Université de Montréal

Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés

par
Patrick Drouin
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 linguistique

Mai 2002

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Acquisition automatique des termes :

l'utilisation des pivots lexicaux spécialisés

présentée par :

Patrick Drouin

a été évaluée par un jury composé des personnes suivantes :

Jean-Claude Corbeil
Président-rapporteur

Monique C. Cormier
Directrice de recherche

Didier Bourigault
Codirecteur

Lynne Bowker
Membre du jury

Pierre Auger
Examinateur externe

Jian-Yun Nie
Représentant du doyen de la FES

Résumé

Les travaux entrepris dans le cadre de la présente thèse se situent dans le domaine de l'acquisition des connaissances à partir de documents ou, plus spécifiquement, dans le domaine de l'*acquisition automatique des termes.* Nous présentons un logiciel d'acquisition automatique des termes utilisé pour tester les hypothèses avancées : TermoStat.

Afin de procéder à l'acquisition automatique des termes en langue anglaise, le logiciel a recours à des tests statistiques qui ont pour objet de comparer le lexique d'un corpus technique (corpus d'analyse) à celui d'un corpus non technique (corpus de référence). Cette comparaison conduit à l'identification des pivots lexicaux spécialisés (PLS) qui constituent les formes les plus représentatives du lexique du corpus d'analyse.

Les PLS sont par la suite utilisés comme point de départ pour l'acquisition automatique des termes. Le processus d'acquisition exploite le concept de frontière de termes afin d'isoler les candidats-termes (CT). Le recours aux PLS permet au logiciel de concentrer son analyse sur les zones du corpus qui ont un comportement lexical particulier. TermoStat peut ainsi se contenter de scruter uniquement le contexte immédiat des PLS en quête de frontières de termes.

Afin de maximiser la qualité des résultats présentés par TermoStat, nous proposons un indice permettant de représenter le caractère terminologique des CT. L'indice *iTer* prend en considération diverses caractéristiques du candidat (fréquence, longueur, etc.)

directement observables dans le corpus d'analyse. Le tri de la liste des candidats retenus par TermoStat a permis d'obtenir une liste de CT dont la précision, mesurée sur la première moitié de la liste, atteint 86,8 %.

Mots clés

Terminologie, linguistique textuelle, linguistique quantitative, acquisition des connaissances, acquisition automatique de la terminologie, acquisition automatique des termes.

Abstract

The research undertaken for this thesis is part of knowledge acquisition from texts; it focuses more particularly on *term acquisition*. Our work led to the development of TermoStat, a piece of software dedicated to testing our methodology for automatic term acquisition in an industrial environment.

To carry out term recognition in English, TermoStat relies on statistical techniques to compare the lexicon of a technical corpus (analysis corpus) to one of a non-technical corpus (reference corpus). The object of this comparison is to establish a list of specialised lexical pivots (SLP). The SLPs correspond to the lexical items that have an abnormally high frequency in the analysis corpus as compared to the reference corpus.

SLPs are used as a starting point for the automatic acquisition of terms, which relies on the concept of term frontiers. Using specialised lexical pivots allows TermoStat to focus its analysis on parts of documents that have a particular behavior. This pinpointing of relevant information allows TermoStat to only look at the immediate context of SLPs.

In order to maximize the quality of the results, we put forward a weighting index to capture the terminological potential of candidate terms (CT). The index, called *iTer*, includes various contextual clues as observed in the corpus (frequency, length, etc.). The first half of the sorted list of CTs obtained from our analysis corpus with TermoStat had a precision of 86.8%.

Keywords

Terminology, corpus linguistics, quantitative linguistics, knowledge acquisition, terminology acquisition, term acquisition.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et acronymes

CA	corpus d'analyse
CG	corpus global
CNRS	Centre national de la recherche scientifique
CR	corpus de référence
CT	candidat-terme
IA	intelligence artificielle
IM	information mutuelle
INIST	Institut de l'information scientifique et technique
ISO	International Standard for Organization
ISR	inventaire des segments répétés
NSE	Nortel Standard English
OLF	Office de la langue française
PLS	pivot lexical spécialisé
RINT	Réseau international de néologie et de terminologie
SN	syntagme nominal
SR	segments répétés
TAL	traitement automatique de la langue
TIA	Terminologie et intelligence artificielle
URI	Unité de recherche et innovation

À ma famille et à Marie-Jo

Remerciements

Les résultats que je soumets dans la présente thèse ont été appuyés financièrement par le Conseil de recherches en sciences humaines du Canada (CRSH), le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche et l'Université de Montréal. Je les en remercie vivement.

Je tiens aussi à remercier les personnes suivantes :

Ma directrice de recherche, Monique C. Cormier, pour avoir cru que cette thèse allait un jour se terminer et pour avoir tenu le coup jusqu'à ce que ça se produise. Sans sa gentillesse, sa franchise, son soutien bienveillant, ses conseils, sa patience et sa minutie, je n'aurais pu mettre un point final à ce manuscrit.

Mon codirecteur, Didier Bourigault, de l'Équipe de recherche en syntaxe et sémantique (ERSS) de l'Université Toulouse-Le Mirail, qui a accepté de jouer le jeu malgré la distance. Je le remercie pour son scepticisme face aux solutions rapides et pour les longues discussions téléphoniques portant sur l'acquisition automatique des termes tenues à des heures plus ou moins acceptables pour son fuseau horaire.

Robert Cléroux, professeur titulaire au Département de mathématiques et de statistique de l'Université de Montréal, pour son aide précieuse pour tout ce qui touche à la statistique. Réussir à m'inculquer les notions de base n'a pas été une mince affaire.

Les chercheurs du Centre interdisciplinaire de recherches sur les activités langagières (CIRAL) de l'Université Laval pour m'avoir

donné le goût de poursuivre mes études aux cycles supérieurs. Ils ont aussi su donner l'encadrement nécessaire à un groupe de jeunes étudiants pour susciter en eux l'étincelle qui fait naître les chercheurs. Je leur en suis reconnaissant.

Les terminologues spécialistes du domaine des télécommunications, Lynne Leslie et Tricia Morgan, qui ont accepté de valider les résultats obtenus à l'aide de TermoStat. Un remerciement particulier à Andy Lauriston, qui a eu la gentillesse de revoir l'ensemble des résultats. Sans ses connaissances techniques, terminologiques et linguistiques, je serais probablement toujours devant les listes de termes à me demander par où commencer.

Mike Milinkovich, de la défunte société Rebel.com, d'avoir si gentiment offert le serveur Linux nécessaire à la réalisation des manipulations informatiques des corpus.

Ma famille, qui a accepté les absences, les sautes d'humeur, les silences, et qui m'a toujours encouragé durant cette longue entreprise qui ressemblait un peu à l'agonie d'un personnage d'opéra. Si j'ai complété ce travail, c'est en partie pour leur remettre une partie de la confiance qu'ils ont eue en moi.

Les amis, qui ont su ramener si souvent cette question si embêtante pour tous les thésards : « Pis, ta thèse? » J'ai enfin la réponse à votre question, que tout ceux et celles qui ont fait des promesses de célébration les tiennent maintenant!

