber, eds. *Advances in Knowledge Acquisition*, Proceedings of the 9th European Knowledge Acquisition Workshop, EKAW'96. 147–162. Nottingham, U.K., May 1996.

Bowker, L. (2003). "Lexical Knowledge Patterns, Semantic Relations, and Language Varieties: Exploring the Possibilities for Refining Information Retrieval in an International Context." *Cataloging & Classification Quarterly* 37(1/2): 153–171.

Bowker, L. and J. Pearson. (2002). *Working with Specialized Corpora*. New York: Routledge.

Buitelaar, P. (2000). "Semantic Lexicons: Between Ontology and Terminology." In K.I. Simov and A. Kiryakov, eds. *Ontologies and Lexical Knowledge Bases. Proceedings of Ontolex 2000*. 16–24. Sozopol, Bulgaria, 8-10 September 2000.

Cabré, M.T. (1992). *La terminologie : Théorie, méthode et applications*. Trans. and adapted by M.C. Cormier and J. Humbley. Ottawa: Presses de l'Université d'Ottawa.

Cabré, M.-T., J. Morel and C. Tebé. (1996). "Las relaciones conceptuales de tipo causal: un caso práctico." *Actas del V Simposio Iberamericano de terminologie RITerm.* Mexico City, 3–8 November 1996. http: //www.unilat.org/dtil/MEXICO/ cabremt.html. Last consulted 6 August 2004.

Cabré, M.T., J. Morel and C. Tebé. (2001). "Propuesta metodológica sobre cómo detectar las relaciones conceptuales en los textos a través de una experimentación sobre la relación causa-efecto." In M.T. Cabré and J. Feliu, eds. *La terminología científico-técnica : Reconocimiento, análisis y extracción de información formal y semántica.* 165–170. Barcelona: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra.

Cabré, M.T., C. Bach, R. Eestopà, J. Feliu, G. Martínez and J. Vivaldi. (2004). "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, vol. 1. 87–90. Lisbon, Portugal, 26-28 May 2004.

Carreño Cruz, S.I. (2005). "Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingues." Master's thesis, Département de linguistique et de traduction, Université de Montréal.

Ceusters, W., B. Smith and J. Flanagan. (2003). "Ontology and Medical Terminology: Why Description Logics Are Not Enough." In *Proceedings of the conference Towards an Electronic Patient Record (TEPR 2003)*. San Antonio, 10–14 May 2003. http://ontology.buffalo.edu/medo/TEPR2003.pdf. Last accessed 23 October 2006.

Chaffin, R. and D.J. Herrman. (1988). "The nature of semantic relations: a comparison of two approaches." In M. Evens, ed. *Relational Models of the Lexicon*. 289–334. Cambridge/New York: Cambridge University Press.

Condamines, A. (2000–2). "*Chez* dans un corpus de sciences naturelles : Un marqueur de relation meronymique?" *Cahiers de lexicologie,* 77: 165-187.

Condamines, A. (2002). "Corpus Analysis and Conceptual Relation Patterns." *Terminology,* 8(1): 141–162.

Condamines, A. and J. Rebeyrolle. (2000). "Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode." In J. Charlet, M. Zacklad, G. Kassel and D. Bourigault, eds. *Ingénierie des connaissances, évolutions récentes et nouveaux défis*. 127–147. Paris: Eyrolles.

Condamines, A. and J. Rebeyrolle. (2001). "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CKTB): Method and Results." In D. Bourigault, C. Jacquemin and M.-C. L'Homme, eds. *Recent Advances in Computational Terminology*. 127–148. Amsterdam/Philadelphia: John Benjamins.

Condamines, A. and P. Amsili. (1993). "Terminology between Language and Knowledge: An Example of Terminological Knowledge Base." In *Proceedings*

of Terminology and Knowledge Engineering, TKE'93. 316–323. Frankfurt: INDEKS-Verlag.

Cruse, D. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Daille, B. (2005). "Variations and application-oriented terminology engineering." *Terminology* 11(1): 181–197.

Daille, B., B. Habert, C. Jacquemin and J. Royauté. 1996. "Empirical observation of term variations and principles for their description." *Terminology* 3(2): 197–257.

Daugaard, J. and S. Kirchmeier-Andersen. (1995). "The Odense Valency Dictionary Programme for Verb Coding." In J. Daugaard, ed. *Odense Working Papers in Language and Communication* 8. 3–35. Odense: Odense University.

de Keizer, N.F., A. Abu-Hanna and J.H.M. Zwetsloot-Schonk. (2000). "Understanding Terminological Systems I: Terminology and Typology." *Methods of Information in Medicine* 39: 16–21.

de Keizer, N.F., A. Abu-Hanna and J.H.M. Zwetsloot-Schonk. (2000). "Understanding Terminological Systems II: Experience with Conceptual and Formal Representation of Structure." *Methods of Information in Medicine* 39: 22–29.

Dictionnaire de la langue française. Lexis. (1992). Paris: Larousse.

Dorland's Illustrated Medical Dictionary, 28th edition. (1994). Philadelphia: W.B. Saunders.

Drouin, P. (2002). "Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés." Doctoral thesis, Département de linguistique et de traduction, Université de Montréal.

Drouin, P. (2003). "Term Extraction using non-technical corpora as a point of leverage." *Terminology* 9(1): 99–115.

Evens, M., ed. (1988). *Relational Models of the Lexicon*. Cambridge/New York: Cambridge University Press.

Faber, P. et al. (2006). "Process-oriented terminology management in the domain of Coastal Engineering." *Terminology* 12(2): 189–213.

Faber, P., C.I. López Rodriguez and M.I. Tercedor Sánchez. (2001). "Utilización de técnicas de corpus en la reprsesentación del conocimento médico." *Terminology,* 7(2): 167–198.

Feliu, J. (2001). "Propuesta de clases conceptuales y de relaciones conceptuales: recopilación y análisis." In M.T. Cabré and J. Feliu, eds. *La terminología científico-técnica: Reconocimiento, análisis y extracción de información formal y semántica*. 143–154. Barcelona: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra.

Feliu, J. (2004). "Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica." Doctoral thesis, Universitat Pompeu Fabra.

Feliu, J., J. Jairo Giraldo, V. Vidal, J. Vivaldi, and M.T. Cabré. (2004). "The GENOMA-KB Project: A Concept-Based Term Enlargement System." In *Proceedings of the Computational and Computer-Assisted Terminology Workshop, LREC 2004.* 32–35. Lisbon, Portugal, 25 May 2004.

Fellbaum, C., ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferrario, C. and W.B. Strawn. (2006). "Role of the renin-angiotensin-aldosterone system and proinflammatory mediators in cardiovascular disease." *American Journal of Cardiology*, 1 July 2006, 98(1):121–128.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition. New York: John Wiley and Sons.

Flowerdew, J. (1992). "Definitions in science lectures." *Applied Linguistics,* 13(2): 202–221.

Flowerdew, J. (1992a). "Salience in the performance of one speech act: The case of definitions." *Discourse Processes* 15: 165–181.

Fodor, J.A. (1979). "Three Reasons for not Deriving 'Kill' from 'Cause to Die'." In D.J. Napoli and E. Norwood, eds. *Syntactic Argumentation*. 211–230. Washington: Georgetown University Press.

Friedman, G.D. (1994). *Primer of Epidemiology*, 4th edition. New York: McGraw Hill Health Professions Division.

Fuchs, C. (1994). *Paraphrase et énonciation*. Paris: Ophrys.

Galen. "Achieving Coherence of Clinical Data with GALEN." http: //www.opengalen.org/technology/GALEN-tech-overview.html. Last consulted 6 August 2004.

Garcia, D. (1996). "COATIS, un outil d'aide à l'acquisition des connaissances causales exprimées dans les textes." In *Actes du Colloque Linguistique et Informatique de Montréal, CLIM'96*. 97–103. Montreal, Canada, 8–10 June 1996.

Garcia, D. (1997). "Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes." In *Actes des deuxièmes rencontres — Terminologie et Intelligence Artificielle, TIA '97*. 7–26. Toulouse, France, 3–4 April 1997.

Gaudin, F. (1993). *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles*. Rouen: Presses de l'Université de Rouen.

Gillam, L. and K. Ahmad. (2002). "Sharing the knowledge of experts." *Fachsprache* 24(1–2): 2–19.

Gillam, L., M. Tariq and K. Ahmad. (2005). "Terminology and the construction of ontology." *Terminology* 11(1): 55–81.

GRADE Working Group. (2004). "Grading quality of evidence and strength of recommendations." *British Medical Journal* 328: 1490–1494. http://bmj.bmjjournals.com/cgi/reprint/328/7454/1490.pdf. Last accessed 11 October 2006.

GRADE Working Group. (2004a). "Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches." *BMC Health Services Research* 4(1): 38. http://www.gradeworkinggroup.org/ publications/index.htm. Last accessed 11 October 2006.

GRADE Working Group. (2004b). "Systems for grading the quality of evidence and the strength of recommendations II: A pilot study of a new system for grading the

quality of evidence and the strength of recommendations." *BMC Health Services Research* 5(1): 25. http://www.gradeworkinggroup.org/publications/index.htm. Last accessed 11 October 2006.

Grand dictionnaire terminologique (GDT). Office québécoise de la langue française. http: //www.granddictionnaire.com/. Last consulted 8 August 2004.

Greenhalgh, T. (2001). *How to Read a Paper: The Basics of Evidence Based Medicine*, 2nd edition. London: BMJ Books.

Gross, G. (1994). "Classes d'objets et descriptions des verbes." *Langages* 115: 15–30.

Habert, B., A. Nazarenko and A. Salem. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

Halloran M. and A.N. Bradford. (1984). "Figures of Speech in the Rhetoric of Science and Technology." In R.J. Connors et al., eds. *Essays on Classical Rhetoric and Modern Discourse*. 179–192. Carbondale/Edwardsville: Southern Illinois University Press.

Hankinson, J. (1998). *Cause and Explanation in Ancient Greek Thought*. Oxford: Clarendon Press.

Hearst, M. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Proceedings of COLING-92*. 539–545. Nantes, France, 23–28 August, 1992.

Hearst, M. (1998). "Automated discovery of WordNet relations." In C. Fellbaum, ed. *WordNet: An Electronic Lexical Database*. 131–151. Cambridge, MA: MIT Press.

Hennekens, C.H. and J.E. Buring. (1987). *Epidemiology in Medicine*. Mayrent, S.L. (ed.) Boston/Toronto: Little, Brown and Co.

Higgins, J.P.T. and S. Green, eds. (2005). "Assessment of study quality." In *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* [updated May 2005]. The Cochrane Library, Issue 3. Chichester (UK): John Wiley & Sons, Ltd.

Hobbs, J.R. (1978). "Resolving pronoun references." *Lingua* 44: 311–338.

Hoffmann, L. (1976). *Kommunikations mittel Fachsprache — Eine Einführung* [Languages for Special Purposes as a Means of Communication: An Introduction]. Sammlung Akademie-Verlag 44. Berlin: Akademie-Verlag.

Hume, D. (1739/1985). *A Treatise of Human Nature*. Ed. and introduction E.C. Mossner. London: Penguin Books.

IMS Textcorpora and Lexicon Group. (1994). Tree Tagger. http: //www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html. Last accessed 28 February 2006.

Jackiewicz, A. (1996). "L'expression lexicale de la relation d'ingrédience (partie-tout)." *Faits de langues*, 7: 53-62.

Jacquemin, C. and P. Zweigenbaum. (2000). "Traitement automatique des langues pour l'accès au contenu des documents." In J. Le Maître, J. Charlet and C. Garbay, eds. *Le document en sciences du traitement de l'information.* 71–109. Toulouse: Cepadues.

Jouis, C. (1993). "Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : Le système Seek." Thèse de doctorat, École des hautes études en sciences sociales de Paris, France.

Jouis, C., I. Biskri, J.P. Desclés, F. Le Priol, J.P. Meunier, W. Mustafa and G. Nault. (1997). "Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques." *Actes de la première journée scientifique et technique du réseau francophone de l'ingénierie de langue de l'Aupelf-Uref.* 427–432. Avignon, France, 15–16 avril 1997.

Kahane, S. and I. Mel'čuk. (forthcoming). "Les sémantèmes de causation en français."

Kant, I. (1781/1993). *Critique of Pure Reason*. V. Politis, ed. Trans. based on that of J.M.D. Meiklejohn. London/Vermont: J.M. Dent/Tuttle Publishing.

Kernbaum, S., ed. (2001). *Dictionnaire de médecine Flammarion*, 7th edition. Paris: Flammarion.

Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Weisbaden: Oscar Branstetter.

L'Homme, M.-C. (1998). "Le statut du verbe en langue de spécialité et sa description lexicographique." *Cahiers de lexicologie,* 73(2): 61-84.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Montreal: Presses de l'Université de Montréal.

L'Homme, M.-C. and E. Marshman. (2006). "Extracting terminological relationships from specialized corpora." In L. Bowker, ed. *Lexicography, Terminology, Translation: Text-Based Studies in Honour of Ingrid Meyer*. 67–80. Ottawa: University of Ottawa Press.

Lakoff, G. (1975). "Hedges: A study in meaning criteria and the logic of fuzzy concepts." In D. Hockney, W. Harper and B. Freed, eds. *Contemporary Research in Philosophical Logic and Linguistic Semantics*. 221–271. Dordrecht: D. Reidel Publishing Company.

Lappin, S. and H.J. Leass. (1994). "An algorithm for pronominal anaphora resolution." *Computational Linguistics* 20(4): 535–561.

Lauriston, A. (1994). "Automatic recognition of complex terms: Problems and the TERMINO solution." *Terminology* 1(1): 147–170.

Lebart, L. and A. Salem (1994). *Statistique textuelle*. Paris: Dunod.

Levin, B. (1993). *English Verb Classes and Alternations, A Preliminary Investigation*. Chicago: University of Chicago Press.

Liberati, A., R. Buzzetti, R. Grilli, N. Magrini and S. Minozzi. (2001). "Which guidelines can we trust? Assessing strength of evidence behind recommendations for clinical practice." *Western Journal of Medicine* 174(4): 262–265.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Lyons, J. (1977). *Semantics: Volume 1*. Cambridge: Cambridge University Press.

Lyons, J. (1991). *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press.

Lysvåg, P. (1975). "Verbs of hedging." In Kimball, J.P., ed. *Syntax and Semantics*, vol. 4. 125–154. New York/San Francisco/London: Academic Press, subsidiary of Harcourt Brace Jovanovich.

Mach, F. (2005). "Inflammation is a crucial feature of atherosclerosis and a potential target to reduce cardiovascular events." *Handbook of Experimental Pharmacology* (170): 697–722.

Mackie, J.L. (1974). *The Cement of the Universe*. Oxford: Clarendon Press.

Madsen, B.N., B. Sandford Pedersen and H.R. Thomsen. (2001). "Defining Semantic Relations for OntoQuery." In A. Jensen and P. Skadhauge, eds. *Proceedings of the First International OntoQuery Workshop, Ontology-based Interpretation of NP's*. Kolding: Department of Business Communication and Information Science, University of Southern Denmark. http://www.ontoquery.dk/publications/docs/Defining.doc. Last accessed 28 November 2005.

Madsen, B.N., B. Sandford Pedersen and H.R. Thomsen. (2002). "Semantic Relations in Content-based Querying Systems: A Research Presentation from the OntoQuery Project." In Simov, K. and A. Kiryakov, eds. *Ontologies and Lexical Knowledge Bases. Proceedings of the 1st International Workshop, OntoLex 2002*. 72–82. Sofia, Bulgaria, 27 May 2002. Sofia: OntoText Lab.

Madsen, B.N., H.E. Thomsen and C. Vikner. (2002). "Data Modelling and Conceptual Modelling in the Domain of Terminology." In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering, TKE'02*. 77–82. Nancy, France, 28–30 August 2002.

Malaisé, V., P. Zweigenbaum and B. Bachimont. (2005). "Mining defining contexts to help structuring differential ontologies." *Terminology* 11(1): 21–53.

Mann, W.C. and S.A. Thompson. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. ISI/RS-87-190, June 1987. Reprint of *The Structure of Discourse*.

Marina del Rey, CA: Information Sciences Institute, University of Southern California.

Marshman, E. (2002). "The Cause Relation in Biopharmaceutical Corpora: English and French Patterns for Knowledge Extraction." Master's thesis, School of Translation and Interpretation, University of Ottawa.

Marshman, E. (2002a). "The Cause Relation in Biopharmaceutical Texts: Some English Knowledge Patterns." In *Proceedings of Terminology and Knowledge Engineering, TKE 2002*. Nancy, France, 28-30 August 2002. pp. 89-94.

Marshman, E. (2004). "The Cause-Effect Relation in a French-Language Biopharmaceuticals Corpus: Some Lexical Knowledge Patterns." In *Proceedings of the Computational and Computer-Assisted Terminology Workshop, LREC 2004.* Lisbon, Portugal, 25 May 2004. pp. 24-7.

Marshman, E. (2004a). "Marqueurs lexicaux de la relation cause-effet : Des essais de désambiguïsation." Presentation at the workshop "Les approches lexicographiques et terminologiques, sont-elles compatibles ?," ACFAS Congress, Montreal, Canada, 13-14 May 2004.

Marshman, E. and M.-C. L'Homme. (2006). "Disambiguating lexical markers of cause and effect using actantial structures and actant classes." In *Proceedings of the 15th European Symposium on Language for Special Purposes, LSP 2005.* Bergamo, Italy, 29 August – 2 September 2005. 261–285.

Marshman, E. and M.-C. L'Homme. (2006a). "Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés." Journées du CRTT, Lyon, France, 28–29 September 2006.

Marshman, E., T. Morgan and I. Meyer. (2002). "French patterns for expressing concept relations." *Terminology* 8(1): 1–29.

Mel'čuk, I. (in preparation). "Lexical functions."

Mel'čuk, I. and A. Polguère. (2005). DiCo-OLST. http://olst.umontreal.ca/dicouebe/. Consulted 2 June 2005.

Mel'čuk, I., A. Clas and A. Polguère. (1995). *Introduction à la lexicologie explicative et combinatoire*. Brussels: AUPELF-UREF / Éditions Duculot.

Meyer, I. (1994). "Linguistic Strategies and Computer Aids for Knowledge Engineering in Terminology." *L'actualité terminologique/Terminology Update* 27(4): 6–10.

Meyer, I. (2001). "Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework." In D. Bourigault, C. Jacquemin and M.-C. L'Homme, eds. *Recent Advances in Computational Terminology.* 279–302. Amsterdam/Philadelphia: John Benjamins.

Meyer, I. and K. Mackintosh. (1996). "The Corpus from a Terminographer's Viewpoint." *International Journal of Corpus Linguistics,* 1(2): 257-285.

Meyer, I. and K. Mackintosh. (1996a). "Refining the Translator's Concept Analysis Methods: How Can Phraseology Help?" *Terminology* 3(1): 1-26.

Meyer, I., Eck, K. and Skuce, D. (1997). "Systematic Concept Analysis within a Knowledge-Based Approach to Terminology." In S.E. Wright and G. Budin, eds. *Handbook of Terminology Management,* vol. 1. Amsterdam/Philadelphia: John Benjamins. pp. 98-118.

Meyer, I., K. Mackintosh, C. Barrière and T. Morgan. (1999). "Conceptual Sampling for Terminographical Corpus Analysis." In *Proceedings of Terminology and Knowledge Engineering TKE '99.* 256–267. Innsbruck, Austria, 23–27 August 1999.

Miller, G.A. (1998). "Nouns in WordNet." In C. Fellbaum, ed. *WordNet: An Electronic Lexical Database.* 23–46. Cambridge, MA: MIT Press.

Mitkov, R. 1998. "Robust pronoun resolution with limited knowledge." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, vol. II. 869–875. Montreal, Canada, 10–14 August 1998.

Morgan, T. (2000). "A Comparative Study of Hypernymic Patterns for Knowledge Extraction." Master's thesis, School of Translation and Interpretation, University of Ottawa.

Morin, E. (1999). "Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique." *Traitement automatique des langues (TAL)* 40(1): 143–166.

Muller, C. (1968). *Initiation à la statistique linguistique*. Paris: Larousse.

Muller, C. (1973). *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette.

National Institutes of Health. (2006). Medical Subject Headings. http: //www.nlm.nih.gov/mesh/. Last consulted 6 November 2006.

Nazarenko, A. (2000). *La cause et son expression en français*. Paris: Ophrys.

Nazarenko, A., P. Zweigenbaum, B. Habert and J. Bouaud. (2001). "Corpus-based extension of a terminological semantic lexicon." In D. Bourigault, C. Jacquemin and M.-C. L'Homme, eds. *Recent Advances in Computational Terminology.* 327–351. Amsterdam/Philadelphia: John Benjamins.

Nazarenko, A., P. Zweigenbaum, J. Bouaud and B. Habert. (1997). "Corpus-based Identification and Refinement of Semantic Classes." *Journal of the American Medical Informatics Association,* 4(suppl.): 585–589.

NCI — National Cancer Institute. (2002). "Genetic Testing for BRCA1 and BRCA2: It's Your Choice." http: //cis.nci.nih.gov/fact/3_62.htm. Last accessed 9 May 2005.

Norman, G.R. and D.L. Streiner. (2003). *PDQ Statistics*, 3rd edition. Hamilton/London: B.C. Decker Inc.

Nuopponen, A. (1994). "Causal Relations in Terminological Knowledge Representation." *Terminology Science and Research*, 5(1): 36-44.

Nuopponen, A. (1994a). "Wüster revisited: On causal concept relationships and causal concept systems." In M. Brekke, Ø. Andersen, T. Dahl and J. Myking (eds.), *Applications and Implications of Current LSP Research, Proceedings of the 9th European Symposium on LSP*, vol. II. 532–539. Bergen: Fagbokforlaget. http://lipas.uwasa.fi/~atn/papers/artikkelit/LinkedDocuments/Nuopponen_Causal_LSP94.pdf. Last accessed 23 October 2006.

Nuopponen, A. (2005). "Concept Relations: An Update of a Concept Relation Classification." In Madsen, B.N. and H.E. Thomsen (eds.). *Terminology and Content Development: Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*. 127–138. Copenhagen, Denmark, 17–18 August 2005.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ohta, T., Y. Tateisi and J.D. Kim. (2002). "The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain." In *Proceedings of Human Language Technology (HLT) 2002*. 73–77. San Diego, United States, 24–27 March 2002.

Ohta, T., Y. Tateisi, J.D. Kim, H. Mima and J. Tsujii. (2001). "Ontology Based Corpus Annotation and Tools." In *Proceedings of the 12th Workshop on Genome Informatics*. 469–470. http: //www-tsujii.is.s.u-tokyo.ac.jp/GENIA/paper/OhtaGIW01.pdf. Last consulted 8 August 2004.

Oster, U. (2006). "Classifying domain-specific intraterm relations: A schema-based approach." *Terminology* 12(1): 1–17.

Otman, G. (1996). "Expression lexicale de la relation partie-tout : Le traitement automatique de la relation partie-tout en terminologie." *Faits de langues* 7: 43–52.

Otman, G. (1996). *Les représentations sémantiques en terminologie*. Paris: Masson.

Ouellet, P. (1984). "La désénonciation : les instances de la subjectivité dans le discours scientifique." *Protée* 18(2): 45-53.

Ouellet, P. (1985). "La vision des choses : La focalisation dans le discours scientifique." *Protée* 19(1): 33-45.

Oxford English Dictionary (OED) Online. (2006). Oxford: Oxford University Press. http://dictionary.oed.com. Accessed 6 November 2006.

Pearson, J. (1998). *Terms in Context.* Amsterdam/Philadelphia: John Benjamins.

Pearson, J. (1999). "Comment accéder aux éléments définitoires dans les textes spécialisés ?" *Terminologies nouvelles* 19: 21-28.

Picoche, J. (1977). *Précis de lexicologie française*. Paris: Nathan.

Polguère, A. (2003). *Lexicologie et sémantique lexicale : Notions fondamentales*. Montreal: Presses de l'Université de Montréal.

Popovic, S. (2004). *Paraphrasage des liens de fonctions lexicales*. Master's Thesis, Département de linguistique et de traduction, Université de Montreal. http: //www.olst.umontreal.ca/FrEng/MAPhD/PopovicMA2004.pdf. Last accessed 6 August 2004.

Popper, Sir K.R. (1990). "Two New Views of Causality." In *A World of Propensities*. 1–26. Bristol: Thoemmes.

Rebeyrolle, J. (2000). "Forme et fonction de la définition en discours." Doctoral thesis, Université de Toulouse II, Toulouse, France.

Rebeyrolle, J. (2000a). "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes." In *Actes des Journées Francophones d'Ingénierie des Connaissances, IC'*2000. 105–114. Toulouse, France, May 2000.

Rector, A.L., W.D. Solomon, W.A. Nowlan and T.W. Rush. (1994). "A Terminology Server for Medical Language and Medical Information Systems." *Methods of Information in Medicine* 34(1–2): 147–157.

Rodríguez Penagos, C. (2004). "Metalinguistic Information Extraction for Terminology." In *Proceedings of the 3rd International Workshop on Computational Terminology*. In association with COLING 2004, Geneva, Switzerland, 29 August 2004. http://arxiv.org/ftp/cs/papers/0504/0504074.pdf. Last accessed 23 October 2006.

Rodríguez Penagos, C. (2004a). "Mining metalinguistic activity in corpora to create lexical resources using Information Extraction techniques: The MOP system." In *Proceedings of ACL 2004*. 215–222. Barcelona, Spain, 21–26 July 2004.

Pearson, J. (1999). "Comment accéder aux éléments définitoires dans les textes spécialisés ?" *Terminologies nouvelles* 19: 21-28.

Picoche, J. (1977). *Précis de lexicologie française*. Paris: Nathan.

Polguère, A. (2003). *Lexicologie et sémantique lexicale : Notions fondamentales*. Montreal: Presses de l'Université de Montréal.

Popovic, S. (2004). *Paraphrasage des liens de fonctions lexicales*. Master's Thesis, Département de linguistique et de traduction, Université de Montreal. http: //www.olst.umontreal.ca/FrEng/MAPhD/PopovicMA2004.pdf. Last accessed 6 August 2004.

Popper, Sir K.R. (1990). "Two New Views of Causality." In *A World of Propensities*. 1–26. Bristol: Thoemmes.

Rebeyrolle, J. (2000). "Forme et fonction de la définition en discours." Doctoral thesis, Université de Toulouse II.

Rebeyrolle, J. (2000a). "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes." In *Actes des Journées Francophones d'Ingénierie des Connaissances, IC'*2000. 105–114. Toulouse, France, May 2000.

Rector, A.L., W.D. Solomon, W.A. Nowlan and T.W. Rush. (1994). "A Terminology Server for Medical Language and Medical Information Systems." *Methods of Information in Medicine* 34(1–2): 147–157.

Rodríguez Penagos, C. (2004). "Metalinguistic Information Extraction for Terminology." In *Metalinguistic Information Extraction for Terminology, Proceedings of the 3rd International Workshop on Computational Terminology*. In association with COLING 2004, Geneva, Switzerland, 29 August 2004. http://arxiv.org/ftp/cs/papers/0504/0504074.pdf. Last accessed 23 October 2006.

Rodríguez Penagos, C. (2004a). "Mining metalinguistic activity in corpora to create lexical resources using Information Extraction techniques: The MOP system." In *Proceedings of ACL 2004*. 215–222. Barcelona, Spain, 21–26 July 2004.

http://acl.ldc.upenn.edu/acl2004/main/pdf/312_pdf_2-col.pdf. Last accessed 23 October 2006.

Rogers, M. and K. Ahmad. (1994). "Computerised Terminology for Translators: The role of text." In M. Brekke, Ø. Andersen, T. Dahl and J. Myking, eds. *Applications and Implications of Current LSP Research, Proceedings of the 9th European Symposium on LSP*, vol. II. 840–860. Bergen: Fagbokforlaget.

Rondeau, G. (1984). *Introduction à la terminologie*, 2nd edition. Chicoutimi (Quebec): Gaëtan Morin.

Rousselot, F., P. Frath, and R. Oueslati. (1996). "Extracting Concepts and Relations from Corpora." In W. Wahlster (ed.) *Proceedings of the 12th European Conference on Artificial Intelligence ECAI'96*. 74–78. Budapest, Hungary, 11–16 August 1996.

Rundell, M. and P. Stock. (1992). "The Corpus Revolution." *English Today*, 30: 9–14.

Rundell, M. and P. Stock. (1992a). "The Corpus Revolution." *English Today*, 31: 21–32.

Rundell, M. and P. Stock. (1992b). "The Corpus Revolution." *English Today*, 32: 45–51.

Russell, B. (1914). *Our Knowledge of the External World*. London: George Allen and Unwin Ltd.

Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Sager, J.C., D. Dungworth and P.F. MacDonald. (1980). *English Special Languages: Principles and Practice in Science and Technology*. Weisbaden: Oscar Brandstetter Verlag KG.

Schaffer, J. (2003). "The Metaphysics of Causation." In E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* (Spring 2003). http://plato.stanford.edu/archives/spr2003/entries/causation-metaphysics/. Last ccessed 6 August 2004.

Schmid, H. (1994). "Probabilistic part-of-speech tagging using decision trees." In D.B. Jones and H. Somers (eds.). *Proceedings of the International Conference on New Methods in Language Processing*. 44–49. Manchester, United Kingdom, 14–16 September 1994.

Schünemann, H.J., D. Best, G. Vist and A.D. Oxman for the GRADE working group. (2003). "Letters, numbers, symbols, and words: How best to communicate grades of evidence and recommendations?" *Canadian Medical Association Journal* 169(7): 677–680.

Scott, M. (1995). *WordSmith Tools.* http: //www.lexically.net/wordsmith/index.html. Last consulted 16 October 2006.

Séguéla, P. (1999). "Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés." *Terminologies nouvelles* 19(1): 52–60.

Séguéla, P. (2001). "Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques." Doctoral thesis, École doctorale Informatique et Télécommunications, Université Toulouse III.

Shibatani, M. (1972). "Three Reasons for not Deriving 'Kill' from 'Cause to Die' in Japanese." In J.P. Kimball, ed. *Syntax and Semantics, Vol. 1.* 125–137. New York/London: Seminar Press, subsidiary of Harcourt Brace Jovanovich.

Shibatani, M. (1976). "The Grammar of Causative Constructions: A Conspectus." In M. Shibatani, ed. *Syntax and Semantics, Vol. 6: The Grammar of Causative Constructions*. 1–40. New York/San Francisco/London: Academic Press, subsidiary of Harcourt Brace Jovanovich.

Skuce, D. and Kavanagh, J. (1999). "A Document-Oriented Knowledge Management System." In *Proceedings of Terminology and Knowledge Engineering TKE '99.* 320–329. Innsbruck, Austria 23–27 August 1999.

Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley.

Stoll, G. and M. Bendszus. (2006). "Inflammation and atherosclerosis: Novel insights into plaque formation and destabilization." *Stroke*, July 2006, 37(7): 1923– 1932.

Streiner, D.L. and G.R. Norman. (1998). *PDQ Epidemiology*, 2nd edition. Hamilton/London: B.C. Decker Inc.

Swan, M. (1995). *Practical English Usage*, 2nd edition. Oxford: Oxford University Press.

Talmy, L. (1975). "Semantics and Syntax of Motion." In J.P. Kimball, ed. *Syntax and Semantics, Vol. 4*. 181–238. New York/San Francisco/London: Academic Press, subsidiary of Harcourt Brace Jovanovich.

Talmy, L. (1976). "Semantic Causative Types." In M. Shibatani, ed. *Syntax and Semantics, Vol. 6: The Grammar of Causative Constructions*. 43–116. New York/San Francisco/London: Academic Press, subsidiary of Harcourt Brace Jovanovich.

Talmy, L. (1985). "Lexicalization patterns: semantic structure in lexical forms." In T. Shopen, ed. *Language Typology and Syntactic Description, Vol. III: Grammatical Categories and Lexicon*. 57–149. Cambridge: Cambridge University Press.

Talmy, L. (1988). "Force Dynamics in Language and Cognition." *Cognitive Science,* 12: 49–100.

Talmy, L. (2000). *Toward a Cognitive Semantics*, vol. 1 and 2. Cambridge, MA: MIT Press.

Tateisi, Y., T. Ohta, N. Collier, C. Nobata, and J. Tsujii. (2000). "Building an Annotated Corpus in the Molecular-Biology Domain." In P. Buitelaar and K. Hasida, eds. *Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*. 28-34. Luxembourg, 5–6 August 2000. http: //www-tsujii.is.s.u-tokyo.ac.jp/GENIA/paper/coling-ws.pdf. Last consulted 8 August 2004.

Temmerman, R. (2000). *Towards New Ways of Terminological Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.

Unified Medical Language System (UMLS) (Release 2004AB, 2005AA). (2004, 2005) http: //umlsks.nlm.nih.gov/kss/servlet/Turbine/template/admin,user,KSS_login.vm. Last accessed 26 April 2005.

University of California at San Francisco (UCSF)–Stanford University Evidence-based Practice Center. (2001). "Evidence-based Review Methodology." *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*. Evidence Report/Technology Assessment Number 43. http://www.ahcpr.gov/CLINIC/PTSAFETY/chap3.htm. Last accessed 11 October 2006.

Upshur, R.J. (2003). "Are all evidence-based practices alike? Problems in the ranking of evidence." *Canadian Medical Association Journal* 169(7): 672–673.

Vandaele, S. (2001). "Utilisation des bases de données bibliographiques spécialisées en traduction médicale." *Meta* 46(1): 103–116.

Vinay, J.P. and J. Darbelnet. (1958). *Stylistique comparée du français et de l'anglais*. Paris: Beauchemin.

Vivaldi, J. 2003. *Sistema de reconocimiento de terminos* Mercedes. *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Weilgaard, L. (2004). "Valency Patterns of Danish Verbs as Terminological Knowledge Patterns." In *Proceedings of the Computational and Computer-Assisted Terminology Workshop, LREC 2004*. 20–23. Lisbon, Portugal, 25 May 2004.

WordNet. A Lexical Database for the English Language. http: //www.cogsci.princeton.edu/~wn/. Last accessed 6 August 2004.

Wright, S.E. and G. Budin, eds. (1997). *Handbook of Terminology Management*, vol. 1. Amsterdam/Philadelphia: John Benjamins.

Wright, S.E. and G. Budin, eds. (2001). *Handbook of Terminology Management*, vol. 2. Amsterdam/Philadelphia: John Benjamins.

Wüster, E. (1974). "Die allgemeine Terminologielehre — ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften."

[The general theory of terminology — a border field between linguistics, logic, ontology, information science and the subject fields]. *Linguistics* 199: 61–106.

Wüster, E. (1981). "L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses." trans. Bureau des traductions, Secrétariat de l'État du Canada. In G. Rondeau and H. Felber, (eds.) under the direction of V.I. Siforov. *Textes choisis de terminologie.* 55–114. Laval: GIRSTERM.

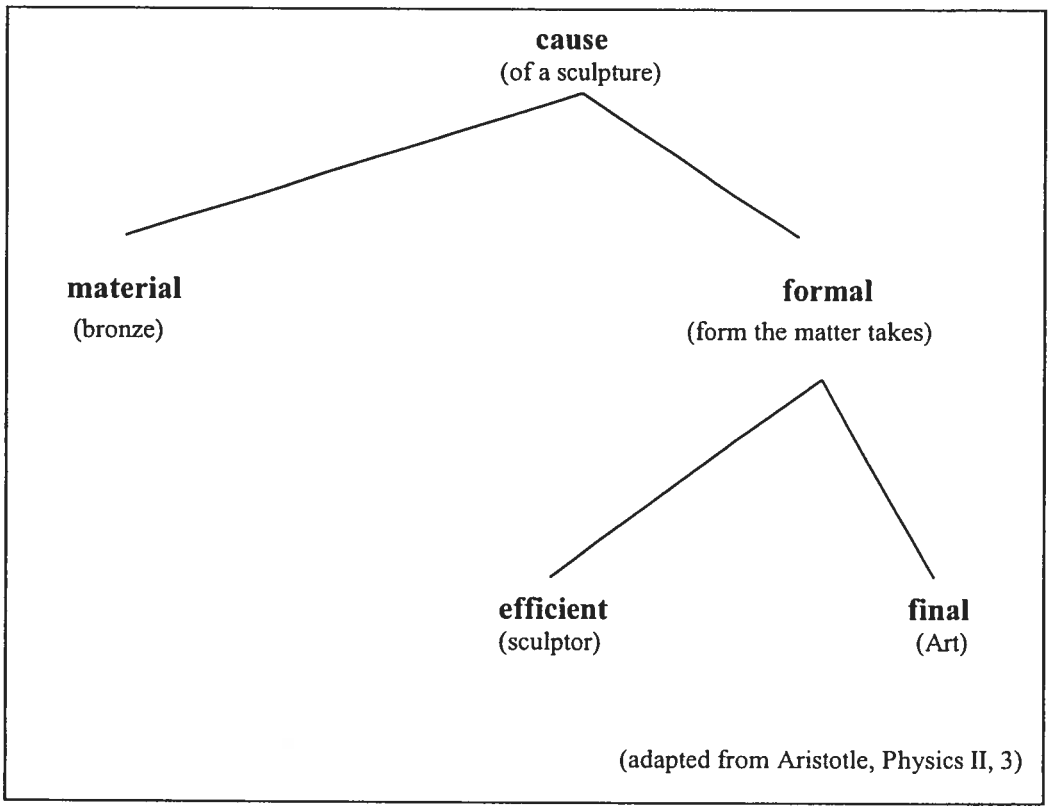
Wüster, E. (2003). "The Wording of the World presented graphically and terminologically." Selected and trans. J.C. Sager. *Terminology* 9(2): 269–297.

Wüster, E. (1985). *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie.* [Introduction to the General Theory of Terminology and terminological lexicography]. Copenhagen: Fashsprachlichen Zentrum, Handelschochschule Kopenhagen.

Zweigenbaum, P. (1994). "MENELAS: An Access System for Medical Records Using Natural Language." *Computer Methods and Programs in Biomedicine* 45: 117–120.

Zweigenbaum, P., B. Bachimont, J. Bouaud, J. Charlet and J.F. Boisvieux. (1995). "A Multi-Lingual Architecture for Building a Normalised Conceptual Representation for Medical Language." In *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care.* 357–361. New Orleans, United States, November 1995.

Appendix A: Aristotle's four causes



Appendix B: Research using knowledge patterns

Citation	Domain	Application	Language(s)	Relation(s)	Methods
Ahmad and Fulford (1992)	Various	Knowledge extraction	English	SYNONYMY HYPERONYMY / HYPONYMY MERONYMY CAUSE–EFFECT MATERIAL	Knowledge probes
Barrière (2001, 2002)	Composting	Information extraction Tool development	English	CAUSE–EFFECT	Lexical knowledge patterns
Bodson (2005)	Medicine Computing	Knowledge extraction	French	HYPERONYMY MERONYMY FUNCTION CAUSE–EFFECT	Lexico-syntactic knowledge patterns Semantic classes
Bowden et al. (1996)	Computer Science	Knowledge extraction	English	DEFINITION EXEMPLIFI-CATION	Lexical knowledge patterns Positive and negative triggers
Bowker (2003)	Infertility	Comparison of patterns in language varieties	French	HYPERONYMY MERONYMY FUNCTION CAUSE–EFFECT	Lexical knowledge patterns
Condamines (2000-2)	Natural sciences	Knowledge extraction	French	MERONYMY	Lexical knowledge pattern
Condamines and Rebeyrolle (2000, 2001)	Software engineering	Knowledge extraction Text modelling	French	VARIOUS	Knowledge patterns Collocations Morphological and semantic similarities

Citation	Domain	Application	Language(s)	Relation(s)	Methods
Feliu (2004)	Human genome	Knowledge extraction Knowledge representation	Catalan	VARIOUS	Lexical knowledge patterns
Garcia (1996, 1997)	Electricity	Knowledge extraction Tool development	French	CAUSE–EFFECT	Lexical knowledge patterns
Gillam et al. (2005)	Nano-technology	Ontology construction	English	HYPERONYMY SYNONYMY	Statistical measures of co-occurrence Lexico-syntactic knowledge patterns
Hearst (1992, 1998)	General language	Knowledge extraction Knowledge representation	English	HYPONYMY	Lexico-syntactic knowledge patterns
Malaisé et al. (2005)	Childhood	Differential ontology construction	French	HYPERONYMY SYNONYMY OTHER RELATIONS IN DEFINITIONS	Lexico-syntactic knowledge patterns Paralinguistic knowledge patterns
Marshman (2002, 2002a, 2004, 2004a)	Pharma-ceuticals (Biophar-maceu-ticals)	Knowledge extraction	English French	CAUSE–EFFECT	Lexical knowledge patterns
Marshman et al. (2002)	Computing Genetics	Knowledge extraction	French	HYPERONYMY MERONYMY FUNCTION	Lexical knowledge patterns
Meyer et al. (1999)	Childbirth	Knowledge extraction	English	CAUSE–EFFECT HYPERONYMY FUNCTION	Knowledge patterns

Citation	Domain	Application	Language(s)	Relation(s)	Methods
Morgan (2000)	Computing Genetics	Knowledge extraction	French	HYPERONYMY	Lexical and grammatical knowledge patterns
Nuopponen (1994)	Medicine	Knowledge extraction	English	CAUSE–EFFECT	Lexical knowledge patterns
Pearson (1998, 1998-9)	Nature Telecom-munications GCSE subjects	Knowledge extraction	English	DEFINITION	Grammatical , lexical and paralinguistic knowledge patterns
Rebeyrolle (2000, 2000a)	Electricity Geomorph-ology Knowledge engineering Software develop-ment	Knowledge extraction Text typology	French	HYPERONYMY HYPONYMY	Lexico-syntactic and paralinguistic knowledge pattern
Rodríguez Penagos (2004, 2004a)	Sociology, Histology, Medline abstracts	Information extraction Metalinguistic Information Database construction	English	META-LINGUISTIC INFORMATION	Lexical and paralinguistic patterns Machine-learning techniques
Séguéla (1999)	Various	Information extraction Tool development	French	HYPONYMY MERONYMY	Lexico-syntactic patterns
Weilgaard (2004)	Hydraulics Popular science	Knowledge extraction	Danish	HYPERONYMY PARTONYMY SYNONYMY …	Lexico-syntactic knowledge patterns Actantial structures Semantic classes

Appendix C: Corpus texts

English Texts

Adams, J., M. White and D. Forman. 2004. "Are there socioeconomic gradients in stage and grade of breast cancer at diagnosis? Cross sectional analysis of UK cancer registry data." *British Medical Journal*, doi:10.1136/BMJ.38114.679387.AE (2 June 2004). http://bmj.com/cgi/content/full/329/7458/142. Consulted 16 October 2006.

Adsay, N.V., K. Merati, H. Nassar, J. Shia, F. Sarkar, C. Pierson, J. Cheng, D.W. Visscher, R.H. Hruban and D.S. Klimstra. 2003. "Pathogenesis of Colloid (Pure Mucinous) Carcinoma of Exocrine Organs: Coupling of Gel-Forming Mucin (MUC2) Production With Altered Cell Polarity and Abnormal Cell-Stroma Interaction May Be the Key Factor in the Morphogenesis and Indolent Behavior." *The American Journal of Surgical Pathology*, May 2003, 27(5): 571–578.

Armitage, J. and L. Bowman. 2004. "Cardiovascular outcomes among participants with diabetes in the recent large statin trials." *Current Opinion in Lipidology*, August 2004, 15(4): 439–446.

Aschenbrenner, D.S. 2004. "HRT Reconsidered: What should you tell patients about it now?" *AJN, American Journal of Nursing*, June 2004, 104(6): 51–53.

Baker, M.K., K. Mikhitarian, W. Osta, K. Callahan, R. Hoda, F. Brescia, R. Kneuper-Hall, M. Mitas, D.J. Cole and W.E. Gillanders. 2003. "Molecular Detection of Breast Cancer Cells in the Peripheral Blood of Advanced-Stage Breast Cancer Patients Using Multimarker Real-Time Reverse Transcription-Polymerase Chain Reaction and a Novel Porous Barrier Density Gradient Centrifugation Technology." *Clinical Cancer Research*, 15 October 2003, 9: 4865–4871.

Baker, S.G., B.S. Kramer and P.C. Prorok. 2004. "Comparing breast cancer mortality rates before-and-after a change in availability of screening in different regions: Extension of the paired availability design." *BMC Medical Research*

Methodology 2004, 4. http://www.biomedcentral.com/1471–2288/4/12. Consulted 15 October 2006.

Balk, E.M., J. Lau, L.C. Goudas, H.S. Jordan, B.Kupelnick, L.U. Kim and R.H. Karas. 2003. "Effects of Statins on Nonlipid Serum Markers Associated with Cardiovascular Disease: A Systematic Review." *Annals of Internal Medicine*, 21 October 2003, 139(8): 670–682.

Bassuk, S.S. and J.E. Manson. 2003. "Physical Activity and Cardiovascular Disease Prevention in Women: How Much Is Good Enough?" *Exercise and Sport Sciences Reviews*, October 2003, 31(4): 176–181.

Beardsley, T. 2000. "Working Under Pressure: Pushing DNA into cells makes a safe form of gene therapy work." *Scientific American*, March 2000. http://www.sciam.com/article.cfm?articleID=000D31C8-56B9-1C75-9B81809EC588EF21&pageNumber=1. Consulted 18 February 2004.

Bellon, J.R., R.B. Livingston, W.B. Eubank, J.R. Gralow, G.K. Ellis, L.K. Dunnwald, D.A. Mankoff. 2004. "Evaluation of the Internal Mammary Lymph Nodes by FDG-PET in Locally Advanced Breast Cancer (LABC)." *American Journal of Clinical Oncology*, August 2004, 27(4): 407–410.

Berra, K. 2003. "The Effect of Lifestyle Interventions on Quality of Life and Patient Satisfaction With Health and Health Care." *The Journal of Cardiovascular Nursing*, September/October 2003, 18(4): 319–325.

Bittner, V. 2003. "Non-high-density lipoprotein cholesterol and cardiovascular disease." *Current Opinion in Lipidology*, August 2003, 14(4): 367–371.

Boyle, P., M. Mezzetti, C. La Vecchia, S. Franceschi, A. Decarli and C. Robertson. 2004. "Contribution of three components to individual cancer risk predicting breast cancer risk in Italy". *European Journal of Cancer Prevention*, June 2004, 13(3): 183–191.

Braidotti, P., P.G. Nuciforo, J. Mollenhauer, A. Poustka, C. Pellegrini, A. Moro, G. Bulfamante, G. Coggi, S. Bosari and G.G. Pietra. 2004. "DMBT1 expression is down-regulated in breast cancer." *BMC Cancer* 2004, 4: 46. http://www.biomedcentral.com/1471–2407/4/46. Consulted 15 October 2006.

"Breast cancer options made clearer." 2000. *Science News*, 22 April 2000, 157(17): 264.

"Breast cancer protein gets lost." 1995. *Science News*, 18 November 1995, 148(21): 334(1).

Brennan, M.-L. and S.L. Hazen. 2003. "Emerging role of myeloperoxidase and oxidant stress markers in cardiovascular risk assessment." *Current Opinion in Lipidology*, August 2003, 14(4): 353–359.

Burstein, H. J. 2003. "Trastuzumab in combination with chemotherapy." *Breast Cancer Research and Treatment* 81 (Suppl. 1): S69–S72. http://journals.kluweronline.com/. Consulted 15 October 2006.

Cabe, D.K. 2000. "Saving Hearts that Grow Old." *Scientific American*, July 2000. http://www.sciam.com/article.cfm?articleID=00055065-0EE8-1C75-9B81809EC588EF21&pageNumber=1. Consulted 18 February 2004.

Campbell, J.B. 2002. "Breast cancer—race, ethnicity, and survival: A literature review." *Breast Cancer Research and Treatment* 74: 187–192. http://www.springerlink.com/. Consulted 15 October 2006.

"Cancer cells on the move." 2000. *Science News*, 25November 2000, 158(22): 348.

Carlson, J.J. and V. Monti. 2003. "The Role of Inclusive Dietary Patterns for Achieving Secondary Prevention Cardiovascular Nutrition Guidelines and Optimal Cardiovascular Health." *Journal of Cardiopulmonary Rehabilitation*, September/October 2003, 23(5): 322–333.

Carrick, S., D. Ghersi, N. Wilcken and J. Simes. 2004. "Platinum containing regimens for metastatic breast cancer." *The Cochrane Database of Systematic Reviews*, 3.

Caslake, M.J. and C.J. Packard. 2003. "Lipoprotein-associated phospholipase A2 (platelet-activating factor acetylhydrolase) and cardiovascular disease." *Current Opinion in Lipidology*, August 2003, 14(4): 347–352.

Chauhan, T.S. 2003. "Gap in male-female cancer deaths narrows." *Canadian Medical Association Journal*, 22 July 2003, 169 (2): 142.

Cheema, A.A. 2004. "Should People on Aspirin Avoid Ibuprofen?: A Review of the Literature." *Cardiology in Review*, May/June 2004, 12(3): 174-176.

"Chemotherapy leads to bone loss." 2001. *Science News*, 11 August 2001, 160(6): 89.

Cody, H.S. III, P.I. Borgen and L.K. Tan. 2004. "Redefining Prognosis in Node-Negative Breast Cancer: Can Sentinel Lymph Node Biopsy Raise the Threshold for Systemic Adjuvant Therapy?" *Annals of Surgical Oncology*, March 2004, 11(3) Supplement: 227S–230S.

Collins, L.C., J.L. Connolly, D.L. Page, R.A. Goulart, E.D. Pisano, L.L. Fajardo, W.A. Berg, D.J. Caudry, B.J. McNeil and S.J. Schnitt. 2004. "Diagnostic Agreement in the Evaluation of Image-guided Breast Core Needle Biopsies: Results from a Randomized Clinical Trial." *The American Journal of Surgical Pathology*, January 2004, 28(1): 126–131.

Colwell, A.S., Ja. Kukreja, K.H. Breuin, S. Lester and D.P. Orgill. 2004. "Occult Breast Carcinoma in Reduction Mammaplasty Specimens: 14-Year Experience." *Plastic and Reconstructive Surgery*, June 2004, 113(7): 1984–1988.

Coresh, J., B. Astor and M.J. Sarnak. 2004. "Evidence for increased cardiovascular disease risk in patients with chronic kidney disease." *Current Opinion in Nephrology and Hypertension*, January 2004, 13(1): 73–81.

Critchley, J.A. and B. Unal. 2004. "Is smokeless tobacco a risk factor for coronary heart disease? A systematic review of epidemiological studies." *European Journal of Cardiovascular Prevention & Rehabilitation*, April 2004, 11(2): 101–112.

Crown, J., and M. Pegram. 2003. "Platinum-taxane combinations in metastatic breast cancer." *Breast Cancer Research and Treatment* 79 (Suppl. 1): S11–S18. http://www.springerlink.com/. Consulted 15 October 2006.

Cuenca, R.E., R.R. Allison, C. Sibata and G.H. Downie. 2004. "Breast Cancer With Chest Wall Progression: Treatment With Photodynamic Therapy." *Annals of Surgical Oncology*, March 2004, 11(3): 322–327.

Davidson, M.H. and P.P. Toth. 2004. "Combination therapy in the management of complex dyslipidemias." *Current Opinion in Lipidology*, August 2004, 15(4): 423–431.

Davignon, J. 2004. "Beneficial Cardiovascular Pleiotropic Effects of Statins." *Circulation: Journal of the American Heart Association*, 15 June 2004, 109(23) Supplement: III-39–III-43.

Davis, C., P. Williams, M. Parle, S. Redman and J. Turner. 2004. "Assessing the Support Needs of Women With Early Breast Cancer in Australia." *Cancer Nursing*, March/April 2004, 27(2): 169–174.

Deutsch, M. and J.C. Flickinger. 2003. "Arm Edema After Lumpectomy and Breast Irradiation." *American Journal of Clinical Oncology*, June 2003, 26(3): 229–231.

DiGiovanna, A.G. and H. Adams. 1999. "Ask the Experts: What Causes Strokes?" *Scientific American*, 14 July 1999. http://www.sciam.com/askexpert_question.cfm?articleID=000DBA9B-791E-1C71-9EB7809EC588F2D7&catID=3. Consulted 18 February 2004.

Doherty, B. 2001. "Extra Insurance against Cancer Return." *Prevention*, March 2001, 53(3): 164.

Dorn, J., J. Vena, J. Brasure, J. Freudenheim and S. Graham. 2003. "Lifetime Physical Activity and Breast Cancer Risk in Pre- and Postmenopausal Women." *Medicine & Science in Sports & Exercise*, February 2003, 35(2): 278–285.

Dranitsaris, G., S. Verma and M. Trudeau. 2003. "Cost Utility Analysis of First-Line Hormonal Therapy in Advanced Breast Cancer: Comparison of Two Aromatase Inhibitors to Tamoxifen." *American Journal of Clinical Oncology*, June 2003, 26(3): 289–296.

Du, C., N. Feng, H. Jin, M. Wang, J.A. Wright and A.H. Young. 2004. "Preclinical efficacy of Virulizin in human breast, ovarian and prostate tumor models." *Anti-Cancer Drugs*, April 2003, 14(4): 289–294.

Duncan, A.M. 2004. "The Role of Nutrition in the Prevention of Breast Cancer." *AACN Clinical Issues: Advanced Practice in Acute and Critical Care*, January/March 2004, 15(1): 119–135.

"Easier radiation." 2003. *Prevention*, February 2003, 55(2): 165(1).

Edwards, D.D. 1988. "Beating breast cancer: Researchers are looking beyond conventional surgeries and chemotherapy." *Science News*, 14 May 1988, 133(20): 314(2).

Eggertson, L. 2004. "MRIs more accurate than mammograms but expensive." *Canadian Medical Association Journal* 12 October 2004, 171(8): 840.

"Exercise after breast cancer extends life." 2004. *Science News*, 10 April 2004, 165(15): 237(1).

"Exercise Rx for breast cancer." 1998. *Prevention*, August 1998, 50(8): 48(1).

Ezzell, C. 1992. "Does gene hike radiation's cancer risk?" *Science News*, 4 January 1992, 141(1): 4(1).

Fackelmann, Kathy A. 1992. "Best time for breast cancer surgery." *Science News*, 11 April 1992, 141(15): 239(1).

Fackelmann, Kathy A. 1992. "Breast cancer risk traced back to the womb." *Science News*, 31 October 1992, 142(18): 293(1).

Fackelmann, Kathy A. 1992. "Motherhood and cancer: Can hormones protect against breast and other cancers?" *Science News*, 31 October 1992, 142(18): 298(3).

Fackelmann, Kathy A. 1994. "Do EMFs pose breast cancer risk?" *Science News*, 18 June 1994, 145(25): 388(1).

Fackelmann, Kathy A. 1995. "A versatile virus: Epstein-Barr virus displays a few new malignant tricks." *Science News*, 18 February 1995, 147(7): 104(2).

Fackelmann, Kathy A. 1998. "Gene therapy for breast, ovarian cancer." *Science News*, 11 April 1998, 153(15): 239(1).

Fibrinogen Studies Collaboration. 2004. "Collaborative meta-analysis of prospective studies of plasma fibrinogen and cardiovascular disease." *European Journal of Cardiovascular Prevention & Rehabilitation*, February 2004, 11(1): 9–17.

Force, T., K. Kuida, M. Namchuk, K. Parang and J.M. Kyriakis. 2004. "Inhibitors of Protein Kinase Signaling Pathways: Emerging Therapies for Cardiovascular Disease." *Circulation: Journal of the American Heart Association*, 16 March 2004, 109(10): 1196–1205.

Franzen, H. 2001. "Eating Soy Boosts Tamoxifen's Powers." *Scientific American*, 28 March, 2001. http://www.sciam.com/print_version.cfm?articleID=000EF0DE-FCFF-1C5A-B882809EC588ED9F. Consulted 16 June 2004.

Garg, A.K., G.N. Hortobagyi, B.B. Aggarwal, A.A. Sahin and T.A. Buchholz. 2003. "Nuclear factor-[kappa]B as a predictor of treatment response in breast cancer." *Current Opinion in Oncology*, November 2003, 15(6): 405–411.

Gibbs, W.W. 2003. "Untangling the Roots of Cancer." *Scientific American*, 9 June 2003. http://www.sciam.com/print_version.cfm?articleID=000C24C1-2210-1EDD-8E1C809EC588EF21. Consulted 15 October 2006.

Graham, S. 2002. "Arsenic in Drinking Water May Accelerate Artery Disease." *Scientific American*, 26 March 2002. http://www.sciam.com/print_version.cfm?articleID=000948AD-FFD5-1CCF-B4A8809EC588EEDF. Consulted 16 June 2004.

Graham, S. 2002a. "Research Reveals Role of Breast Cancer Gene in Repairing Damaged DNA." *Scientific American*, 13 September 2002. http://www.sciam.com/print_version.cfm?articleID=000E3EF1-0688-1D81-90FB809EC5880000. Consulted 15 October 2006.

Graham, S. 2003. "Ozone Linked to Hardening Arteries." *Scientific American*, 7 November 2003. http://www.sciam.com/print_version.cfm?articleID=0003CCAB-DF13-1FAA-9F1383414B7F0000. Consulted 16 June 2004.

Graham-Rowe, D. 2002. "The intelligent way to treat breast cancer." *New Scientist*, 27 July 2002, 175(2353): 18(1).

Granger, D.N., T. Vowinkel and T. Petnehazy. 2004. "Modulation of the Inflammatory Response in Cardiovascular Disease." *Hypertension: Journal of the American Heart Association*, May 2004, 43(5): 924–931.

Grann, A., J.C. Abdou, N. Dragman and R. Goodman. 2004. "The Value of Postexcision Preradiation Mammography in Patients With Early-Stage Breast Cancer." *American Journal of Clinical Oncology*, June 2004, 27(3): 285–288.

Griendling, K.K. and G.A. FitzGerald. 2003. "Oxidative Stress and Cardiovascular Injury: Part I: Basic Mechanisms and In Vivo Monitoring of ROS." *Circulation: Journal of the American Heart Association*, 21 October 2003, 108(16): 1912–1916.

Griendling, K.K. and G.A. FitzGerald. 2003a. "Oxidative Stress and Cardiovascular Injury: Part II: Animal and Human Studies." *Circulation: Journal of the American Heart Association*, 28 October 2003, 108(17): 2034–2040.

Grimes, D.A. and R.A. Lobo. 2002. "Perspectives on the Women's Health Initiative Trial of Hormone Replacement Therapy." *Obstetrics & Gynecology*, December 2002, 100(6): 1344–1353.

Grundy, S.M., B. Hansen, S.C. Smith Jr, J.I.Cleeman and R.A. Kahn. 2004. "Clinical Management of Metabolic Syndrome: Report of the American Heart Association/National Heart, Lung, and Blood Institute/American Diabetes Association Conference on Scientific Issues Related to Management." *Arteriosclerosis, Thrombosis, and Vascular Biology: Journal of the American Heart Association*, February 2004, 24(2): e19–e24.

Grunfeld, E., D. Coyle, T. Whelan, J. Clinch, L. Reyno, C.C. Earle, A. Willan, R. Viola, M. Coristine, T. Janz and R. Glossop. 2004. "Family caregiver burden: results of a longitudinal study of breast cancer patients and their principal caregivers." *Canadian Medical Association Journal*, 8 June 2004, 170(12): 1795–1801.

Gupta, M., C.Vavasis and W.H. Frishman. 2004. "Heat Shock Proteins in Cardiovascular Disease: A New Therapeutic Target." *Cardiology in Review*, January/February 2004, 12(1): 26–30.

Halyard, M.Y., K.E. McCombs, W.W. Wong, E.W. Buchel, B.A. Pockaj, S.A. Vora, R.J. Gray and S.E. Schild. 2004. "Acute and Chronic Results of Adjuvant Radiotherapy After Mastectomy and Transverse Rectus Abdominis Myocutaneous (TRAM) Flap Reconstruction for Breast Cancer." *American Journal of Clinical Oncology*, August 2004, 27(4): 389–394.

Harris, K.F. and K.A. Matthews. 2004. "Interactions Between Autonomic Nervous System Activity and Endothelial Function: A Model for the Development of

Cardiovascular Disease." *Psychosomatic Medicine*, March/April 2004, 66(2): 153–164.

Haskell, W.L. 2003. "Cardiovascular Disease Prevention and Lifestyle Interventions: Effectiveness and Efficacy." *The Journal of Cardiovascular Nursing*, September/October 2003, 18(4) :245–255.

Hausmaninger, H., G. Morack, B. Heinrich, D. Wallwiener, K. Hoffken, S. Buksmaui, K.Krejcy, M.A. Miller and K. Possinger. 2004. "Gemcitabine Combined With Epirubicin in the Treatment of Patients With Locally Advanced or Metastatic Breast Cancer: A Phase II Study." *American Journal of Clinical Oncology*, August 2004, 27(4): 429–435.

Heitman, L.K., CS. 2004. "Social Support and Cardiovascular Health Promotion in Families." *The Journal of Cardiovascular Nursing*, January/February 2004, 19(1): 86–91.

"Hot flash cooler for breast cancer survivors." 2002. *Prevention*, July 2002, 54(7): 159(1).

Humphries, S.E., P.M. Ridker and P.J. Talmud. 2004. "Genetic Testing for Cardiovascular Disease Susceptibility: A Useful Clinical Management Tool or Possible Misinformation?" *Arteriosclerosis, Thrombosis, and Vascular Biology: Journal of the American Heart Association*, April 2004, 24(4): 628–636.

Jaffer, F.A. and R. Weissleder. 2004. "Seeing Within: Molecular Imaging of the Cardiovascular System." *Circulation Research: Journal of the American Heart Association*, 5 March 2004, 94(4): 433–445.

Kabat, G.C., E.S. O'Leary, E.R. Schoenfeld, J.M. Greene, R. Grimson, K. Henderson, W.T. Kaune, M.D. Gammon, J.A. Britton, S.L. Teitelbaum, A.I. Neugut and M.C. Leske. 2003. "Electric Blanket Use and Breast Cancer on Long Island." *Epidemiology*, September 2003, 14(5): 514–520.

Kang, D.-H. 2002. "Oxidative Stress, DNA Damage, and Breast Cancer." *AACN Clinical Issues: Advanced Practice in Acute and Critical Care*, November 2002, 13(4): 540–549.

Kannel, W.B., R.S. Vasan and D. Levy. 2003. "Is the Relation of Systolic Blood Pressure to Risk of Cardiovascular Disease Continuous and Graded, or Are There Critical Values?" *Hypertension: Journal of the American Heart Association*, October 2003, 42(4): 453–456.

Karas, RH. 2004. "Current Controversies Regarding the Cardiovascular Effects of Hormone Therapy." *Clinical Obstetrics and Gynecology*, June 2004, 47(2): 489–499.

Karvellas, C.J., M. Sawyer, M. Hamilton and J.R. Mackey. 2004. "Effect of Capecitabine on Mean Corpuscular Volume in Patients With Metastatic Breast Cancer." *American Journal of Clinical Oncology*, August 2004, 27(4): 364–368.

Khoury-Collado, F. and A.T. Bombard. 2004. "Hereditary Breast and Ovarian Cancer: What the Primary Care Physician Should Know." *Obstetrical & Gynecological Survey*, July 2004, 59(7): 537–542.

Klett, E.L. and S. Patel. 2003. "Genetic defenses against noncholesterol sterols." *Current Opinion in Lipidology*, August 2003, 14(4): 341–345.

Kocjan, T. and G.M. Prelevic. 2003. "Hormone replacement therapy update: who should we be prescribing this to now?" *Current Opinion in Obstetrics and Gynecology*, December 2003, 15(6): 459–464.

Koil, C.E., J.N. Everett, L. Hoechstetter, R.E. Ricer and K.M. Huelsman. 2003. "Differences in physician referral practices and attitudes regarding hereditary breast cancer by clinical practice location." *Genetics in Medicine*, September/October 2003, 5(5): 364–369.

Kranitz, L.J.D. and P. Lehrer. 2004. "Biofeedback Applications in the Treatment of Cardiovascular Diseases." *Cardiology in Review*, May/June 2004, 12(3): 177–181.

La Vecchia, C. 2003. "Menopause, hormone therapy and breast cancer risk." *European Journal of Cancer Prevention*, October 2003, 12(5): 437–438.

Lambe, M., D. Trichopoulos, C.-C. Hsieh, J. Wuu, H.-O. Adami and L. Wide. 2003. "Ethnic Differences in Breast Cancer Risk: A Possible Role for Pregnancy Levels of Alpha-Fetoprotein?" *Epidemiology*, January 2003, 14(1): 85–89.

Legorreta, A.P., H.O. Chernicoff, J.B. Trinh and R.G. Parker. 2004. "Diagnosis, Clinical Staging, and Treatment of Breast Cancer: A Retrospective Multiyear Study of a Large Controlled Population." *American Journal of Clinical Oncology*, April 2004, 27(2): 185–190.

Leidenius, M.H., E.A. Leppanen, L.A. Krogerus and K.A. Smitten. 2004. "The impact of radiopharmaceutical particle size on the visualization and identification of sentinel nodes in breast cancer." *Nuclear Medicine Communications*, March 2004, 25(3): 233–238.

Lerwill, M.F. 2004. "Current Practical Applications of Diagnostic Immunohistochemistry in Breast Pathology." *The American Journal of Surgical Pathology*, August 2004, 28(8): 1076–1091.

Leutwyler, K. 2001. "New Breast Cancer Gene." *Scientific American*, 26 January 2001. http://www.sciam.com/print_version.cfm?articleID=0009BC5B-21B5-1C5A-B882809EC588ED9F. Consulted 16 June 2004.

Leutwyler, K. 2001a. "Portable Scanner for Breast Cancer." *Scientific American*, 15 March 2001. http://www.sciam.com/print_version.cfm?articleID=000CE745-EF37-1C5A-B882809EC588ED9F. Consulted 16 June 2004.

Levine, P.H., X.-J. Wei, J.-P. Gagner, H. Flax, K. Mittal and S.V. Blank. 2003. "Pleomorphic Liposarcoma of the Uterus: Case Report and Literature Review." *International Journal of Gynecological Pathology*, October 2003, 22(4): 407–411.

Li, Y., R.C Millikan, D.A Bell, L. Cui, C.-K.J. Tse, B. Newman and K. Conway. 2004. "Polychlorinated biphenyls, cytochrome P450 1A1 (CYP1A1) polymorphisms, and breast cancer risk among African American women and white women in North Carolina: A population-based case-control study." *Breast Cancer Research*, 7(1). http://breast-cancer-research.com/content/7/1/R12. Consulted 15 October 2006.

Liu, W., Y. Chen, W. Wang, P. Keng, J. Finkelstein, D. Hu, L. Liang, M. Guo, B. Fenton, P. Okunieff and I. Ding. 2003. "Combination of Radiation and Celebrex

(Celecoxib) Reduce Mammary and Lung Tumor Growth." *American Journal of Clinical Oncology*, August 2003, 26(4) Supplement 2: S103–S109.

Loecher, B. 2001. "Get the Best Breast Care." *Prevention*, March 2001, 53(3): 34.

MacIsaac, R.J., G. Jerums and M.E. Cooper. 2004. "New insights into the significance of microalbuminuria." *Current Opinion in Nephrology and Hypertension*, January 2004, 13(1): 83–91.

MacKenzie, J.R. 2004. "Predicting CAD Events: C-Reactive Protein a Marker for Atherosclerotic Risk." *The Nurse Practitioner*, June 2004, 29(6): 14–27.

Mackness, M., P. Durrington and B. Mackness. 2004. "Paraoxonase 1 activity, concentration and genotype in cardiovascular disease." *Current Opinion in Lipidology*, August 2004, 15(4): 399–404.

Madjid, M., I. Aboshady, I. Awan, S. Litovsky and S.W. Casscells. 2004. "Influenza and Cardiovascular Disease: Is There a Causal Relationship?" *Texas Heart Institute Journal*, 31(1): 4–13.

Madjid, M., M. Naghavi, S. Litovsky and S.W. Casscells. 2003. "Influenza and Cardiovascular Disease: A New Opportunity for Prevention and the Need for Further Studies." *Circulation: Journal of the American Heart Association*, 2 December 2003, 108(22): 2730–2736.

Major, M.A. 2003. "Clinical Trials Update: Medical Management of Advanced Breast Cancer." *Cancer Nursing*, December 2003, 26(Supplement 6S): 10S–15S.

Malek, K., A.K. Fink, S.S. Thwin, J. Gurwitz, P.A. Ganz and R.A. Silliman. 2004. "The Relationship Among Physicians' Specialty, Perceptions of the Risks and Benefits of Adjuvant Tamoxifen Therapy, and Its Recommendation in Older Patients With Breast Cancer." *Medical Care*, July 2004, 42(7): 700–706.

Manne, S., J. Ostroff, G. Winkel, L. Goldstein, K. Fox and G. Grana. 2004. "Posttraumatic Growth After Breast Cancer: Patient, Partner, and Couple Perspectives." *Psychosomatic Medicine*, May/June 2004, 66(3): 442–454.

Marcus, J.N., P. Watson, D.L. Page, S.A. Narod, P. Tonin, G.M. Lenoir, O. Serova and H.T. Lynch. 1997. "BRCA2 hereditary breast cancer pathophenotype." *Breast Cancer Research and Treatment*, 44: 275–277.

Mason, R.P., P. Marche and T.H. Hintze. 2003. "Novel Vascular Biology of Third-Generation L-Type Calcium Channel Antagonists: Ancillary Actions of Amlodipine." *Arteriosclerosis, Thrombosis, and Vascular Biology: Journal of the American Heart Association*, December 2003, 23(12): 2155–2163.

McCance, K.L. and R.E. Jones. 2003. "Estrogen and Insulin Crosstalk: Breast Cancer Risk Implications." *The Nurse Practitioner*, May 2003, 28(5): 12–23.

McGrath, K.G. 2003. "An earlier age of breast cancer diagnosis related to more frequent use of antiperspirants/deodorants and underarm shaving." *European Journal of Cancer Prevention*, December 2003, 12(6): 479–485.

Meric-Bernstam, F. 2004. "Breast conservation in breast cancer: surgical and adjuvant considerations." *Current Opinion in Obstetrics and Gynecology*, February 2004, 16(1): 31–36.

Micheli-Tzanakou, E. and T. Cooley. 1997. "A Mobile Automated Mammography System." *Journal of Medical Systems*, 21(5). http://www.springerlink.com/. Consulted 15 October 2006.

Miller, K.A. 2002. "New Light on Breast Cancer: Laser light and thermal heat could help improve the accuracy of mammograms." *Scientific American*. http://www.sciam.com/article.cfm?articleID=00049F59-4A54-1D48-90FB809EC5880000&pageNumber=1. Consulted 18 February 2004.

Minkel, J.R. 2001. "Drug Delivers Longer-Lasting Isotopes to Tumors." *Scientific American*, 16 November 2001. http://www.sciam.com/print_version.cfm?articleID=000EE4B5-0F80-1C68-B882809EC588ED9F. Consulted 16 June 2004.

Minkin, M.J. and T. Hanlon. 1999. "When to Worry." *Prevention*, December 1999, 51(12): 95.

Mirsky, S. 2001. "A Host with Infectious Ideas." *Scientific American*, 17 May 2001. http://www.sciam.com/print_version.cfm?articleID=000DBEA5-16B4-1C70-84A9809EC588EF21. Consulted 16 June 2004.

Moore, T.D., J.J. Nawarskas and J.R. Anderson. 2003. "Eplerenone: A Selective Aldosterone Receptor Antagonist for Hypertension and Heart Failure." *Heart Disease*, September/October 2003, 5(5): 354–363.

Munson, M. and G. Gutfeld. 1994. "Breast-saving implants: Radioactive 'seeds' may cut treatment time." *Prevention*, April 1994, 46(4): 30(2).

Naik, A.M., J. Fey, M. Gemignani, A. Heerdt, L. Montgomery, J. Petrek, E. Port, V. Sacchini, L. Sclafani, K. VanZee, R.Wagman, P.I. Borgen and H.S. Cody III. 2004. "The Risk of Axillary Relapse After Sentinel Lymph Node Biopsy for Breast Cancer Is Comparable With That of Axillary Lymph Node Dissection: A Follow-up Study of 4008 Procedures." *Annals of Surgery*, September 2004, 240(3): 462–471.

Napoli, A., D. Fleischmann, F.P. Chan, C. Catalano, J.C. Hellinger, R. Passariello and G.D. Rubin. 2004. "Computed Tomography Angiography: State-of-the-Art Imaging Using Multidetector-Row Technology." *Journal of Computer Assisted Tomography*, July/August 2004, 28 Supplement 1: S32–S45.

Newman, L.A., N.L. Pernick, V. Adsay, K.A. Carolin, P.I. Philip, S. Sipierski, D.L. Bouwman, M.A. Kosir, M. White and D.W. Visscher. 2003. "Histopathologic Evidence of Tumor Regression in the Axillary Lymph Nodes of Patients Treated With Preoperative Chemotherapy Correlates With Breast Cancer Outcome." *Annals of Surgical Oncology*, August 2003, 10(7): 734–739.

Nguyen, H.Q., V. Carrieri-Kohlman, S.H. Rankin, R. Slaughter and M.S. Stulbarg. 2004. "Supporting Cardiac Recovery Through eHealth Technology." *The Journal of Cardiovascular Nursing*, May/June 2004, 19(3): 200–208.

Orians, C.E., J. Erb, K.L.Kenyon, P.M. Lantz, E.B.Liebow, J.R. Joe and L. Burhansstipanov. 2004. "Public Education Strategies for Delivering Breast and Cervical Cancer Screening in American Indian and Alaska Native Populations." *Journal of Public Health Management and Practice*, January/February 2004, 10(1): 46–53.

Oz, M. 2004. "Emerging Role of Integrative Medicine in Cardiovascular Disease." *Cardiology in Review*, March/April 2004, 12(2): 120–123.

Pantaleo, A. and J. Zonszein. 2003. "Using Insulin as a Drug Rather Than as a Replacement Hormone During Acute Illness: A New Paradigm." *Heart Disease*, September/October 2003, 5(5): 323–333.

Penckofer, S.M., D. Hackbarth and D.W. Schwertz. 2003. "Estrogen Plus Progestin Therapy: The Cardiovascular Risks Exceed the Benefits." *The Journal of Cardiovascular Nursing*, November/December 2003, 18(5): 347–355.

Pennisi, E. 1993. "Dissecting breast cancer treatments." *Science News*, May 29, 1993, 143(22): 344(1).

Perlmutter, C. 1992. "Why women reject the kinder cut: and what you should know about your breast-cancer treatment options." *Prevention*, November 1992, 44(11): 50(14).

"Pill boosts cancer risk in some women." 2000. *Science News*, 28 October 2000, 158(18): 280.

Pistoi, S. 2001. "Breast Cancer: Knocking Out a Killer." *Scientific American*, http://www.sciam.com/print_version.cfm?articleID=000B0D50-9D97-1C75-9B81809EC588EF21. Consulted 18 February 2004.