Marie-Jo d'avoir été là et d'avoir partagé les moments difficiles, tout comme les moments d'euphorie. C'est d'ailleurs au cours d'une séance de motivation que mon logiciel a été baptisé TermoStat : c'est

son idée. Je lui dois la persévérance et la tranquillité d'esprit nécessaires à la réalisation de la présente recherche. Elle pourra désormais dire qu'elle n'a plus un « chum virtuel »...

1. INTRODUCTION

La présente recherche est effectuée dans un contexte d'entreprise privée. Elle vise à répondre aux besoins de la société Nortel Networks face à un problème précis : le dépouillement terminologique d'un volume important de documentation en langue anglaise. Cette société est une multinationale qui œuvre dans le domaine des télécommunications et qui est présente sur presque tous les continents. Depuis quelques années, la mondialisation des marchés, particulièrement dans le domaine des télécommunications, a fait sentir une demande croissante pour la traduction de documents techniques vers des langues diverses.

Au sein de Nortel Networks, on évalue à plusieurs milliers le nombre total de pages qui sont rédigées chaque année et ce, uniquement pour les documents techniques. Afin de satisfaire aux besoins de traduction de cette masse de documents, la société se doit d'envisager la mise en place de processus facilitant sa gestion, sa manipulation, sa rédaction et sa traduction. Une stratégie étudiée par la société pour faciliter la traduction de ses documents est l'établissement d'une langue contrôlée qui vient assurer une certaine uniformité de la qualité de l'anglais dans sa documentation technique. Cet anglais contrôlé est nommé *Nortel Standard English* (*NSE*).

Un des défis qui se posent à cette entreprise, dans le cadre du projet *NSE*, est la conversion en masse de documents déjà rédigés vers une langue contrôlée. Cette conversion manuelle, assistée par un outil de rédaction, touche bien des niveaux de la langue anglaise. Cependant, son aspect le plus important demeure celui de la

normalisation de la terminologie[1]. Afin d'y parvenir, les rédacteurs sont mandatés pour identifier les termes nécessaires à la réalité des télécommunications. Ce travail manuel est donc long et fastidieux et empêche souvent le rédacteur de se livrer à son activité principale.

Afin de pallier ce problème, la société Nortel Networks tente de mettre en place des systèmes d'acquisition des termes. Ces systèmes permettent d'extraire au préalable la liste des termes contenus dans un document et de les utiliser pour alimenter le dictionnaire de l'outil de rédaction assistée. Le travail du rédacteur est d'autant plus simplifié que l'outil ne sollicite plus son expertise à tout moment et le rédacteur n'a plus à accomplir le travail du terminologue. La présente thèse s'attaque à cette problématique et a pour objectif ultime de concevoir une stratégie d'acquisition automatique de termes. Cette dernière est testée à l'aide d'un logiciel (TermoStat) élaboré dans le cadre du présent travail.

Ce système d'acquisition automatique des termes élaboré se doit cependant de ne pas imposer une charge additionnelle de travail aux terminologues. Il doit présenter des résultats de haute qualité même si la liste de termes dressée ne constitue qu'un sous-ensemble de tous les termes présents dans le corpus. Cette recherche de qualité se fait donc au détriment d'un dépouillement systématique des documents. Ainsi, l'objectif principal du prototype mis en place n'est pas de procéder à un dépouillement complet des textes analysés, mais à un dépouillement dont le résultat sera le plus pertinent possible.

---

[1] Par *normalisation*, nous entendons la recherche d'une seule et unique dénomination pour chaque concept. Cette élimination des synonymes du processus de rédaction est à la base de l'élaboration d'une langue contrôlée.

Depuis quelques années, l'acquisition automatique des connaissances connaît une popularité sans cesse grandissante auprès de la communauté qui s'intéresse à l'intelligence artificielle (IA). La mise sur pied d'un groupe de travail en terminologie et intelligence artificielle, le groupe TIA, en est un bon exemple (Otman 1995, Toussaint *et al.* 1997, Condamines et Enguehard 1999, URI-INIST-CNRS 2001). Les travaux de ce groupe ont débouché sur la tenue de colloques en 1995, 1997, 1999 et en 2001. La synergie créée par ce groupe a su attirer des chercheurs de nombreux domaines (linguistique, informatique, bibliothéconomie, etc.).

La mise en circulation d'une grande quantité de textes en format électronique a relancé l'intérêt pour la linguistique de corpus. Un des objectifs de cette discipline consiste à mettre de l'avant des techniques permettant de cerner le contenu des corpus et de distinguer, d'un point de vue terminologique, le pertinent du non-pertinent dans cette masse de données. Afin d'être efficace, ce filtrage nécessite un recours à des connaissances et c'est à ce niveau que les spécialistes du groupe TIA interviennent.

La question qui se pose est donc de savoir comment identifier les connaissances ou leurs manifestations dans les textes. Dans un texte, le spécialiste expose ses connaissances qui se réalisent en discours. Cette manifestation passe principalement par le biais des termes. Il s'agit d'ailleurs ici de l'hypothèse de départ exprimée par Toussaint *et al.* (1997 : 28) dans leurs travaux sur l'acquisition automatique des termes. Dans le cadre de la présente recherche, c'est cette manifestation textuelle que nous cherchons à isoler et à cerner automatiquement.

De concert avec l'intérêt sans cesse grandissant pour la linguistique de corpus, nous assistons à une augmentation du nombre de recherches ayant pour objet de pousser plus loin les frontières de l'informatisation du travail du terminologue (voir David et Plante 1990; Bourigault 1992a, 1992b, 1993, 1994a et 1994b; Auger 1994; Daille 1994a et 1994b; Jacquemin 1996, 1997 et 2001; Bourigault *et al.* 2001). Nous pouvons désormais envisager l'utilisation de techniques qui dépassent la simple attestation de termes et qui rendent possible l'identification automatique d'information terminologique de toute nature dans les corpus.

Les travaux entrepris dans le cadre de la présente thèse se situent dans le domaine de l'acquisition des connaissances à partir de documents ou, plus spécifiquement, dans le domaine de l'*acquisition automatique de la terminologie*. Ce domaine d'activité a pour objectif de recenser l'information habituellement utilisée en terminologie : les contextes, les définitions, les synonymes, etc. On peut donc envisager que chacun des éléments d'information utilisés en terminologie peut donner naissance à un domaine de recherche plus restreint (acquisition automatique des définitions, des contextes, des termes, etc.).

Parmi toutes les activités qui composent l'éventail des tâches du terminologue, c'est au défi que constitue la recherche des unités terminologiques que nous désirons consacrer le présent travail. L'*acquisition automatique des termes* consiste à extraire automatiquement, sans connaissance préalable, des termes d'un corpus spécialisé. Nous reprenons ici la terminologie utilisée dans Bourigault (1994a), Enguehard (1994) et reprise par Christian Jacquemin plus récemment (1997 et 2001).

Les algorithmes mis en œuvre dans TermoStat reposent sur l'hypothèse que l'on peut utiliser la thématique principale d'un corpus pour en identifier la terminologie. Cette thématique, en relation étroite avec le domaine dont est tiré le document, se traduit au sein d'un corpus par une forte spécificité lexicale (Kittredge 1982 : 111, Picht 1987 : 149). Nous croyons que l'identification de cette thématique, à l'aide de tests statistiques, permet d'accéder aux termes du corpus. L'approche que nous adoptons repose donc sur la comparaison des particularités lexicales de deux corpus : un corpus de référence et un corpus d'analyse. Le corpus de référence est composé d'une gamme de textes journalistiques qui traitent de divers domaines. Pour sa part, le corpus d'analyse est formé de documents techniques circonscrits à un domaine particulier du savoir : les télécommunications.

Afin de procéder à l'acquisition automatique des termes dans le corpus d'analyse, le logiciel construit une liste de fréquences des unités lexicales qu'il contient. Les fréquences observées sont par la suite comparées par TermoStat à celles tirées du corpus de référence. Une analyse probabiliste permet ainsi d'identifier les unités lexicales qui possèdent une fréquence qui se démarque significativement de celle observée dans le corpus de référence : les pivots lexicaux spécialisés (PLS). Ces derniers constituent le point de départ du processus d'acquisition des termes.

La technique d'acquisition des termes que nous mettons de l'avant se fonde sur les travaux effectués sur l'acquisition par frontières de termes (Bourigault 1992a, 1992b, 1993, 1994a et 1994b). Ces recherches utilisent les connaissances au sujet des termes de façon négative, c'est-à-dire qu'ils exploitent une description informatique de ce qui ne peut pas constituer un terme d'un point de

vue morphosyntaxique. Nous exploitons donc ces connaissances négatives afin de décrire les formes potentielles que peuvent prendre les termes en discours. Au cours du processus d'acquisition, seuls les PLS identifiés à l'étape précédente sont utilisés à titre d'éléments admis au sein des termes potentiels, les candidats-termes (CT).

Les résultats issus du processus d'acquisition sont ensuite triés de façon à placer les CT les plus pertinents en tête de la liste des CT. Nous cherchons ainsi à présenter en priorité aux terminologues les CT dont le statut terminologique est le plus probable. En cherchant à atteindre une liste qui serait triée par ordre de pertinence, nous soulevons le problème de l'évaluation du statut terminologique du CT et de son approximation par le terminologue ou par le logiciel.

L'évaluation de ce caractère terminologique des CT passe par l'observation des indices contenus dans le corpus d'analyse afin de tenter de distinguer les CT qui sont des termes de ceux qui n'en sont pas. Dans le cadre de la présente recherche, nous tentons de déterminer une façon de qualifier ou de quantifier ce statut, une façon de représenter le caractère terminogénique d'un CT. Afin d'y parvenir, nous proposons l'indice terminogénique *iTer*.

En résumé, la thèse vise donc à concevoir une stratégie d'acquisition automatique de termes. Pour ce faire, nous précédons à :

- l'identification des particularités lexicales d'un corpus technique : les pivots lexicaux spécialisés,
- la mise sur pied d'une stratégie d'acquisition automatique des termes fondée sur les PLS,

- la description d'un indice terminogénique représentant l'intérêt terminologique des CT,
- l'intégration des stratégies précédentes au sein d'un logiciel : TermoStat.

2. ÉTAT DE LA QUESTION

2.1 Terminologie

2.1.1 Rappel historique

La paternité de la terminologie, en tant que discipline ou méthodologie, est généralement attribuée à l'ingénieur autrichien Eugen Wüster (Rondeau 1984 : 6; Sager 1990 : 2; Cabré 1998 : 22; Pearson 1998 : 10; Cabré 1999 : 17). Les travaux de Wüster seront poursuivis par ce qu'il convient maintenant de nommer l'École de Vienne (Cabré 1998 : 37), dont Felber (1984) et Picht et Draskau (1985) sont de bons exemples. La théorie issue de cette école a eu un impact si important sur le domaine de la terminologie que l'on qualifie l'approche de Wüster et de ses disciples de *théorie générale de la terminologie* (Cabré 1998 : 30, Pearson : 1998 : 10) ou de *théorie traditionnelle de la terminologie* (Pearson : 1998 : 10; Kageura 1999 : 21; Temmerman 1999 : 77).

En tant qu'ingénieur, Wüster était confronté aux difficultés de la communication entre professionnels dues à la multiplication des termes dans son domaine d'expertise. La solution à cette situation passait nécessairement par la normalisation de la terminologie employée. La publication de son *Dictionnaire multilingue de la machine-outil* (Wüster 1968) constitue une première tentative pratique pour pallier cette situation.

Les motivations pratiques qui sont à la base de la démarche de Wüster ont eu un impact direct sur la forme que prendra la théorie qu'il va élaborer. En effet, son objectif de désambiguïsation de la communication professionnelle l'amènera à adopter une vision très normalisatrice, au centre de laquelle se situe la notion :

> « Le plus important pour elle [la terminologie] dans une langue, c'est le système de notions sur lequel elle repose. »
> Wüster (1981 : 64)

Non seulement la notion est centrale à la théorie, mais elle s'insère dans un système qui est tout aussi important. Cette dernière caractéristique fera ajouter à Wüster :

> « Je le répète : non seulement la recherche terminologique prend-elle les notions comme point de départ, mais elle étudie de plus les liens qui unissent toutes les notions d'un domaine spécialisé, autrement dit, elle étudie les notions en tant que parties d'un système de notions. »
> Wüster (1981 : 70)

Dans l'esprit de Wüster, les termes sont donc secondaires et ils ne sont intéressants que parce qu'ils servent à dénommer les notions. Cette structuration du monde ou des domaines d'expertise qui le composent en notions normalisées, fixes et statiques, implique nécessairement une recherche de la normalisation des étiquettes associées aux notions. Cette démarche, selon Wüster, doit se faire à l'échelle internationale afin d'être fructueuse :

> « Dans la langue commune, le dirigisme linguistique n'est guère possible, si ce n'est dans une très petite mesure. En terminologie, en revanche, non seulement est-il un fait courant, mais il ne tient pas compte, de surcroît, des frontières géographiques et linguistiques. On ne normalise pas les notions et les termes sur le seul plan national, mais également à l'échelle internationale. »
> Wüster (1981 : 67)

Ce principe de normalisation conceptuelle, indépendante des langues qui sont concernées, est aussi repris par Drodz qui insiste sur l'importance de la normalisation :

> « L'activité de structuration et de classification devra alors aboutir à des activités de codification (normalisation) dont la nature consiste justement à éliminer les différences créées par les langues naturelles. »
> Drodz (1981 : 124)

Il s'agit cependant d'un travail d'envergure, facilité, selon Wüster, par le fait qu'il se limite à la langue spécialisée et non à la langue de tous les jours. Son optimisme face au dirigisme linguistique dans les domaines spécialisés serait peut-être aujourd'hui contesté à la lumière des travaux ayant pour objet l'évaluation du succès des efforts de normalisation (voir Loubier et Rousseau 1994 et Boulanger 1995) et la mesure de l'implantation terminologique (notamment Loubier 1993; Quirion 1996 et 2000; Chansou 1997; Cholette 1994; Fossat 1997; Gouadec 1997). Avec une approche essentiellement normalisatrice fondée sur l'aspect notionnel, on court le risque de s'éloigner de l'usage de façon considérable :

> « Traditional TT [Terminology Theory] stipulates that a specialist can describe the concept before paying any attention to the term. [...] The fact that terms already exist to communicate about knowledge in the specialised domain under consideration is conveniently overlooked. »
> Temmerman (1999 : 80)

On ne peut cependant considérer que les tenants de la terminologie traditionnelle suggèrent une approche qui oblitère entièrement le terme en laissant toute la place à la notion. En effet,

Wüster était entièrement conscient du double rôle de la terminologie qui consiste à décrire les notions et les termes qui les dénomment. Il les considère cependant comme deux choses indépendantes (Wüster 1981 : 63). Drodz (1981 : 121) reprendra plus tard cette idée puisque, selon ce dernier, l'analyse des unités terminologiques est à caractère binaire et elle doit se faire tant d'un point de vue linguistique, que d'un point de vue conceptuel.