Pockaj, B.A., G.D. Basu, L.B. Pathangey, R.J. Gray, J.L. Hernandez, S.J. Gendler and P. Mukherjee. 2004. "Reduced T-Cell and Dendritic Cell Function Is Related to Cyclooxygenase-2 Overexpression and Prostaglandin E2 Secretion in Patients With Breast Cancer." *Annals of Surgical Oncology*, March 2004, 11(3): 328–339.

Polednak, A.P. 2004. "Chemotherapy of nonelderly breast cancer patients by poverty-rate of area of residence in Connecticut." *Breast Cancer Research and Treatment* 83: 245–248.

"Portrait of a cancer drug at work." 2003. *Science News*, 8 March 8, 163(10): 157(1).

Pritchard, K.I. 2004. "Is exercise effective in reducing the risk of breast cancer in postmenopausal women?" *Canadian Medical Association Journal*, 2 March 2004, 170(5): 787.

Rackley, C.E. 2004. "New Clinical Markers Predictive of Cardiovascular Disease: The Role of Inflammatory Mediators." *Cardiology in Review*, May/June 2004, 12(3): 151–157.

Raloff, J. 1992. "Breast cancer therapy's leukemia risks." *Science News*, 27 June 1992, 141(26): 420(1).

Raloff, J. 1998. "Does light have a dark side?: Nighttime illumination might elevate cancer risk." *Science News*, 17 October 1998, 154(16): 248(3).

Ram, C. and S. Venkata. 2003. "Possible Therapeutic Role of Endothelin Antagonists in Cardiovascular Disease." *American Journal of Therapeutics*, November/December 2003, 10(6): 396–400.

Rayson, D., D. Chiasson and R. Dewar. 2004. "Elapsed time from breast cancer detection to first adjuvant therapy in a Canadian province, 1999–2000." *Canadian Medical Association Journal*, 16 March 2004, 170(6): 957–961.

Rich, S. and V.V. McLaughlin. 2003. "Endothelin Receptor Blockers in Cardiovascular Disease." *Circulation: Journal of the American Heart Association*, 4 November 2003, 108(18): 2184–2190.

Ridker, P.M. 2003. "Rosuvastatin in the Primary Prevention of Cardiovascular Disease Among Patients With Low Levels of Low-Density Lipoprotein Cholesterol and Elevated High-Sensitivity C-Reactive Protein: Rationale and Design of the JUPITER Trial." *Circulation: Journal of the American Heart Association*, 11 November 2003, 108(19): 2292–2297.

Rigby, J.E., J.A. Morris, J. Lavelle, M. Stewart and A.C. Gatrell. 2002. "Can physical trauma cause breast cancer?" *European Journal of Cancer Prevention*, June 2002, 11(3): 307–311.

Saba, J.D. and T. Hla. 2004. "Point-Counterpoint of Sphingosine 1-Phosphate Metabolism." *Circulation Research: Journal of the American Heart Association*, 2 April 2004, 94(6): 724–734.

Sangiorgio, M., G. Gutfeld and L. Rao. 1992. "Good for yew." *Prevention*, February 1992, 44(2): 18(1).

Sawka, C. 2004. "Is antibiotic use associated with an increased risk of breast cancer?" *Canadian Medical Association Journal*, 22 June 2004, 170(13): 1912.

Schneeweiss, A., S. Kolay, S. Aulmann, G. von Minckwitz, J. Torode, M. Koehler and G. Bastert. 2004. "Induction of remission in a patient with metastatic breast cancer refractory to trastuzumab and chemotherapy following treatment with gefitinib ('Iressa', ZD1839)." *Anti-Cancer Drugs*, March 2004, 15(3): 235–238.

Schondorf, T., M. Hoopmann, M. Breidenbach, D.T. Rein, U.-J. Gohring, M. Becker, P. Mallmann and C.M. Kurbacher. 2004. "Dysregulation of protein kinase C activity in chemoresistant metastatic breast cancer cells." *Anti-Cancer Drugs*, March 2004, 15(3): 265–268.

Schootman, M. and D. Sun. 2004. "Small-Area Incidence Trends in Breast Cancer." *Epidemiology*, May 2004, 15(3): 300–307.

Schwartz, A.R., W. Gerin, K.W. Davidson, T.G. Pickering, J.F. Brosschot, J.F. Thayer, N. Christenfeld and W. Linden. 2003. "Toward a Causal Model of Cardiovascular Responses to Stress and the Development of Cardiovascular Disease." *Psychosomatic Medicine*, January/February 2003, 65(1): 22–35.

Schwartz, J.B. 2003. "Gender-Specific Implications for Cardiovascular Medication Use in the Elderly: Optimizing Therapy for Older Women." *Cardiology in Review*, September/October 2003, 11(5): 275–298.

Seed, M. and R.H. Knopp. 2004. "Estrogens, lipoproteins, and cardiovascular risk factors: an update following the randomized placebo-controlled trials of hormone-replacement therapy." *Current Opinion in Lipidology*, August 2004, 15(4): 459–467.

Seppa, N. 2002. "Breast surgeries yield same survival rate." *Science News*, 19 October 2002, 162(16): 243(1).

Shaaban, A.M., J.P. Sloane, C.R. West, F.R. Moore, C. Jarvis, E. Williams and C.S. Foster. 2002. "Histopathologic Types of Benign Breast Lesions and the Risk of Breast Cancer: Case-Control Study." *The American Journal of Surgical Pathology*, April 2002, 26(4): 421–430.

Shah, S.H. and L.K. Newby. 2003. "C-Reactive Protein: A Novel Marker of Cardiovascular Risk." *Cardiology in Review*, July/August 2003, 11(4): 169–179.

Shenkier, T., L. Weir, M. Levine, I. Olivotto, T. Whelan and L. Reyno. 2004. "Clinical practice guidelines for the care and treatment of breast cancer: 15. Treatment for women with stage III or locally advanced breast cancer." *Canadian Medical Association Journal*, 16 March 2004, 170(6): 983–994.

Sicinski, P., and R.A. Weinberg. 1997. "A Specific Role for Cyclin D1 in Mammary Gland Development." *Journal of Mammary Gland Biology and Neoplasia* 2(4). http://www.springerlink.com/. Consulted 15 October 2006.

Siegelmann-Danieli, N., A. Tamir, H. Zohar, M.Z. Papa, L.L. Chetver, Z. Gallimidi, M.E. Stein and A. Kuten. 2003. "Breast Cancer in Women With Recent Exposure to Fertility Medications is Associated With Poor Prognostic Features." *Annals of Surgical Oncology*, November 2003, 10(9): 1031–1038.

Slomovitz, B.M., C.C. Sun, P.T. Ramirez, D.C. Bodurka, P. Diaz and K.H. Lu. 2004. "Does Tamoxifen Use Affect Prognosis in Breast Cancer Patients Who Develop Endometrial Cancer?" *Obstetrics & Gynecology*, August 2004, 104(2): 255–260.

Spiegel, A.J. and C.E. Butler. 2003. "Recurrence following Treatment of Ductal Carcinoma in Situ with Skin-Sparing Mastectomy and Immediate Breast Reconstruction." *Plastic and Reconstructive Surgery*, February 2003, 111(2): 706–711.

Stevens, L.A. and A. Levin. 2003. "Anaemia, cardiovascular disease and kidney disease: Integrating new knowledge in 2002." *Current Opinion in Nephrology and Hypertension*, March 2003, 12(2): 133–138.

Stix, G. 2003. "Signal Jammer: An academic experiment leads to a new class of drug for attacking heart disease." *Scientific American*, July 2003. http://www.sciam.com/article.cfm?articleID=0001F400-70D6-1ED9-8E1C809EC588EF21&pageNumber=1&catID=2. Consulted 18 February 2004.

Susnik, B., S. Frkovic-Grazio and M. Bracko. 2004. "Occult Micrometastases in Axillary Lymph Nodes Predict Subsequent Distant Metastases in Stage I Breast

Cancer: A Case-Control Study with 15-Year Follow-Up." *Annals of Surgical Oncology*, June 2004, 11(6): 568–572.

Szmitko, Paul E., C.-H. Wang, R.D. Weisel, J.R. de Almeida, T.J. Anderson and S. Verma. 2003. "New Markers of Inflammation and Endothelial Cell Activation: Part I." *Circulation: Journal of the American Heart Association*, 21 October 2003, 108(16): 1917–1923.

Taniyama, Y. and K.K. Griendling. 2003. "Reactive Oxygen Species in the Vasculature: Molecular and Cellular Mechanisms." *Hypertension: Journal of The American Heart Association*, December 2003, 42(6): 1075–1081.

Thomas, R. 2003. "Examining Quality of Life Issues in Relation to Endocrine Therapy for Breast Cancer." *American Journal of Clinical Oncology*, August 2003, 26(4) Supplement 1: S40–S44.

Torres, J.L. and P.M. Ridker. 2003. "Clinical use of high sensitivity C-reactive protein for the prediction of adverse cardiovascular events." *Current Opinion in Cardiology*, November 2003, 18(6): 471–478.

Toth-Fejel, S., P. Muller, B. Ham, K. Esvelt, N. Dumas, K. Calhoun and R. Pommier. 2004. "DNA Fingerprints Provide a Patient-Specific Breast Cancer Marker." *Annals of Surgical Oncology*, June 2004, 11(6): 560–567.

Tran, D. and J. Lawson. 2004. "Rates of Estrogen Receptor-[alpha] Expression Are No Different in Low-risk (Vietnam) and High-risk (Australian) Breast Cancer." *Applied Immunohistochemistry & Molecular Morphology*, June 2004, 12(2): 139–141.

Travis, J. 2004. "Pill puzzle: do antibiotics increase breast cancer risk?" *Science News*, 21 February 2004, 165(8): 118(1).

"Treatment. (Breast Cancer, part 2)." 1991. *Health News*, August 1991, 9(4): 1(4).

Truong, P.T., I.A. Olivotto, T.J. Whelan and M. Levine. 2004. "Clinical practice guidelines for the care and treatment of breast cancer: 16. Locoregional post-mastectomy radiotherapy." *Canadian Medical Association Journal*, 13 April 2004, 170(8): 1263–1273.

Umehara, H., E.T. Bloom, T. Okazaki, Y. Nagano, O. Yoshie and T. Imai. 2004. "Fractalkine in Vascular Biology: From Basic Research to Clinical Disease." *Arteriosclerosis, Thrombosis, and Vascular Biology: Journal of the American Heart Association*, January 2004, 24(1): 34–40.

Vallance, P. and J. Leiper. 2004. "Cardiovascular Biology of the Asymmetric Dimethylarginine:Dimethylarginine Dimethylaminohydrolase Pathway." *Arteriosclerosis, Thrombosis, and Vascular Biology: Journal of the American Heart Association*, June 2004, 24(6): 1023–1030.

van der Hel, O.L., P. Peeters, D.W. Hein, M.A. Doll, D.E. Grobbee, D. Kromhout and H. Bas Bueno de Mesquita. 2003. "NAT2 slow acetylation and GSTM1 null genotypes may increase postmenopausal breast cancer risk in long-term smoking women." *Pharmacogenetics*, July 2003, 13(7): 399–407.

Vick, G.W. III. 2003. "Recent advances in pediatric cardiovascular MRI." *Current Opinion in Pediatrics*, October 2003, 15(5): 454–462.

Vogel, C.L. 2003. "Update on the current use of hormonals as therapy in advanced breast cancer." *Anti-Cancer Drugs*, April 2003, 14(4): 265–273.

Wang, L., A. Hoque, R.Z. Luo, J. Yuan, Z. Lu, A. Nishimoto, J. Liu, A.A. Sahin, S.M. Lippman, R.C. Bast, Jr. and Y. Yu. 2003. "Loss of the Expression of the Tumor Suppressor Gene ARHI Is Associated with Progression of Breast Cancer." *Clinical Cancer Research*, 1 September 2003, 9: 3660–3666.

Watkins, D. 2003. "Hormone Hysteria?" *Scientific American*, 15 September 2003. http://www.sciam.com/print_version.cfm?articleID=000C63B6-2F80-1F62-905980A84189EEDF. Consulted 16 June 2004.

Weaver, C. 2002. "Breast cancer breakthroughs." *Nursing Management*, November 2002, 33(11): 27–33.

Weinfurt, K.P., L.D. Castel, Y. Li, J.W. Timbie, G.A. Glendenning and K.A. Schulman. 2004. "Health-Related Quality of Life Among Patients With Breast Cancer Receiving Zoledronic Acid or Pamidronate Disodium for Metastatic Bone Lesions." *Medical Care*, February 2004, 42(2): 164–175.

Weiss, L.K. et al. 2002. "Hormone Replacement Therapy Regimens and Breast Cancer Risk." *Obstetrics & Gynecology*, December 2002, 100(6): 1148–1158.

Weiss, M. and B. Loecher. 2003. "10 biggest breast cancer myths: if you believe any of them, you've been duped--or worse." *Prevention*, October 2003, 55(10): 110(7).

Whelan, S.A. and G.W. Hart. 2003. "Proteomic Approaches to Analyze the Dynamic Relationships Between Nucleocytoplasmic Protein Glycosylation and Phosphorylation." *Circulation Research: Journal of the American Heart Association*, 28 November 2003, 93(11): 1047–1058.

Willerson, J.T. and P.M. Ridker. 2004. "Inflammation as a Cardiovascular Risk Factor." *Circulation: Journal of the American Heart Association*, 1 June 2004, 109(21) Supplement: II-2–II-10.

Witters, L.M., J. Crispino, T. Fraterrigo, J. Green and A. Lipton. 2003. "Effect of the Combination of Docetaxel, Zoledronic Acid, and a COX-2 Inhibitor on the Growth of Human Breast Cancer Cell Lines." *American Journal of Clinical Oncology*, August 2003, 26(4) Supplement 2: S92–S97.

Wong, K. 2000. "Beating Breast Cancer." *Scientific American*. 19 September 2000. http://www.sciam.com/article.cfm?articleID=000EAAB5-74EE-1C61-B882809EC588ED9F&pageNumber=1. Consulted 18 February 2004.

Wong, K. 2001. "How Breast Cancer Starts and Spreads." *Scientific American*. 2 February 2001. http://www.sciam.com/article.cfm?articleID=000C09FF-BB7A-1C5A-B882809EC588ED9F&pageNumber=1. Consulted 18 February 2004.

Xydakis, A.M. and C.M. Ballantyne. 2003. "Role of non-high-density lipoprotein cholesterol in prevention of cardiovascular disease: Updated evidence from clinical trials." *Current Opinion in Cardiology*, November 2003, 18(6): 503–509.

Yan, S.F., R. Ramasamy, Y. Naka and A.M. Schmidt. 2003. "Glycation, Inflammation, and RAGE: A Scaffold for the Macrovascular Complications of Diabetes and Beyond." *Circulation Research: Journal of the American Heart Association*, 12/26 December 2003, 93(12): 1159–1169.

Yang, X., and M.E. Lippman. 1999. "BRCA1 and BRCA2 in breast cancer." *Breast Cancer Research and Treatment* 54: 1–10.

Zambon, A., S.S. Deeb, P. Pauletto, G. Crepaldi and J.D. Brunzell. 2003. "Hepatic lipase: a marker for cardiovascular disease risk and response to therapy." *Current Opinion in Lipidology*, April 2003, 14(2): 179–189.

Zhang, S.M. 2004. "Role of vitamins in the risk, prevention, and treatment of breast cancer." *Current Opinion in Obstetrics and Gynecology*, February 2004, 16(1): 19–25.

Zheng, T., T.R. Holford, S.T. Mayne, J. Luo, P.H. Owens, B. Zhang, W. Zhang and Y. Zhang. 2002. "Radiation exposure from diagnostic and therapeutic treatments and risk of breast cancer." *European Journal of Cancer Prevention*, June 2002, 11(3): 229–235.

"Zip out cancer." 2003. *New Scientist*, 4 January 2003, 177(2376): 20(1).

French Texts

Abrial, C., F. Kwiatkowski, R. Chevrier, F. Gachon, H. Curé and P. Chollet. 2005. "Potentiel thérapeutique de la mélatonine dans la prise en charge de la pathologie cancéreuse." *Pathologie Biologie* 53(5): 265–268. http://france.elsevier.com/direct/PATBIO/. Consulted 15 October 2006.

Adel, M., V. Guis and M. Rasigni. 2004. "Étude de la faisabilité du scorage automatique de fantômes mammographiques par traitement d'image." *ITBM-RBM*. http://www.sciencedirect.com/. Consulted 15 October 2006.

Alecu, C., P. Abraham, C. Ternisien, B. Enon and J.-L. Saumet. 1999. "Thrombocytémie essentielle et accident ischémique cérébral: à propos de deux observations." *Journal des maladies vasculaires*, 24(4): 300–202.

Angèle, S., P. Tanière and J. Hall. 2001. "Que savons-nous de l'expression de la protéine ATM dans le tissu mammaire?" *Bulletin du Cancer*, July 2001, 88(7): 671–675.

Arnal, J.-F., P. Gourdy, B. Garmy-Susini, É. Delmas and F. Bayard. 2003. "Effets vasculaires des oestrogènes." *Médecine/Sciences*, December 2003, 19(12). http:

//www.erudit.org/revue/ms/2003/v19/n12/007398ar.html. Consulted 18 May 2004.

Asmar R., M. Safar and P. Queneau. 2003. "Placebo et nouvelles perspectives thérapeutiques dans l'artériopathie des membres inférieurs." *Journal des maladies vasculaires*, 28(3): 117–120.

Aubard, Y., J. Mollard and V. Fermeaux. 2004. "Comment éviter les aléas de l'examen extemporané du ganglion sentinelle dans le cancer du sein ?" *Gynécologie Obstétrique & Fertilité* 32: 981–984.

Augros, M., A. Buénerd, M. Devouassoux-Shisheboran and G. Berger. 2004. "Variant histiocytoïde du carcinome lobulaire infiltrant du sein. À propos de trois cas." *Ann Pathol*, 24: 259–263.

Azoulay, C. 2004. "Ménopause en 2004: le traitement hormonal substitutif n'est plus ce qu'il était." *La revue de médecine interne* 25 (2004): 806–815.

Bachelot, T., I. Ray-Coquard, D. Coeffic, H. Barletta, M.-C. Gouttebel, B. Mayer, T. Muron, G. de Laroche, M. Vincent, F. Farsi, T. Philip. 2002. "Traitement adjuvant du cancer du sein: création d'un arbre de décision thérapeutique au sein d'un réseau de soins." *Bulletin du Cancer*, October 2002, 89(10): 897–903.

Balaton, A.J. et al. 1999. "Recommandations pour l'évaluation immunohistochimique des récepteurs hormonaux sur coupes en paraffine dans les carcinomes mammaires. Mise à jour 1999." *Annales de pathologie* 19: 336.

Baril, J.-G. and P. Junod. 2004. "Les effets indésirables des traitements antirétroviraux sont-ils importants ?" *Le Médecin du Québec*, January 2004, 39(1): 65–77.

Barranger, E., D. Grahek, M. Antoine, Y. Benchimol, J.N. Talbot and S. Uzan. 2002. "Le ganglion sentinelle dans le cancer du sein: Étude de faisabilité et expérience initiale." *Gynécologie Obstétrique & Fertilité* 30: 492–497.

Basuyau, J.-P., M.-P. Blanc-Vincent, J.-M. Bidart, A. Daver, L. Deneux, N. Eche, G. Gory-Delabaere, M.-F. Pichon and J.-M. Riedinger. 2000. "Standards, Options et Recommandations (SOR): marqueurs tumoraux sériques du cancer du sein." *Bulletin du Cancer*, October 2000, 87(10): 723–737.

Bauduceau, B., O. Dupuy, H. Mayaudon, L. Bordier, J. Margery and J.-P. Le Berre. 2004. "Athérosclérose: du beurre sur les artères ?" *EMC-Médecine* 1 (2004): 27–36.

Béliveau, L. and L. Léger. 2004. "L'évaluation de la condition physique où, quand, comment, pourquoi ?" *Le Médecin du Québec*, April 2004, 39(4): 61–71.

Bénard, J. 1997. "TSG101 et cancer du sein: un gène suppresseur de tumeur bien nommé ?" *Bulletin du Cancer*, December 1997, 84(12): 1141–1142.

Benchellal, Z., A. Wagner, Y. Harchaoui, N. Huten and G. Body. 2002. "Cancer du sein chez l'homme: à propos de 19 cas." *Annales de Chirurgie* 127: 619–623.

Benoit, L., C. Franceschini, A. Margarot, L. Arnould, J. Fraisse and J. Cuisenier. 2002. "Traitement conservateur du cancer du sein: Proposition d'une incision cutanée combinée pour un meilleur résultat esthétique." *Annales de chirurgie*, 129: 310–312.

Bertucci, F, V. Nasser, R. Houlgatte and D. Birnbaum. 2002. "Profils d'expression génique et puces à ADN dans le cancer du sein: intérêt pronostique." *Bulletin du Cancer*, June 2002, 89(6): 571–574.

Bertucci, F., R. Houlgatte, C. Nguyen, A. Benziane, V. Nasser, S. Granjeaud, B. Tagett, B. Loriod, A. Giaconia, J. Jacquemier, P. Viens and D. Birnbaum. 2001. "Typage moléculaire du cancer du sein: Transcriptome et puces à ADN" *Bulletin du Cancer*, March 2001, 88(3): 277–286.

Beuzeboc, P., S. Scholl, X. Sastre Garau, A. Vincent-Salomon, P. de Cremoux, J. Couturier, T. Palangié and P. Pouillart. 1999. "Anticorps humanisé anti-HER2 (Trastuzumab ou Herceptin): une avancée thérapeutique majeure dans le cancer du sein surexprimant cet oncogène ?" *Bulletin du Cancer*, June 1999, 86(6): 544–549.

Blais, C. 2001. "Suivre une diète... est-ce vraiment nécessaire, docteur ?" *Le Médecin du Québec*, April 2001, 36(4): 49–51.

Blais, C. 2001a. "Mais quelle diète, docteur... ?" *Le Médecin du Québec*, April 2001, 36(4): 53–58.

Blais, C. and N. Rivard-Gervais. 2001. "Questions sur le cholestérol: comment répondre à vos patients ?" *Le Médecin du Québec*, April 2001, 36(4): 83–92.

Blanchard, J.-M. 2003. "Des oncogènes aux régulateurs de la mitose: Un changement de perspective dans notre vision des processus cancéreux." *Médecine/Sciences*, February 2003. 19(2). http: //www.erudit.org/revue/ms/2003/v19/n2/000688ar.html. Consulted 18 May 2004.

Blandy, C., S. Schwab, D. Stoppa-Lyonnet and A. Dazord. 1998. "Impact psychosocial d'une première consultation d'oncogénétique dans un contexte familial de cancers du sein et de l'ovaire." *Bulletin du Cancer*, July 1998, 85(7): 637–643.

Blot, E, O. Delastre and H. Lévesque. 1999. "La Modulation de l'angiogenèse: Un nouvel outil thérapeutique pour les maladies vasculaires." *Journal des maladies vasculaires* 24(3): 189–193.

Blottière, L. 2002. "Cancer du sein: des puces à la rescousse." *La Recherche*, May 2002, 353: 14–5.

Bobin, J.-Y., C. Zinzindohoue and C. Faure Virelizier. 2001. "Evolution actuelle des techniques chirurgicales dans le traitement des cancers invasifs du sein." *Bulletin du Cancer*, January 2001, 88(1): 45–53.

Bobin, J.-Y., N. Guiochet and S. Saez. 2002. "Guérie d'un cancer du sein." *Bulletin du Cancer*, June 2002, 89(6): 579–587.

Boisseau, M.-R. 2004. "Hémorhéologie clinique. Concept, physiopathologie et applications aux maladies vasculaires." *EMC-Cardiologie Angéiologie* 1 (2004): 364–381.

Bonadona, V. and C. Lasset. 2003. "Prédispositions héréditaires au cancer du sein: après BRCA1 et BRCA2, quel(s) autre(s) gène(s) ?" *Bulletin du Cancer*, July 2003, 90(7): 587–594.

Bonnefont-Rousselot, D., J. Peynet, J.-L. Beaudeux, P. Thérond, A. Legrand and J. Delattre. 2002. "Stress oxydant, fonctions vasculaires et athérosclérose." *Nutrition clinique et métabolisme* 16 (2002): 260–267.

Bonnier, P., R. Sakr, F. Bessenay, C. Lejeune, C. Charpin, P.M. Martin and L. Piana. 2000. "Effets des traitements hormonaux substitutifs de la ménopause sur les facteurs pronostiques des cancers du sein." *Gynécologie Obstétrique & Fertilité* 28: 745–753.

Borella, L., S. Finkel, N. Crapeau, P. Peuvrel, M. Sauvage, L. Perrier, E. Lepage, J. Villeminot and B. Garrigues. 2002. "Volume et coût de la prise en charge hospitalière du cancer en France en 1999." *Bulletin du Cancer*, September 2002, 89(9): 809–821.

Bouchard, R. 2001. "L'hormonothérapie à la ménopause pourquoi ? pour qui ?" *Le Médecin du Québec*, May 2001, 36(5): 103–111.

Brain, E.G.C. 2000. "Chimiothérapie néoadjuvante dans le traitement du cancer du sein." *Annales de médecine interne* 151: 215–219.

Briffod, M., V. Le Doussal and F. Spyratos. 2002. "Détermination des récepteurs hormonaux par immunohistochimie sur cytoblocs de cytoponctions des cancers du sein." *Bulletin du Cancer*, October 2001, 88(10): 1028–1035.

Broët, P., M.-F. Pichon, H. Magdelenat, J.-C. Delarue, F. Spyratos, J.-P. Basuyau, S. Saez, A. Rallet, P. Courrière, R. Millon and B. Asselain. 1998. "Rôle pronostique à long terme des récepteurs stéroïdiens dans le cancer du sein." *Bulletin du Cancer*, April 1998, 85(4): 347–352.

Bussières, E., B. Barreau, B. Doche de la Quintane, C. Tunon de Lara, O. Le Touze, C. Henriquès, G. Mac Grogan and M.H. Dilhuydy. 2003. "Les prélèvements mammaires en stéréotaxie: macrobiopsies avec aspiration et biopsies chirurgicales stéréotaxiques." *Gynécologie Obstétrique & Fertilité* 31: 256–264.

Buttarelli, M., G. Houvenaeghel, M. Martino, I. Rossi, I. Ronda, F. Ternier, A. Tallet and J. Jacquemier. 2004. "Prélèvement de ganglions sentinelles dans les carcinomes intracanalaires du sein (micro-invasion)." *Annales de chirurgie* 129: 508–512.

Caligiuri, G. 2004. "Rôle de l'immunité dans l'athérosclérose et dans les syndromes coronariens aigus." *Médecine/Sciences*, February 2004, 20(2). http:

//www.erudit.org/revue/ms/2004/v20/n2/007677ar.html. Consulted 18 May 2004.

Calvet, P., V. Chabbert, P. Chemla, S. Moussouni, L. Bouchard, F. Joffre and H. Rousseau. 2001. "Traitement par voie endoluminale des anévrismes périphériques par endoprothèse couverte." *Journal des maladies vasculaires* 26(5): 299–306.

"Cancer du sein: Que fait la protéine BRCA1?" 1997. *La Recherche*, April 1997, 297: 17.

Capron, L. 1997. "Peut-on attraper un infarctus ?" *La Recherche*, January 1997, 294: 32–34.

Carpentier, A. 2004. "Les cent tours du tour de taille !" *Le Médecin du Québec*, February 2004, 39(2): 83–89.

Catros-Quemener, V., F. Bouet and N. Genetet. 2003. "Immunité anti-tumorale et thérapies cellulaires du cancer." *Médecine/Sciences*, January 2003, 19(1). http: //www.erudit.org/revue/ms/2003/v19/n1/000756ar.html. Consulted 18 May 2004.

Chailleux, C., C. Giamarchi, V. Morales, F. Moro and H. Richard-Foy. 2000. "Remodelage de la structure chromatinienne des régions régulatrices de deux gènes oestrogéno-régulés dans des lignées de cellules cancéreuses mammaires humaines." *Annales d'endocrinologie* 61: 130–135.

Chalès, G. and P. Guggenbuhl. 2004. "Glycogénoses, hyperoxaluries, aminoacidopathies et hyperlipidémies." *EMC-Rhumatologie Orthopédie* 1 (2004): 423–435.

Charafe-Jauffret, E., J. Jacquemier, B. Lelong, P. Meynard, M.-P. Mathoulin-Portier, G. Houvenhaegel and J. Hassoun. 2001. "Carcinome mammaire développé sur une adénose microglandulaire: à propos d'un cas." *Ann Pathol* 21: 435–438.

Chène, P. 1999. "Une molécule au cœur des mécanismes du cancer." *La Recherche*, September 1999, 323: 46–50.

Cherfils, J. and P. Pacaud. 2004. "L'activation des protéines G en 3 dimensions: un pas vers l'inhibition thérapeutique des facteurs d'échange nucléotidique."

Médecine/Sciences, April 2004, 20(4). http: //www.erudit.org/revue/ms/2004/v20/n4/008110ar.html. Consulted 18 May 2004.

Chevallier, P., M. Haïdopoulos and D. Mantovani. 2003. "Docteur, mon spécialiste m'a proposé l'implantation d'une prothèse artérielle synthétique. Qu'est-ce que vous en pensez?" *Le Médecin du Québec*, June 2003, 38(6): 115–123.

Chidiac, C. and E. Braun. 2002. "Athérosclérose, sclérose en plaque et maladie d'Alzheimer: Quel rôle pour les Herpesviridae ?" *Pathologie Biologie* 50 (2002): 463–468.

Clavel-Chapelon, F. and C. Hill. 2000. "Traitement hormonal de la ménopause et risque de cancer du sein." *La Presse Médicale*, 21 October 2000, 29(31): 1688–1693.

Colleau, M., G. Magalon and P. Bonnier. 2005. "Cancer du sein diagnostiqué par la réduction mammaire. Étude rétrospective sur une période de trois ans." *Annales de chirurgie plastique et esthétique*, April 2005, 50(2): 127–133. http://www.sciencedirect.com/. Consulted 15 October 2006.

Collet, J.-P., R. Choussat and G. Montalescot. 2004. "L'agrégation plaquettaire et ses inhibiteurs dans les syndromes coronariens aigus." *Médecine/Sciences*, March 2004, 20(3). http: //www.erudit.org/revue/ms/2004/v20/n3/007848ar.html. Consulted 18 May 2004.

Colonna, M., F. Ménégoz, C. Exbrayat, M.-F. Veran-Peyret, T. Philip and R. Schaerer. 1997. "Estimation de la prévalence des cancers colorectaux et du sein dans la région Rhône-Alpes." *Bulletin du Cancer*, February 1997, 84(2): 162–168.

Constance, C. and N. Pranno. 2002. "Le traitement de l'hypertension: quelles sont les nouveautés?" *Le Clinicien*, November 2002, 85–93.

Constans, J., P. Gosse, C. Conri and J. Clémenty. 2002. "La mesure de la distensibilité artérielle par la méthode du QKd: un nouveau marqueur vasculaire." *Rev Méd Interne* 23: 308–311.

Cornez, N. and M.J. Piccart. 2002. "Cancer du sein et Herceptin(r)." *Bulletin du Cancer*, November 2000, 87(11): 847–858.

Côté, G. 2002. "Contraceptifs oraux et lipides: l'ABC pour s'y retrouver." *Le Médecin du Québec*, January 2002, 37(1): 49–53.

Cottu, P.-H. and M. Espié. 1999. "Chimioprévention du cancer du sein: que penser après Eurocancer 1999 ?" *Bulletin du Cancer*, October 1999, 86(10): 870–873.

Coupier, I. and D. Stoppa-Lyonnet. 2002. "CHEK2 et risque de cancer du sein." *Bulletin du Cancer*, November 2002, 89(11): 921–922.

Cracowski, J.-L. 2004. "Les isoprostanes: un rôle physiopathologique potentiel en pathologie vasculaire." *La revue de médecine interne* 25 (2004): 459–463.

Cracowski, J.-L., O. Berdeaux and T. Durand. 2005. "Les isoprostanes, biomarqueurs de peroxydation lipidique chez l'homme. Partie 3: biomarqueurs et médiateurs en physiologie et pathologie vasculaire." *Pathologie Biologie*, July 2005, 53(6) : 364–368. http://www.sciencedirect.com/. Consulted 15 October 2006.

Dauplat, J., F. Guillemin, G. Depadt and L. Borella. 2002. "Analyse de l'offre de chirurgie cancérologique en France." *Bulletin du Cancer*, February 2002, 89(2) Numéro spécial: 28–32.

de Crémoux, P. 2000. "Les inhibiteurs de l'aromatase: aspects pharmacologiques." *Bulletin du Cancer*, December 2000, 87(12) Numéro spécial: 23–30.

Debourdeau, P., T. Bachelot, C. Zammit, M. Aletti, C. Gallineau and J. Gligorov. 2004. "Traitement des bouffées de chaleur associées au cancer du sein." *Bulletin du Cancer*, April 2004, 91(4): 339–349.

Delcourt, C. 1999. "Les agissements des radicaux libres." *La Recherche*, July/August 1999, 322: 62–5.

Deléglise, A. 2002. "Traitement hormonal substitutif mis à mal." *CyberSciences*, 18 July 2002. http: //www.cybersciences.com/cyber/3.0/n2868.asp. Consulted 18 February 2004.

Delozier, T., O. Switsers, J.-Y. Génot, J.-M.l Ollivier, M. Héry, M. Namer, M. Frenay, P. Kerbrat, J.-P. Julien, A. Naja, M. Janvier and J. Macé-Lesec'h. 1997. "Tamoxifène adjuvant retardé dans le cancer du sein curable. Résultats d'un essai coopératif randomisé." *Bulletin du Cancer*, January 1997, 84(1): 25–30.

Deniaud-Alexandre, E., B. Lauratet, J.P. Lefranc, C. Genestié, D. Lerouge, L. Moureau-Zabotto and E. Touboul. 2004. "Rechute locale isolée après traitement conservateur pour un carcinome mammaire de stade I ou II, à propos de 57 patientes." *Cancer/Radiothérapie* 8: 95–107.

Dequanter, D., D. Hertens, I. Veys, J.M. Nogaret, D. Larsimont and P. Bourgeois. 2001. "Envahissement du ganglion sentinelle dans les cancers mammaires T0-T1." *Annales de Chirurgerie* 126: 654–658.

Dequanter, D., P. Michel, P. De Wilde, J. Ferreira and J.-P. Dereume. 2003. "Anévrisme mycotique de la carotide interne extra-crânienne." *Journal des maladies vasculaires* 28(3): 151–154.

Desauw, C., E. Hachulla, Y. Boumbar, J. Bouroz-Joly, D. Ponard, J. Arvieux, S. Dubucquoi, A.L. Fauchais, P.Y. Hatron and B. Devulder. 2002. "Syndrome des antiphospholipides avec anticorps antiphosphatidyléthanolamine isolés: À propos de 20 cas." *Rev Méd Interne* 23: 357–363.

D'Hondt, L., M. André, J.-L. Canon, T. Guillaume, C. Doyen, A.-M. Feyens, B. Chatelain, A. Dromelet, Y. Humblet, J. Longueville and M. Symann. 1997. "Efficacité d'une chimiothérapie FEC à forte dose pour la mobilisation des cellules souches hématopoïétiques dans le sang périphérique." *Bulletin du Cancer*, July 1997, 84(7): 729–733.

Dodin, S. C. Blanchet and I. Marc. 2003. "Phytooestrogènes chez la femme ménopausée." *Médecine/Sciences*, October 2003, 19(10). http://www.erudit.org/revue/ms/2003/v19/n10/007179ar.html. Consulted 18 May 2004.

Domont, J., M. Namer, D. Khayat and J.-P. Spano. 2004. "Hormonothérapie néoadjuvante dans le cancer du sein: revue de la littérature." *Bulletin du Cancer*, January 2004, 91(1): 55–62.

Drouet, L. 2004. "La résistance à l'aspirine existe-t-elle ?" *La revue de médecine interne* 25 (2004): 101–103.

Drouet, L. and C. Bal Dit Sollier. 2002. "Fibrinogène et risque d'accident cardio-vasculaire: L'exemple des molécules à la fois marqueurs et facteurs de risque." *Journal des maladies vasculaires* 27(3): 143–156.

Dufour, R. 2001. "Le futur: au-delà du cholestérol." *Le Médecin du Québec*, April 2001, 36(4): 77–82.

Dufresne, J. 2003. "Le cancer du sein: que faire après le traitement?" *Le Clinicien*, November 2003, 41–45.

Duong, M.,Y. Cottin, M. Froidure, J.M. Petit, L. Piroth, M. Zeller, I. L'huillier, A. Fargeot, M. Mahrousseh, P. Chavanet, J.E.Wolf and H. Portier. 2003. "Les patients infectés par le virus de l'immunodéficience humaine sous traitement antirétroviral ont-ils un risque cardiovasculaire accru ?" *Annales de Cardiologie et d'Angiologie* 52 (2003): 302–307.

Duriez, P. 2004. "Mécanismes de formation de la plaque d'athérome." *La revue de médecine interne* 25 (2004): S3–S6.

Dussault, S. 1997. "BRCA1: l'ennemi des seins et des ovaires." *CyberSciences*, 25 November 1997. http: //www.cybersciences.com/cyber/1.0/1%5F171%5F210.asp. Consulted 18 February 2004.

Eisinger, F. and A. Noizet. 2002. "Cancer du sein et grossesse: Modalités de décision et point de vue de la mère." *Bulletin du Cancer*, September 2002, 89(9): 755–757.