C'est probablement cette dualité qui permet à Wüster de situer la terminologie dans le cadre de la linguistique appliquée (Wüster 1981 : 60). En effet, sans ce recours au penchant linguistique du signe, la terminologie demeure essentiellement conceptuelle et en dehors des sciences du langage. Les termes, pour leur part, se retrouvent aussi confinés à un rôle d'étiquette qui ne leur convient pas très bien (à l'exception des nomenclatures) et leur fonctionnement observable dans la langue naturelle devient difficile à expliquer.

Après quelques décennies de travail pratique en terminologie et à la lumière des nouvelles technologies, les fondements de la théorie générale de la terminologie sont désormais, et pour des raisons diverses, remis en question (Sager 1990 et 1998; Gaudin 1993; Kageura 1995 et 1999; Bourigault et Slodzian 1999; Rey 1999; Temmerman 1999).

Les critiques portent autant sur l'autonomie de la terminologie en tant que science distincte (Sager 1998; Kageura 1999; Rey 1999) que sur des propositions de rattachement à une (Gaudin 1993; Kageura 1995; Bourigault et Slodzian 1999) ou plusieurs autres disciplines (Cabré 1999, 2000a, 2000b). Le débat sur le sujet est loin d'être clos et nous devrons encore attendre quelques années avant que la question ne soit tranchée. Un consensus semble par contre

vouloir émerger des articles parus au cours de la dernière décennie indépendamment de la position théorique et de l'approche adoptée par les auteurs : le travail terminologique doit prendre en considération le fonctionnement textuel des termes (Auger et L'Homme 1994 : 17; Filipec 1994 : 352; Boulanger 1995 : 195; Cabré 1999 : 17-18 et 2000a : 45; Kageura 1999 : 29; Temmerman 1999 : 80-81).

2.1.2 Terme et textualité

Comme nous l'avons mentionné, les appels pour que les textes trouvent une place plus importante en terminologie sont nombreux et ils émanent des articles portant sur divers sujets. Les chercheurs qui aimeraient voir la terminologie intégrée dans le cadre de la linguistique sont directement intéressés par l'aspect linguistique du terme et par son fonctionnement dans les textes.

Qu'on associe la terminologie à la sociolinguistique (Gaudin 1993; Loubier et Rousseau 1994; Boulanger 1995; Condamines 1995), à la linguistique appliquée (Kageura 1995) ou à la linguistique de corpus (Bowker 1996; Bourigault et Slodzian 1999), le fonctionnement réel du terme au sein du discours et, plus particulièrement des textes, est d'importance capitale. En fait, l'écart entre la réalité terminologique décrite dans la théorie générale de la terminologie et le comportement réel des termes est ce qui a servi de catalyseur à ces études. Jean-Claude Boulanger attribue à la socioterminologie le mérite d'avoir su revaloriser le discours en terminologie :

> « La socioterminologie fait surgir de l'ombre le concept d' "usage" et le ramène dans l'environnement du terme. Qui dit usage dit aussi somme de discours dans lesquels s'enchâssent les unités lexicales que des interactions spécifiques agitent, comme des atomes, pour révéler la vraie nature de la terminologie. »
> Boulanger (1995 : 204-205)

Cette valorisation du discours se fait aussi sentir auprès des chercheurs qui n'associent pas nécessairement la terminologie à la linguistique. C'est le cas, entres autres, de Temmerman (1999) qui propose une vision sociocognitive de la terminologie. Cette approche, bien que fondée sur les connaissances sous forme de prototypes et de catégories, prend en considération et décrit les variations de sens que subissent les termes dans le discours. La notion n'est donc pas finement délimitée, mais elle se construit à partir des occurrences des termes dans les textes.

Ainsi, on ne cherche pas à mettre aux oubliettes la face linguistique du terme. Les chercheurs continuent de reconnaître la dualité terme – sens des unités terminologiques. On ne remet donc pas en question cette dualité, mais le rapport de force est inversé : puisque le terme ne peut être observé qu'en contexte, il devient le point de départ du travail terminologique (Temmerman 1999 : 80-81). Comme l'illustrent si bien Loubier et Rousseau (1994 : 75), l'acte de langage devient ainsi source et fin de la terminologie.

Comme le fait remarquer Kageura, l'élaboration d'un produit terminologique sans recours au discours est une tâche qui tient de l'utopie :

> «All these empirical phenomena relating to terms are observed within texts, actual discourse or the broader context of communication. Generally speaking, one cannot make a terminological dictionary of a domain without papers or articles [...]. This is because terms are at the level of parole [...] ».
> Kageura (1999 : 28-29)

Lorsque ces produits sont élaborés et mis à la disponibilité des utilisateurs, il demeure essentiel de valider leur contenu à l'aide d'enquêtes terminologiques ou de retours fréquents aux textes. Une telle façon de procéder aura nécessairement comme conséquence d'éviter la situation décrite par Jean-Claude Boulanger :

> « Une fois extirpées des discours textuels, en vertu de l'opération dite de *dépouillement*, les unités terminologiques n'étaient pas vraiment réinsérées dans les discours actualisés sur le terrain, ni observées dans ce même milieu ambiant. Traités et cantonnés dans les réservoirs dictionnairiques, les amas de termes étaient soumis aux préceptes de la canonisation. Pendant ce temps, les discours réels évoluaient, s'éloignaient. »
> Boulanger (1995 : 195)

En plus de l'inadéquation à représenter l'usage courant, l'élaboration de produits terminologiques, dans le contexte actuel de l'explosion des besoins terminologiques, a fait surgir de nouvelles problématiques. Les besoins en entreprise se multiplient et se diversifient. Dans le cadre des entreprises, on ne cherche pas nécessairement à établir une terminologie systématique d'un domaine ou de l'entreprise, mais à répondre rapidement à des besoins précis.

L'avènement d'outils d'aide à la rédaction, d'aide à la traduction et à la terminologie a eu une influence marquante sur le type de

produits terminologiques que les terminologues ont à mettre en place. Le langagier est maintenant appelé à élaborer des lexiques, des thésaurus, des dictionnaires pour des systèmes de traduction automatique, des glossaires dans une optique de vulgarisation, ou des index (Bourigault et Slodzian 1999 : 29).

En fonction des visées du travail, les termes retenus seront différents puisque l'utilisation qu'on compte en faire est différente. On ne peut donc plus parler de la terminologie d'un domaine particulier, mais d'une terminologie de ce domaine (Bourigault et Slodzian 1999 : 30). Cette constatation bouleverse la conception classique de la terminologie selon laquelle le dépouillement d'un corpus conduirait systématiquement à la même liste de termes. La dépendance des termes ne s'articule donc plus uniquement autour d'un domaine, mais aussi autour du corpus pris en ligne de compte et de l'objectif du terminologue lors de son dépouillement des documents. Un même corpus peut donc conduire à de nombreuses terminologies dont la pertinence est déterminée par le produit terminologique visé.