Eisinger, F., B. Bressac, D. Castaigne, P.-H.Cottu, J. Lansac, J.-P. Lefranc, A. Lesur, C. Noguès, J. Pierret, S. Puy-Pernias, H. Sobol, A. Tardivon, H. Tristant and R. Villet. 2004. "Identification et prise en charge des prédispositions héréditaires aux cancers du sein et de l'ovaire (mise à jour 2004)." *Bulletin du Cancer*, March 2004, 91(3): 219–37.

"En abrégé." 2003. *Le Clinicien*, July 2003, 18(7): 24. http: //www.stacommunications.com/journals/pdfs/clinicien/clijuly03/enabrege.pdf. Consulted 15 October 2006.

Espié, M. 1997. "Le traitement du cancer du sein." *La Presse médicale* 27(26): 1332.

Essiambre, R. 2003. "Maladie coronarienne et athéromatose périphérique: Une seule et même maladie, qu'on se le dise !" *Le Médecin du Québec*, May 2003, 38(5): 77–83.

Félétou, M., R. Busse, G. Edwards, I. Fleming, A.H. Weston and P.M. Vanhoutte. 2003. "Dialogue entre cellules endothéliales et cellules musculaires lisses." *Médecine/Sciences*, December 2003, 19(12). http: //www.erudit.org/revue/ms/2003/v19/n12/007400ar.html. Consulted 18 May 2004.

Ferrero, J.-M., N. Magné, C. Foa, R. Largillier and M. Namer. 2003. "Tolérance cardiaque des associations paclitaxel-anthracyclines dans le cadre de la prise en charge du cancer du sein." *Bulletin du Cancer*, March 2003, 90(3): 219–226.

Ferrières, J. 2004. "Facteurs de risque, lipoprotéines et activité physique et sportive." *Science & Sports* 19 (2004): 118–123.

FNCLCC. 2004. "Recommandations pour la pratique clinique: Mise à jour 2003 des Standards, Options et Recommandations pour l'utilisation de la TEP-FDG dans la prise en charge des cancers gynécologiques et cancers du sein." *Gynécologie Obstétrique & Fertilité* 32: 352–371.

Fondriner, E., O. Guérin and G. Lorimier. 1997. "Etude comparative de l'évolution métastatique des carcinomes canalaires et lobulaires du sein à partir de deux séries appariées (376 patientes)." *Bulletin du Cancer*, December 1997, 84(12): 1101–7.

Fontana, X, I. Peyrotte, E. Valente, C. Rossi, F. Ettore, M. Namer and F. Bussière. 1997. "Glutathion S-transférase mu 1 (GSTM1): gène de susceptibilité du cancer du sein." *Bulletin du Cancer*, January 1997, 84(1): 35–40.

Fournier, A., C. Hill and F. Clavel-Chapelon. 2003. "Traitement hormonal substitutif de la ménopause et risque de cancer du sein." *Bulletin du Cancer*, October 2003, 90(10): 821–831.

Frebourg, T. et al. 2001. "Le syndrome de Li-Fraumeni: Mise au point, données nouvelles et recommandations pour la prise en charge." *Bulletin du Cancer*, June 2001, 88(6): 581–587.

Fredenrich, A., P.-J. Bouillanne and M. Batt. 2004. "Artériopathie diabétique des membres inférieurs." *EMC-Endocrinologie* 1 (2004): 117–132.

Fréneaux, P., C. Nos, J.-Y. Charvolin, A. Vincent-Salomon, B. Zafrani, R.J. Salmon, K.B. Clough and X. Sastre-Garau. 2000. "Intérêt de l'examen macroscopique pour l'authentification des ganglions sentinelles axillaires repérés par le Bleu Patenté seul au cours de la chirurgie des cancers du sein." *Annales de pathologie* 20: 545–548.

Fricker, J.-P., D. Muller, B. Cutuli, J.-F. Rodier, J.-C. Janser, G.-M. Jung, R. Mors, T. Petit, P. Haegele and J. Abecassis. 2000. "Mutations germinales du gène BRCA1 dans le Nord-Est de la France." *Bulletin du Cancer*, October 2000, 87(10): 739–744.

Garbay, J.-R., O. Picone, G. Baron-Merle, S. Yacoub, S. Lasry, M.-C. Missana, L. Barreau-Pouhaer, V. Fourchotte, A. Cavalcanti and A. Thoury. 2004. "Le curage de l'aisselle avec capitonnage musculaire sans drainage." *Gynécologie Obstétrique & Fertilité* 32: 1039–1046.

Garnier, E. 2002. "Prévention cardiovasculaire primaire et si l'HTS avait quand même un effet protecteur ?" *Le Médecin du Québec*, September 2002, 37(9): 91–92.

Garnier, E. 2002a. "Hormonothérapie à faible dose est-ce la solution ?" *Le Médecin du Québec*, September 2002, 37(9): 87–89.

Garnier, E. 2002b. "Les phytoestrogènes rien ne remplace l'hormonothérapie." *Le Médecin du Québec*, September 2002, 37(9): 92–94.

Garnier, E. 2002c. "Étude LIFE un ARA plus efficace qu'un bêta-bloquant." *Le Médecin du Québec*, August 2002, 37(8): 126–140.

Garnier, E. 2002d. "Nouvelles études sur l'hormonothérapie substitutive. Qu'en est-il maintenant ?" *Le Médecin du Québec*, August 2002, 37(8): 14, 127.

Gauthier, N., L. Arnould, A. Chantôme, D. Reisser, A. Bettaieb, S. Reveneau and J.-F. Jeannin. 2004. "Faut-il inhiber ou stimuler la production de monoxyde d'azote des cellules cancéreuses mammaires ?" *Bulletin du Cancer*, September 2004, 91(9): 705–712.

Gendreau, R. 2003. "Le traitement pharmacologique de l'angine stable." *Le Médecin du Québec*, May 2003, 38(5): 63–69.

Gerber, M. 2001. "Micronutriments et microconstituants végétaux protecteurs dans le cancer du sein." *Bulletin du Cancer*, October 2001, 88(10): 943–953.

Gervois, P. and J.-C. Fruchart. 2003. "PPARγ: un récepteur nucléaire majeur de l'adipogenèse." *Médecine/Sciences*, January 2003, 19(1). http: //www.erudit.org/revue/ms/2003/v19/n1/000751ar.html. Consulted 18 May 2004.

Gisserot , O., G. Cellarier, J. Bonal, H. Thouard and G.-V. Dussarat. 1999. "Anévrisme poplité juvénile." *Journal des maladies vasculaires* 24(4): 306–308.

Goggin, P. 2002. "Dépistage du cancer du sein: où en sommes-nous?" *Le Médecin du Québec*, October 2002, 37(10): 85–90.

Gonzalez, M. 2004. "Avons-nous peur de traiter l'hypertension?" *Le Médecin du Québec*, August 2004, 39(8): 71–77.

Gonzalez, M. and J. Palardy. 2004. "Hypertension artérielle et diabète." *Le Médecin du Québec*, November 2004, 39(11): 99–105.

Gorins, A., M. Espié, N. Bedairia, F. Perret, B. Tournant, H. Novak, E. Lucchi-Angelier and M. Marty. 2003. "Thérapeutique hormonale substitutive après cancer du sein: une étude de 230 patientes, avec cas-témoins." *Gynécologie Obstétrique & Fertilité* 31: 614–619.

Guastalla, J.-P., T. Bachelot and I. Ray-Coquard. 2004. "Cyclo-oxygénase 2 et cancer du sein. Des concepts biologiques aux essais thérapeutiques." *Bulletin du Cancer*, January 2004, 91(1): 99–108.

Hermans, J., F. Bodart, D. François, P. Merlo, J.-P. Fauconnier, A. Schmitz, R. Gérard, D. de Ruyver and L. Eeckhoudt. "Apport de la mammoscintigraphie 99m Tc MiBi dans le diagnostic et le suivi des cancers du sein." *Bulletin du Cancer*, April 2000, 87(4): 334–340.

Hill, C. 2003. "Intérêt du dépistage systématique du cancer du sein chez les femmes de plus de 70 ans." *Bulletin du Cancer*, December 2003, 90(12): 1035–1037.

"Hormone et pression artérielle." 1999. *La Recherche*, November 1999, 325.

Houvenaeghel, G., M. Martino, J. Jacquemier, V. Moutardier, A. Tallet, P. Viens, B. Puig and V.J. Bardou. 2003. "Du risque de sous-traitement des cancers du sein en utilisant la technique du ganglion sentinelle." *Bulletin du Cancer*, May 2003, 90(5): 467–473.

Kirkiacharian, S. 2000. "Les modulateurs de l'activité estrogénique." *Annales pharmaceutiques françaises* 58: 383–391.

Kolb, J.-P. 2001. "Rôle pro- et anti-apoptotique du monoxyde d'azote, NO." *Sciences de la vie / Life Sciences* 324: 413–424.

Krulik, M. 1997. "Tumeurs de Krükenberg et métastases ovariennes du cancer du sein." *La Presse médicale*, 29 March 1997, 26(10): 452–454.

L'Allier, P.L. 2003. "L'athérosclérose: quelles sont les nouvelles stratégies d'intervention?" *Le Clinicien*, March 2003, 113–120.

Lambert, M. 2002. "La prévention des maladies cardiovasculaires chez les jeunes: mieux vaut prévenir que guérir!" *Le Clinicien*, October 2002, 65–72.

Larsen, C.-J. 2001. "Cycline D1 et cancer du sein: la nouvelle de l'été 2001 ?" *Bulletin du Cancer*, August 2001, 88(8): 717–718.

Launois, R.J., J.M. Reboul-Marty and J. Bonneterre. 1997. "Evaluation médico-économique de la chimiothérapie de 2e ligne dans le cancer du sein métastatique: comparaison du docétaxel, du paclitaxel et de la vinorelbine." *Bulletin du Cancer*, July 1997, 84(7): 709–721.

Laurian, C., C. Saliou, L. Guillemot, J.M. Clerget and X. de Kerangal. 2004. "Pathologie athéroscléreuse des troncs supra-aortiques." *EMC-Cardiologie Angéiologie* 1 (2004): 426–436.

Lavelle, F. and A. Jehanno. 1998. "Actualités en pharmacologie antitumorale: Réalités et perspectives en 1998." *Bulletin du Cancer*, January 1998, 85(1): 83–88.

Leblond, J. 2001. "La rosiglitazone (AvandiaTM)." *Le Médecin du Québec*, December 2001, 36(12): 129–138.

Lefort, E., O. Groussard, J. Bouquet de la Jolinière, C. Degott and B. Zafrani. 1999. "Carcinome mammaire à cellules fusiformes et différenciation neuro-endocrine.

Une entité rare pouvant simuler une tumeur bénigne." *Annales de pathologie* 19: 309.

Lerebours, F. 2006. "Traitements néoadjuvants du cancer du sein: Marqueurs géno- et phénotypiques de la réponse thérapeutique et du pronostic." *Pathologie Biologie*, May 2006, 54(4): 209–214. http://www.sciencedirect.com/. Consulted 15 October 2006.

Leriche, N. and J. Bonneterre. 1997. "Progestatifs et métastases osseuses dans les cancers du sein." *Bulletin du Cancer*, September 1997, 84(9): 891–894.

Lermusiaux, P., R. Martinez, A. Donadey, F. Bleuet and L. Castellani. 2000. "Échographie endo-artérielle: Limites et perspectives." *Journal des maladies vasculaires* 25(4): 229–236.

Lerouge, D., E. Touboul, J.P. Lefranc, C. Genestie, L. Moureau-Zabotto and J. Blondon. 2004. "Cancer du sein localement évolué non inflammatoire traité par association de chimiothérapie et de radiothérapie à dose préopératoire: réactualisation des résultats d'une série de 120 patientes." *Cancer/Radiothérapie* 8: 155–167.

Levenson, J., M. Del-Pino and A. Simon. 2000. "Rhéologie du sang, de la paroi artérielle et facteurs de risque cardiovasculaire." *Journal des maladies vasculaires* 25(4): 237–240.

Levesque, H. 2004. "Les effets non lipidiques des statines au cours de l'athérosclérose." *La revue de médecine interne* 25 (2004): 783–785.

Lilliu, H., D. Stevens, C. Brun, J. Morel, C. Le Pen, J. Bonastre, F. Bachelot, C. Davesne, A. Gentile, É. Hirlimann, J.-C. Sabourin, J. Berlie and J. Rouëssé. 2002. "Evaluation rétrospective du coût du cancer du sein dans un Centre de lutte contre le cancer." *Bulletin du Cancer*, June 2002, 89(6): 635–642.

Lizard G. and P. Gambert. 2001. "Implication et modes d'action des agents infectieux dans la formation de la plaque d'athérome. Infection et athérosclérose." *Pathologie Biologie* 49: 824–829.

Mallat, Z. and A. Tedgui. 2004. "Apoptose et syndromes coronariens aigus." *Médecine/Sciences*, March 2004, 20(3). http:

//www.erudit.org/revue/ms/2004/v20/n3/007849ar.html. Consulted 18 May 2004.

Martin, G. 2003. "Le cancer du sein: Quoi de neuf en hormonothérapie ?" *Le Clinicien*, April 2003, 90–98.

Martínez, M.C., C. Kunzelmann and J.-M. Freyssinet. 2004. "Remodelage de la membrane plasmique et stimulation cellulaire." *Médecine/Sciences*, February 2004, 20(2). http: //www.erudit.org/revue/ms/2004/v20/n2/007679ar.html. Consulted 18 May 2004.

Masse, R. 2000. "Rayonnements ionisants." *Sciences de la vie / Life Sciences* 323: 633–640.

Maublant, J., F. Cachin, D. Mesta and B. Geissler. 2002. "Le repérage du ganglion sentinelle en médecine nucléaire." *Bulletin du Cancer*, July – August 2002, 89(7): 671–680.

Mayoussi, C., H. Akoudad, L. Villalba, C. Dauphin, J.R. Lusson, S. Ztot and J. Cassagnes. 2004. "Thrombus flottant de la crosse de l'aorte: une cause rare d'embolies artérielles périphériques. À propos d'un cas clinique." *Journal des maladies vasculaires* 29(2): 94–98.

Mazeron, J.J. 2005. "Compte rendu de la 23e réunion de l'European Society for Therapeutic Radiology and Oncology (ESTRO). Amsterdam, 24–28 October 2004." *Cancer/Radiothérapie*, March 2005, 9(2): 122–126. http://www.sciencedirect.com/. Consulted 15 October 2006.

Méchine-Neuville, A., M.-P. Chenard, B. Gairard, C. Mathelin and J.-P. Bellocq. 2000. "Les coupes larges en pathologie mammaire de routine. Une technique adaptée à la chirurgie conservatrice." *Annales de pathologie* 20: 275–279.

"Médecine: cancer du sein et obésité." 2000. *La Recherche*, June 2000, 332: 12.

Mertz, L., C. Mathelin, C. Marin, B. Gairard, M.-P. Chenard, J.-P. Brettes, J.-P. Bellocq and A. Constantinesco. 1999. "Injection sous-aréolaire de sulfocolloïdes technétiés pour l'identification des ganglions sentinelles dans les cancers du sein invasifs à foyers multiples." *Bulletin du Cancer*, November 1999, 86(11): 939–945.

Meyer, O. 2001a. "Athérosclérose et connectivites." *Rev Rhum* 68: 931–943.

Michel, J.-B. 2004. "Système rénine-angiotensine et remodelage vasculaire." *Médecine/Sciences*, April 2004, 20(4). http: //www.erudit.org/revue/ms/2004/v20/n4/008114ar.html. Consulted 18 May 2004.

Mignot, L. 2002. "Cancer du sein et grossesse: le point de vue du sénologue." *Bulletin du Cancer*, September 2002, 89(9): 772–778.

Moutardier, V. and G. Houvenaeghel. 2001. "Biopsie du ganglion sentinelle dans le cancer du sein: courbe d'apprentissage et contrôle qualité." *Bulletin du Cancer*, December 2001, 88(12): 1246–1247.

Nalbone, G., D. Bernot, F. Peiretti, M.-C. Alessi and I. Juhan-Vague. 2002. "Les statines en thérapeutique cardiovasculaire." *Médecine/Sciences*, December 2002, 18(12). http: //www.erudit.org/revue/ms/2002/v18/n12/000602ar.html. Consulted 18 May 2004.

Namer, M. and A. Ramaioli. 2000. "État actuel du traitement adjuvant des cancers du sein de la femme non ménopausée et premiers résultats des castrations médicales par analogues de la LH-RH." *Bulletin du Cancer*, February 2000, 87(2): 139–144.

Nguyen, T.D., S.Tahiri and A. Cauchois. 2002. "Témoignages de patientes maghrébines traitées pour cancer du sein." *Bulletin du Cancer*, July – August 2002, 89(7): 733–736.

Noël, G., L. Feuvret, M. Gasowski, A. Bernard and P. Cappelaere. 1998. "Traitement hormonal substitutif de la ménopause et cancer du sein." *Bulletin du Cancer*, December 1998, 85(1): 997–1014.

Noël, J.-C., I. Fayt and S. Fernandez-Aguilar. 2004. "Apport de la protéine p63 dans le diagnostic du carcinome tubuleux mammaire." *Annales de Pathologie* 24: 319–323.

Normand, J., J.-C. Lasry, R.G. Margolese, J.C. Perry and D. Fleiszer. 2004. "Communication conjugale et symptômes dépressifs dans des couples dont la

femme est atteinte de cancer du sein." *Bulletin du Cancer*, February 2004, 91(2): 193–199.

Nos, C., D. Bourgeois, C. Darles, B. Asselain, F. Campana, B. Zafrani, J.-C. Durand and K. Clough. 1999. "Traitement conservateur des cancers du sein multifocaux: étude à propos de 56 cas traités à l'Institut Curie de 1983 à 1989." *Bulletin du Cancer*, February 1999, 86(2): 184–188.

Nowak, M. 2002. "Le dépistage du cancer du sein est-il efficace?" *La Recherche*, April 2002, 352: 55.

Papadopoulo, D. and E.Moustacchi. 2002. "FANCD1 et BRCA2, un seul et même gène ?" *Médecine/Sciences*, November 2002, 18(11). http: //www.erudit.org/revue/ms/2002/v18/n11/000456ar.html. Consulted 18 May 2004.

Papon, X., M. Eudo, C. Brillu, F. Villapadierna, J. Picquet and B. Enon. 1999. "Ischémie distale sévère du membre supérieur: Bilan d'un suivi chirurgical de 11 ans chez 34 patients." *Journal des maladies vasculaires* 24(5): 368–372.

Pecking, A.-P., C. Corone Méchélany, J.L. Albérini, F. Bertrand Kermorgant, C. Pallud, J.L. Floiras, A. Goupil and M.F. Pichon. 2002. "Tomographie d'émission de positrons (tep) au 18FDG et maladie occulte en cancérologie: un nouvel argument justifiant l'emploi des marqueurs tumoraux sériques." *Immuno-analyse & Biologie spécialisée* 17: 287–292.

Penault-Llorca, F., A. Etessami and J. Bourhis. 2002. "Principales utilisations thérapeutiques des anticorps monoclonaux en cancérologie." *Cancer/Radiothérapie* 6 Suppl 1: 24s–28s.

Peyrat, J.-P., P. Vennin, L. Hornez and J. Bonneterre. 1997. "Mutations germinales de BRCA1 chez 36 patientes atteintes de cancer du sein et/ou de l'ovaire et appartenant à des familles à risque du Nord de la France." *Bulletin du Cancer*, January 1997, 84(1): 41–46.

Philip, T., J. Clavier, C. Maylin and D. Serin. 2002. "La chirurgie reste l'arme numéro 1 de lutte contre le cancer: place de la chirurgie dans le plan cancer. La réflexion

du cercle des cancérologues français." *Bulletin du Cancer*, February 2002, 89(2) Numéro spécial: 5–6.

Piette, J.-C. 2004. "Antiphospholipides, lupus systémique et athérosclérose: aspects cliniques." *La revue de médecine interne* 25 (2004): S12–S13.

Pilon, D. and L. Lanthier. 2003. "La maladie vasculaire artérielle périphérique: cette grande oubliée." *Le Clinicien*, March 2003, 81–90.

Pinet, F. 2004. "À quoi sert le système endothéline ?" *Médecine/Sciences* 20(3), March 2004. http: //www.erudit.org/revue/ms/2004/v20/n3/007855ar.html. Consulted 18 May 2004.

Pocard, M. and R.-J. Salmon. 1997. "Résection hépatique pour métastase du cancer du sein. Le concept de chirurgie adjuvante." *Bulletin du Cancer*, January 1997, 84(1): 47–50.

Poirier, P. and J.-P. Després. 2003. "Obésité et maladies cardiovasculaires." *Médecine/Sciences*, October 2003, 19(10). http: //www.erudit.org/revue/ms/2003/v19/n10/007164ar.html. Consulted 18 May 2004.

Pourquier, D., C. Lemanski, P. Faurous, H. Couty, R. Delard, P. Rouanet and J.-B. Dubois. 1998. "Cancer du sein: le retour de la lymphoscintigraphie ?" *Bulletin du Cancer*, August 1998, 85(8): 675–684.

"Les premiers anticorps humanisés." 2002. *La Recherche*, April 2002, 352: 14.

Prud'homme, D. and S.J. Weisnagel. 2004. "Traitement de l'obésité: comme médecin faites-vous le poids ?" *Le Médecin du Québec*, February 2004, 39(2): 51–59.

Pujol, P., P. This, M. Noruzinia, D. Stoppa-Lyonnet and T. Maudelonde. 2004. "Les formes héréditaires de cancer du sein liées à BRCA1 et BRCA2 sont-elles sensibles aux œstrogènes ?" *Bulletin du Cancer*, July–August 2004, 91(7): 583–591.

Racadot, S., C. Marchal, C. Charra-Brunaud, J.-L.Verhaeghe, D. Peiffert and P. Bey. 2003. "Ré-irradiation pariétale après mastectomie de rattrapage pour récidive d'un cancer du sein après traitement conservateur: Étude rétrospective sur 20 patientes (Nancy: 1988–2001)." *Cancer/Radiothérapie* 7: 369–379.

Reynaud, P., S. Abbey-Toby, E. Albuisson, A. Wattiez, F. Suzanne and P. Déchelotte. 1999. "Comparaison histologique entre curage axillaire endoscopique après lipo-aspiration et technique chirurgicale classique." *Annales de pathologie* 19: 289.

Rivard-Gervais, N. 2001. "Aliments fonctionnels et produits nutraceutiques – II les acides gras." *Le Médecin du Québec*, April 2001, 36(4): 71–76.

Roemer-Becuwe, C., I. Krakowski and T. Conroy. 2003. "Bisphosphonates, douleurs et qualité de vie dans le cancer du sein métastatique: Une revue de la littérature." *Bulletin du Cancer*, December 2003, 90(1): 1097–1105.

Rossignol, P., A. La Batide Alanore, S. Roueff, G. Bobrie and P.-F. Plouin. 2002. "Prise en charge des sténoses athéroscléreuses des artères rénales." *Journal des maladies vasculaires* 27(1): 7–11.

Rousseau, F. and N. Laflamme. 2003. "Génétique moléculaire humaine: des maladies monogéniques aux maladies complexes." *Médecine/Sciences*, October 2003, 19(10). http: //www.erudit.org/revue/ms/2003/v19/n10/007165ar.html. Consulted 18 May 2004.

Rozenbaum, H. 2004. "Actualités sur la ménopause. Données de la médecine factuelle (evidence based medecine)." *EMC-Endocrinologie*, February 2005, 2(1) : 90–101. http://www.sciencedirect.com/. Consulted 15 October 2006.

Saliou, C. and C. Laurian. 2001. "Artériopathie oblitérante des membres inférieurs et lésions proximales aorto-iliaques: quelles techniques ? Quelles indications ?" *Ann Cardiol Angéiol* 50: 101–111.

Salmon, R.J. 1998. "Evolution de la chirurgie du cancer du sein." *Bulletin du Cancer*, June 1998, 85(6): 539–543.

Salmon, R.J., C. Nos, F. Lojodice, O. Languille, Y. Remvikos, J.R. Vilcoq and K.B. Clough. 2000. "Ganglion sentinelle et cancer opérable du sein: utilisation du bleu patent. Étude pilote." *Annales de Chirurgerie* 125: 253–258.

Sancho-Garnier, H. 2000. "Part des comportements humains, et de l'environnement dans la prévention des cancers." *Sciences de la vie / Life Sciences* 323: 597–601.

Sasco, A.J. 2000. "Actualités dans le dépistage des cancers." *Bulletin du Cancer*, March 2000, 87(3): 239–243.

Sasco, A.J., R. Ah-Song, I. Gendre, P. Zlatoff, J.-Y. Bobin, P. Hallonet and B. Leduc. 1997. "Cancer de l'endomètre et tamoxifène. Discussion à partir d'une série de cas." *Bulletin du Cancer*, January 1997, 84(1): 51–60.

Schoch, C. 2002. "Le dépistage du cancer du sein généralisé: une priorité de santé publique en France." *Bulletin du Cancer*, December 2002, 89(12): 1079–1080.

Séradour, B., H. Allemand and P. Schaffer. 1997. "Programme français de dépistage du cancer du sein. Résultats de cinq départements (1989–1994)." *Bulletin du Cancer*, August 1997, 84(8): 822–828.

Serin, D. and M. Escoute. 1998. "Actualités en sénologie." *Bulletin du Cancer*, January 1998, 85(1): 31–34.

Serin, D., L. Aimard, S. Kirscher, Y. Brewer, C. Félix-Faure, P. Vincent, B. Chauvet and F. Reboul. 1997. "Traitement adjuvant des cancers du sein par radiochimiothérapie concomitante: Étude de faisabilité d'une nouvelle alternative stratégique dans les stades I et II." *Bulletin du Cancer*, March 1997, 84(3): 247–253.

Simard, A.-M. 1997. "La prise de poids favoriserait le cancer du sein." *CyberSciences*, 10 November 1997. http: //www.cybersciences.com/cyber/3.0/n429.asp. Consulted 18 February 2004.

Simard, A.-M. 1998. "Tamoxifène: vraiment préventif contre le cancer du sein?" *CyberSciences*, 13 July 1998. http: //www.cybersciences.com/cyber/3.0/n815.asp. Consulted 18 February 2004.

Simard, A.-M. and S. Dussault. 1997. "Les grands bonds de la médecine du coeur." *Québec Sciences*. http: //www.cybersciences.com/cyber/4.0/juin97/coeu9706.asp. Consulted 11 January 2005.

Spyckerelle, Y., C. Kuntz, J.-P. Giordanella and R. Ancelle-Park. 2002. "Pratiques de la mammographie chez les femmes de 35 à 75 ans: Étude descriptive dans la

population consultant les centres d'examens de santé." *Bulletin du Cancer*, November 2002, 89(11): 957–962.

Stoppa-Lyonnet, D. and M. Jeanpierre. 2004. "BRCA1: de l'identification du gène à l'estimation des risques tumoraux." *Médecine/Sciences*, March 2004, 20(3). http://www.erudit.org/revue/ms/2004/v20/n3/007839ar.html. Consulted 18 May 2004.

Stoppa-Lyonnet, D., C. Blandy and F. Eisinger. 1997. "Cancer du sein: évaluer le risque." *La Recherche*, January 1997, 294: 72–76.

"Tamoxifène et cancer du sein." 1998. *La Recherche*, September 1998, 312: 16.

Teiger, E. 2001. "Physiopathologie de l'angor instable." *Ann Cardiol Angéiol* 50: 359–365.

Thiébaut, A.C.M. and F. Clavel-Chapelon. 2001. "Consommation de graisses et cancer du sein: Résultats préliminaires de la cohorte E3N-Epic." *Bulletin du Cancer*, October 2001, 88(10): 954–958.

Trahan, J. 2002. "L'ail et la santé." *Le Clinicien*, August 2002, 29–32.

Trop, I. 2003. "Experts-conseils." *Le Clinicien*, October 2003, 25.

Trop, I. 2003. "Faut-il cesser l'hormonothérapie avant une mammographie?" *Le Clinicien*, January 2003, 51–56.

Trop, I. 2003. "Le dépistage du cancer du sein: Principes et controverses." *Le Clinicien*, May 2003, 53–58.

Trunet, P. and M. Marty. 1999. "Nouveaux développements dans l'hormonothérapie du cancer du sein chez la femme ménopausée." *Bulletin du Cancer*, October 1999, 86(10): 815–820.

Tubiana, M. 2001a. "La santé et la ville: santé physique et santé mentale." *Sciences de la vie / Life Sciences* 324: 757–767.

Tubiana, M. 2002. "Le vieillissement: aspects médicaux et sociaux." *C. R. Biologies* 325: 699–717.

Tubiana-Hulin, M., P. Beuzeboc, L. Mauriac, N. Barbet, M. Frenay, A. Monnier, J.-M. Pion, O Switsers, J.-L.Misset, S.Assadourian and E. Bessa. 2001. "Essai comparatif randomisé en double aveugle clodronate oral 1 600 mg/j versus

placebo chez des patientes avec métastases osseuses de cancer du sein." *Bulletin du Cancer*, July 2001, 88(7): 701–707.

Vadeboncoeur, A. 2004. "La douleur thoracique atypique: L'évaluation est la clé." *Le Clinicien*, June 2004, 77–83.

Vandhuick, O., B. Guias, L. De Saint Martin and L. Bressollette. 2004. "Traitement antirétroviral et risque cardio-vasculaire." *Journal des maladies vasculaires* 29(4): 192–199.

Vasseur, S. and J.L. Iovanna. 2003. "Le gène p8 est nécessaire au développement tumoral." *Médecine/Sciences* 19(12), December 2003. http: //www.erudit.org/revue/ms/2003/v19/n12/007402ar.html. Consulted 18 May 2004.

Viens, P. and D. Maraninchi. 2001. "Chimiothérapie à haute dose avec greffe de cellules souches hématopoïétiques dans les cancers du sein." *Bulletin du Cancer*, September 2001, 88(9): 835–841.

Vilain, M.-O., A. Delobelle-Deroide, F. Bloget, V. Cabaret, J.-P. Peyrat, C. Fournier, B. Hecquet, J. Fournier, P. Vennin and J. Bonneterre. 1997. "Détection immunohistochimique des récepteurs de l'oestradiol et de la progestérone sur coupes en paraffine après traitement par micro-ondes. Comparaison au dosage biochimique des récepteurs sur une série de 123 carcinomes mammaires avec détermination." *Ann. Pathol* 17(2): 82–88.

Vinatier, D. and G. Orazi. 2003. "Peut-on prévenir chimiquement le cancer du sein en 2003 ?" *Gynécologie Obstétrique & Fertilité* 31: 327–336.

Virag, R. 2002. "Vasodilatation postocclusive des artères caverneuses. Un test potentiel de la réserve en NO du pénis." *Journal des maladies vasculaires* 27(4): 214–217.

Weisnagel, S.J. and D. Prud'homme. 2004. "Un examen proportionnel à l'IMC." *Le Médecin du Québec*, February 2004, 39(2): 43–49.

Wyplosz, B. and L. Capron. 2004. "Aspects infectieux de l'athérosclérose." *Médecine/Sciences*, February 2004, 20(2). http: //www.erudit.org/revue/ms/2004/v20/n2/007676ar.html. Consulted 18 May 2004.

Xhignesse, M. and P. Grand'Maison. 2001. "Taux de cholestérol total élevé un problème de LDL-C ou de triglycérides ?" *Le Médecin du Québec*, April 2001, 36(4): 37–40.

Appendix D: Samples of TermoStat candidate terms

English Breast Cancer Corpus

Rank	Candidate	Frequency	Score	Variants
1	patient	1824	176.239538038895	patients
2	breast cancer	1119	145.88589976288	breast cancer, breast cancers
3	woman	1168	108.788840154425	women, woman
4	study	785	94.3047934904199	studies, study
5	tumour	302	77.1254456545308	tumor, tumors, tumour, tumours
6	breast	352	76.2463109044819	breast, breasts
7	datum	287	75.0355093102698	data
8	chemotherapy	285	72.0415847581739	chemotherapy, chemotherapies
9	table	391	69.1740855677445	table, tables
10	risk	423	65.7652255830385	risk, risks
11	cell	258	64.9668656404107	cells
12	fig	208	62.0138175369551	fig, figs
13	diagnosis	235	60.8739616002897	diagnosis, diagnoses
14	ci	174	58.4668198494263	ci
15	tamoxifen	167	57.2714111122373	tamoxifen
16	situ	162	56.2119577080129	situ
17	cancer	329	55.76751799266	cancer, cancers
18	tamoxifen	152	54.6218075761259	tamoxifen
19	mastectomy	153	54.4132566691411	mastectomy, mastectomies
20	estrogen	151	52.3743904439555	estrogen, estrogens

English Heart Disease Corpus

Rank	Candidate	Frequency	Score	Variants
1	patient	1015	131.744051920077	patients
2	study	646	86.5136195839876	study, studies
3	crp	315	83.1925845582329	crp
4	cvd	312	82.7941448079654	cvd
5	diabetes	354	82.6261574724528	diabetes
6	ldl	297	80.4782807257173	ldl
7	c	323	79.8894487158858	c
8	atherosclerosis	265	75.3572115527315	atherosclerosis
9	risk	458	74.0110446323065	risk, risks
10	hdl	230	70.5445799645403	hdl

Rank	Candidate	Frequency	Score	Variants
11	statins	224	70.1064515824974	statins
12	datum	220	69.3036274593833	data
13	hrt	216	68.8371805261588	hrt
14	mg	241	67.9395788578486	mg
15	non	198	65.89199879345	non
16	cardiovascular disease	202	65.6780795432247	cardiovascular disease, cardiovascular diseases
17	chd	174	61.7469834520221	chd
18	hypertension	176	61.1633444113279	hypertension
19	crp level	158	58.8216128160411	crp level, crp levels
20	effect	393	58.2519181877572	effects

English Corpus[192]

Rank	Candidate	Frequency	Score	Variants
1	be	14268	373.445831785241	is, were, was, are, am
2	patient	2839	159.695960580705	patients
3	use	1487	107.483002920471	used, using, uses
4	breast cancer	1153	105.15363241013	breast cancer, breast cancers
5	study	1431	99.5719406485045	studies, study
6	woman	1664	97.7761166513834	women, woman
7	associate	845	91.7968063779598	associated, associating
8	show	1056	87.6904443467637	shows, showed, showing, shown
9	compare	753	85.7073719527371	compared, comparing, compares
10	report	823	83.9865353929805	reported, reporting, reports
11	increase	742	83.8820043917688	increases, increased, increasing
12	follow	696	79.9286130857794	follows, followed, following
13	risk	881	77.121628580089	risk, risks
14	include	1131	75.4446533530109	included, including, includes
15	reduce	555	73.4316569471329	reduced, reduces, reducing
16	suggest	730	73.2452171415376	suggest, suggests, suggested, suggesting
17	datum	507	70.9905282087879	data
18	find	732	68.6525692351978	found, find, finding, finds
19	base	482	67.8029157000071	based, bases, basing
20	demonstrate	455	66.2899971451623	demonstrating, demonstrated, demonstrates

[192] Due to some technical issues in the part-of-speech tagging and comparisons of the corpora in English, many verbs were identified as specific to the corpus. Because of these issues, the verbs were not considered for the purposes of this research.

French Breast Cancer Corpus

Rank	Candidate	Frequency	Score	Variants
1	mélatonine	50	332.502554768934	mélatonine
2	atm	25	230.487978961058	atm
3	microcalcifications	19	203.596449471151	microcalcifications
4	tau_taux	15	179.595814291336	taux
5	protéine	65	174.041818005985	protéine, protéines
6	fig	14	173.078490893893	fig
7	protéine atm	13	166.307226221013	protéine atm, protéines atm
8	p53	13	160.25208681221	p53
9	tumeur	32	149.475154846428	tumeur, tumeurs
10	carcinome	11	139.686049515207	carcinomes
11	athérosclérose	9	128.933550788178	athérosclérose
12	patient atteindre	8	127.199085240091	patients atteints, patientes atteintes
13	strie lipidique	8	127.199085240091	strie lipidique
14	œstradiol	8	119.917503953711	oestradiol
15	cancer du sein	17	119.240457779977	cancer du sein, cancers du sein
16	foi_fois	7	117.85014278672	fois
17	mmic	7	117.85014278672	mmic
18	niveau de gris	7	117.85014278672	niveaux de gris, niveau de gris
19	rehaussement de contraste	7	117.85014278672	rehaussement de contraste
20	segmentation	14	114.414391550123	segmentation

French Heart Disease Corpus

Rank	Candidate	Frequency	Score	Variants
1	tau_taux	435	227.594195844988	taux
2	patient	784	225.131772701218	patient, patiente, patients, patientes
3	athérosclérose	358	206.126268280915	athérosclérose
4	cholestérol	343	197.685752561504	cholestérol
5	artère	358	178.33203690295	artères
6	facteur	651	171.428413854167	facteurs
7	plaque	451	166.262972794589	plaque, plaques
8	cour_cours	170	142.021128467512	cours

Rank	Candidate	Frequency	Score	Variants
9	fibrinogène	167	140.754894996576	fibrinogène, fibrinogènes
10	sténose	167	139.905585269512	sténoses
11	lésion	292	139.437605143521	lésions
12	diabète	227	137.854309427887	diabète, diabètes
13	foi_fois	159	137.321205502368	fois
14	cellule	594	136.155774203418	cellules
15	cellule endothélial	148	132.454629133259	cellules endothéliales, cellule endothéliale
16	hypertension	154	122.410244369794	hypertension
17	moi_mois	125	121.65162242132	mois
18	traitement	579	120.227946893754	traitement, traitements
19	endothélium	122	120.170963853626	endothélium
20	activation	133	118.872356419911	activation

French Corpus

Rank	Candidate	Frequency	Score	Variants
1	cancer	3697	333.169271258705	cancer, cancers
2	patient	2944	297.147014982997	patients, patientes
3	tau_taux	1337	222.229722774642	taux
4	tumeur	1382	220.13391968066	tumeur, tumeurs
5	traitement	2245	203.853184346005	traitement, traitements
6	sein	2599	174.756942567812	sein, seins
7	cancer du sein	785	167.229581333366	cancer du sein, cancers du sein, cancers des seins
8	cellule	1545	163.602171127413	cellules
9	chimiothérapie	711	158.738246201005	chimiothérapie, chimiothérapies
10	étude	2570	157.797370005786	étude, études
11	ganglion	680	157.092506316242	ganglions
12	facteur	1190	150.251827499501	facteurs
13	gène	892	147.118827494809	gène, gènes
14	moi_mois	499	135.675340708472	mois
15	cour_cours	471	131.805742805043	cours
16	récepteur	487	129.083280712833	récepteurs
17	tamoxifène	450	128.676847587802	tamoxifène
18	mg	517	128.109055104211	mg
19	lésion	606	126.616587389685	lésions
20	risque	1849	125.345191252419	risque, risques

Appendix E: Candidate terms for concordances

English[193]

Term	*UMLS Semantic Type*	*Frequency (in full corpus)*[194]	*Specificity F//C//H*[195]	*Sample size*[196]
chemotherapy/ chemotherapies	**Activity**	540	51//72//n/a	100
HRT/ HRTs/ hormone replacement therapy/ hormone replacement therapies	**Activity**	516	53//37//69	101
patient/ patients	**Entity**/ conceptual entity	3992	160//176//132	100
cell/ cells	**Entity**/ physical object	2143	58//65//44	106
CRP/ CRPs/ C-reactive protein/ C-reactive proteins	**Entity**/ physical object	562	56//n/a//83	101

[193] Bold in the Term field indicates the form of the term suggested as a candidate by TermoStat. In most cases, this is the lemmatized form of the term. However, all forms of the term indicated in the field were included in generating the concordances. In the case of abbreviations, both the forms of the abbreviation and the full forms of the term were included. In the case of processes, only the singular form was used (although in rare cases, plurals did occur in the corpus).

[194] This value is the frequency in the corpus as calculated by TermoStat, which due to technical differences may vary slightly from frequencies calculated using other software, e.g., WordSmith Tools.

[195] Specificities are indicated in the full corpus (F), breast cancer corpus (C) and heart disease corpus (H). *n/a* indicates that no specificity value was available for the term in the corpus in question.

[196] The sampling was done using WordSmith's "At random" feature, which allows user to define the chances of each hit for a given string appearing in the results. Description in the Help file for settings of 100 Entries Wanted and 1 in 3 at random: "Entries Wanted: The maximum is 16,368 lines. This feature is useful if you're doing a number of searches and want, say, 100 examples of each. In that case, the 100 entries will be the first 100 found. [A]t random is a feature which allows you to randomise the search. Here Concord goes through the text files and gets the 100 entries by giving each hit a random one-in-three chance of being selected. To get 100 entries Concord will have found around 250-350 hits. You can set the randomiser anywhere from 1 in 2 to 1 in 1,000."

breast cancer/ breast cancers	**Phenomenon or process**/ Natural phenomenon or process	2533	105//146//n/a	99
tumour/ tumours/ tumor/ tumors	**Phenomenon or process**/ Natural phenomenon or process	1325	55//77//8	100
diabetes	**Phenomenon or process**/ Natural phenomenon or process	425	56//n/a//83	92
atherosclerosis	**Phenomenon or process**/ Natural phenomenon or process	410	51//3//75	85
CHD/ CHDs/ coronary heart disease/ coronary heart diseases	**Phenomenon or process**/ Natural phenomenon or process	373	42//n/a//62	77
expression	**Phenomenon or process**/ Natural phenomenon or process	592	30//21//30	100
development	**Phenomenon or process**/ Natural phenomenon or process	367	25//8//30	99
activation	**Phenomenon or process**/ Natural phenomenon or process	266	35//21//47	107
oxidation	**Phenomenon or process**/ Natural phenomenon or process	84	18//n/a//26	84
pathogenesis	**Phenomenon or process**/ Natural phenomenon or process	61	23//9//33	61

French

Term	*UMLS Semantic Type*	*Frequency (in full corpus)*	*Specificity F//H*[197]	*Sample size*
traitement/ traitements	**Activity**	2357	203//120	100
chimiothérapie/ chimiothérapies	**Activity**	738	159//n/a	100
patient/ patiente/ patients/ patientes	**Entity/** Conceptual entity	3504	297//225	100
cellule/ cellules	**Entity/** physical object	1678	163//136	100
cholestérol/ cholestérols	**Entity/** physical object	356	112//198	100
cancer du sein/ cancers du sein/ cancer des seins/ cancers des seins	**Phenomenon or process/** Natural phenomenon or process	2092	167//37	96
tumeur/ tumeurs	**Phenomenon or process/** Natural phenomenon or process	1481	220//19	99
athérosclérose	**Phenomenon or process/** Natural phenomenon or process	392	119//206	100
récidive	**Phenomenon or process/** Natural phenomenon or process	272	99//19	100
diabète	**Phenomenon or process/** Natural phenomenon or process	233	78//138	100

[197] A technical problem with the coding of the documents made it impossible to rely on the specificity of the terms as indicated in the TermoStat results for the French breast cancer sub-corpus alone. The specificity obtained from the full corpus, in comparison with the heart disease corpus, was used to support the term choice.

activation	**Phenomenon or process**/ Natural phenomenon or process	237	90//118	100
prolifération	**Phenomenon or process**/ Natural phenomenon or process	138	41//21	101
transcription	**Phenomenon or process**/ Natural phenomenon or process	101	36//16	101
oxydation	**Phenomenon or process**/ Natural phenomenon or process	54	41//68	54
coagulation	**Phenomenon or process**/ Natural phenomenon or process	41	37//65	41

Appendix F: Candidate terms and their definitions

English

Term	Definition
activation[198]	1. The act or process of rendering active. 2. The transformation of a proenzyme into an active enzyme by the action of a kinase or another proenzyme. … 4. The process by which the central nervous system is stimulated into activity through the mediation of the reticular activating system. (Dorland's 28th) (Biochemistry) DEF – The act or process of rendering active, as in the transformation of pre-enzyme into an active enzyme by the action of a kinase or another pre-enzyme,... (Physical chemistry) DEF – The process of treating a substance or a molecule or atom by heat or radiation or the presence of another substance so that the first mentioned substance, atom or molecule will undergo chemical or physical change more rapidly or completely. (TERMIUM)
atherosclerosis	An extremely common form of arteriosclerosis in which deposits of yellowish plaques (atheromas) containing cholesterol, lipoid material, and lipophages are formed within the intima and inner media of large and medium-sized arteries. (Dorland's 28th) (Vessels, Medicine) DEF – Thickening and hardening of the walls of the arteries, associated with atheroma. (TERMIUM)
breast cancer	(malignant neoplasm of breast) A breast neoplasm with metastatic potential arising from the breast parenchyma or the nipple. The most common are breast carcinomas. Malignant breast neoplasms occur more frequently in females than in males. -- 2003 (NCI Thesaurus)
cell	1. any one of the minute protoplasmic masses that make up organized tissue, consisting of a nucleus which is surrounded by cytoplasm which contains the various organelles and is enclosed in the cell or plasma membrane. A cell is the fundamental, structural, and functional unit of living organisms. (Dorland's 28th)
chemotherapy	(Chemotherapy, Pharmacodynamics) Chemotherapy DEF – The treatment of a disease by means of chemical substances or drugs. OBS – The term chemotherapy has been applied over the centuries to a variety of therapies, including the treatment of malaria with herbs and the use of mercury for syphilis. In modern usage, chemotherapy usually refers to the use of chemicals to destroy cancer cells on a selective basis. (TERMIUM)

[198] The presence of multiple specialized senses for candidate forms was not considered to reduce their value for use in the context of this research; rather, these were seen as good examples of the type of candidate terms that a terminologist might find difficult to describe, and for which techniques for facilitating conceptual analysis would be particularly useful.

coronary heart disease	An imbalance between myocardial functional requirements and the capacity of the coronary vessels to supply sufficient blood flow. It is a form of MYOCARDIAL ISCHEMIA (insufficient blood supply to the heart muscle) caused by a decreased capacity of the coronary vessels. (MeSH) (Vessels, Medicine) CONT – Coronary heart disease (CHD), also called coronary artery disease ... develops when fatty material (plaque) builds up in the heart arteries. Coronary arteries supply blood and oxygen to the heart muscle (myocardium). The plaque may slow the flow of blood. This slowing causes chest pain, or angina. (TERMIUM)
C-reactive protein	A plasma protein that circulates in increased amounts during inflammation and after tissue damage. (MeSH) A globulin that forms a precipitate with the somatic C-polysaccharide of the pneumococcus *in vitro*; the most predominant of the acute phase proteins. Abbreviated CRP. (Dorland's 28th)
development	The process of growth and differentiation (Dorland's 28th) /.../ the act, process, /.../ of developing. (GDT)
diabetes	A general term referring to disorders characterized by excessive urine excretion (polyuria), as in diabetes mellitus and diabetes insipidus. When used alone, the term refers to diabetes mellitus. (= a chronic syndrome of impaired carbohydrate, protein and fat metabolism owing to insufficient secretion of insulin or to target tissue insulin resistance. ...) (Dorland's 28th) (The Pancreas) (d~ mellitus) DEF – A chronic disorder characterized by impaired metabolism of glucose and other energy-yielding fuels, as well as the late development of vascular and neuropathic complications. Diabetes mellitus consists of a group of disorders involving distinct pathogenic mechanisms with hyperglycemia as the common denominator. Regardless of cause, the disease is associated with insulin deficiency, which may be total, partial, or relative when viewed in the context of coexisting insulin resistance. (TERMIUM)
expression	(gene expression) The phenotypic manifestation of a gene or genes by the processes of GENETIC TRANSCRIPTION and GENETIC TRANSLATION. (MeSH) (protein biosynthesis) The biosynthesis of PEPTIDES and PROTEINS on RIBOSOMES, directed by MESSENGER RNA, via TRANSFER RNA that is charged with standard proteinogenic AMINO ACIDS. (MeSH)
hormone replacement therapy	Therapeutic use of hormones to alleviate the effects of hormone deficiency. (MeSH)
oxidation	The act of oxidizing or state of being oxidized. Chemically it consists in the increase of positive charges of an atom or the loss of negative charges. Most biological oxidations are accomplished by the removal of a pair of hydrogen atoms (dehydrogenation) from a molecule. Such oxidations must be accompanied by reduction of an acceptor molecule. (Dorland's 28th) (oxidation-reduction) A chemical reaction in which an electron is transferred from one molecule to another. The electron-donating molecule is the reducing agent or reductant; the electron-accepting molecule is the oxidizing agent or oxidant. Reducing and oxidizing agents function as conjugate reductant-

	oxidant pairs or redox pairs (Lehninger, Principles of Biochemistry, 1982, p471). (MeSH) (Industrial chemistry processes and operations) DEF – Chemical reaction of a compound with oxygen or a reaction that causes an atom or a group of atoms to lose one or more electrons. OBS – The term "oxidation" originally meant a reaction in which oxygen combines chemically with another substance, but its usage has long been broadened to include any reaction in which electrons are transferred. Oxidation and reduction always occur simultaneously (redox reactions), and the substance which gains electrons is termed the oxidizing agent. OBS – The opposite of reduction. (TERMIUM)
pathogenesis	The development of morbid conditions or of disease; more specifically the cellular events and reactions and other pathologic mechanisms occurring in the development of disease. (Dorland's 28th)
patient	Individuals participating in the health care system for the purpose of receiving therapeutic, diagnostic, or preventive procedures. (MeSH)
tumour[199]	(neoplasm) New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms. (MeSH) An abnormal tissue growth resulted from uncontrolled cell proliferation. Benign neoplastic cells resemble normal cells without exhibiting significant cytologic atypia, while malignant ones exhibit overt signs such as dysplastic features, atypical mitotic figures, necrosis, nuclear pleomorphism, and anaplasia. Representative examples of benign neoplasms include papillomas, cystadenomas, and lipomas; malignant neoplasms include carcinomas, sarcomas, lymphomas, and leukemias. -- 2004 (NCI Thesaurus)

[199] While the choice was made to follow the classification of the UMLS in this case for the sake of consistency, it seems important to note that linguistically, the co-occurrents of this term seem to indicate that a tumour is often considered as a concrete entity rather than as a phenomenon or process (e.g., *large tumour*, *tumour is located*, *tumour can be observed*). This alternate classification of the term is also reflected in the definition given in the UMLS Metathesaurus. Thus it may be observed that the term is at best viewed from different perspectives, and at worst is mis-classified from a semantic point of view. However, given the goals of associating terms with classes in the research, and the presence of this phenomenon in both corpora (as it may also be observed in the case of *tumeur*) was not considered to preclude the use of this term in the study.

French

Term	Definition
activation	(Physical chemistry) DEF – Passage d'une molécule, d'un atome ou d'un ion, de sa forme normale à une forme activée. (Biological sciences) CONT – Principaux modes de régulation de l'activité enzymatique : - La rétroinhibition. (...) - L'activation d'un enzyme par un précurseur du substrat ou par le substrat lui-même. - L'activation par un produit de dégradation du métabolite terminal, permettant d'élever à nouveau la concentration de ce métabolite (qui peut être une substance à haut potentiel énergétique par exemple). - L'activation d'un enzyme d'une suite métabolique conduisant à un métabolite A par un métabolite B, qui est synthétisé par une suite indépendante, lorsque A et B sont tous deux nécessaires à la synthèse des mêmes macromolécules, ce qui permet une production coordonnée des précurseurs. (Biochemistry) CONT – Les activateurs des enzymes sont les substances qui exaltent de façon plus ou moins spécifique l'activité du biocatalyseur. (...) L'activation enzymatique par des ions est bien différente de l'activation des proenzymes ou zymogènes, (...), et qui implique une modification de la protéine enzymatique, souvent accompagnées d'une variation du poids moléculaire (...) (TERMIUM) ... 2. Augmentation de l'énergie d'une molécule ou d'un atome (énergie d'activation). 3. Accroissement de la perméabilité membranaire lié à la dépolarisation. 4. Dans un sens plus large, dépolarisation d'une fibre myocardique lors de la propagation de l'excitation auriculaire ou ventriculaire. (*Flammarion*, Kernbaum 2001)
athérosclérose	(Vessels, Medicine) DEF – Sclérose artérielle caractérisée par l'accumulation de lipides amorphes dans la tunique interne du vaisseau (athérome). OBS – L'athérosclérose siège surtout sur les vaisseaux coronariens, l'aorte et ses principales collatérales, plus rarement sur les vaisseaux pulmonaires. (TERMIUM) V. athérome. = Lésion très fréquente entraînant dans le cadre de l'artériosclérose, frappant essentiellement les artères de type élastique (aorte et gros vaisseaux) et caractérisée, initialement, par une altération dégénérative de l'intima avec dépôts lipidiques, réaction histiocytaire de type lipophagique et sclérose périfocale. Secondairement, la nécrose lipophagique libre des cristaux lipoïdiques et de cholestérine avec réaction à corps étranger, sclérose et éventuellement calcification. (*Flammarion*, Kernbaum 2001)
cancer du sein	C'est avec le cancer de l'utérus, la plus fréquente des néoplasies de la femme. Parmi les facteurs prédisposants, on peut retenir, outre l'âge, une ménopause tardive, la nulliparité, un poids, et une taille élevés, un facteur génétique. Le cancer du sein peut évoluer sous différents aspects, dont il faut individualiser la mastite aiguë carcinomateuse de la jeune accouchée, en raison de sa haute fréquence. /.../ Il se révèle le plus souvent par une tuméfaction indolore, découverte par hasard, ou par un écoulement du mamelon. (GDT)
cellule	1. (Histol.) Masse de protoplasme limitée par une membrane et renfermant un noyau, correspondant à la plus petite quantité de matière vivante structurée,

	douée de vie autonome et susceptible de se reproduire. (*Flammarion*, Kernbaum 2001)
chimiothérapie	Terme générique désignant tout traitement par des agents chimiques. Le mot s'applique plus particulièrement à certains traitements antinéoplasiques et anti-infectieux. (*Flammarion*, Kernbaum 2001) (1) à l'origine, administration d'un produit chimique spécifique qui peut stériliser l'organisme en le libérant de la présence d'agents infectieux pour lesquels il a une affinité particulière, sans qu'il en résulte des phénomènes toxiques notables pour le malade lui-même ; (2) actuellement, administration d'un produit chimique spécifique afin de guérir une maladie cliniquement reconnaissable ou d'enrayer sa progression (EURODICAUTOM, from Manuila, vol.4, p.558)
cholestérol	Stérol synthétisé par de nombreux tissus de l'organisme, et surtout le tissu hépatique, à partir d'acétyl-coenzyme A. Le cholestérol et ses esters entrent dans la constitution des lipoprotéines sériques et sont présents dans de nombreux produits de sécrétion. Le cholestérol est également le précurseur des hormones stéroïdes et des acides biliaires. (*Flammarion*, Kernbaum 2001)
coagulation	Ensemble des processus biochimiques permettant l'élaboration du caillot de fibrine. La coagulation, phénomène plasmatique, complète l'hémostase primaire pour assurer l'arrêt des hémorragies. (*Flammarion*, Kernbaum 2001)
diabète	Terme générique englobant un certain nombre d'affections dont le dénominateur commun est l'association d'une polyurie et d'une polydipsie. Le terme diabète, sans épithète, désigne, le plus souvent, le diabète sucré. (= 1. Stricto sensu : passage anormal de sucre dans les urines lié à une élévation anormale du taux de glucose dans le sang. 2. Cette définition très restrictive du diabète n'est plus acceptable. On considère actuellement que le diabète sucré est une affection chronique, caractérisée par une insuffisance absolue ou relative de la sécrétion en insuline, dont l'une des conséquences est l'hyperglycémie (permanent dans le nycthémère ou seulement post-prandiale) qui peut s'accompagner ou non de glycosurie.) (*Flammarion*, Kernbaum 2001)
oxydation	V. oxyréduction. = Ensemble comprenant des réactions couplées d'oxydation (ou perte d'électrons) et de réduction (ou gain d'électrons) et dont l'équilibre peut varier selon les circonstances. Ces réactions comportent toujours un agent oxydant (accepteur d'électrons) et un agent réducteur (donneur d'électrons). (*Flammarion*, Kernbaum 2001) (Industrial chemistry processes and operations) DEF – Réaction au cours de laquelle un composé chimique (dit «réducteur») perd des électrons au profit d'un autre (appelé «oxydant»). CONT – Il ne peut y avoir oxydation d'un composé sans qu'il y ait réduction simultanée d'un autre composé. OBS – Contraire de réduction.
patient	Personne qui a recours aux services médicaux ou paramédicaux, qu'elle soit malade ou non. Note(s) : Le terme patient n'est plus réservé aux malades qui souffrent, comme le voudrait son étymologie latine. Il peut désigner une personne soumise à un examen médical, suivant un traitement ou subissant une intervention chirurgicale de même que les femmes enceintes. (GDT)

prolifération	Multiplication rapide de cellules ou de micro-organismes. Note(s) : La prolifération cellulaire s'accompagne parfois de l'apparition d'une certaine anarchie de structure, pouvant aller jusqu'à la perte de la forme et de propriétés caractéristiques. Des proliférations cellulaires s'observent notamment au cours de processus de nature inflammatoire ou tumorale. Ce phénomène, normal au cours du développement et de la croissance pour la plupart des tissus et d'une façon permanente pour certaines lignées cellulaires (éléments figurés du sang, lignée spermatique, etc.), devient anormal dans certaines conditions; il conduit à la formation de tissus néoformés ou néoplasiques. La prolifération d'un agent pathogène est sa multiplication au sein d'un organisme réceptif ou d'une culture. (GDT)
récidive	Réapparition d'une maladie antérieurement guérie. (*Flammarion*, Kernbaum 2001) Réapparition d'une maladie, habituellement infectieuse, après une période de santé complète. Note(s) : Il ne faut pas confondre les termes « rechute », « récidive », « récurrence » et « recrudescence ». .... La récurrence se définit comme la reprise d'une maladie infectieuse apparemment guérie. De plus, elle apparaît plus tardivement que la rechute. Le terme « récurrence » est rendu en anglais par « récurrence ». (GDT)
traitement	Ensemble de prescriptions médicamenteuses et hygiénodiététiques employées pour guérir une maladie ou combattre ses effets. (*Flammarion*, Kernbaum 2001)
transcription	(Genetics) DEF – Processus par lequel la séquence d'un gène est copiée en ARN. (TERMIUM) Passage de l'information génétique de l'ADN à l'ARN, sous forme de ribonucléotides complémentaires, survenant durant la synthèse de l'ARN simple brin à partir d'une matrice d'ADN par l'action d'une ARN polymérase. (GDT) En génétique, opération de copie d'un gène en un messager, ou chez des eucaryotes, en un précurseur du messager. (*Flammarion*, Kernbaum 2001)
tumeur	1. Anciennement, toute lésion provoquant une augmentation de volume localisée. Cette définition correspond maintenant à celle du terme tuméfaction. 2. Actuellement, synonyme de néoplasme : « toute néo-formation tissulaire (plus ou moins volumineuse) qui ressemble (plus ou moins) au tissu normal homologue (adulte ou embryonnaire) aux dépons duquel elle s'est développée, qui a tendance à persister et à s'accroître et qui échappe aux règles biologiques de la croissance et de la différenciation cellulaire ». (*Flammarion*, Kernbaum 2001)

Appendix G: Statistical tests

Chi-square (χ^2) test

As used in this thesis, this statistical test of association compares the number of cases in which a given phenomenon was present in the results (i.e., the number of contexts in which a given characteristic was noted, *C+*) and the number of cases in which it was absent (i.e., the rest of the annotated contexts, *C-)*, in each of two groups (i.e., languages, *EN* and *FR*). These data can be represented in a 2 x 2 table such as Table 129, (which is similar to Table 6, presented in the description of the ASSOCIATION relation).

Table 129. Illustration of a 2 x 2 table as used for the Chi-square test

	EN	FR
C+	V1	V3
C-	V2	V4

The test involves the calculation of how much the values (*V*) — i.e., numbers of occurrences — observed in each group deviate from an expected value (*EV*), which is predicted for each cell in the table using proportions based on the combined data for the two groups, in formulae such as those below (adapted from Norman and Streiner 2003: 87):

$$EV1 = \frac{V1+V3}{V1+V2+V3+V4} * (V1+V2)$$

$$EV2 = \frac{V2+V4}{V1+V2+V3+V4} * (V1+V2)$$

$$EV3 = \frac{V1+V3}{V1+V2+V3+V4} * (V3+V4)$$

$$EV4 = \frac{V2+V4}{V1+V2+V3+V4} * (V3+V4)$$

For calculation of the Chi-square (χ^2) statistic, these differences between the observed and expected values for each cell are then squared, divided by the expected value for that cell, and summed, as shown below:

$$\chi^2 = \frac{(V1 - EV1)^2}{EV1} + \frac{(V2 - EV2)^2}{EV2} + \frac{(V3 - EV3)^2}{EV3} + \frac{(V4 - EV4)^2}{EV4}$$

The result of this calculation provides a measure of how much the observations differ from the expected values. This result can then be interpreted using standard tables (such as the one provided in Fleiss 1981: 258) in order to identify the *p* value with which it is associated (i.e., a measure of the probability that a difference at least as large could have occurred strictly by chance).[200] The lower the *p* value, the lower the probability that the results could have been observed by chance, and thus the more confident one can be that the difference observed is real. The Chi-square values reported in this thesis were computed using Microsoft Excel (v. 2003).

In statistics, a *p* value of less than or equal to 0.05 is generally considered to be statistically significant (Norman and Streiner 2003: 32). This test may be used when expected values of *V* (i.e., *EV*) are greater than or equal to 5 (Norman and Streiner 2003: 88); below this threshold, the test is considered inaccurate.

Another form of this test that adds additional rows to the table allows for the comparison of more than two categories of occurrences (e.g., the measurement of variation of pattern marker distribution among several part of speech classes, rather than of the presence or absence of a single characteristic).[201]

---

[200] Use of such standard tables requires specification of the number of degrees of freedom (*df*) in the data. The number of degrees of freedom = (no. rows - 1) * (no. columns - 1). Therefore, in the case of a 2 x 2 table the appropriate *p* value is that corresponding to one degree of freedom.

[201] This involves an increase in the degrees of freedom, which is taken into account in the identification of the appropriate *p* value in the standard table.

Example of a Chi-square test

An example based on the data on the number of occurrences of verbal marker occurrences observed in the passive voice (as described in Section 4.8.1.1, Table 63) is presented below. In the English corpus, 175 occurrences of verbal markers were observed, 24 of which were in the passive voice; in the French data, 140 verbal marker occurrences, 5 in the passive voice, were identified. This data is represented in Table 130, where for reasons of clarity, cells containing the totals for each row and column have been added to the basic 2 x 2 table.

Table 130. Comparison of the proportions of verbal marker occurrences in passive and active voice

	EN	FR	Total
passive	24	5	29
active	151	135	286
Total	175	140	315

The proportion of occurrences of markers in the passive voice is thus calculated based on the ratio of the total number of occurrences of markers in the passive voice divided by the total number of verbal marker occurrences, or 28/315 = 0.092; the proportion of verbal marker occurrences in the active voice is then 286/315 = 0.908.

Thus, the expected values for English passive occurrences would be 0.092 * 175 = 16.10, French passive occurrences 0.092 * 140 = 12.88, English active occurrences 0.908 * 175 = 158.90, and French active occurrences 0.908 * 140 = 127.12.

This provides the information required for the Chi-square test, as shown below:

$$\chi^2 = \frac{(24-16.10)^2}{16.10} + \frac{(5-12.88)^2}{12.88} + \frac{(151-158.90)^2}{158.90} + \frac{(135-127.12)^2}{127.12}$$

$$\chi^2 = \frac{(7.90)^2}{16.10} + \frac{(-7.88)^2}{12.88} + \frac{(-7.90)^2}{158.90} + \frac{(7.88)^2}{127.12}$$

$$\chi^2 = \frac{62.41}{16.10} + \frac{62.09}{12.88} + \frac{62.41}{158.90} + \frac{62.09}{127.12}$$

$$\chi^2 = 3.88 + 4.82 + 0.39 + 0.49$$

$$\chi^2 = 9.58$$

In Fleiss (1981: 258) and Arkin and Colton (1963: 126), the p value for a Chi-square result of 9.58 with one degree of freedom is found to be less than 0.01 ($p = 0.01$ at a Chi-square value of 6.63 and $p = 0.001$ at a value of 10.83).[202] Since the p value is clearly less than 0.05 (as $p = 0.05$ corresponds to a Chi-square value of 3.84), the difference observed between the two data sets is considered to be statistically significant, and thus supports the rejection of the null hypothesis that there is no difference in the proportions of passive and active verbal marker occurrences between the languages.

[202] The value returned by Excel for this calculation is p = 0.00197495.

Appendix H: Complete list of pattern markers observed in the sample

English

Association (33 markers, 18 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms[203]	Occur-rences in sample	Total occur-rences in corpus[204]	Occur-rences per 1,000 corpus tokens[205]	Sample context
associated	ppl.a.	X [conjunction] associated Y	17	848	1.47	Abnormal endothelium-dependent vasomotor responses predict the long-term progression of atherosclerosis and **associated** coronary events… (Davignon 2004)

[203] Pattern forms reflect the basic standard structure of the pattern as observed in the occurrences analyzed, including the relative positions of pattern components (markers and related elements), as well as regular insertions within pattern forms (excluding the presence of structures involving relative pronouns, which are considered separately). In cases in which regular interruptions by ASSOCIATION or CAUSE–EFFECT markers were observed, the nature of these elements is noted in the pattern form. Irregular interruptions were eliminated from these structures to more clearly reflect the pattern structures. Optional but relatively regular elements appearing within pattern structures are indicated by parentheses, and alternatives by forward slashes (/). Copula verbs appearing within pattern structures are represented by the indicator *[copula]*, except in cases in which a specific form that was considered to be potentially important to the analysis of the structures or the nature of the marker itself was observed.

[204] Due to technical restrictions in WordSmith Tools, a maximum of 16,001 occurrences of a given string can be extracted from each corpus. In the few cases in which this ceiling was reached in the calculation of the total number of occurrences of a given marker (generally in the case of prepositions), this value is indicated.

[205] These figures were calculated by dividing the number of occurrences observed by the number of tokens in the appropriate corpus and multiplying the result by 1,000.

	ppl.a. + prep.	X [copula] associated with ([causal marker]) Y X associated with Y				Diabetes was **associated with** accelerated atherosclerosis at both 14 and 20 weeks of age… (Yan et al. 2003) …cardiovascular dysfunction **associated with** diseases such as hyperlipidemia, diabetes mellitus, hypertension, ischemic heart disease, and chronic heart failure. (Taniyama and Griendling 2003)
risk	n.	[causal marker] X [preposition] [causal marker]Y risk	14	2961	5.13	Most recent studies indicate a particularly harmful effect of combined estrogen/progestin regimens in terms of increased breast cancer **risk**. (Kocjan and Prelevic 2003)
	n. + prep.	X ([causal marker]) risk for Y				There is good evidence that HRT increases the **risk for** VTE… (Kocjan and Prelevic 2003) In univariate analysis, genotype CHD **risk for** APOE was P =0.01. (Humphries et al. 2004)
	n. + prep.	([causal marker] [preposition] [article]) risk of X [association marker] Y X ([verb])[causal marker] [article] risk of Y				…analysis… demonstrated a modest increase in **risk of** breast cancer with increasing duration of use. (Weiss et al. 2002) Pike argues that oral contraceptives may slightly increase the **risk of** breast cancer, a contention disputed by a number of other researchers. (Fackelmann 1992) Users of combined HT, however, had an overall 1.7-fold (95% CI 1.3-2.2) increased **risk of** breast cancer… (La Vecchia 2003)
	n. + prep.	[causal marker] X risk in relation to Y				Indirect evidence of a biologically important role of AFP comes from epidemiologic findings of a

						reduced maternal breast cancer **risk in relation to** pregnancy conditions, such as hypertension, 15 preeclampsia 16,17 and multiple births… (Lambe et al. 2003)
risk factor	n. + n. n. + n. + prep.	X risk factor, [conjunction] Y X [copula] ([article]) risk factor for Y X as ([article/quantifier]) risk factor for Y	10	421	0.72	Development of CVD and its **risk factors,** including HTN and atherosclerosis… (Schwartz 2003) Hyperhomocysteinaemia is a **risk factor for** the development of CHD. (Mackness et al. 2004) … unable to establish low-dose diagnostic X-ray exposure or therapeutic treatment **as risk factors for** female breast cancer (Zheng et al. 2002)
marker	n. n. + prep.	X [copula] [article/quantifier] Y marker X marker Y X Y marker X [copula] [article] marker for Y X [copula] [article] marker of Y	9	339	0.59	CRP, sCD40L, and IL-18 are three inflammatory **markers** that result in endothelial activation. (Torres and Ridker 2003) The inflammatory **marker** C-reactive protein (CRP) can indicate low-grade chronic inflammation… (MacKenzie 2004) However, a recent analysis from the Women's Health Initiative Observational Study suggests that although hs-CRP and other inflammatory **markers** such as IL-6 may independently predict adverse cardiovascular events WT-1 is not a general **marker for** ovarian surface epithelial-stromal tumors… (Lerwill 2004) As carotid IMT is a good early **marker of** atherosclerosis and risk of cerebrovascular ischemic events… (Zambon et al. 2003)

		X [noun] as [article] marker of Y marker of X: Y				Soluble P-selectin originates from both endothelial cells and platelets, 13 thereby limiting its utility as a **marker of** endothelial cell activation. (Granger et al. 2004) … the human subjects had a modest but significant reduction in key **markers of** blood vessel inflammation: C-reactive protein, tumor necrosis factor, and the interleukins IL-1 and IL-6… (Cabe 2000)
relationship	n. n. + prep. + conj.	X-Y relationship relationship between X and Y	9	300	0.52	The influenza-atherosclerosis **relationship** is analogous to some other relationships between CHD risk factors and atherosclerosis. (Madjid et al. 2004) … additional randomized clinical trials are necessary to further elucidate the **relationship between** CRP **and** CHD. (Rackley 2004)
in	prep.	X in Y X [causal marker/modifier] in Y	8	15453	26.76	Although Ras is not often mutated **in** breast cancer, physiological activation of Ras is frequently associated with malignant progression. (Wang et al. 2003) Moreover, these processes are exaggerated **in** diabetes… (Yan et al. 2003) Because of the traditional view that ERa expression is low **in** breast cancer in women from developing countries… (Tran and Lawson 2004)
association	n. + prep. + conj. n. + prep. + prep.	association between X and Y association of X with Y	7	456	0.40	Overall, results of our investigation indicate that the **association between** risk of breast cancer **and** HRT varies by regimen. (Weiss et al. 2002) **Associations of** plasma fibrinogen **with** coronary heart disease (CHD) have been investigated in

						many long-term observational studies… (Fibronogen Studies Collaboration 2004)
and	conj.	X and Y	5	16001	27.70	CRP **and** Acute Myocardial Infarction The first association between CRP and cardiovascular disease was in the context of… (Shah and Newby 2003)
link	v. + prep.	X has been linked to Y [study, finding] links X to Y [pathway] links X with Y	5	130	0.23	Oxidative stress **has been linked to** the activation of both NF-[kappa]B and AP-1. (Granger et al. 2004) LDL-C remains the primary target of lipid-lowering therapy based on a robust database of studies **linking** LDL-C **to** atherosclerosis and cardiovascular events… (Bittner 2003) These factors appear to converge with known pathways that **link** oxidative stress **with** adhesion molecule expression (Granger et al. 2004)
with	prep.	X ([causal marker]) with Y [causal marker] X with Y	5	7007	12.12	The mammographic density does not increase **with** tibolone, unlike with HRT. (Kocjan and Prelevic 2003) Odds ratios for VTE were increased **with** oral HRT as compared with controls; the same applied for oral as opposed to transdermal HRT. (Seed and Knopp 2004) A further aspect of the change of atherogenicity of lipoproteins **with** HRT was tackled by Wakatsuki et al. … (Seed and Knopp 2004)
related to	ppl.a. + prep.	X [copula] related to Y X related to Y	4	150	0.26	… psychosocial factors may be **related to** the development of CVD. (Harris and Matthews 2004) … the risk of mortality from breast cancer **related to** HRT could not be determined. (Watkins 2003)

correlate	v. + prep. v. + conj.	X correlate with Y X has been correlated with Y [data] correlates X and Y	3	118	0.20	… increased circulating IGF-1 concentrations **correlate** very closely **with** the relative risk for the development of several common cancers, including breast, prostate, colon, and lung. (McCance and Jones 2003) Cell adhesion molecules **have also been correlated with** CHD. (Rackley 2004) To date, only limited data **correlate** Bcl-XL expression **and** breast cancer treatment response in humans… (Garg et al. 2003)
relevant to	adj. + prep.	X [copula] relevant to Y X relevant to Y	3	18	0.03	… lipid-independent effects of statins on various signaling pathways that are potentially **relevant to** the pathogenesis of atherosclerosis. (Balk et al. 2003) The endothelium contributes to the regulation of vascular tone, platelet aggregation, and other processes **relevant to** atherosclerosis. (Schwartz 2003)
find… in	v. + prep.	X was found in Y [researcher] found X in Y	2	106	0.18	… strong expression of cyclin D1, p21WAF1/CIP1, and Ki-67 **was found in** a DCIS lesion… (Wang et al. 2003) Barbareschi et al. (15) **found** high expression of p21WAF1/CIP1 **in** DCIS and invasive breast cancer. (Wang et al. 2003)
link between… and	n. + prep. + conj.	link between X and Y	2	42	0.07	Part 1 will provide a brief overview of the **link between** inflammation, endothelial dysfunction, **and** atherosclerosis… (Szmitko et al. 2003)
predict	v.	X predicts Y	2	195	0.34	In addition, baseline renal function **predicted** development of CHF. (Coresh et al. 2004)
prediction of	n. + prep.	X [conjunction/ association marker]	2	22	0.04	High-sensitivity C-reactive protein and the **prediction of** coronary events among patients with

		[article] prediction of Y				renal disease (Torres and Ridker 2003)
relation	n. + prep. + conj. n. + prep. + prep.	relation between X and Y relation of X to Y	2	110	0.19	… the exact nature of the **relation between** hepatic lipase **and** atherosclerosis remains controversial (Zambon et al. 2003) **Relation of** Nonhypertensive BP Categories **to** Development of Hard CVD in Framingham Study Subjects, Ages 35-90 (Kannel et al. 2003)
connect… to	v. + prep.	[data] connecting X to Y	1	2	0.003	The first data **connecting** the ubiquitous Epstein-Barr virus **to** smooth muscle tumors surfaced… (Fackelmann 1995)
detect… in	v. + prep.	X was detected in Y	1	36	0.06	Microalbuminuria (urinary ACR>2 mg/mmol) was **detected in** 32.2% of patients with diabetes and **in** 14.7% of patients without diabetes. (MacIsaac et al. 2004)
identify	v.	X identifies Y	1	408	0.71	In this clinical setting higher levels of albuminuria most likely **identified** patients with an exaggerated inflammatory response. (MacIsaac et al. 2004)
indicate	v.	X indicates Y	1	338	0.59	The inflammatory marker C-reactive protein (CRP) can **indicate** low-grade chronic inflammation… (MacKenzie 2004)
indication of	n. + prep.	X [copula] [article] indication of Y	1	20	0.03	Low expression of co-stimulatory molecules on circulating DCs is an **indication of** immaturity. (Pockaj et al. 2004)
note… in	v. + prep.	X have been noted in Y	1	12	0.02	Although improved CRP levels **have been noted in** acutely ill patients receiving insulin… (Pantaleo and Zonszein 2003)
observe… in	v. + prep.	[researcher] observed [causal marker] X in Y	1	94	0.16	...we **observed** further down-regulation of ARHI expression **in** the invasive carcinoma component compared with adjacent normal epithelia (P 0.000001; Table 3). (Wang et al. 2003)
occur with	v. + prep.	X occurs with Y	1	12	0.02	… syndrome of high insulin levels that **occur with** type II diabetes and atherosclerosis. (McCance and