2.1.3 Le terme

On peut oser affirmer sans se tromper que la présence des termes dans les textes spécialisés est inévitable (Kageura 1999 : 28, Cabré 2000b : 15). En effet, afin de véhiculer les concepts reliés à leur discipline, les auteurs doivent faire appel à la terminologie. L'approche terminologique consiste donc à chercher les termes là où ils sont susceptibles de se manifester : au sein de documents traitant d'un ou de plusieurs domaines de l'activité humaine, de documents spécialisés.

Il faut cependant se rendre à l'évidence que les documents spécialisés ne contiennent pas que des termes et qu'on y trouve aussi des mots[2]. Ces derniers agissent, notamment, comme des charnières nécessaires à l'articulation du discours du spécialiste. Le défi qui se présente au terminologue est donc de réussir à distinguer les termes des mots dans l'ensemble des unités lexicales qui se présentent à lui.

Selon Kageura (1995 : 251, 1999 : 27), afin de pouvoir décrire correctement le terme, il est primordial de décrire ce qui le démarque du non-terme. Si cet objectif n'est pas atteint, il sera, selon le même auteur, toujours impossible d'affirmer que les observations faites au sujet des termes ne sont valides que pour les termes :

> « An observation which is not only relevant to terms but also to words does not become an observation about terms and only about terms simply because it is made on the basis of terminological data. Science is not theology. »
> Kageura (1999: 26)

Cette prise de position, très intéressante et légèrement radicale[3], soulève la question suivante : quels sont les critères à utiliser pour distinguer les termes des mots? Cette question n'est cependant pas

---

[2] Notre position relève d'une conception *a priori* de l'opposition entre *terme* et *mot*.

[3] Selon Kageura (1999), une distinction claire entre termes et mots n'a jamais été établie. Étant donné l'absence de critères permettant de distinguer ces deux concepts, on ne peut pas affirmer, à ce jour, que les observations qui ont été faites en terminologie portent essentiellement sur la terminologie et sur les termes puisque les mêmes observations s'appliquent généralement aussi aux mots. Il en sera ainsi tant que la distinction entre les termes et les non-termes n'aura pas été établie de façon convaincante.

nouvelle et elle préoccupe les terminologues depuis que les premiers travaux pratiques de terminologie ont été mis de l'avant dans les années 1970. Une table ronde sur le thème du découpage du terme a d'ailleurs été organisée au Québec à la fin de la même décennie (OLF 1979). Le point qui suit présente un aperçu des travaux qui ont depuis lors abordé ce sujet.

2.1.3.1 Le découpage du terme

Bien que de très bonnes définitions soient disponibles pour le terme, il n'en demeure par moins que son recensement dans les textes demeure un défi de taille (Sager 1990 : 61). En effet, le terminologue est bien souvent confronté à des unités syntagmatiques qui semblent constituer des termes, mais dont le statut terminologique est difficile à établir. Ce problème a d'ailleurs été soulevé très tôt dans le cadre des travaux en terminologie.

Ainsi, au milieu des années 60, Louis Guilbert (1965) et Émile Benveniste (1966) élaborent des descriptions relativement semblables des particularités des unités terminologiques complexes. Ces deux textes, et les éléments de description qu'ils renferment, sont, encore aujourd'hui, utilisés en terminologie. Dans sa thèse, Guilbert tente de discerner l'*unité lexicale complexe* du groupement syntagmatique accidentel. Il fait l'affirmation suivante :

> « Le passage du statut de groupement syntagmatique du discours au statut d'unité lexicale suppose la réalisation d'un certain nombre de conditions : la stabilité du rapport syntagmatique au plan du discours, la stabilité du rapport de signification entre l'unité syntagmatique et un signifié unique, la fréquence d'emploi

> qui stabilise à la fois le lien syntagmatique et le rapport de signification. »
> Guilbert (1965 : 275)

Benveniste, pour sa part, présente une nouvelle notion qu'il nomme *synapsie* et il en décrit les particularités ainsi :

> «Ce qui caractérise la synapsie est un ensemble de traits dont les principaux sont : 1° – la nature syntaxique (non morphologique) de la liaison entre les membres; – 2° l'emploi de joncteurs à cet effet, notamment *de* et *à*; – 3° l'ordre déterminé + déterminant des membres; – 4° leur forme lexicale pleine, et le choix de tout substantif ou adjectif; – 5° l'absence d'article devant le déterminant; – 6° la possibilité d'expansion pour l'un ou l'autre membre; – 7° le caractère unique et constant du signifié. »
> Benveniste (1966 : 172-173)

Des deux descriptions, on peut retenir, sans risquer de trahir la pensée des auteurs, les éléments suivants : l'aspect syntagmatique, le recours à des joncteurs qui assurent la stabilité syntagmatique, la linéarité de construction, le caractère constant de la référence, la fréquence d'occurrence.

Ces caractéristiques ont été reprises par des chercheurs en terminologie afin d'élaborer une liste de critères qui permettraient aux terminologues de distinguer le terme du syntagme de discours. Roger Goffin (1979 : 157-168) propose divers critères afin de distinguer le terme du mot. À l'exception du critère taxinomique[4], nous reprenons

---

[4] De par sa nature sémantique, nous avons cru bon de regrouper le critère taxinomique présenté par Goffin (1979 : 166-167) sous le critère sémantique.

l'ensemble de ses propositions puisqu'elles constituent une bonne synthèse des discussions sur le sujet. Aux critères établis par l'auteur, nous ajoutons un critère pragmatique.

2.1.3.1.1 Le critère formel

L'aspect syntagmatique des unités complexes est décrit à la fois par Guilbert (1965 : 275) et par Benveniste (1966 : 172). Cette cohésion syntaxique des unités lexicales qui composent l'unité complexe est telle que celle-ci fonctionne, en discours, comme une unité simple (Martinet 1979 : 185)[5].

Autre fait marquant, les unités se construisent selon des règles qui relèvent de la syntaxe (Guilbert 1965 : 255; Benveniste 1966 : 172; Kocourek 1991 : 139) et non de la morphologie. Afin d'y parvenir, elles ont recours à des joncteurs qui assurent la stabilité syntagmatique de l'unité. On peut donc ainsi décrire les structures potentielles des termes complexes en fonction des unités lexicales qui peuvent (Guilbert 1965 : 256-257) ou non (Dubuc 1979 : 55; Kocourek 1991 : 139) les composer et des joncteurs (Benveniste 1966 : 175-176; Lotte 1981 : 21) qui relient ces unités lexicales. Le recours aux joncteurs n'est cependant pas généralisé à toutes les langues et le français semble en faire un usage plus important que l'anglais, qui semble lui préférer la juxtaposition (voir Sager 1990 : 71-79) .

---

[5] Les observations de Guilbert (1965) et de Benveniste (1966) ne portent cependant pas uniquement sur les termes mais sur les unités nominales complexes. L'extension de leurs affirmations aux termes complexes doit cependant se faire avec précaution puisque ces derniers font parfois l'objet, en contexte, d'élision et de variations.

Les terminologues ont exploité cet aspect afin de mettre à la disposition des terminologues une liste de structures potentielles. Auger (1979 : 16-17) présente sept matrices qui décrivent la formation des unités complexes en français[6]; une telle description est aussi disponible pour la langue anglaise dans Sager *et al.* (1980 : 265-276). Comme le faisait remarquer Sager dans un article précédent (1979 : 47) au sujet des formes issues de la juxtaposition en anglais (*fire extinguishing method*) et en allemand (*Feuerlöschverfahren*), les unités n'ont pas toujours recours à des modes de formation qui respectent la syntaxe de la langue. Cette particularité n'empêche cependant pas leur description sous forme de matrices.

Il est cependant important de noter que cet aspect syntagmatique et la possibilité de décrire les structures potentielles des termes ne résolvent pas le problème du découpage du terme. En effet, comme nous l'avons déjà mentionné, c'est justement ce recours à des constructions syntagmatiques régulières qui ne permet pas de distinguer le terme du non-terme d'un point de vue de la syntaxe. La délimitation en contexte à gauche de l'unité terminologique est relativement simple. Comme le fait remarquer Boulanger :

> « La présence de déterminants (articles, adjectifs possessifs ou démonstratifs, etc.) ou d'adjectifs préposés écarte les possibilités d'expansion du syntagme de ce côté [gauche] de la chaîne du discours. Ils marquent la limite sénestre du découpage du syntagme.

---

[6] Kocourek (1985 : 97, 1991 : 139) propose une liste de matrices plus exhaustive, mais ces dernières sont couvertes par la description de Auger (1979 : 16-17).

> À noter toutefois quelques exceptions : haute tension, basse pression, etc.»
> Boulanger (1979 : 173)

On peut donc avoir recours, pour le découpage à gauche, à la description des unités complexes à l'aide de matrices pour identifier les unités qui ne peuvent faire partie du terme complexe et identifier leur extrémité gauche. Vers la droite, le problème est plus complexe puisqu'« *on ne peut guère limiter de la même manière cette expansion vers la droite* » (Boulanger 1979 : 174).

Ce fait, mis en lumière par Louis Guilbert dans sa thèse (1965 : 272), est démontré par Pierre Auger (1979). Ce dernier illustre que les matrices sont récursives et qu'on ne peut limiter l'expansion vers la droite à l'aide de critères formels (Auger 1979 : 18-20). Sager *et al.* (1980 : 273) présentent aussi quelques exemples de matrices récursives où des termes complexes sont repris dans la construction de termes encore plus complexes : *(((IBM System/360) (operating system)) assembler programme).*

La linéarité, en syntaxe, de la construction déterminé-déterminant[7] (Benveniste 1966 : 172; Kocourek 1991 : 141) ne laisse pas nécessairement présager de la structure hiérarchique

---

[7] Cette généralisation est valide pour le français; toutes les langues ne respectent pas nécessairement cet ordre. L'anglais, par exemple, utilise dans la majorité des cas une structure inverse déterminant-déterminé (Sager *et al.* 1980 : 268; Sager 1990 : 73). Lotte (1981 : 23) suggère qu'en russe les deux constructions existent, mais que la construction déterminé-déterminant est plus fréquente. Par contre, la linéarité du processus de détermination est tout de même respectée.

des rapports internes au syntagme. Comme le démontre Kocourek (1991 : 141), qui reprend l'exemple de *gardien d'asile de nuit* de Benveniste (1966 : 173), la linéarité observée en surface ne laisse rien paraître de l'ambiguïté potentielle et des problèmes de découpage de l'unité. On peut ainsi proposer deux découpages potentiels pour la forme précédente : [[*gardien d'asile*] *de nuit*] ou [*gardien* [*d'asile de nuit*]]. Dans un cas comme celui-ci, le recours à des critères uniquement formels ne rend pas possible l'identification de la hiérarchie sous-jacente.

D'un point de vue de la réalisation des unités en discours, Boulanger (1979 : 182), Vinay (1979 : 91) et Kocourek (1991 : 138) font remarquer que le terminologue peut utiliser les marques graphiques (guillemets, ponctuation, etc.) ou typographiques (italique, gras, souligné, etc.) à titre d'indices lui permettant de délimiter le terme complexe dans l'ensemble de la phrase.

2.1.3.1.2 Le critère sémantique

De tous les critères énoncés, le critère sémantique semble celui qui caractérise le plus l'unité terminologique, ce qui fera dire à Benveniste :

> « C'est toujours et seulement la nature du désigné qui permet de décider si la désignation syntagmatique est ou n'est pas une synapsie : *valet de chambre* en est une, mais non *coin de chambre* »
> Benveniste (1966 : 173)

Guilbert fait la même constatation en insistant, à l'instar de Benveniste (1966 : 173), sur le caractère permanent de la relation

entre l'unité syntagmatique et un signifié unique (1965 : 275-276). C'est donc l'existence de ce lien qui distingue le syntagme de discours du syntagme terminologique. Dans une optique terminologique, Sager suggère que seul le recours à des connaissances, générales ou spécialisées, permet de distinguer les termes complexes :

> « In practice, terminologists face difficulties with the recognition of terminological units in running text, which can generally only be resolved by general or special purpose knowledge. »
> Sager (1990 : 61)

Afin de permettre aux terminologues d'utiliser ce critère dans leur travail quotidien, Kocourek (1991 : 148-149) suggère de vérifier l'existence d'une définition pour le syntagme dont le statut terminologique est douteux. L'existence d'une telle définition viendrait confirmer la permanence du rapport de signification.

Selon Guilbert (1965 : 276), il est impossible d'insérer un élément lexical nouveau entre les éléments constituants du syntagme. Cette impossibilité constitue, selon cet auteur, un bon indice de la lexicalisation (Guilbert 1965 : 276; Boulanger 1979 : 176; Martinet 1979 : 186) d'une unité complexe. Martinet (1979 : 186) illustre ce phénomène[8] avec la forme *chemin de fer*, qui ne peut devenir *chemin creux de fer forgé* sans que le « synthème » s'en retrouve « cassé ». La modification d'un des éléments du terme complexe rendrait donc impossible le rétablissement du lien avec la définition. Cette cohésion

---

[8] Il est important de signaler que les remarques de Martinet (1979) portent sur des unités qu'il nomme *synthèmes* et non sur les unités terminologiques.

sémantique de l'unité reflète la cohésion syntaxique de l'unité (Sager *et al.* 1980 : 267).

Pour sa part, Boulanger (1979 : 179) suggère la possibilité d'une substitution en contexte de l'unité complexe par une forme synonyme simple : la substitution référentielle. Cette technique est d'ailleurs employée par Cruse (1986 : 102) pour démontrer l'importance de la cohésion sémantique entre les unités lexicales. La technique de la substitution référentielle, reprise par Kocourek (1991 : 145), permettrait au terminologue, dans les cas où un synonyme existe, de vérifier la cohésion du syntagme et de le délimiter dans la phrase. Dans un même ordre d'idées, Rondeau (1984 : 80) suggère un critère additionnel, le recours à la traduction. Les unités complexes représentées par une seule unité dans une autre langue peuvent être considérées comme des termes.