						Jones 2003)
parallel	v.	X parallels Y	1	3	0.005	… frequency rises rapidly in middle age and **parallels**, with some lag time, the development of obesity in the population. (Grundy et al. 2004)
predictor of	n. + prep.	X [copula] [article] predictor of Y	1	134	0.23	In this regard, epidemiologic and laboratory evidence suggest that hs-CRP may be a powerful **predictor of** vascular risk among such patients. (Torres and Ridker 2003)
prognostic factor in	adj. + n. + prep.	X [copula] [article] prognostic factor in Y	1	11	0.02	Lymph node status at the time of diagnosis is a major **prognostic factor in** breast cancer… (Susnik et al. 2004)
relate… to	v. + prep.	[mechanism] relates X to Y	1	36	0.06	… endothelial function may provide a testable model for exploring a novel mechanism **relating** psychosocial factors **to** CVD development. (Harris and Matthews 2004)
report… in	v. + prep.	X has been reported in Y	1	92	0.16	As is the case for chemotherapy, radiation-induced NF-[kappa]B activation **has been reported in** a variety of cancer cell types… (Garg et al. 2003)
specific to	adj. + prep.	X [copula] specific to Y	1	11	0.02	It is also seen in this study that MUC2 expression is highly **specific to** CC, regardless of the location of the tumor in the breast or pancreas. (Adsay et al. 2003)
susceptibil-ity to	n. + prep.	X [causal marker] [noun]'s susceptibility to Y	1	16	0.03	While heredity can influence a person's **susceptibility to** development of the disease, a sedentary lifestyle and long-term obesity are key triggering events for most people. (Haskell 2003)
Mean			3.76	1390.97	0.26	

CAUSE–EFFECT: CREATION (51 markers, 26 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
role	v. + art. + n. + prep.	X plays ([article]) role in ([article])Y	33	596	1.03	Endothelial dysfunction **plays a** central **role in** the pathogenesis of CVD in patients with abnormal carbohydrate metabolism. (Pantaleo and Zonszein 2003)
		X in which Y plays a role				It has been recognized that atherosclerosis is an inflammatory disease **in** which various cytokines **play a** significant **role**… (Taniyama and Griendling 2003)
	n. + prep. + prep.	role for X in Y				William Osler 3 was one of the first to propose a major **role for** acute infection **in** the pathogenesis of atherosclerosis. (Madjid et al. 2004)
		role of X in ([article]) Y				As with heart failure, the **role of** aldosterone **in** the pathogenesis of hypertension has also been studied for decades. (Moore et al. 2003)
	n. + prep.	X role in Y				Endothelial function is important because of its **role in** CVD development. (Harris and Matthews 2004)
contribute to	v. + prep.	X contributes to Y	13	206	0.36	By studying the normal function of BRCA2, we can understand how changes in the protein **contribute to** the development of cancer… (Graham 2002)
induce	v.	X induces Y	11	338	0.58	hs-CRP has also been reported to **induce** the expression of plasminogen activator inhibitor-1…

		X [verb phrase] by inducing Y				(Torres and Ridker 2003) Other preclinical studies show that CRP may facilitate the development of atherosclerosis by 1) **inducing** foam cell formation necessary for plaque development … (Rackley 2004)
lead to	v. + prep.	X leads to Y	9	352	0.61	While the ADH3 [gamma]1 allele **leads to** rapid oxidation of ethanol, the [gamma]2 allele results in slow ethanol oxidation. (Humphries et al. 2004)
involved in	ppl.a. + prep.	X involved in Y X [copula] involved in [article] Y	8	105	0.18	…the relationship between ANS and endothelial function provides a testable model for examining processes **involved in** CVD development. (Harris and Matthews 2004) Recently, accumulating evidence has shown that fractalkine is **involved in** the pathogenesis of various clinical disease states or processes, such as atherosclerosis… (Umehara et al. 2004)
implicate	v. + prep.	[evidence] implicates X in [article] Y X is/have been implicated in Y	7	31	0.05	There is a large body of evidence that **implicates** inflammation and adhesion molecules **in** the pathogenesis of CVD, including atherosclerosis, stroke, and myocardial infarction. (Granger et al. 2004) MMPs **have been** broadly **implicated in** a number of cardiovascular diseases, including atherosclerosis, 90,94 aortic aneurysms, 95 and heart failure… (Jaffer and Weissleder 2004)
result	v. + prep.	X results from Y X results in Y	7	302	0.52	The response to injury hypothesis developed by Russell Ross in the late 1970s suggested that atherosclerosis, at least, **resulted from** an initial injury to endothelial cells… (Griendling and FitzGerald 2003a) While the ADH3 [gamma]1 allele leads to rapid

						oxidation of ethanol, the [gamma]2 allele **results in** slow ethanol oxidation. (Humphries et al. 2004)
mediated	ppl.a. ppl.a. + prep.	X-mediated Y X ([copula]) mediated by Y	6	155	0.27	Endothelial dysfunction and the subsequent changes in blood flow promote CD40-**mediated** endothelial activation by decreasing the intracellular expression of a CD40 signaling blocker. (Szmitko et al. 2003) Platelet activation and aggregation ensue, **mediated by** interactions with thrombin, TF, and von Willebrand factor. (Szmitko et al. 2003)
cause	v.	X causes Y	5	229	0.40	Preoperative chemotherapy often **caused** shrinkage of the tumour… (Shenkier et al. 2004)
importance of… in	n. + prep. + prep.	importance of X in ([article]) Y	5	27	0.05	Third, researchers increasingly recognize the **importance of** nonlipid factors **in** the pathogenesis of atherosclerosis. (Balk et al. 2003)
important in	adj. + prep.	X [copula] important in Y [copula] X important in Y X [causal marker] [article] important [noun] in Y	5	52	0.09	We now appreciate that the fractalkine/CX3CR1 system is **important in** various clinical diseases, such as atherosclerosis, cardiovascular disease, graft rejection, HIV infection, and inflammatory diseases. (Umehara et al. 2004) … the following critical question remains unanswered: 'is oxidation **important in** human atherosclerosis?'. (Brennan and Hazen 2003) Aldosterone has been implicated for many years as an **important** substance **in** the pathogenesis of heart disease. (Moore et al. 2003)
pathway	n. + prep.	X [copula]/[verb + preposition] [article] pathway for Y X [copula] [article] pathway in Y	4	268	0.46	… mitochondrial oxidative stress is a major central **pathway in** the pathogenesis of diabetic complications. (MacIsaac et al. 2004) Endothelial dysfunction is a new **pathway in** cardiovascular disease (CVD) development.

		X as[article] pathway of Y				However, many questions remain regarding interactions between ANS and endothelial function as a **pathway of** CVD development. (Harris and Matthews 2004)
due to	adj. + prep.	X due to Y X [copula] due to Y	3	181	0.31	… persons scoring higher on a scale of spirituality or religious participation have lower mortality **due to** CHD… (Haskell 2003) Therefore, the anti-tumor efficacy of Virulizin observed in these models is likely **due to** activation of macrophages. (Du et al. 2003)
mediate (by/through/ via)	v.	X ([copula]) mediated by Y X ([copula]) mediated through Y X ([copula]) mediated via Y	3	215	0.37	MMP overexpression and activation within the plaque **are mediated by** IL-1[beta], TNF-[alpha], oxLDL, and CD40L. (Szmitko et al. 2003) These effects appear to **be mediated** in part **through** activation of the c-Src, p38 mitogen-activated protein kinase, and the cell survival kinase (Akt) … (Griendling and FitzGerald 2003a) The chemopreventive effects of retinoic acids might **be mediated via** PKC-[delta] activation. (Schondorf et al. 2004)
produce	v.	X produces Y X [copula] produced by Y	3	248	0.43	Activation of these receptors **produces** endothelium-dependent relaxation of human coronary arteries. (Harris and Matthews 2004) Reactive oxygen species **are produced** continuously **by** all cells in normal and pathological aerobic metabolism, from xenobiotics to ionizing radiation. (Kang 2002)
cause of	n. + prep.	cause of X [copula] Y	2	95	0.16	The most common **cause of** brain infarction is hardening of the arteries (atherosclerosis). (DiGiovanna and Adams 1999)

		X [copula] [article] cause of Y				Atherosclerosis is the leading **cause of** morbidity and mortality in developed countries.
drive	v.	X drives Y	2	22	0.04	It is presumed that aberrant cyclin D1 expression **drives** the phosphorylation and functional inactivation of pRB in tumor cells. (Sicinski and Weinberg 1997)
implicated in	ppl.a. + prep.	X implicated in Y	2	41	0.07	... we recently tested whether statins decrease formation of nitric oxide-derived oxidants in vivo [22**], species **implicated in** development of atherosclerosis. (Brennan and Hazen 2003)
induced	ppl.a. ppl.a. + prep.	X-induced Y X induced by Y	2	207	0.37	As is the case for chemotherapy, radiation-**induced** NF-[kappa]B activation has been reported in a variety of cancer cell types... (Garg et al. 2003) Rho belongs to a family of small GTP-binding proteins that mediate intracellular signaling **induced by** activation of heterotrimeric G protein-coupled receptors and growth factor receptors. (Force et al. 2004)
initiate	v.	X initiates Y	2	88	0.15	Thus, other triggers--including diabetes, high blood pressure, or chemicals in cigarette smoke--can also **initiate** the signals... (Stix 2003)
key... in	adj. + prep.	X [copula] [article] key [noun] in Y	2	29	0.05	Oxidation of LDL is a **key** process **in** atherogenesis. (Mason et al. 2003)
mechanism of	n. + prep.	([causal marker]) X [as] [article] mechanism of Y	2	73	0.13	Further, recent studies implicating translocation of SK1 to the membrane as a **mechanism of** activation have not been demonstrated for SK2. (Saba and Hla 2004) Activation of nuclear factor-[kappa]B as a **mechanism of** resistance to chemotherapy (Garg et al. 2003)

participate in	v. + prep.	X participates in Y	2	73	0.13	… is consistent with this heme protein **participating in** the development of atherosclerosis and its thrombotic complications. (Brennan and Hazen 2003)
product of	n. + prep.	X, [article] product of Y	2	34	0.06	AGEs, the **products of** nonenzymatic glycation and oxidation of proteins and lipids, accumulate in the vessel wall… (Yan et al. 2003)
trigger	v.	X triggers Y X [causal marker] trigger Y	2	72	0.12	This enhances retention of the lipoprotein and possibly **triggers**, along with oxidation, the formation of a recognizably foreign substance… (Caslake and Packard 2003) This is in stark contrast with the properties of fibroblasts where the ectopic expression of cyclin D1 shortens the G1 phase but is not sufficient to **trigger** S-phase entry. (Sicinski and Weinberg 1997)
via	prep.	X via Y	2	86	0.15	Lipid oxidation **via** reactive nitrogen species (Brennan and Hazen 2003)
as a result of	prep. + art. + n. + prep.	X as a result of Y	1	49	0.08	… coronary arteries that are becoming blocked **as a result of** atherosclerosis. (Beardsley 2000)
behind	prep.	X behind Y	1	20	0.03	The mechanisms **behind** aberrant expression of NF-[kappa]B are beginning to be understood. (Garg et al. 2003)
caused by	ppl.a. + prep.	X caused by Y	1	31	0.05	The association between antibiotic use and death **caused by** breast cancer was similar to that observed for incident breast cancer… (Sawka 2004)
complication of	n. + prep.	X [copula] as [article] complication of Y	1	12	0.02	Although atherosclerosis is a multifactorial disease, often occurring as a **complication of** hypertension, obesity, and diabetes… (Umehara et al. 2004)
confer	v.	X confers Y	1	32	0.06	… oxidation of LDL **conferred** these apparent 'atherogenic' properties to an otherwise

						'nonatherogenic' native LDL particle... (Brennan and Hazen 2003)
consequence of	n. + prep.	consequence of X [copula] Y	1	37	0.06	One of the **consequences of** increased ROS production is oxidation of LDL, which modifies its bioactivity extensively in vitro... (Griendling and FitzGerald 2003a)
contributor to	n. + prep.	X [copula] ([article]) contributor to Y	1	10	0.02	Platelets, angiotensin II, and the CD40/CD40 ligand signaling system are gaining recognition as **contributors to** the pathogenesis of CVD. (Granger et al. 2004)
create	v.	X [causal marker] create Y	1	54	0.09	Endothelial cells help **create** this antithrombogenic surface. (Granger et al. 2004)
dependent	adj.	X-dependent Y	1	237	0.41	Witztum's group 40,41 has developed a range of antibodies directed against oxidation-**dependent** epitopes in LDL (anti-oxLDL)...
effect of	n. + prep.	effect of X (Y)	1	417	0.72	... they are important targets of the biological **effects of** fractalkine (ie, chemotaxis, adhesion, and activation) while also having cytoplasmic granules containing perforin and granzyme B. (Umehara et al. 2004)
elicit	v.	X elicits Y	1	20	0.03	Activated platelets can release and/or activate a variety of inflammatory molecules (Table 2) that can **elicit** endothelial activation. (Granger et al. 2004)
explain	v.	X explains Y	1	149	0.26	... common tumor pathogenesis might **explain** similarities in tumor features in the FDT and PABC groups. (Siegelmann-Danieli et al. 2003)
explanation for	n. + prep.	explanation for X [copula] Y	1	29	0.05	The **explanation for** this discrepancy could be the incidental activation of the TP by lipid peroxidation products, including the iPs... (Griendling and FitzGerald 2003)
for	prep.	X (for Y)	1	6076	10.53	There was no consistency among these patients with respect to prior chemotherapy (1 **for** metastatic disease, 6 adjuvant, 3 chemonaive).

from	prep.	X from Y	1	2246	3.89	In 2000 there were over an estimated over 1 million new cases and approximately 373,000 deaths **from** breast cancer worldwide... (Carrick et al. 2004)
generate	v.	X [copula] generated by Y	1	129	0.22	MCa-35 and A549 tumors **were generated** by injecting 1 x 106 or 10 x 106 cell s.c. in the right thighs of C3H/He or nu/nu NCR mice... (Liu et al. 2003)
involve	v.	X involves Y	1	179	0.31	... a highly coordinated and well-regulated process that **involves** the expression and/or activation of adhesion molecules on endothelial cells and circulating inflammatory cells. (Granger et al. 2004)
involvement of... in	n. + prep. + prep.	involvement of X in [article] Y	1	13	0.02	... in accordance with the **involvement of** a localized process, such as clustering of mononuclear infiltrates, **in** the pathogenesis of plaque rupture. (Willerson and Ridker 2004)
mediator of	n. + prep.	X [copula] [article] mediator of Y	1	15	0.03	ONOO[middle dot]- is an important **mediator of** lipid peroxidation and protein nitration, including oxidation of LDL, which has dramatic proatherogenic effects. (Griendling and FitzGerald 2003a)
participant in	n. + prep.	X [copula] [article] participant in Y	1	25	0.04	Collectively, this has raised questions about the validity of the hypothesis that oxidation is a critical **participant in** the atherosclerotic disease process. (Brennan and Hazen 2003)
production of	n. + prep.	X production of Y	1	110	0.19	The presence of TNF- [alpha], IL-6, and other cytokines cause hepatic **production of** C-reactive protein (CRP)... (Pantaleo and Zonszein 2003)
prompt	v.	X prompts Y	1	24	0.04	The R +enantiomer of amlodipine, once present in the plasma membrane, **prompts** the production of NO, ultimately through the activation of eNOS. (Mason et al. 2003)

provoked	ppl.a.	X-provoked Y	1	1	0.002	… the potential for recall-**provoked** activation may be sustained over significant periods of time. (Schwartz 2003)
responsible for	adj. + prep.	X, responsible for Y	1	61	0.11	The 26S proteosome, **responsible for** the degradation of the inhibitory I[kappa]B[alpha] protein and subsequent activation of NF-[kappa]B (Fig. 2), has been the subject of intense study… (Garg et al. 2003)
spur	v.	X spurs Y	1	11	0.02	Dickman and his colleagues found that the tumor cells contained an Epstein-Barr protein known to **spur** certain immune cells to divide in laboratory experiments. (Fackelmann 1995)
Mean			**3.31**	**276.71**	**0.20**	

CAUSE–EFFECT: DESTRUCTION (5 markers, 2 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
anti-	affix	X [verb] anti-Y [causal marker]	3	880	1.52	Administration of Virulizin showed **anti**-tumor efficacy in the treatment of human pancreatic cancers and melanoma… (Du et al. 2003)
against	prep.	X [preposition] [causal marker] against Y X against Y	2	188	0.33	As the armamentarium of CTX combinations with effectiveness **against** breast tumors expands, additional crossover regimens for patients with resistant disease will be more readily available… (Newman et al. 2003) COX-2 inhibition combined with immune-based therapy that would induce cytotoxic T-lymphocyte activity **against** tumor cells is a novel

						concept that needs further exploration in preclinical animal models and in clinical settings. (Pockaj et al. 2004)
destroy	v.	X [copula] [causal marker] destroy Y	1	11	0.02	Others skipped radiation treatment, which is used to **destroy** any cancer cells left behind after surgery. (Loecher 2001)
loss of... through	n. + prep. + prep.	loss of X through Y	1	1	0.002	... allele has been found to be imprinted, methylated, and silenced in 7 of 9 informative cases (3), consistent with a **loss of** expression **through** imprinting and a LOH of the nonimprinted allele. (Wang et al. 2003)
kill off	v. + prep.	X [copula] killed off by [article] Y	1	2	0.003	Long before that, cancer cells make their way into the bloodstream and lymphatic system, where they're either **killed off** by the body's own immunity... (Perlmutter 1992)
Mean			**1.6**	**216.40**	**0.16**	

CAUSE–EFFECT: MAINTENANCE/PERMISSION (11 markers, 1 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
required for	ppl.a. + prep.	X [copula] required for Y	2	23	0.04	Therefore, it is currently suggested that ER[alpha] function may be **required for** maximum activation of IGF-signaling pathways. (McCance and Jones 2003)
allow for	v. + prep.	X allows for [article] X	1	34	0.06	Furthermore, the fact that cancer cells express proteins different from those of normal healthy cells **has allowed for** the development of targeted molecular therapies such as the use of HER2 monoclonal antibody trastuzumab. (Major 2003)

critical for	adj. + prep.	X [copula] critical for Y	1	12	0.02	S1P1 is particularly **critical for** Rac activation in endothelial cells… (Saba and Hla 2004)
dependent	adj.	X-dependent Y	1	237	0.41	First, the estrogen-**dependent** step in mammary gland development, the ductal elongation that takes place during puberty (27,28), proceeds normally in cyclin Dl-deficient mice… (Sicinski and Weinberg 1997)
enable	v.	X enables [article] Y	1	32	0.06	With faster scans, greater coverage, and improved spatial and temporal resolution, MDCT **has enabled** the development of fundamentally new applications of CTA… (Napoli et al. 2004)
necessary for	adj. + prep.	X [copula] necessary for Y	1	20	0.03	It is possible that basal IGF activation of the ER may be **necessary for** maximal estrogen-mediated activation… (Mason et al. 2003)
permit	v.	X permits [article] Y	1	25	0.04	Preoperative chemotherapy often caused shrinkage of the tumour and **permitted** the performance of breast-conserving surgery (BCS)… (Shenkier et al. 2004)
perpetuate	v.	X perpetuates Y	1	9	0.01	This suggests that cognitions or emotions induced after the stressor **perpetuate** cardiovascular activation. (Schwartz 2003)
pivotal in	adj. + prep.	X [copula] pivotal in [article] Y	1	2	0.003	Because T cells and DCs are **pivotal in** the development of antitumor immunity… (Pockaj et al. 2004)
require	v.	X requires Y	1	234	0.41	… effects of oxidized LDL on vascular smooth muscle cells, which contribute to the atherogenic process appear to **require** the activation of SK. (Saba and Hla 2004)
support	v.	X supports [article] Y	1	220	0.38	A LARGE AREA of suspicious new blood-vessel growth **supporting** a cancer tumor is clear in an image made with computed tomography laser mammography (CTLM). (Miller 2002)
Mean			**1.09**	**77.09**	**0.13**	

CAUSE–EFFECT: PREVENTION (6 markers, 3 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
prevent	v.	X prevents Y	6	194	0.34	Normally, HDL **prevents** LDL oxidation. (Cabe 2000)
		[causal marker] X to prevent Y				… it would have important implications for the ability of PON1 to **prevent** atherosclerosis. (Mackness et al. 2004)
		X [causal marker] to prevent Y				… in future tamoxifen may even help to **prevent** cancer development. (Health News 1991)
prevention	n.	X as Y prevention	6	322	0.56	It is believed that this effect is insufficient to recommend HRT **as** a diabetes **prevention** strategy in women with CHD. (Kocjan and Prelevic 2003)
	n. + prep.	X in ([article]) prevention of Y				These studies provide a scientific basis for further trials of HRT **in prevention** and amelioration **of** type 2 diabetes… (Seed and Knopp 2004)
		X ([ppl.a.]/[noun]/ [copula] [adjective]) for ([article]) prevention of Y				HRT is effective **for prevention** or treatment **of** osteoporosis… (Kocjan and Prelevic 2003)
suppressor	n.	X suppressor Y	4	48	0.08	This effect has now been seen for more than a dozen tumor **suppressor** genes, and investigators expect to find many more like them. (Gibbs 2003)
block	v.	X block Y	1	86	0.15	As one caveat, PD98059 and U0126 also **block** activation of MEK5… (Force et al. 2004)

role of... in	n. + prep. + prep.	role of X in Y	1	113	0.20	Although the **role of** dietary and vitamin antioxidants **in** the development of breast cancer is not conclusive in human studies... (Kang 2002)
suppression of	n. + prep.	X, [causal marker] suppression of Y	1	21	0.04	... the inhibitory effect of statins on promoter IV of MHC-II transactivating factor, leading to **suppression of** T-lymphocyte activation. (Davignon 2004)
Mean			**3.17**	**130.67**	**0.23**	

CAUSE–EFFECT: MODIFICATION (20 markers, 6 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
effect	n. + prep.	X effects of Y	12	1465	2.54	If so, how could prior assessments of the health **effects of** hormone replacement therapy (HRT) have been so different? (Grimes and Lobo 2002)
	n. + prep.	X's effect on Y				No study evaluated the associations between statins' **effects on** LDL oxidation and lipid levels... (Balk et al. 2003)
	n. + prep. + prep.	X has ([article]/[quantifier]) effect on Y effect of X on Y				Unlike combination HRT, therapy with estrogen alone did not appear to have any **effect** (either favorable or adverse) **on** heart disease... (Aschenbrenner 2004) Recognition of the **effects of** influenza **on** CHD provides the medical community with a valuable opportunity to further reduce cardiovascular death and morbidity. (Madjid et al. 2004)

affect	v.	X affects Y X [copula] affected by Y	7	198	0.34	In addition, interactions between dihydropyridines and these pathways **affect** lipid oxidation and cholesterol metabolism and can thereby reduce atherosclerosis development. (Mason et al. 2003) It is well known that multiple aspects of one's QOL can **be affected by** the development of coronary artery disease. (Berra 2003)
respond to	v. + prep.	X responds to Y	6	47	0.08	… among ER-positive tumors, nearly 70% of those that are also progesterone receptor (PR)-positive and 25-30% of PR-negative tumors **will respond to** hormonal therapy. (Vogel 2003) Patients with stage IIIB disease who **respond to** chemotherapy should receive surgery plus locoregional radiotherapy. (Shenkier et al. 2004)
response	n. + prep. + prep. n. + prep.	response of X to Y X response to Y X [verb] [article] response to Y	5	726	1.29	Koh [36**] has argued that the **response of** endothelium **to** HRT depends on the presence of estrogen receptors. The conceptual advantage of in vivo assessment of primary tumor **response to** the selected CTX regimen is another benefit derived from the neoadjuvant CTX approach. (Newman et al. 2003) … colorectal cancer patients homozygous for the triple repeat (3R/3R) had a poorer **response to** 5-FU chemotherapy. (Karvellas et al. 2004)
influence	v.	X influences [article] Y	2	124	0.21	Emerging data reveals that a large number of additional proteins (i.e., growth factors) **influence** the transcriptional activation of ER[alpha] and possibly ER[beta]. (McCance and Jones 2003)
regulated	ppl.a.	X-regulated Y	2	35	0.06	TNF-[alpha]-**regulated** SK activation is likely to be important in nuclear factor-[kappa]B (NF-[kappa]B) activation and inhibition of apoptosis.

						(Saba and Hla 2004)
act on	v. + prep.	X acts on Y	1	10	0.02	The balance between these opposing forces **acts on** the vascular smooth muscle cells to maintain the appropriate vessel tone. (Harris and Matthews 2004)
address	v.	X addresses Y	1	118	0.20	Further testing of other therapies, particularly ones **addressing** oxidation and thrombosis, is needed. (Coresh et al. 2004)
change in	n. + prep.	X [causal marker] ([article]) change in Y	1	252	0.44	Receptor-mediated leukocyte activation leads to conformational **changes in** LFA-1 structure… (Granger et al. 2004)
complicatio n of… in	n. + prep. + prep.	complication of X in Y	1	12	0.02	Ultimately, these pathways synergize to construct a scaffold on which the **complications of** diabetes in the vasculature and heart may be built. (Yan et al. 2003)
control	v.	X for controlling Y	1	166	0.29	… it warrants extensive research, which may pave the way for new, more efficient methods for preventing and **controlling** CHD. (Madjid et al. 2004)
for	prep.	X for Y	1	6051	10.48	… 28% beginning new medications **for** cholesterol, blood pressure, or diabetes and 12% changing current medication dosages. (Berra 2003)
impact on	n. + prep.	impact on X [copula] [causal marker] [article] Y	1	41	0.07	It should be remembered that mouse Lp-PLA2 cannot bind to LDL and the **impact on** atherosclerosis in these models is the result of the action of the enzyme carried in the HDL reservoir. (Caslake and Packard 2003)
importance of… upon	n. + prep. + prep.	importance of X upon [article] Y	1	1	0.002	Recent research has emphasized the **importance of** the intra-uterine environment **upon** the subsequent development of a number of adult diseases (Barker, 1992). (Rigby et al. 2002)
influence on	n. + prep.	X [verb] [article] influence on Y	1	12	0.02	… no account was made of differing methods of laboratory analysis, grade (which has a profound **influence on** ERa expression), type of breast

						cancer, and threshold value… (Tran and Lawson 2004)
manifestation of	n. + prep.	X manifestation of Y	1	14	0.02	… can occur well before the structural **manifestation of** atherosclerosis… (Szmitko et al. 2003)
modulate	v.	X modulates Y	1	51	0.09	Recently, an exciting report provided evidence for a new pathway by which hepatic lipase may **modulate** atherosclerosis. (Zambon et al. 2003)
regulation of… by	n. + prep. + prep.	regulation of X by Y	1	6	0.01	The **regulation of** VSMC growth **by** oxidized LDLs operates through the activation of the Ras/Raf/MEK/MAPK signaling pathway by a Pertussis toxin-sensitive, G protein-coupled receptor. (Mason et al. 2003)
regulator of	n. + prep.	X [copula] [article] regulator of Y	1	18	0.03	Although promising, this kinase is a critical **regulator of** many basic cellular processes, including development, cardiac growth and hypertrophy, and tumorigenesis. (Shenkier et al. 2004)
role of… for	n. + prep. + prep.	role of X for Y	1	4	0.01	the **role of** radiation therapy **for** invasive breast cancer treated with BCS is now well accepted. (Meric-Bernstam 2004)
Mean			**2.4**	**467.55**	**0.16**	

CAUSE–EFFECT: INCREASE (14 markers, 7 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
promote	v.	X promotes Y	10	159	0.28	IL-18 also **promotes** adhesion molecule expression on the endothelium and promotes plaque instability by enhancing MMP secretion.

		[causal marker] to promote X, Y				(Szmitko et al. 2003) In addition to having direct effects **to promote** EC activation, CRP appears to function in a fashion that inhibits bone marrow-derived endothelial progenitor cell survival… (Szmitko et al. 2003)
increase	v.	X increases Y X, increasing Y	9	741	1.28	Several recent reports have demonstrated that estrogen therapy **increases** expression of MMP. (Karas 2004) … all of which are reported to interact, **increasing** the development of insulin resistance syndrome. (McCance and Jones 2003)
enhance	v.	X enhances Y X has been enhanced by Y	2	130	0.23	Lp(a) also **enhances** oxidation of LDL. (Cabe 2000) The development of spatial statistics **has been** greatly **enhanced** by rapidly evolving computational tools… (Schootman and Sun 2004)
facilitate	v.	X facilitates Y	2	73	0.13	Other preclinical studies show that CRP may **facilitate** the development of atherosclerosis… (Rackley 2004)
increased	ppl.a.	increased X [causal marker] Y X [causal marker] increased Y	2	538	0.93	This **increased** thrombogenesis is caused by several mechanisms that include platelet activation and hyperaggregability, as well as elevated levels of procoagulants such as fibrinogen and von Willebrand factor. (Pantaleo and Zonszein 2003) Receptor-mediated leukocyte activation leads to … **increased** adhesiveness…: (Granger et al. 2004)
stimulate	v.	X stimulates Y	2	99	0.17	Adiponectin decreases postprandial FFA levels and **stimulates** myocellular fatty acid oxidation… (Pantaleo and Zonszein 2003)
upregulate	v.	X upregulates Y	2	6	0.01	Because LDL **upregulates** angiotensin II receptor type 1 (AT1) receptor expression… (Griendling

						and FitzGerald 2003)
augment	v.	X can be augmented by Y	1	33	0.06	Plasma levels of PAI-1 are regulated on a genetic basis, and its expression can **be augmented** by insulin resistance and other factors such as abnormal adiposity… (Pantaleo and Zonszein 2003)
catalyse	v.	X [copula] catalysed by Y	1	32	0.06	(In)activation of aromatic amine carcinogens is **catalysed by** metabolic enzymes including N-acetyltransferase 1… (Van der Hel et al. 2003)
catalyst	n.	X as [article] catalyst of Y	1	6	0.01	Myeloperoxidase as an enzymatic **catalyst** of lipid oxidation: formation of the lipid-laden plaque (Brennan and Hazen 2003)
elevate	v.	X elevates Y	1	30	0.05	By inhibiting ACC, AMPK **elevates** fat oxidation. (Force et al. 2004)
enhanced	ppl.a.	X, [causal marker] enhanced Y	1	65	0.11	Impaired ANS regulation is associated with greater platelet activation, contributing to **enhanced** aggregation and adhesion to vessel walls. (Harris and Matthews 2004)
pro-	affix	X [verb] pro-Y [causal marker]	1	136	0.24	… oxidation of LDL, which has dramatic **pro**atherogenic effects. (Griendling and FitzGerald 2003a)
stimulated by	ppl.a. + prep.	X stimulated by Y	1	7	0.01	NO[middle dot], which, because of its ability to scavenge radicals, can act as an antioxidant and reduces VCAM-1 expression **stimulated by** TNF-[alpha], possibly by inhibiting the formation of peroxy-fatty acids. (Taniyama and Griendling 2003)
Mean			**2.57**	**146.79**	**0.26**	

CAUSE–EFFECT: DECREASE (13 markers, 7 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
reduce	v.	X reduces Y	6	554	0.96	CRP was recently shown to **reduce** synthesis of the vasodilator nitric oxide in cultured endothelial cells. (Rackley 2004)
		by X, Y [copula] reduced				The interesting thing they found was that by combining soy and tamoxifen, the tumors **were reduced** even further, by an impressive 62 percent. (Franzen 2001)
		X [causal marker] reduce Y				... chemotherapy,... tamoxifen,... and RT... all act to **reduce** LR independently of surgery. (Naik et al. 2004)
inhibit	v.	X inhibits Y	5	175	0.30	Hydroxy metabolites of atorvastatin, but not the parent compound, **inhibit** oxidation of both LDL and very-low-density lipoprotein as well as high-density lipoprotein. (Davignon 2004)
decrease	v.	X decreases Y	2	297	0.51	NO is an important vasodilator that **decreases** LDL oxidation and smooth muscle cell proliferation. (Torres and Ridker 2003)
downsizing	n.	X for Y downsizing	2	7	0.01	… the role of preoperative chemotherapy **for** tumor **downsizing**… (Meric-Bernstam 2004)
	n. + prep.	X downsizing with Y				… locally advanced tumors may also become eligible for breast-conserving surgery after tumor **downsizing with** preoperative chemotherapy… (Meric-Bernstam 2004)

inhibition of	n. + prep.	X [causal marker] inhibition of Y	2	96	0.17	… free radical-scavenging abilities that may contribute to **inhibition of** lipoprotein oxidation. (Davignon 2004)
lower	v.	X lowers Y X [preposition] lowering Y	2	236	0.41	… studies showed that HRT **lowered** low-density lipoprotein (LDL) cholesterol levels… (Aschenbrenner 2004) In randomized, controlled trials, 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase inhibitors, in the form of statins, have been shown to provide effective therapy for **lowering** CRP, in conjunction with their lipid-lowering effects. (Willerson and Ridker 2004)
reduced	ppl.a.	X [causal marker] reduced Y	2	108	0.19	Loss of ER[alpha] in MCF-7 cells causes **reduced** expression of IGF-signaling molecules, diminished IGF signaling, and failure to proliferate in response to estrogen or IGF-1. (McCance and Jones 2003)
attenuate	v.	X attenuates Y	1	30	0.05	Preliminary observations also suggest that CRP upregulates nuclear factor [kappa]B (NF[kappa]B) signaling in ECs while **attenuating** endothelial progenitor cell survival… (Szmitko et al. 2003)
decreased	ppl.a.	X, [causal marker] decreased Y	1	174	0.30	As is the case for chemotherapy, radiation-induced NF-[kappa]B activation has been reported in a variety of cancer cell types, including breast cancer, leading to **decreased** apoptosis… (Garg et al. 2003)
diminished	ppl.a.	X [causal marker] diminished Y	1	9	0.02	… this phenomenon contributed, at least in part, to **diminished** atherosclerosis… (Yan et al. 2003)
impaired	ppl.a.	X [causal marker] impaired Y	1	68	0.12	These results indicate that SNS activation may contribute to **impaired** endothelial function, possibly because of activation of [beta]-adrenergic receptors. (Harris and Matthews 2004)

inhibitor of	n. + prep.	X, [article] inhibitor of Y	1	60	0.10	… formation of malonyl-CoA, a potent **inhibitor of** fatty acid oxidation. (Force et al. 2004)
target	v.	X targets Y	1	101	0.17	These findings have implications for therapies **targeting** insulin resistance and diabetes as well as CVD. (Willerson and Ridker 2004)
Mean			**2.08**	**147.31**	**0.25**	

CAUSE–EFFECT: PRESERVATION (1 marker, none with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
sustain	v.	X sustains Y	1	27	0.04	Rumination may be an example of a psychological process that tends **to sustain** cardiovascular activation. (Schwartz 2003)
Mean			**1**	**27**	**0.04**	

French

ASSOCIATION (30 markers, 13 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
et	conj.	X et ([association marker]) Y	10	16001	23.00	CMV **et** athérosclérose (Chidiac and Braun 2002)

						Traitement hormonal substitutif **et** risque de cancer du sein (Serin and Escoute 1998)
lié à	ppl.a. + prep.	X lié à [article] Y X, lié à [article] Y	7	90	0.13	L'hypertension artérielle exacerbe les complications **liées au** diabète, telles que les complications microvasculaires (néphropathie et rétinopathie)… (Gonzalez and Palardy 2004) Cette implication des LDL oxydées nous amène à envisager une autre voie du dysfonctionnement endothélial, **liée à** l'oxydation des LDL et à l'athérosclérose. (Bonnefont-Rousselot et al. 2002)
facteur de risque	n. + prep. + n. + prep. n. + prep. + n. n. + prep. + n. + prep.	facteur de risque X tels que [article] Y facteur de risque X (Y) facteur de risque de X (Y)	6	386	0.56	… les **facteurs de risques** cardiovasculaires traditionnels tels que l'hypertension, les dyslipidémies, le diabète et l'obésité tronculaire sont en effet observés avec une fréquence croissante chez les patients VIH+… (Duong et al. 2003) … enfants démontrant d'autres **facteurs de risque** cardiovasculaire (obésité, tabagisme, hypertension, diabète, consommation d'aliments riches en matières grasses, prise de médicaments augmentant les lipides plasmatiques…)… (Lambert 2002) Nous avons recherché chez tous les patients les **facteurs de risque** d'athérosclérose (diabète, hypertension artérielle, tabagisme, hormonothérapie, intoxication alcoolique, dyslipidémie, hérédité)… (Desauw et al. 2002)
caractérisé par	ppl.a. + prep.	X [copula] caractérisé par [article] Y	5	12	0.02	L'hypercholestérolémie familiale est **caractérisée par** un cholestérol sérique élevé, des xanthomes tendineux, xanthélasmas, arcs cornéens et une athérosclérose précoce. (Chalès and Guggenbuhl 2004)