Un autre aspect sémantique intéressant des unités terminologiques complexes est leur imprévisibilité ou l'impossibilité de déduire leur sens à partir de l'addition des sens de leurs parties (Lotte 1981 : 10). Kocourek suggère les exemples suivants (1991 : 146-147): *chou de Bruxelles*, *gardien de nuit*. Après un long exposé des facteurs contribuant à l'imprévisibilité des unités complexes, il dresse une liste de facteurs contribuant à ce phénomène : structure hiérarchique du syntagme, rapports sémantiques sous-jacents entre les constituants, ambiguïté ou emploi figuré des mots constitutifs, caractère exocentrique du syntagme et emploi vraiment idiomatique (unique) d'un constituant (Kocourek 1991 : 148). Il conclut sa démonstration par la constatation suivante :

> « On peut constater que, toutes choses étant égales par ailleurs, plus un syntagme est imprévisible, plus de chances il a d'être considéré comme lexicalisé. »
> Kocourek (1991 : 148)

Guilbert (1965 : 278; 1981 : 219) souligne que le procédé de dénomination à l'aide de l'unité syntagmatique complexe répond aux besoins de l'inventeur de situer son invention dans un domaine tout en mettant en évidence ses caractéristiques. Cette idée est reprise par Roger Goffin (1979 : 166) et Drodz (1981 : 125) qui proposent un critère taxinomique qui repose sur la possibilité d'insérer la notion dans un réseau de notions sans conflit.

Cette constatation n'est pas sans rappeler celle d'Auger (1979 : 23-24) et de Sager (1990 : 57), qui font remarquer que les séries de termes voisins[9], du point de vue des unités lexicales utilisées, ont pour objet de mettre en évidence les similarités et les différences entre les notions (*band select switch, tape select switch*). On rejoint ici aussi l'idée derrière la compatibilité systémique de Koucourek (1991 : 146), selon laquelle on vérifie le rapport entre le syntagme et les autres unités non synonymes, mais apparentées, d'un même domaine. Pierre Lerat (1995 : 52) fait aussi mention du même phénomène d'insertion du terme dans une série morphologiquement apparentée au sein d'un domaine spécialisé.

Un autre phénomène, celui de la cooccurrence, qui s'exprime sur le plan de la syntaxe, est avant tout régi par des phénomènes sémantiques. Sager (1990 : 19) et Lerat (1995 : 52) illustrent le fait

---

[9] Auger (1979 : 23) qualifie ces séries de « cohérentes sur le plan onomasiologique ».

que les cooccurrences des termes sont sujettes à des restrictions sémantiques beaucoup plus fortes que celles des mots.

À titre d'exemple, Lerat (1995 : 52) cite le cas du mot *action*, qui entretient une relation morphologique et sémantique avec *agir* dans la langue générale. Ce lien devient sémantiquement non significatif lorsqu'on oppose des formes de la langue juridique comme *agir en justice* et *acte notarié*. Le mot *action* accepte aussi des cooccurrences avec une liste non fermée de verbes en langue générale; par contre les cooccurrents du terme sont prévisibles : *exercer, intenter*, etc.

2.1.3.1.3 Le critère quantitatif

Repris sous le nom de *critère d'usage* par Kocourek (1991 : 149-150), ce critère vise essentiellement à vérifier la « consécration du syntagme par la communauté terminologique » (Boulanger 1979 : 181). Ce critère n'est d'aucune utilité par lui-même, il doit être combiné aux précédents :

> « Le terminologue doit donc ajouter aux critères formels et aux critères sémantiques un autre instrument qui lui permettra, dans une certaine mesure, de distinguer les unités de discours et les hapax des unités terminologiques lexicalisées ou en voie de lexicalisation.
>
> Cet instrument est l'étude de la fréquence des occurrences d'une unité et de la répartition de ces occurrences dans des sources différentes et représentatives sur le plan du discours de l'ensemble du domaine qui fait l'objet de la recherche. »
> (Rousseau 1979 : 35)

On se doit donc de considérer deux choses : le nombre d'occurrences d'une forme complexe et la répartition de ces occurrences dans l'ensemble d'un domaine. Sans une étude de la répartition des occurrences, des observations sur la fréquence sont d'une utilité très limitée. On est ainsi à la merci des tics d'auteurs qui pourraient permettre d'attester, de façon erronée, une unité terminologique.

Le recours à la répartition a pour objet de pallier cette lacune et de s'assurer qu'une forme est effectivement employée dans de nombreuses sources indépendantes les unes des autres. Les occurrences de l'unité complexe se doivent cependant de respecter une uniformité dans l'ensemble du corpus. Comme le soulignait Guilbert (1965 : 276), « [l]a lexicalisation a peu de chances de se produire si les limites de l'unité varient d'un énoncé à l'autre et si l'unité se réalise sous des formes grammaticales diverses ».

Non seulement la fréquence et la répartition d'une unité terminologique peuvent être des indices intéressants, mais une étude de la fréquence de certaines matrices de formation syntagmatique peut aussi s'avérer utile (Pearson 1998 : 125-128). Par exemple, la proportion du recours à la composition savante et à la création d'unités syntagmatiques est élevée dans les domaines scientifique et technique. Ces indices peuvent être traduits sous forme d'heuristiques et pris en charge par l'ordinateur afin de vérifier si certaines constructions sont plus fréquentes. Pearson (1998 : 125-128) tire d'ailleurs profit de ces modes de formation fréquents en

utilisant quelques matrices de formation types pour procéder au recensement des termes dans trois corpus[10].

2.1.3.1.4 Le critère pragmatique

Selon Cabré (1998 : 76), l'aspect le plus marquant qui distingue les termes des mots est la pragmatique. Les termes sont principalement employés par les spécialistes d'un domaine[11] alors que les restrictions d'usage sur les mots sont moins importantes. Dans le même ordre d'idées, Cabré (2000b : 15) insiste sur le fait que plus un document est spécialisé, plus il est intéressant, d'un point de vue terminologique, puisque la terminologie y est systématique, concise et précise. Le texte spécialisé constitue donc une source privilégiée d'information terminologique :

> « Un linguiste concerné par la description des unités terminologiques doit les chercher dans les productions orales et écrites des spécialistes. »
> Cabré (2000b : 14)

Cette observation rejoint celle faite par Pearson (1998 : 36-39), selon qui les termes ont plus de chances d'être présents dans certaines situations de communication. Elle regroupe les situations de

---

[10] L'auteure postule cependant que la répartition des matrices de formation en fonction des corpus n'est pas uniforme et elle met de l'avant une liste de matrices par corpus analysé. Ces listes sont établies manuellement.

[11] Cette affirmation mérite d'être nuancée puisque les termes sont aussi employés par des non-spécialistes. Cependant, on peut affirmer que les termes, avant d'être utilisés par des non-spécialistes, ont été proposés par des spécialistes d'un domaine qui ont répandu leur usage dans la collectivité.

communications sous quatre catégories : (1) expert/expert, (2) expert/initié, (3) pseudo-expert/non-initié et (4) enseignant/élève[12]. Toujours selon l'auteure, les situations 1, 2 et 4 sont les situations privilégiées pour l'utilisation d'une terminologie bien définie.