		X caractérisé par [article] Y				Dans l'adénose sclérosante, affection bénigne du sein **caractérisée par** une prolifération des cellules épithéliales et myoépithéliales, Clarke et al. [13] ont montré, à l'aide de l'anticorps CT-1, que… (Angèle et al. 2001)
risque de	n. + prep.	risque de X [causal marker] [article]/[preposition] Y X [causal/association marker] risque de ([article]) Y	5	695	1.00	Pour 17 patientes, il y avait un haut **risque de** récidive pariétale du fait de la présentation clinique… (Racadot et al. 2003) Par ailleurs, les **risques de** diabète de type 2, de maladie coronarienne et d'hypertension s'accroissent si le tour de taille dépasse 88 cm pour les femmes… (Béliveau and Léger 2004) L'obésité, le syndrome métabolique et le diabète accroissent notablement le **risque de** maladies cardiovasculaires. (Lambert 2002)
associé à	ppl.a. + prep.	X [copula] associé à [article] Y X associé à [article] Y	4	123	0.18	La prolifération des cellules myoépithéliales est donc **associée à** une néosynthèse de la protéine ATM au niveau nucléaire… (Angèle et al. 2001) Bien que primordial, nous ne parlerons pas du traitement de la dyslipidémie ou des autres troubles fréquemment **associés à** l'athérosclérose (notamment le diabète et l'hypertension). (Gendreau 2003)
lien entre… et	n. + prep. + conj.	lien(s) ([verb]) entre ([article]) X et ([association marker]) ([article]) Y	4	18	0.03	Il semble exister un **lien** très étroit **entre** le syndrome de lipodystrophie, l'hyperlipidémie, l'intolérance au glucose **et** le diabète, bien que chacun de ces troubles puisse survenir isolément. (Baril and Junod 2004)
au cours de	prep. + n. + prep.	X [causal marker] au cours de Y	2	302	0.52	La vitesse de l'onde de pouls est significativement altérée **au cours du** vieillissement, de l'hypertension artérielle, du diabète et de

						l'athérosclérose. (Levenson et al. 2000)
corrélé avec	ppl.a. + prep.	X [copula] corrélé avec [article] Y	2	25	0.04	… ses changements peuvent être **corrélés avec** une activation ou une répression de la transcription. (Chailleux et al. 2000)
en cas de	prep. + n. + prep.	en cas de X, Y	2	353	0.51	**En cas de** diabète équilibré, TG et LDL sont quasi normaux, cependant on peut noter un taux de HDL… (Fredenrich et al. 2004)
observé	ppl.a. + prep.phr.	X observé au niveau de/dans [article] Y	2	111	0.19	Par ailleurs, les anomalies qualitatives des lipoprotéines sont similaires à celles **observées** dans le diabète de type 2. (Fredenrich et al. 2004)
prédisposition	n. prep. + n. + prep.	prédispositions X [causal marker/association marker] Y X de prédisposition à [article] Y	2	102	018	Les gènes BRCA1 et BRCA2 sont impliqués dans deux tiers des **prédispositions** génétiques à l'origine d'un risque majeur de cancer du sein. (Coupier and Stoppa-Lyonnet 2002) Nous présentons ici une mise au point des connaissances sur les gènes de **prédisposition** héréditaire **au** cancer du sein, les perspectives de recherche et leurs implications dans la pratique du conseil génétique. (Bonadona and Lasset 2003)
retrouvé dans	ppl.a. + prep.	X retrouvé dans [article] Y X [copula] retrouvé dans [article] Y	2	68	0.10	Le profil lipidique le plus fréquemment **retrouvé dans** le diabète de type 2 associe une élévation du taux plasmatique des triglycérides (TG)… (Fredenrich et al. 2004) … Cox1 est localisée dans les cellules du stroma et n'est pas **retrouvée dans** les cellules tumorales… (Guastalla et al. 2004)
accompagner	v.	X accompagne [article] Y	1	16	0.03	La protéine C réactive (CRP) a longtemps été considérée comme un marqueur de l'état inflammatoire **accompagnant** l'athérosclérose. (Nalbone et al. 2002)
associer… à	v. + prep.	[person] associe [article] X à [article] Y	1	16	0.02	[O]n **associe** maintenant une faible capacité aérobie et une mauvaise composition corporelle **aux**

						maladies cardiovasculaires et au diabète, un manque de souplesse aux maux de dos, etc. (Béliveau and Léger 2004)
avec	prep.	X avec [article] [causal marker] [article] Y	1	2605	3.74	Les hyperlipoprotéinémies familiales primitives regroupent des maladies héréditaires **avec** une augmentation plasmatique du cholestérol et des triglycérides, résultant d'un déficit dans l'une des étapes du métabolisme lipidique… (Chalès and Guggenbuhl 2004)
corrélation de… et	n. + prep. + conj.	corrélation de X et Y	1	1	0.002	Il y a une **corrélation** significative **de** la transcription de l'ARNm de Cox2 **et** la progression depuis le tissu normal témoin, normal voisin des tumeurs, carcinomateux et… (Guastalla et al. 2004)
dans	prep.	dans X, Y [causal marker]	1	7621	10.94	**Dans** les cellules AT exposées aux rayonnements ionisants, l'induction de p53 est réduite et très retardée… (Angèle et al. 2001)
de susceptibilité à	prep. + n. + prep.	X de susceptibilité à Y	1	8	0.01	En raison de la concordance de plusieurs études en faveur de l'implication d'autre(s) gène(s) **de susceptibilité au** cancer du sein… (Bonadona and Lasset 2003)
élevé chez	adj. + prep.	X [copula] élevé chez Y	1	32	0.05	La prolifération lymphocytaire en réponse aux extraits protéiques des lésions a tendance à être plus **élevée chez** les patients atteints d'angor instable que chez les patients stables. (Caligiuri 2004)
fréquemment chez	adv. + prep.	X [verb] fréquemment chez [article] Y	1	2	0.003	Par ailleurs, le diabète est apparu moins **fréquemment chez** les patients qui étaient traités avec le losartan plutôt qu'avec l'aténolol. (Garnier 2002b)
fréquent dans	adj. + prep.	X [copula] fréquent dans [article] Y	1	26	0.037	L'inactivation des gènes codant pour p16 et p15 est très **fréquente dans** les lignées cellulaires cancéreuses et les tumeurs dont elles sont issues. (Blanchard 2003)

marqueur de	n. + prep.	X, marqueur de [article] Y	1	66	0.10	Les MP, **marqueurs de** l'activation cellulaire in vivo
présent dans	adj. + prep.	X [copula] présent dans [article/quantifier] Y	1	61	0.09	Sur 44 tumeurs du sein étudiées par immunoblot et immunohistochimie, une expression de Cox1 est **présente dans** 30 cas… (Guastalla et al. 2004)
rapport entre… et	n. + prep. + conj.	rapport entre [article] X et [association marker] [article] Y	1	11	0.02	… les **rapports entre** le traitement hormonal subtitutif [sic] de la ménopause **et** le risque de cancer du sein. (Serin and Escoute 1998)
relier à	v. + prep.	X a été relié à [article] Y	1	22	0.03	… la perfusion d'angiotensine II induit la formation d'anévrismes, qui a été **reliée à** l'activation des leucocytes circulants. (Michel 2004)
sans	prep.	sans X, [article] Y	1	872	1.25	**Sans** traitement, le taux de récidives spontanées est de 21 % dans la série du NSABP 04.
suivant	prep.	suivant [article] X, [article] Y	1	9	0.01	**Suivant** le type de chimiothérapie utilisé, un taux plus ou moins important d'aménorrhées va survenir dans le bras qui reçoit la chimiothérapie… (Namer and Ramaioli 2000)
sur	prep.	X [causal marker] sur [article] Y	1	3545	5.09	De plus, l'activation et l'agrégation des plaquettes sont inhibées **sur** une prothèse modifiée par rapport à une prothèse vierge… (Chevallier et al. 2003)
témoigner de	v. + prep.	X témoigne de [article] Y	1	21	0.03	… elle double pour une HbA1c à 9 %, valeur **témoignant d'**un diabète déséquilibré. (Fredenrich et al. 2004)
Mean			**2.33**	**1107.33**	**0.17**	

CAUSE–EFFECT: CREATION (54 markers, 25 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
conduire à	v. + prep.	X conduit à [article] Y	8	190	0.27	Cette oxydation **conduit à** la déplétion des LDL en antioxydants, en phosphatidylcholines et en esters de cholestérol… (Bonnefont-Rousselot et al. 2002)
entraîner	v.	X entraîne [article] Y X, entraînant [article] Y	7	326	0.47	Cette activation **entraîne** de nombreuses réponses cellulaires avec stimulation de la croissance et de la division cellulaire… (Penault-Llorca et al. 2002) Les ERO formées par la NADPH oxydase des cellules musculaires lisses sont également impliquées dans l'activation par la thrombine du facteur de transcription hypoxia-inducible factor-1 (HIF-1), **entraînant** l'expression de l'inhibiteur de l'activateur du plasminogène (PAI-1) et du vascular endothelial growth factor (VEGF)… (Bonnefont-Rousselot et al. 2002)
induire	v.	X induit [article] Y	7	232	0.33	… l'engagement de Fas **induit** la dénitrosylation de la caspase 3 et son activation. (Kolb 2001)
induit par	ppl.a. + prep.	X induit par [article] Y X (induit par Y)	7	92	0.13	Une hypothèse est que l'activation des récepteurs TP **induite par** les isoprostanes est responsable des effets indépendants des cyclooxygénases. (Cracowski 2004) Augmentation de la survie globale et sans récidive par la suppression ovarienne (**induite** ou non **par** chimiothérapie) et la prescription de tamoxifène (Debourdeau et al. 2004)

participer à	v. + prep.	X participe à [article] Y X, participant à [article] Y	6	63	0.09	Cette prolifération musculaire lisse **participe à** la constitution de la plaque athéroscléreuse… (Teiger 2001) … un signal de transduction capable d'activer directement la NADPH oxydase, **participant** en retour **à** l'oxydation des LDL. (Bonnefont-Rousselot et al. 2002)
résulter de	v. + prep.	X résulte de [article] Y il X résulterait [article] Y	6	54	0.08	Dans les conditions d'études in vitro et in vivo chez le rat, la formation d'adduits hépatiques **résulte de** l'activation des microsomes hépatiques. (Sasco 2000) … en altérant le statut réducteur de la cellule lié aux groupements thiols [36]. Il **en résulterait** une activation de NFjB, peut-être par l'intermédiaire de radicaux libres issus de l'homocystéine… (Bonnefont-Rousselot et al. 2002)
activer	v.	X active Y [causal marker] [article] X : activer [article] Y X [copula] [causal marker] activer [article] Y	5	115	0.20	En conséquence, la caténine ß n'est plus dégradée, diffuse dans le noyau, déplace Groucho et **active** la transcription sous le contrôle de LEF/Tcf. (Blanchard 2003) Ce dernier résultat illustre le double rôle que joue le complexe Cdk4-cycline D1: **activer** la transcription du gène cycline E, et limiter la quantité libre de l'inhibiteur p27Kip1… (Blanchard 2003) Les AINS sont capables d'**activer** la transcription de leur propre enzyme cible, notamment Cox2 (mais pas Cox1)… (Guastalla et al. 2004)
exprimer	v.	X ([affix])exprime ([article]) Y	5	243	0.35	La cellule transfectée produisant du NO endogène **exprimerait** Fas et produirait du FasL autotoxique. (Gauthier et al. 2004)

facteur de	n. + prep.	facteur de X ([conjunction]) Y X [causal marker] facteur de Y	4	533	0.92	À l'opposé, le **facteur de** transcription c-Jun, en se fixant sur le promoteur de son propre gène, contribue à amplifier sa production. (Blanchard 2003) … des protéines G comme H-ras ou K-ras; des kinases cytoplasmiques comme raf/mil, mos ou pim-1; des **facteurs de** transcription comme myc, jun, fos ou erbA (récepteur de l'hormone thyroïdienne T3). (Blanchard 2003) L'ensemble des données recueillies suggère que la protéine p8 agit en tant que **facteur de** transcription dans la voie conduisant à la tumorigenèse. (Vasseur and Iovanna 2003)
impliqué dans	ppl.a. + prep.	X [copula] impliqué dans [article] Y X impliqué dans [article] Y	4	151	0.22	Les ERO formées par la NADPH oxydase des cellules musculaires lisses sont également **impliquées dans** l'activation par la thrombine du facteur de transcription hypoxia-inducible factor-1 (HIF-1)… (Bonnefont-Rousselot et al. 2002) … un autre groupe de gènes surexprimés dans les lignées (figure 3, C) contenait essentiellement des gènes **impliqués dans** la prolifération cellulaire (cyclines, CDK, PCNA, tubulines...)… (Bertucci et al. 2002)
provoquer	v.	X provoque [article] Y X [causal marker] provoque [article] Y	4	116	0.17	La présence de bactéries et/ou de virus dans les cellules endothéliales et/ou musculaires lisses **provoquerait** d'abord une activation des cellules infectées avec pour conséquence une augmentation de l'expression de molécules d'adhérences… (Lizard and Gambert 2001) Le dimère ainsi formé se lie au PPRE et **provoque** l'activation de la transcription du gène cible.

						(Gervois and Fruchart 2003)
rôle	v. + art. + n. + prep.	X joue [article] rôle dans/lors de [article] Y	4	203	0.34	La NADPH oxydase **jouerait** donc **un rôle** majeur **lors des** premières étapes du processus athéromateux (oxydation des LDL, adhésion monocytaire, accumulation de cellules spumeuses). (Bonnefont-Rousselot et al. 2002)
	n. + ppl.a. + prep. + prep.	rôle joué par [article] X dans [article] Y				Des études de plus en plus nombreuses mettent en évidence le **rôle joué par** l'activation des NADPH oxydases **dans** des modèles expérimentaux de pathologies cardiovasculaires telles que l'hypertension… (Bonnefont-Rousselot et al. 2002)
	n. + prep. + prep.	rôle de [article] X dans [article] Y				Le **rôle des** estrogènes **dans** la prolifération des tumeurs mammaires hormonodépendantes a été montré depuis de nombreuses années [1]. (De Crémoux 2000)
stimuler	v.	X stimule [article] Y	4	85	0.12	NO **stimule** l'activation de caspases et l'apoptose dans les RAW 264. (Kolb 2001)
conséquence de	n. + prep.	X [copula] [article] conséquence de [article] Y	3	52	0.07	… les maladies métaboliques qui en découlent, c'està- dire [sic] le diabète, les dyslipidémies et l'hypertension artérielle, sont les **conséquences du** mode de vie adopté par les humains… (Essiambre 2003)
		conséquence de [article] X [copula] [article] Y				La première **conséquence** fonctionnelle majeure **de** l'activation des plaquettes est le changement de conformation des glycoprotéines GP IIb/IIIa présentes à leur surface… (Collet et al. 2004)
déclenche-ment de	n. + prep.	X [causal marker] [article] déclenchement de [article] Y	3	9	0.01	L'événement essentiel responsable du **déclenchement de** la coagulation après une lésion vasculaire est l'externalisation de molécules de facteur tissulaire à la surface de l'adventice… (Mallat and Tedgui 2004)
	n. + prep.	déclenchement de				

	+ prep.	[article] X par [article] Y				Lorsque la plaque est rompue, le **déclenchement de** la coagulation **par** les cellules inflammatoires aboutit à la thrombose… (Collet et al. 2004)
déclencher	v.	X déclenche [article] Y X doit être déclenché par [article] Y	3	53	0.08	L'oxydation exagérée des acides gras de ces lipoprotéines modifiées **déclenche** une réaction inflammatoire… (Ferrières 2004) Comme toute réponse immunitaire, la réponse anti-tumorale doit **être déclenchée par** des cellules présentatrices d'antigènes. (Catros-Quemener et al. 2003)
pour	prep.	X pour ([article]) Y	3	4806	6.91	… mastectomies subtotales **pour** tumeur maligne… (Lilliu et al. 2002) ou in situ (intracanalaire). (Martin 2003)
produire	v.	X produit [article] Y	3	100	0.14	La distribution de l'apoptose dans la plaque est hétérogène: elle est ainsi plus fréquente dans les régions riches en cellules **produisant** des cytokines pro-inflammatoires. (Mallat and Tedgui 2004)
responsable de	adj. + prep.	X [copula] responsable de [article] Y	3	147	0.21	L'activation des ostéoclastes est **responsable de** l'hyperrésorption osseuse et de la libération de facteurs de dégradation… (Tubiana-Hulin et al. 2001)
à l'origine de	prep. + n. + prep.	X [copula] à l'origine de [article] Y	2	51	0.07	L'athérosclérose est **à l'origine de** la plupart des maladies coronaires. (Ferrières 2004)
important	adj. + prep.	X important dans/pour [article] Y	2	109	0.16	… visent d'abord une restriction en lipides totaux ainsi qu'en graisses saturées et en cholestérol alimentaire, deux facteurs **importants dans** l'apparition de l'athérosclérose. (Blais_2001a) … les c-jun kinases (JNK), **importantes pour** la croissance et la prolifération cellulaire… (Bonnefont-Rousselot et al. 2002)
intervenir dans	v. + prep.	X intervient dans [article] Y	2	35	0.05	… de nombreuses autres cellules peuvent les synthétiser, en particulier d'autres cellules

						vasculaires **intervenant dans** la pathologie thrombotique, principalement les monocytes… (Drouet 2004)
médié par	ppl.a. + prep.	X [copula] médié par [article] Y	2	4	0.006	L'ensemble des données disponibles suggère que l'effet vasculaire de la 15- F2t-IsoP **est médié par** une activation du récepteur TP (récepteur commun à la prostaglandine H2 et au thromboxane)… (Cracowski 2004)
par	prep.	X par ([article]) Y	2	6116	8.79	… peut réduire de façon significative la mortalité **par** cancer du sein [6]. (Spyckerelle et al. 2002)
réponse à	n. + prep. prep. + n. + prep.	X [verb] [preposition] [article] réponse à [article] Y X en réponse à [article] Y	2	152	0.22	L'athérosclérose est considérée actuellement comme une **réponse** inflammatoire **aux** lésions de la paroi artérielle. (Duriez 2004) La prolifération lymphocytaire en **réponse aux** extraits protéiques des lésions a tendance à être plus élevée chez les patients… (Caligiuri 2004)
à cause de	prep. + n. + prep.	X à cause de [article] Y	1	46	0.07	… les femmes pensent que cette séquelle est due à un traitement plus important **à cause d'**une tumeur agressive. (Bobin et al. 2002)
activateur de	adj. + prep.	X [causal marker] activateur de Y	1	4	0.007	AF-1 (Activating Function-1), en position N terminale de la section A-B, qui exerce une fonction **activatrice de** la transcription… (Kirkiacharian 2000)
activateur de	n. + prep.	[verb] [preposition] X [article] activateur de [article] Y	1	13	0.02	C/EBPß est indispensable pour l'expression du gène p8, faisant du facteur de transcription C/EBPß un **activateur** majeur **de** la transcription du gène. (Vasseur and Iovanna 2003)
cause de	n. + prep.	X [copula] [article] cause de Y	1	254	0.37	Le diabète est en France la première **cause de** cécité, chez des malades souvent jeunes. (Blot et al. 1999)
causé par	ppl.a. + prep.	X causé par [article] Y	1	9	0.01	… la castration chimique **causée par** la chimiothérapie… (Dufresne 2003)

causer	v.	X cause [article] Y	1	33	0.05	Tout particulièrement, l'activation du récepteur AT1 de l'angiotensine, en plus de **causer** de la vasoconstriction, participe au développement... (Constance and Pranno 2002)
complica-tion de... dans	n. + prep. + prep.	complication de [article] X dans [article] Y	1	1	0.001	Les **complications de** l'ostéolyse maligne **dans** le cancer du sein engagent rarement le pronostic vital immédiat, mais sont source d'une morbidité importante.
découler de	v. + prep.	X Y découle	1	7	0.01	... le cumul de certains gènes prédisposants et, bien sûr, les maladies métaboliques qui **en découlent**, c'est-à-dire le diabète, les dyslipidémies et l'hypertension artérielle... (Essiambre 2003)
dérivé de	n. + prep.	X [copula] [article] dérivé de Y	1	20	0.03	Les espèces lipidiques oxydées responsables de ces effets sont essentiellement des **dérivés d'**oxydation des phospholipides tels que le POVPC. (Bonnefont-Rousselot et al. 2002)
dû à	adj. + prep.	X [copula] dû à [article] Y	1	38	0.05	Même très modéré, il fait craindre la rechute, car les femmes pensent que cette séquelle est **due à** un traitement plus important à cause d'une tumeur agressive. (Bobin et al. 2002)
du fait de	prep. + n. + prep.	X du fait de [article] Y	1	101	0.15	Elle conduit à la surcharge des monocytes en lipoprotéines **du fait de** l'activation de leur récepteur " scavenger ". (Boisseau 2004)
engager... vers	v. + prep.	X engage Y1 vers [article] Y2	1	2	0.003	... des effecteurs impliqués dans la mort cellulaire peuvent entrainer soit l'apoptose, soit **engager** la cellule **vers** la prolifération ou la différenciation. (Kolb 2001)
engendrer	v.	X [causal marker] engendrer [article] Y	1	39	0.06	Cette activation directe permet d'**engendrer** une réponse cellulaire cytotoxique protectrice. (Catros-Quemener et al. 2003)
expression par... de	n. + prep. + prep.	expression par [article] X de [article] Y	1	13	0.02	Le premier concerne l'**expression par** des cellules tumorales **du** ligand de la molécule Fas. (Sasco 2000)

forcer	v.	X, forçant [article] Y	1	3	0.005	Mois après mois après mois d'assauts oestrogéniques inondant les récepteurs cellulaires, **forçant** la transcription et finalement l'erreur. (Bouchard 2001)
implication dans	n. + prep.	X implication dans [article] Y	1	7	0.01	Par ailleurs, leur **implication dans** l'athérosclérose mériterait d'être approfondie.
impliquer… dans	v. + prep.	[(pro)noun] implique X dans [article] Y	1	8	0.01	… ce qui **implique** Cox2 **dans** la prolifération tumorale. (Guastalla et al. 2004)
induction de… par	n. + prep. + prep.	induction de X par [article] Y	1	16	0.02	L'**induction de** tumeurs bénignes ou malignes ovariennes **par** une stimulation continue des ovaires est une hypothèse qui a déjà été soulevée... (Sasco et al. 1997)
initiateur de	n. + prep.	X [copula] [article] initiateur de [article] Y	1	7	0.01	Le facteur tissulaire est l'**initiateur de** la coagulation. (Duriez 2004)
intervention de… dans	n. + prep.	intervention de [article] X dans [article] Y	1	3	0.004	Les débats sur l'**intervention** possible **du** cytomégalovirus **dans** l'athérosclérose commencent à peine à s'allumer que déjà d'autres microbes candidats pointent à l'horizon. (L'Allier 2003)
médiateur de	n. + prep.	médiateur de [article] X [preposition] [article] Y	1	11	0.02	En recherchant des **médiateurs de** l'activation du gène p8 parmi les cytokines et les facteurs de croissance, nous avons observé que le TGFß induit… (Vasseur and Iovanna 2003)
production de… par	n. + prep. + prep.	production de X par [article] Y	1	20	0.03	De cette interaction résulte une **production d'**IL-1 et de prostaglandines **par** les cellules endothéliales, source d'entretien d'une réaction inflammatoire… (Meyer 2001)
produit de	n. + prep.	X, produit de Y	1	57	0.08	À des degrès [sic] divers l'estrone 2, **produit d'**oxydation de l'hydroxyle en 17 de l'estradiol et son dérivé 17-a-éthinylé possèdent aussi les propriétes [sic] biologiques de… (Kirkiacharian 2000)
produit par	ppl.a. + prep.	X produit par Y	1	41	0.06	Dans ce modèle, c'est le NO **produit par** les cellules tumorales elles-mêmes, et non par les cellules stromales, qui a une activité antitumorale.

						(Gauthier et al. 2004)
réalisé par	ppl.a. + prep.	X réalisé par [article] Y	1	53	0.08	… les économies **réalisées par** le traitement sur le coût des complications évitées… (Launois et al. 1997)
réaliser	v.	X réalise [article] Y	1	603	0.87	L'adénose microglandulaire (AMG) **réalise** une prolifération glandulaire bénigne qui peut être prise pour un adénocarcinome… (Charafe-Jauffret et al. 2001)
résultat de	n. + prep.	résultat de X [copula] [article] Y	1	291	0.50	… le **résultat de** cette cascade d'activations est la transcription du gène de Cox2, une forte concentration de son ARNm, de sa protéine Cox2 et de PGE2. (Guastalla et al. 2004)
secondaire à	adj. + prep.	X secondaire à [article] Y	1	26	0.04	De façon analogue, les inhibiteurs du système rénine-angiotensine diminuent la prolifération intimale des cellules musculaires lisses **secondaire à** une angioplastie. (Michel 2004)
synthétiser	v.	X Y synthétise	1	15	0.02	… de nombreuses autres cellules peuvent les **synthétiser**, en particulier d'autres cellules vasculaires intervenant dans la pathologie thrombotique… (Drouet 2004)
Mean			**2.41**	**292.19**	**0.17**	

CAUSE–EFFECT: DESTRUCTION (7 markers, 2 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
anti-	affix	X antiY	2	1223	1.76	… nécessité d'une radiothérapie antalgique ou recours à un nouveau traitement **anti**tumoral en raison d'une progression osseuse. (Tubiana-Hulin et al. 2001)

destruction de	n. + prep.	X [causal marker] [article] destruction de [article] Y	2	9	0.01	... plusieurs protéines qui sont impliquées dans des processus qui conduisent à la **destruction de** la cellule. (Chène 1999)
arrêt de	n. + prep.	X [causal marker] [article] arrêt de [article] Y	1	84	0.12	... une toxicité non négligeable ayant conduit à l'**arrêt du** traitement dans trois cas (2 pour somnolence et 1 pour anxiété) et à la diminution des doses dans un cas [31]. (Debourdeau et al. 2004)
curatif	adj.	X [verb] [article] [causal marker] curatif [preposition] Y	1	12	0.02	Des essais précliniques montrent que les cellules dendritiques présentent un pouvoir **curatif** et préventif à l'égard de tumeurs greffées. (Catros-Quemener et al. 2003)
détruire	v.	X [causal marker] détruisant [article] Y	1	5	0.007	La chimiothérapie agit **en détruisant** les cellules en multiplication (cellules tumorales potentiellement en circulation). (Martin 2003)
élimination de	n. + prep.	X [copula] [causal marker] [article] élimination de [article] Y	1	17	0.02	Les protéines mutées sont incapables de provoquer l'**élimination des** cellules ayant, par exemple, un ADN endommagé par les UV. (Chène 1999)
stopper	v.	X stoppe [article] Y	1	3	0.004	In vitro, les anticorps se fixent sur ce récepteur, le bloquent et **stoppent** la prolifération cancéreuse. (La Recherche 2002)
Mean			**1.29**	**193.29**	**0.13**	

CAUSE–EFFECT: MAINTENANCE/PERMISSION (10 markers, 3 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
permettre	v.	X permet ([article]) Y	5	911	1.31	Par ailleurs, l'exercice physique **permet** l'oxydation mitochondriale des acides gras au

		X, permettant [article] Y				niveau des muscles… (Ferrières 2004) Ainsi, les éléments de base nécessaires au développement et à l'évolution de l'artériosclérose seront en place, **permettant** l'oxydation des particules de LDL et leur incorporation dans les macrophages (monocytes)… (Essiambre 2003)
nécessaire à	adj. + prep.	X nécessaire à [article] Y X [copula] nécessaire à [article] Y	4	53	0.08	… est en effet capable de stimuler le recrutement et l'assemblage des sous-unités p47phox et p67phox, étape **nécessaire à** l'activation de la NADPH oxydase. (Bonnefont-Rousselot et al. 2002) … l'activation du protéasome est, au contraire, **nécessaire à** l'accomplissement du processus apoptotique… (Kolb 2001)
dépendant	adj. adj. + prep.	X Y dépendant X [copula] dépendant de [article] Y X dépendant de Y	3	133	0.21	Inhibition de la transcription REα **dépendante** de gènes de la prolifération par BRCA1 (Pujol et al. 2004) …à l'inverse de celle-ci, il est **dépendant de** l'activation des caspases. (Kolb 2001) Dans des lignées de cancer du sein et de la prostate, BRCA1 inhibe la transcription **dépendante du** REα de gènes impliqués dans la prolifération cellulaire. (Pujol et al. 2004)
assurer	v.	X pour assurer [article] Y	1	89	0.13	… le récepteur ET-A devait être activé par ET-1 au stade embryonnaire X0 pour **assurer** une prolifération normale des cellules de la crête neurale. (Pinet 2004)
crucial dans	adj. + prep.	X cruciale dans [article] Y	1	5	0.007	L'extravasation des lipoprotéines de basse densité (LDL) et leur oxydation dans l'espace sous-endothélial pourraient constituer l'étape

						cruciale dans la formation de la plaque. (Caligiuri 2004)
essentiel dans	adj. + prep.	[noun] essentiel de [article] X dans [article] Y	1	16	0.02	Le caractère **essentiel** de la PS **dans** la coagulation est illustré par le syndrome de Scott. (Martin 2003)
nécessiter	v.	X nécessite [article] Y	1	142	0.20	L'expression du gène cycline E est alors directement sous la dépendance des signaux extrinsèques, et ne **nécessite** plus une activation préalable de la cycline D1. (Blanchard 2003)
passer par	v. + prep.	X passe par [article] Y	1	25	0.04	Cette protection **passe** vraisemblablement **par** une activation de PKG, protéines kinases dépendantes du GMPc… (Ferrières 2004)
sous la dépendance de	prep. + article + n. + prep.	X [copula] sous la dépendance de [article] Y	1	8	0.01	L'assemblage et l'activation d'un tel complexe sont **sous la dépendance**, d'une façon qui n'est pas encore claire, **des** kinases Cdk2-cycline E… (Blanchard 2003)
valeur… de	n. + prep.	valeur X de [article/quantifier] Y	1	25	0.04	Néanmoins, la définition classique du syndrome de Li-Fraumeni ne tient pas compte de la **valeur** diagnostique **de** certaines tumeurs exceptionnelles dans la population générale, comme les corticosurrénalomes… (Frebourg et al. 2001)
Mean			**1.90**	**140.70**	**0.10**	

CAUSE–EFFECT: PREVENTION (11 markers, 4 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
suppresseur de	adj. + prep.	X suppresseur de Y	4	14	0.02	Avec les gènes RB p53, WTA ou APC, est apparue une première génération de gènes

		[causal marker] suppresseur de X [preposition] Y				**suppresseurs de** tumeurs. (Bénard 1997) … cette observation ouvre plusieurs perspectives très intéressantes sur le rôle **suppresseur de** tumeur de Ptc1. (Blanchard 2003)
prévention	prep. + n. n. + prep.	X de prévention [verb] [article] Y X [verb] [preposition] [article] prévention de [article] Y [causal marker] X [copula] [article] prévention de [article] Y	3	328	0.47	Les stratégies hormonales **de prévention** pourraient ainsi concerner à la fois les tumeurs sporadiques et les tumeurs génétiques. Le THS n'est recommandé qu'en cas d'intolérance à un autre traitement indiqué dans la **prévention de** l'ostéoporose et après une évaluation individuelle précise et soigneuse… (Rozenbaum 2004) Le but du traitement du syndrome métabolique reste la **prévention du** diabète et des maladies cardiovasculaires. (Gonzalez and Palardy 2004)
bloquer	v.	X bloque [article] Y	2	71	0.10	Le tamoxifène **bloque** la prolifération cellulaire qui est rétablie par l'addition d'estrogènes. (Vinatier and Orazi 2003)
préventif	adj.	X [verb] [article] [causal marker] préventif [preposition] Y X Y, préventif	2	62	0.09	Des essais précliniques montrent que les cellules dendritiques présentent un pouvoir curatif et **préventif** à l'égard de tumeurs greffées. (Catros-Quemener et al. 2003) … la résistance biologique ainsi définie a une signification clinique justifiant sa recherche et son traitement, au moins **préventif**.
blocage de	n. + prep.	X ([causal marker] [article] blocage de [article] Y)	1	18	0.03	Cependant, une inactivation de BRCA1 par méthylation de la région promotrice (aboutissant au **blocage de** la transcription de BRCA1) a été décrite dans des cancers sporadiques. (Pujol et al. 2004)

bloquant	adj.	X [causal marker] bloquant [preposition] [article] Y	1	22	0.03	L'inhibition par NO du protéasome pourrait donc rendre compte de son effet **bloquant** sur l'activation de NF-jB. (Kolb 2001)
empêcher	v.	X [verb phrase] en empêchant [article] Y	1	51	0.07	Il inhibe l'induction de Cox2 par le phorbol-ester (PMA) dans les celules mammaires humaines **en empêchant** l'activation de la protéine kinase C par PMA et l'activation du promoteur de Cox2 par c-Jun. (Guastalla et al. 2004)
garantir contre	v. + prep.	X garantit contre [article] Y	1	1	0.001	Bien qu'une RCH ne **garantit** pas définitivement **contre** une récidive, sa valeur puissante pronostique est confirmée dans de nombreuses analyses multifactorielles. (Brain 2000)
protecteur contre	adj. + prep.	X [verb] [article] [causal marker] protecteur contre [article] Y	1	7	0.01	Les estrogènes ont un effet **protecteur contre** l'athérosclérose et l'hypertension artérielle [19] découlant de leur action sur le métabolisme des lipides... (Kirkiacharian 2000)
suppresseur de	n. + prep.	suppresseur de X Y	1	18	0.03	Interaction entre le **suppresseur de** tumeur Ptc et la cycline B1. (Blanchard 2003)
verrouiller	v.	X vérouille [article] Y	1	1	0.002	... un autre type d'altération, telle une hyperméthylation **verrouillant** la transcription du gène... (Frebourg et al. 2001)
Mean			**1.64**	**53.91**	**0.07**	

CAUSE–EFFECT: MODIFICATION (32 markers, 9 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
effet	n.	X [verb] [article] effet Y	7	1578	2.27	... de nombreux traitements ont des **effets** rhéologiques... (Boisseau 2004)

	n. + prep. n. + prep. + prep.	X [verb] [article] effet sur [article] Y effet de/[preposition] [article] X sur [article] Y				Dans ce travail [12], l'exercice physique n'a pas eu d'**effet sur** le cholestérol total ou le LDL cholestérol. (Ferrières 2004) … il est nécessaire de connaître les **effets du** NO **sur** les cellules cancéreuses, les tumeurs et l'hôte.
régulation (de/entre… et)	n. + prep. n. + prep. + conj.	X [causal marker] [article] régulation de [article] Y régulation entre X et Y	4	74	0.11	Ce complexe migre vers le noyau, où il intervient dans la **régulation de** la transcription… (Vasseur and Iovanna 2003) Sachant qu'il existe une **régulation** étroite **entre** apoptose **et** prolifération cellulaire, un point important à considérer très rapidement dans le développement clinique des inhibiteurs… (Lavelle and Jehanno 1998)
moduler	v.	X module [article] Y	3	58	0.08	Les molécules qui **modulent** sélectivement l'activation des récepteurs hormonaux (SERM)… (Vinatier and Orazi 2003)
anti-	affix	X antiY	2	1223	1.76	… une hormonothérapie (anti-aromatase) ou une chimiothérapie **anti**tubuline… (Guastalla et al. 2004)
commander	v.	X commande [article] Y	2	8	0.01	L'activation de ces récepteurs **commande** la transcription des gènes insulinosensibles… (Leblond 2001)
complica-tion … de	n. + prep. + prep.	complication X de [article] Y	2	30	0.04	Les interactions entre système rénine-angiotensine et **complications** vasculaires **du** diabète constituent un autre exemple de l'implication du TGF-ß. (Michel 2004)
contrôler	v.	X est contrôlé par [article] Y	2	114	0.16	La prolifération des cellules cancéreuses mammaires **est contrôlée par** les oestrogènes et les facteurs de croissance… (Chailleux et al. 2000)

nuire à	v. + prep.	X nuit à [article] Y	2	8	0.01	Plus besoin non plus du coeur-poumon artificiel, qui dégrade le sang et **nuit à** sa coagulation. (Simard and Dussault 1997)
agir sur	v. + prep.	X, agissant sur [article] Y	1	29	0.04	… facteurs de croissance, **agissant sur** la prolifération des cellules musculaires lisses… (Caligiuri 2004)
altération de	n. + prep.	X [causal marker] [article] altération de [article] Y	1	64	0.09	Cette oxydation serait susceptible d'entraîner l'**altération de** diverses structures nerveuses. (La Recherche 1997)
altérer	v.	X est altéré [preposition] [article] Y	1	57	0.08	La vitesse de l'onde de pouls **est** significativement **altérée** au cours du vieillissement, de l'hypertension artérielle, du diabète et de l'athérosclérose. (Levenson et al. 2000)
amélioration de… par	n. + prep.	amélioration de [article] X par [article] Y	1	97	0.14	… on peut donc proposer avec un objectif de 5 % d'**amélioration de** la SSR **par** la chimiothérapie… (Bachelot et al. 2002)
améliorer	v.	X améliore [article] Y	1	154	0.22	… l'utilisation de trastuzumab, en association avec la chimiothérapie, **améliore** le taux de survie globale. (Penault-Llorca et al. 2002)
contrôle	n. + prep.	X [verb] [noun] de [article] contrôle de [article] Y	1	303	0.44	Proto-oncogènes et gènes suppresseurs de tumeurs ont ainsi constitué le yin et le yang du **contrôle de** la prolifération. (Blanchard 2003)
dans	prep.	[causal marker] [verb] [preposition] [article] X dans [article] Y	1	7887	11.34	Les résultats obtenus avec le paclitaxel en monothérapie **dans** le cancer du sein métastatique ont tout naturellement conduit à associer ce médicament aux anthracyclines… (Ferrero et al. 2003)
dépendre de	v. + prep.	X dépend de [article] Y	1	107	0.15	… l'effet ambivalent du NO sur la croissance tumorale **dépend de** la cellule productrice et de la quantité de NO produite. (Gauthier et al. 2004)
déterminer	v.	X détermine [article] Y	1	109	0.19	… ce qui **détermine** la transcription de Cox2 via l'élément de réponse AMP cyclique-dépendant… (Guastalla et al. 2004)

gouverner	v.	X gouverne [article] Y	1	4	0.001	… l'oncostatine M module le système fibrinolytique qui **gouverne** l'évolution des thrombus intravasculaires et l'activation des MMP… (Drouet and Bal Dit Sollier 2002)
impact de… sur	n. + prep. + prep.	impact de X sur [article] Y	1	57	0.08	Valantine [26] a évalué l'**impact** du ganciclovir administré en prophylaxie immédiate après transplantation cardiaque **sur** l'athérosclérose du transplant… (Chidiac and Braun 2002)
influencer	v.	X influence [article] Y	1	73	0.10	Après analyse multifactorielle, seul le traitement chirurgical de la rechute **a influencé** le contrôle local après le traitement de la rechute locale isolée. (Deniaud-Alexandre 2004)
influer sur	v. + prep.	X [causal marker] en influant sur [article] Y	1	7	0.01	Nous verrons plus loin que NO peut également exercer un effet indirect sur la perméabilité mitochondriale **en influant sur** la transcription ou la dégradation des molécules des familles bcl-2 et bax… (Kolb 2001)
maîtrise de	n. + prep.	X [causal marker] [article] maîtrise de [article] Y	1	18	0.02	… lorsque l'association de metformine et d'une sulfonylurée ne permet pas une **maîtrise** optimale **du** diabète ou ne peut être utilisée en raison d'une contre-indication ou de l'intolérance à l'un de ces médicaments… (Leblond 2001)
maîtrisé par	ppl.a. + prep.	X [copula] maîtrisé par [article] Y	1	1	0.001	… qui est atteint d'un "petit diabète" qui semble bien **maîtrisé par** le régime alimentaire. (Gendreau 2003)
modification par	n. + prep.	X modification par Y	1	2	0.003	… leur rétention dans la matrice sousendothéliale et leur **modification par** oxydation radicalaire ou fixation aux glycosaminoglycanes de la matrice extracellulaire. (Meyer 2001)
modifier	v.	X modifie [article] Y	1	180	0.26	Les rétinoïdes règlent la croissance cellulaire …, **modifient** la prolifération …, inhibent l'ornithine décarboxylase …, facilitent la différentiation et l'apoptose. (Vinatier and Orazi 2003)

modulateur de	n. + prep.	X [copula] [article] modulateur de [article] Y	1	6	0.009	Comme NO est par ailleurs un **modulateur de** l'activation des caspases, ceci fournit un moyen, via la PARP de contrôler finement les processus... (Kolb 2001)
régler	v.	X règle [article] Y	1	12	0.02	... la phosphorylation des protéines Smad qui, en migrant dans le noyau, **règlent** la transcription des gènes cibles du TGFß, dont p8. (Vasseur and Iovanna 2003)
régulateur de	n. + prep.	régulateur de [article] X [conjunction] [article] Y	1	7	0.01	... une incidence accrue de cancers avec un dérèglement de l'expression de plusieurs **régulateurs du** cycle cellulaire et de la prolifération tels que les cyclines D1 et A et les protéines Mdm2 et c-myc. (Blanchard 2003)
réguler	v.	X régule [article] Y	1	20	0.03	... la p53 qui **régule** la transcription de diverses molécules impliquées dans l'apoptose (bax, inhibiteurs de kinases dépendantes de cyclines)... (Kolb 2001)
répercussion de... dans	n. + prep. + prep.	répercussion de [article] X dans [article] Y	1	2	0.003	L'idée initiale restait ici d'approcher les **répercussions** psychosociales **du** cancer du sein **dans** un groupe particulier de la population française, les femmes immigrées d'origine maghrébine... (Nguyen et al. 2002)
réponse à	n. + prep.	réponse X à [article] Y	1	152	0.22	Cependant, la **réponse** osseuse **au** traitement reste toujours difficile à évaluer de par la faible spécificité de la scintigraphie osseuse... (Leriche and Bonneterre 1997)
retentisse-ment de... sur	n. + prep. + prep.	retentissement de [article] X sur Y	1	2	0.003	... le **retentissement** métabolique moindre **des** associations " à la française ", en particulier **sur** la coagulation... (Azoulay 2004)
Mean			**1.50**	**388.84**	**0.13**	

CAUSE–EFFECT: INCREASE (10 markers, 5 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
favoriser	v.	X favorise [article] Y X, favorisant [article] Y	7	188	0.27	…l'expression de Cox2 **favorise** la prolifération tumorale en inhibant l'apoptose… (Guastalla et al. 2004) …une activation de NFjB, peut-être par l'intermédiaire de radicaux libres issus de l'homocystéine, **favorisant** la prolifération des cellules musculaires lisses vasculaires. (Bonnefont-Rousselot et al. 2002)
augmenta-tion de	n. + prep.	X [causal marker] [article] augmentation de [article] Y augmentation de [article] X, [causal marker] [article] Y	4	497	0.71	L'activation de récepteurs endothéliaux produit une **augmentation de** [Ca]i dans les cellules endothéliales… (Feletou et al. 2003) Les hyperlipoprotéinémies familiales primitives regroupent des maladies héréditaires avec une **augmentation** plasmatique **du** cholestérol et des triglycérides, résultant d'un déficit dans l'une des étapes du métabolisme lipidique… (Chalès and Guggenbuhl 2004)
augmenter	v.	X augmente Y augmenter X, [causal	3	325	0.47	En revanche, les deux études randomisées les plus récentes ont désormais démontré qu'une chimiothérapie d'induction peut **augmenter** les possibilités de chirurgie sans diminuer significativement les taux de survie… (Lerouge et al. 2004) Le but essentiel de la radiochimiothérapie

		marker] Y				concomitante est d'**augmenter** l'activité cytotoxique de la radiothérapie, et donc le contrôle local, grâce à une chimiothérapie radiosensibilisante... (Serin_1997)
accroître	v.	X accroît [article] Y X [causal marker] accroître Y	2	92	0.13	Le traitement hormonal substitutif **accroît** l'incidence du cancer du sein. (Noël et al. 1998) Les monocytes sont alors activés en macrophages (Ma) ce qui contribue probablement à **accroître** l'oxydation des LDL. (Arnal et al. 2003)
faciliter	v.	X facilite [article] Y	2	53	0.08	Dans le cas des tumeurs, l'expression de p8 **faciliterait** la transcription de gènes indispensables à la progression tumorale. (Vasseur and Iovanna 2003)
accru	ppl.a.	X [causal marker] [article] Y accru	1	86	0.15	Lorsqu'il est activé, il induit une cascade de phosphorylations intracellulaires, conduisant à une transcription de protéines et à une croissance cellulaire **accrues**. (Cornez and Piccart 2002)
augmenté	ppl.a.	X [causal marker] [article] Y augmenté	1	117	0.17	… une athérosclérose prématurée responsable d'une mortalité coronarienne et neurovasculaire **augmentée**… (Meyer 2001)
catalyser	v.	X [verb phrase] en catalysant [article] Y	1	18	0.03	Les macrophages ont une fonction d'épuration des lipides de l'intima mais initient un cercle vicieux en **catalysant** l'oxydation des lipides de la plaque et en augmentant la perméabilité endothéliale aux lipides circulants. (Teiger 2001)
pro-	affix	[causal marker] pro-X de [article] Y	1	121	0.17	Dans ce cas également, le potentiel **pro**-coagulant des cellules athéroscléreuses est lié au nombre élevé de cellules en apoptose… (Mallat and Tedgui 2004)

promoteur	adj.	[causal marker…] promoteur […causal marker…] X […causal marker] Y	1	6	0.01	… si on admet la possibilité d'un effet **promoteur** du THS sur certains cancers du sein, il est difficilement concevable que cet effet cesse dès la 1re année suivant l'arrêt de celui-ci… (Rozenbaum 2004)
Mean			**2.30**	**150.30**	**0.22**	

CAUSE–EFFECT: DECREASE (12 markers, 5 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
inhiber	v.	X inhibe [article] Y X peut être inhibé par [article] Y	7	173	0.25	… si l'on bloque la synthèse de NO dans ces cellules ou si l'on **inhibe** l'activation de la guanylate cyclase par NO, on induit l'apoptose… (Kolb 2001) À droite : l'activation de la protéine G Arf par ses GEF à domaine Sec7 peut **être inhibée par** stabilisation de complexes abortifs Arf-GD… (Cherfils and Pacaud 2004)
diminuer	v.	X diminue Y X [causal marker] diminuer Y	3	310	0.45	De façon analogue, les inhibiteurs du système rénine-angiotensine **diminuent** la prolifération intimale des cellules musculaires lisses secondaire à une angioplastie. (Michel 2004) La chimiothérapie et l'hormonothérapie sont des traitements systémiques qui ont pour but de **diminuer** la récidive, surtout systémique. (Martin 2003)

réduire	v.	X réduit Y X [causal marker] réduire Y	3	291	0.42	Ces médicaments non seulement **réduisent** le cholestérol plasmatique et ses dérivés, mais aussi ont des effets " pléïotropes "… (Asmar et al. 2003) Cette stratégie à [sic] pour effet de **réduire** le cholestérol total et le cholestérol LDL (ou mauvais cholestérol) de l'ordre de 10 à 20 % et de 12 à 16 % respectivement. (Blais 2001a)
inhibiteur	adj. + prep.	X inhibiteur de [article] Y [causal marker…] inhibiteur de X [...causal marker] [article] Y	2	249	0.36	L'activité paraoxanase 1 **inhibitrice de** l'oxydation des LDL est très diminuée chez les patients ayant des antiphospholipides. (Meyer 2001) …ce qui contribuerait à l'effet **inhibiteur** de NO sur la prolifération cellulaire. (Kolb 2001)
inhibition de (… par)	n. + prep.	X [causal marker] [article] inhibition de [article] Y inhibition de X par Y	2	92	0.13	… le récepteur AT2, dont l'activation oppose les effets de l'activation du récepteur AT1 et cause donc, entre autres, une **inhibition de** la croissance et de la prolifération cellulaire… (Constance and Pranno 2002) **Inhibition de** la transcription REα dépendante de gènes de la prolifération **par** BRCA (Pujol et al. 2004)
baisse de	n. + prep.	X [causal marker] [article] baisse de Y	1	87	0.12	Un traitement de 3 mois par la pravastatine entraîne une **baisse du** contenu en lipides (et de leur oxydation) des plaques carotidiennes humaines… (Nalbone et al. 2002)
contre	prep.	X contre [article] Y	1	564	0.81	Il y a quelques mois, les médias du monde entier chantaient les vertus d'un "médicament-miracle" **contre** le cancer du sein. (Simard 1998)
contrer	v.	X pour contrer [article] Y	1	1	0.001	… n'a mis en évidence aucun bénéfice attribuable à l'HTR pour **contrer** l'athérosclérose….

						(Bouchard 2001)
diminution de... par	n. + prep. + prep.	diminution de [article] X par [article] Y	1	9	0.02	... une **diminution de** la prolifération de l'épithélium mammaire normal **par** une carence estrogénique réduirait l'incorporation d'une mutation... (Vinatier and Orazi 2003)
freiner	v.	X pour freiner [article] Y	1	11	0.02	... ce qui aidera le mélange gène-cholestérol à se coller assez longtemps sur la paroi pour **freiner** la prolifération des cellules... (Simard and Dussault 1997)
opposer	v.	X oppose [article] Y	1	58	0.08	... le récepteur AT2, dont l'activation **oppose** les effets de l'activation du récepteur AT1... (Constance and Pranno 2002)
réprimer	v.	X réprime [article] Y	1	7	0.01	Par contre, p53 **réprime** la transcription de gènes anti-apoptotiques comme bcl-2 et comme la NOSi elle-même. (Kolb 2001)
Mean			**2.0**	**154.33**	**0.21**	

CAUSE–EFFECT: PRESERVATION (1 marker, none with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Total occur-rences in corpus	Occur-rences per 1,000 corpus tokens	Sample context
limiter	v.	X limite [article] Y	1	94	0.16	... ces substances stimulaient les défenses immunitaires de l'organisme qui, à leur tour, **limitaient** la prolifération cancéreuse. (Catros-Quemener et al. 2003)
Mean			**1.0**	**94.00**	**0.16**	

Appendix I: Part of speech classes of pattern markers observed

Table 131. Parts of speech of English markers

POS	ASSOCIATION	CAUSE	CREATION	DESTRUCTION	MAINTENANCE / PERMISSION	PREVENTION	MODIFICATION	INCREASE	DECREASE	PRESERVATION	Total
adj.	2	9	5	0	4	0	0	0	0	0	11
adv.	0	0	0	0	0	0	0	0	0	0	0
affix	0	2	0	1	0	0	0	1	0	0	2
conj.	1	0	0	0	0	0	0	0	0	0	1
n.	12	36	16	1	0	4	11	1	3	0	48
ppl.a.	2	15	6	0	1	0	1	3	4	0	17
prep.	2	6	4	1	0	0	1	0	0	0	8
v.	14	53	20	2	6	2	7	9	6	1	67
Total	33	121	51	5	11	6	20	14	13	1	154

Table 132. Parts of speech of English marker occurrences

POS	ASSOCIATION	CAUSE	CREATION	DESTRUCTION	MAINTENANCE / PERMISSION	PREVENTION	MODIFICATION	INCREASE	DECREASE	PRESERVATION	Total
adj.	4	17	13	0	4	0	0	0	0	0	21
adv.	0	0	0	0	0	0	0	0	0	0	0
affix	0	4	0	3	0	0	0	1	0	0	4
conj.	5	0	0	0	0	0	0	0	0	0	5
n.	59	100	55	1	0	14	24	1	5	0	159
ppl.a.	21	36	23	0	2	0	2	4	5	0	57
prep.	13	8	5	2	0	0	1	0	0	0	21
v.	23	152	71	2	6	6	19	30	17	1	175
Total	125	317	167	8	12	20	46	36	27	1	442

Table 133. Parts of speech of French markers

POS	ASSOCI-ATION	CAUSE	CREATION	DESTRUC-TION	MAINTENANCE / PERMISSION	PREVEN-TION	MODIFICA-TION	INCREASE	DECREASE	PRESERVA-TION	Total
adj.	3	16	5	1	4	4	0	1	1	0	19
adv.	1	0	0	0	0	0	0	0	0	0	1
affix	0	3	0	1	0	0	1	1	0	0	3
conj.	1	0	0	0	0	0	0	0	0	0	1
n.	8	46	20	3	2	3	14	1	3	0	54
ppl.a.	6	9	6	0	0	0	1	2	0	0	15
prep.	7	5	3	0	0	0	1	0	1	0	12
v.	4	58	20	2	4	4	15	5	7	1	62
Total	30	137	54	7	10	11	32	10	12	1	167

Table 134. Parts of speech of French marker occurrences

POS	ASSOCI-ATION	CAUSE	CREATION	DESTRUC-TION	MAINTENANCE / PERMISSION	PREVEN-TION	MODIFICA-TION	INCREASE	DECREASE	PRESERVA-TION	Total
adj.	3	34	8	1	12	9	1	1	2	0	37
adv.	1	0	0	0	0	0	0	0	0	0	1
affix	0	5	0	2	0	0	2	1	0	0	5
conj.	10	0	0	0	0	0	0	0	0	0	10
n.	21	75	33	4	1	4	25	4	4	0	96
ppl.a.	23	20	17	0	0	0	1	2	0	0	43
prep.	8	9	7	0	0	0	1	0	1	0	17
v.	4	136	68	2	8	5	20	15	17	1	140
Total	70	279	133	9	21	18	50	23	24	1	349

Appendix J: Analysis of pattern variation

Summary

	English	French	Difference
Total markers (occurrences ≥ 2)	70	65	5
Total occurrences	360	240	120
Total marker forms	106	85	21
Total pattern structures	134	112	22
ASSOCIATION markers (occurrences ≥ 2)	18	13	5
ASSOCIATION occurrences	109	53	56
ASSOCIATION marker forms	32	17	15
ASSOCIATION pattern structures	43	25	18
CAUSE–EFFECT markers (occurrences ≥ 2)	52	52	0
CAUSE–EFFECT occurrences	251	187	64
CAUSE–EFFECT marker forms	74	68	6
CAUSE–EFFECT pattern structures	91	87	4
Marker occ. to form ratio	3.4	2.8	0.6
ASSOCIATION marker occ. to form ratio	3.4	3.1	0.3
CAUSE–EFFECT marker occ. to form ratio	3.4	2.8	0.6
Pattern occ. to structure ratio	2.7	2.1	0.6
ASSOCIATION pattern occ. to structure ratio	2.5	2.1	0.4
CAUSE–EFFECT pattern occ. to structure ratio	2.8	2.1	0.7
Marker form to marker ratio	1.5	1.3	0.2
ASSOCIATION marker form to marker ratio	1.8	1.3	0.5
CAUSE–EFFECT marker form to marker ratio	1.4	1.3	0.1
Pattern structure to marker ratio	1.9	1.7	0.2
ASSOCIATION pattern structure to marker ratio	2.4	1.9	0.5
CAUSE–EFFECT pattern structure to marker ratio	1.8	1.7	0.1
Mean marker forms per occ.	0.41	0.43	-0.02
Mean ASSOCIATION marker forms per occ.	0.38	0.43	-0.05
Mean CAUSE–EFFECT marker forms per occ.	0.42	0.43	-0.01
Mean pattern structures per occ.	0.50	0.54	-0.04
Mean ASSOCIATION pattern structures per occ.	0.48	0.56	-0.08
Mean CAUSE–EFFECT pattern structures per occ.	0.51	0.53	-0.02

English

Association (33 markers, 18 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms[206]	Occur-rences in sample	Marker forms[207]			Pattern forms[208]	
				Total	Details	Forms per occ.	Total	Forms per occ.
associated	ppl.a. ppl.a. + prep.	X [conjunction] associated Y X [copula] associated with ([causal marker]) Y X associated with Y	17	2	associated, associated with	0.12	4	0.24
risk	n. n. + prep. n. + prep.	[causal marker] X [preposition] [causal marker]Y risk X ([causal marker]) risk for Y ([causal marker] [preposition] [article]) risk of X [association	14	4	risk, risk for, risk of, risk in relation to	0.29	5	0.36

[206] The conventions for representing the pattern structures described in the full lists of markers have also been followed in these tables.

[207] Marker forms are distinguished on the basis of changes or additions of marker components, based on the identification of the longest possible form in each context analyzed. Morphological variations are not taken into account in the analysis, nor are variations in articles that may occasionally appear within marker forms.

[208] These counts of pattern forms reflect differences linked to the order of pattern elements (i.e., markers, related elements) as well as insertions of additional items within pattern forms. They also reflect variations linked to the transition from active to passive voice or verbal markers. They do not, however, take into account the presence or absence of articles within structures.

	n. + prep.	marker] Y X ([verb])[causal marker] [article] risk of Y [causal marker] X risk in relation to Y						
risk factor	n. + n. n. + n. + prep.	X risk factor, [conjunction] Y X [copula] ([article]) risk factor for Y X as ([article/quantifier]) risk factor for Y	10	3	risk factor, risk factor for, as a risk factor for	0.33	3	0.30
marker	n. n. + prep.	X [copula] ([article/quantifier]) Y marker X marker Y X Y marker X [copula] [article] marker for Y X [copula] [article] marker of Y X [noun] as [article] marker of Y marker of X: Y	9	4	marker, marker for, marker of, as a marker of	0.44	6	0.67
relationship	n. n. + prep. + conj.	X-Y relationship relationship between X and Y	9	2	relationship, relationship between… and	0.22	2	0.22
in	prep.	X in Y X [causal marker/modifier] in Y	8	1		0.12	2	0.25

association	n. + prep. + conj. n. + prep. + prep.	association between X and Y association of X with Y	7	2	association between… and, association of… with	0.28	2	0.28
and	conj.	X and Y	5	1		0.20	1	0.20
link	v. + prep.	X [copula] linked to Y [study, finding] links X to Y [pathway] links X with Y	5	2	link to, link with	0.40	2	0.40
with	prep.	X ([causal marker]) with Y [causal marker] X with Y	5	1		0.20	3	0.6
related to	ppl.a. + prep.	X [copula] related to Y X related to Y	4	1		0.25	2	0.50
correlate	v. + prep. v. + conj.	X correlates with Y X has been correlated with Y [data] correlates X and Y	3	2	correlate with, correlate… and	0.67	3	1.00
relevant to	adj. + prep.	X [copula] relevant to Y X relevant to Y	3	1		0.33	2	0.67
find… in	v. + prep.	X was found in Y [researcher] found X in Y	2	1		0.50	2	1.00
link between… and	n. + prep. + conj.	link between X and Y	2	1		0.50	1	0.50
predict	v.	X predicts Y	2	1		0.50	1	0.50

prediction of	n. + prep.	X [conjunction/association marker] [article] prediction of Y	2	1		0.50	1	0.50
relation	n. + prep. + conj. n. + prep. + prep.	relation between X and Y relation of X to Y	2	2	relation between... and, relation of... to	1.00	1	0.50
Total			109	32			43	

CAUSE–EFFECT: CREATION (51 markers, 26 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
role	v. + art. + n. + prep. n. + prep. + prep. n. + prep.	X plays ([article]) role in ([article])Y X in which Y plays a role role for X in Y role of X in ([article]) Y X role in Y	33	5	play a role in, in which... plays a role, role for... in, role of... in, role in	0.15	4	0.12
contribute to	v. + prep.	X contributes to Y	13	1		0.08	1	0.08
induce	v.	X induces Y X [verb phrase] by inducing Y	11	2	induce, by inducing	0.18	2	0.18
lead to	v. + prep.	X leads to Y	9	1		0.11	1	0.11

involved in	ppl.a. + prep.	X involved in Y X [copula] involved in [article] Y	8	1		0.12	2	0.25
implicate in	v. + prep.	[evidence] implicates X in Y X [copula] implicated in Y	7	1		0.14	2	0.29
result	v. + prep.	X results from Y X results in Y	7	2	result in, result from	0.29	1	0.14
mediated	ppl.a. ppl.a. + prep.	X-mediated Y X ([copula]) mediated by Y	6	2	mediated, mediated by	0.33	3	0.50
cause	v.	X causes Y	5	1		0.20	1	0.20
importance of… in	n. + prep. + prep.	importance of X in ([article]) Y	5	1		0.20	1	0.20
important in	adj. + prep.	X [copula] important in Y [copula] X important in Y X [causal marker] [article] important [noun] in Y	5	1		0.20	3	0.60
pathway	n. + prep.	X [copula]/[verb + preposition] [article] pathway for Y X [copula] [article] pathway in Y	4	3	as a pathway of, pathway in, pathway for	0.75	2	0.50

		X [copula] [article] pathway of Y						
due to	adj. + prep.	X due to Y X [copula] due to Y	3	1		0.33	2	0.67
mediate	v.	X ([copula]) mediated by Y X ([copula]) mediated through Y X ([copula]) mediated via Y	3	3	mediated by, mediated through, mediated via	1.00	2	0.67
produce	v.	X produces Y X [copula] produced by Y	3	2	produce, produced by	0.67	2	0.67
cause of	n. + prep.	cause of X [copula] Y X [copula] [article] cause of Y	2	1		0.50	2	1.00
drive	v.	X drives Y	2	1		0.50	1	0.50
implicated in	ppl.a. + prep.	X implicated in Y	2	1		0.50	1	0.50
induced	ppl.a. ppl.a. + prep.	X-induced Y X induced by Y	2	2	induced, induced by	1.00	2	0.50
initiate	v.	X initiate Y	2	1		0.50	1	0.50
key… in	adj. + prep.	X [copula] [article] key [noun] in Y	2	1		0.50	1	0.50
mechanism of	n. + prep.	([causal marker]) X as [article] mechanism of Y	2	1		0.50	2	0.50

participate in	v. + prep.	X participating in Y	2	1		0.50	1	0.50
product of	n. + prep.	X, [article] products of Y	2	1		0.50	1	0.50
trigger	v.	X triggers Y X [causal marker] trigger Y	2	1		0.50	2	1.00
via	prep.	E1 via E2	2	1		0.50	1	0.50
Total			144	39			44	

CAUSE–EFFECT: DESTRUCTION (5 markers, 2 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ
anti-	affix	X [verb] anti-Y [causal marker]	3	1		0.33	1	0.33
against	prep.	X [preposition] [causal marker] against Y X against Y	2	1		0.50	2	1.00
Total			5	2			3	

CAUSE–EFFECT: MAINTENANCE/PERMISSION (11 markers, 1 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
required for	ppl.a. + prep.	X [copula] required for Y	2	1		0.50	1	0.50
Total			2	1			1	

CAUSE–EFFECT: PREVENTION (6 markers, 3 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
prevent	v.	X prevents Y [causal marker] X to prevent Y X [causal marker] to prevent Y	6	1		0.17	3	0.50
prevention	n.	X as Y prevention X in ([article]) prevention of Y X ([participial adjective]/[noun]/ [copula][adjective]) for ([article]) prevention of Y	6	3	as… prevention, in prevention of, for prevention of	0.50	5	0.83
suppressor	n.	X suppressor Y	4	1		0.25	1	0.25
Total			16	5			9	

CAUSE–EFFECT: MODIFICATION (20 markers, 6 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms Total	Details	Forms per occ.	Pattern forms Total	Forms per occ.
effect	n. + prep. n. + prep. + prep. n. + prep.	X effect of Y effect of X on Y X's effect on Y X has ([article]/[quantifier]) effect on Y	12	3	effect, effect of… on, effect on	0.25	4	0.33
affect	v.	X affects Y X [copula] affected by Y	7	2	affect, affected by	0.28	2	0.28
respond to	v. + prep.	X responds to Y	6	1		0.17	1	0.17
response	n. + prep. + prep. n. + prep.	response of X to Y X response to Y X [verb] [article] response to Y	5	2	response of… to, response to	0.40	3	0.60
influence	v.	X influences [article] Y	2	1		0.50	1	0.50
regulated	ppl.a.	X-regulated Y	2	1		0.50	1	0.50
Total			34	10			12	

CAUSE–EFFECT: INCREASE (14 markers, 7 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
promote	v.	X promotes Y X [preposition] promoting Y	10	1		0.10	2	0.20
increase	v.	X increases Y X, increasing Y	9	1		0.11	2	0.22
enhance	v.	X enhances Y X has been enhanced by Y	2	2	enhance, enhanced by	1.00	2	1.00
facilitate	v.	X facilitate Y	2	1		0.50	1	0.50
increased	ppl.a.	increased X [causal marker] Y X [causal marker] increased Y	2	1		0.50	2	1.00
stimulate	v.	X stimulates Y	2	1		0.50	1	0.50
upregulate	v.	X upregulates Y	2	1		0.50	1	0.50
Total			29	8			11	

CAUSE–EFFECT: DECREASE (13 markers, 7 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms Total	Details	Forms per occ.	Pattern form Total	Forms per occ.
reduce	v.	X reduces Y by X, Y [copula] reduced X [causal marker] to reduce Y	6	2	reduce, reduce by	0.33	3	0.50
inhibit	v.	X inhibits Y	5	1		0.20	1	0.20
decrease	v.	X decreases Y	2	1		0.50	1	0.50
downsizing	n. n. + prep.	X for Y downsizing X downsizing with Y	2	2	for downsizing, downsizing with	1.00	2	1.00
inhibition of	n. + prep.	X [causal marker] inhibition of Y	2	1		0.50	1	0.50
lower	v.	X lowers Y X [preposition] lowering Y	2	1		0.50	2	1.00
reduced	ppl.a.	X [causal marker] reduced Y	2	1		0.50	1	0.50
Total			21	9			11	

French

ASSOCIATION (30 markers, 13 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
et	conj.	X et ([association marker]) Y	10	1		0.10	2	0.20
lié à	ppl.a. + prep.	X lié à [article] Y X, lié à [article] Y	7	1		0.14	2	0.28
facteur de risque	n. + prep. + n. + prep. n. + prep. + n. + prep. n. + prep. + n.	facteur de risque X tels que [article] Y facteur de risque de X (Y) facteur de risque X (Y)	6	3	facteur de risque, facteur de risque de, facteur de risque… tel que	0.50	3	0.50
caractérisé par	ppl.a. + prep.	X [copula] caractérisé par [article]/[preposition] Y X caractérisé par [article] Y	5	1		0.20	3	0.60

risque de	n. + prep.	risque de X [causal marker] [article] Y X [causal/association marker] risque de ([article]) Y	5	1		0.20	2	0.40
associé à	ppl.a. + prep.	X [copula] associé à [article] Y X associé à [article] Y	4	1		0.25	2	0.50
lien entre… et	n. + prep. + conj.	lien ([verb]) entre ([article]) X et ([association marker]) ([article]) Y	4	1		0.25	3	0.75
au cours de	prep. + n. + prep.	X [causal marker] au cours de Y	2	1		0.50	1	0.50
corrélé avec	ppl.a. + prep.	X [copula] corrélé avec [article] Y	2	1		0.50	1	0.50
en cas de	prep. + n. + prep.	en cas de X, Y	2	1		0.50	1	0.50
observé	ppl.a. + prep.phr.	X observé au niveau de/dans [article] Y	2	2	observé dans, observé au niveau de	1.00	1	0.50
prédisposition	n. prep. + n. + prep.	prédisposition X [causal marker/association marker] Y X de prédisposition à [article] Y	2	2	prédisposition, de prédisposition à	1.00	2	1.00
retrouvé dans	ppl.a. + prep.	X retrouvé dans [article] Y X [copula] retrouvé dans [article] Y	2	1		0.50	2	1.00
Total			53	17			25	

CAUSE–EFFECT: CREATION (54 markers, 25 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
conduire à	v. + prep.	X conduit à [article] Y	8	1		0.12	1	0.12
entraîner	v.	X entraîne [article] Y X, entraînant [article] Y	7	1		0.28	2	0.28
induire	v.	X induit [article] Y	7	1		0.14	1	0.14
induit par	ppl.a. + prep.	X induit par [article] Y X (induit par Y)	7	1		0.14	2	0.28
participer à	v. + prep.	X participe à [article] Y X, participant à [article] Y	6	1		0.17	2	0.33
résulter de	v. + prep.	X résulte de [article] Y il X résulte [article] Y	6	2	résulte de, il en résulterait	0.33	2	0.33
activer	v.	X active Y [causal marker] [article] X : activer [article] Y X [copula] [causal marker] activer [article] Y	5	1		0.20	3	0.60
exprimer	v.	X ([affix])exprime ([article]) Y	5	2	exprimant, surexprimant	0.40	1	0.20

facteur de	n. + prep.	facteur de X ([conjunction]) Y X [causal marker] facteur de Y	4	1		0.25	3	0.75
impliqué dans	ppl.a. + prep.	X [copula] impliqué dans [article] Y X impliqué dans [article] Y	4	1		0.25	2	0.50
provoquer	v.	X provoque [article] Y X [causal marker] provoquer [article] Y	4	1		0.25	2	0.50
rôle	v. + art. + n. + prep. n. + v. + prep. + prep. n. + prep. + prep.	X joue [article] rôle dans/lors de [article] Y rôle joué par [article] X dans [article] Y rôle de [article] X dans [article] Y	4	4	jouer un rôle dans, jouer un rôle lors de, rôle joué par, rôle de.... dans	1.00	3	0.75
stimuler	v.	X stimule [article] Y	4	1		0.25	1	0.25
conséquence de	n. + prep.	X [copula] [article] conséquence de [article] Y conséquence de [article] X [copula] [article] Y	3	1		0.33	2	0.67
déclenchement de	n. + prep.	X [causal marker] [article] déclenchement de [article] Y	3	2	déclenchement de, déclenchement de... par	0.66	2	0.67

	n. + prep. + prep.	déclenchement de [article] X par [article] Y						
déclencher	v.	X déclenche [article] Y X doit être déclenché par [article] Y	3	2	déclenche, déclenché par	0.67	2	0.67
pour	prep.	X pour ([article]) Y	3	1		0.33	1	0.33
produire	v.	X produit [article] Y	3	1		0.33	1	0.33
responsable de	adj. + prep.	X [copula] responsable de [article] Y	3	1		0.33	1	0.33
à l'origine de	prep. + n. + prep.	X [copula] à l'origine de [article] Y	2	1		0.50	1	0.50
important	adj. + prep.	X important dans/pour [article] Y	2	2	important dans, important pour	1.00	1	0.50
intervenir dans	v. + prep.	X intervient dans [article] Y	2	1		0.50	1	0.50
médié par	ppl.a. + prep.	X [copula] médié par [article] Y	2	1		0.50	1	0.50
par	prep.	X par ([article]) Y	2	1		0.50	1	0.50
réponse à	n. + prep. prep. + n. + prep.	X [verb] [preposition] [article] réponse à [article] Y X en réponse à [article] Y	2	2	réponse à, en réponse à	1.00	2	1.00
Total			101	34			41	

CAUSE–EFFECT: DESTRUCTION (7 markers, 2 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
anti-	affix	X antiY	2	1		0.50	1	0.50
destruction de	n. + prep.	X [causal marker] [article] destruction de [article] Y	2	1		0.50	1	0.50
Total			4	2			2	

CAUSE–EFFECT: MAINTENANCE/PERMISSION (10 markers, 3 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
permettre	v.	X permet ([article]) Y X, permettant [article] Y	5	1		0.20	2	0.40
nécessaire à	adj. + prep.	X nécessaire à [article] Y X [copula] nécessaire à [article] Y	4	1		0.25	2	0.50
dépendant	adj. adj. + prep.	X Y dépendant X [copula] dépendant de [article] Y X dépendant de Y	3	2	dépendant, dépendant de	0.67	3	1.00
Total			12	4			7	

CAUSE–EFFECT: PREVENTION (11 markers, 4 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
suppresseur de	adj. + prep.	X suppresseur de Y [causal marker] suppresseur de X [preposition] Y	4	1		0.25	2	0.50
prévention	prep. + n. n. + prep.	X de prévention [verb] [article] Y X [verb] [preposition] [article] prévention de [article] Y [causal marker] X [copula] [article] prévention de [article] Y	3	2	de prévention, prévention de	0.33	3	1.00
bloquer	v.	X bloque [article] Y	2	1		0.50	1	0.50
préventif	adj.	X [verb] [article] [causal marker] préventif [preposition] Y X Y, préventif	2	1		0.50	2	1.00
Total			11	5			8	

CAUSE–EFFECT: MODIFICATION (32 markers, 8 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occur-rences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
effet	n. n. + prep. n. + prep. + prep.	X [verb] [article] effet Y X [verb] [article] effet sur [article] Y effet de/[preposition] [article] X sur [article] Y	7	3	effet, effet sur, effet de… sur	0.43	2	0.28
régulation de	n. + prep. n. + prep. + conj.	X [causal marker] [article] régulation de [article] Y régulation entre X et Y	4	2	régulation de, régulation entre… et	0.50	2	0.50
moduler	v.	X module [article] Y	3	1		0.33	1	0.33
anti-	affix	X antiY	2	1		0.50	1	0.50
commander	v.	X commande [article] Y	2	1		0.50	1	0.50
complication … de	n. + prep. + prep.	complication X de [article] Y	2	1		0.50	1	0.50
contrôler	v.	X est contrôlé par [article] Y	2	1		0.50	1	0.50
nuire à	v. + prep.	X nuit à [article] Y	2	1		0.50	1	0.50
Total			24	11			10	

CAUSE–EFFECT: INCREASE (10 markers, 5 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
favoriser	v.	X favorise [article] Y X, favorisant [article] Y	7	1		0.14	2	0.28
augmentation de	n. + prep.	X [causal marker] [article] augmentation de [article] Y augmentation de [article] X, [causal marker] [article] Y	4	1		0.25	2	0.50
augmenter	v.	X augmente Y augmenter X, [causal marker] Y	3	1		0.33	2	0.67
accroître	v.	X accroît [article] Y X [causal marker] accroître Y	2	1		0.50	2	1.00
faciliter	v.	X faciliter [article] Y	2	1		0.50	1	0.50
Total			18	5			9	

CAUSE–EFFECT: DECREASE (12 markers, 5 with 2 or more analyzed occurrences)

Marker	Marker POS	Pattern Forms	Occurrences in sample	Marker forms			Pattern forms	
				Total	Details	Forms per occ.	Total	Forms per occ.
inhiber	v.	X inhibe [article] Y X peut être inhibé par [article] Y	7	2	inhiber, inhibé par	0.28	2	0.28

diminuer	v.	X diminuent Y X [causal marker] diminuer Y	3	1		0.33	2	0.67
réduire	v.	X réduit Y X [causal marker] réduire Y	3	1		0.33	2	0.67
inhibiteur	adj. + prep.	X inhibiteur de [article] Y [causal marker…] inhibiteur de X […causal marker] [article] Y	2	1		0.50	2	1.00
inhibition de	n. + prep. n. + prep. + prep.	X [causal marker] [article] inhibition de [article] Y inhibition de X par Y	2	2	inhibition de, inhibition de… par	1.00	2	1.00
Total			17	7			10	