Dans ces contextes, l'utilisation de la terminologie est régie par les connaissances des acteurs et les variations de sens en contexte sont minimes. Par contre, dans la situation 3, la terminologie peut être présente, mais un flottement dans la signification et la précision des unités terminologiques est beaucoup plus probable. Selon Pearson (1998 : 39), à l'exception de la dernière situation évoquée, il s'agit de sources riches en termes potentiels.

2.1.3.1.5 Récapitulatif

Déjà en 1979, Roger Goffin affirmait au sujet des critères qu'il venait de présenter :

> « Les quatre critères énoncés – formel, sémantique, quantitatif et taxinomique – se complètent et se conjuguent. Aucun de ces critères ne permet à lui seul de faire le départ entre l'unité synaptique et le syntagme occasionnel; aucun ne permet à lui seul de déterminer de manière objective où se place la césure dans le syntagme complexe non figé. »
> Goffin (1979 : 167)

Les critères pris séparément ne sont donc que d'une utilité limitée; c'est dans la mise en commun des diverses propositions que l'on peut songer à différencier le non-terme du terme. Par contre, et

---

[12] La terminologie proposée par l'auteure est *expert/expert, expert/initiate, relative expert/uninitiated* et *teacher/pupil* (1998 : 36-39).

c'est ce que Goffin (1979 : 167) ne mentionne pas, la simple addition de tous ces critères ne peut permettre d'affirmer sans hésitation qu'une unité lexicale complexe possède un statut terminologique.

On peut donc se demander, plus de deux décennies plus tard, si les critères permettant de distinguer le terme du non-terme sont mieux définis. Selon Kyo Kageura, le débat reste entier :

> « Although there seems to be no consensus what a term is or about the criteria or procedures for actually recognising terms, it seems that the viewpoint from which terms are understood is basically shared. »
> Kageura (1995: 245)

Il est intéressant de noter que Jean-Claude Boulanger est du même avis :

> « La première praxis de la terminologie fut donc une tâche à dominante lexicographique. La démarche consignatrice répondait aux tentatives pour résoudre l'énigme du terme, associé à son inévitable revers, le non-terme. Plusieurs efforts ont été consentis afin de tracer une ligne de démarcation nette entre les deux, sans jamais réellement aboutir. »
> Boulanger (1995 : 195)

Comme l'affirmera Kageura, quelques années plus tard, les travaux en ce sens ne sont cependant pas inutiles :

> « In general, concrete studies dealing in some way or other with the borderline between terms and non-terms, based on viewpoints applicable both to terms and to non-terms, are important for a "theory of terminology".

> But yet, even in studies concerned with the "borderline", we are still faced with the problem to what extent the observation can be generalised with respect to terms. »
> Kageura (1999 : 27)

Les recherches ayant pour but la description de la frontière entre le terme et le non-terme sont donc utiles pour le travail terminologique. Le danger réside dans le fait que les observations effectuées sur les termes ne sont parfois valables que pour ces derniers.

Les critères présentés dans la présente section ne sont pas infaillibles; il faut donc les considérer comme des indices du statut terminologique d'une unité lexicale complexe. Dès que l'on considère le problème sous ce jour, on peut aller de l'avant avec des méthodologies qui utilisent de tels critères tout en étant prudents et pleinement conscients que la qualité des résultats obtenus peut varier considérablement. C'est d'ailleurs ce qui a permis aux travaux de terminographie de se mettre en branle sans une description entièrement adéquate de la différence entre le terme et le non-terme.

2.1.3.2 Le découpage du terme et TAL

Les paragraphes qui suivent abordent l'utilisation des critères présentés précédemment dans le cadre du traitement automatique de la langue. Les avantages et les limites des critères sont exposés à la lumière des possibilités et des restrictions qu'impose le recours à l'ordinateur.

2.1.3.2.1 Le critère formel

Comme nous l'avons vu, le critère formel ne permet pas, de façon catégorique, de distinguer le terme du non-terme. Il est donc douteux, voire impossible, que l'ordinateur, en utilisant les mêmes connaissances que l'humain, puisse réussir là où le terminologue échoue.

Cependant, la systématicité de l'ordinateur favorise le recensement des unités nominales complexes dont la structure syntagmatique correspond à celle qu'adoptent les termes. Le problème non résolu, d'un point de vue terminologique et informatique, demeure l'imposition d'une contrainte sur la longueur des expansions potentielles.

Tel que le faisait remarquer Boulanger (1979 : 173) au sujet du français, on peut aisément identifier la limite extrême gauche d'une unité nominale complexe. Le grand problème réside dans le fait que les matrices qui ont été utilisées pour décrire les règles de formation syntagmatique des termes sont récursives et qu'il est encore aujourd'hui impossible de limiter cette récursivité à l'aide de critères purement formels (Auger 1979 : 18-20). Cette constatation, valide pour le français, s'applique aussi à l'anglais et à d'autres langues, puisque le seul critère permettant de déterminer où se termine l'expansion en est un de compréhension de l'unité en discours (Kocourek 1991 : 140).

Une description des matrices de formation syntagmatique des termes peut par contre être utilisée dans le cadre du traitement automatique de la langue. Cependant, lorsque ces descriptions sont en place, on doit artificiellement et arbitrairement limiter la

récursivité des matrices afin d'éviter de recenser des chaînes dont la cohésion interne serait, pour reprendre le terme de Martinet (1979 : 186), « cassée ».

La cohérence onomasiologique (Auger 1979 : 23-24) observée dans un domaine a pour conséquence que les unités terminologiques complexes ont, d'un point de vue formel, une tendance générale à la réutilisation des matériaux déjà existants. On observe donc, dans un domaine particulier, la naissance d'unités syntagmatiques qui réutilisent des unités qui appartiennent déjà à ce même domaine pour créer des unités plus longues[13]. Les néologismes ont aussi recours aux modes de formation plus classiques de la morphologie en faisant appel à la dérivation et à la composition. Cette dernière a une tendance très marquée pour l'utilisation des racines « savantes » et elle est donc, dans une certaine mesure, prévisible.

Cette prévisibilité peut être exploitée dans le cadre d'un processus d'acquisition automatique des termes et de nombreux chercheurs[14] ont d'ailleurs proposé des techniques de décomposition qui reposent sur des heuristiques. Cette étape de décomposition des termes a comme objectif, dans la majorité des cas, de structurer les résultats de l'acquisition en fonction des notions de tête et d'expansion.

---

[13] Il s'agit d'ailleurs d'une observation faite par Nakagawa et Mori pour la langue japonaise (1998 : 64).

[14] Voir à ce sujet les descriptions, dans ce même chapitre, des travaux de Bourigault (1994b), de Lauer (1994), d'Assadi et Bourigault (1996), d'Oueslati (1999) et de Plante *et al.* (2000).

2.1.3.2.2 Le critère sémantique

Le recours à la sémantique demeure toujours un grand défi pour les ordinateurs. Contrairement à l'analyse syntaxique, dont les techniques informatiques sont relativement bien maîtrisées, la prise en charge de l'information sémantique en est encore à ses balbutiements (Gazdar et Mellish 1989 : 279). Il y a, bien sûr, quelques travaux prometteurs entrepris dans le cadre de l'acquisition automatique des termes (Lauer 1994; Jacquemin 1997 et 2001, Maynard et Ananiadou 2001).

Le développement de ressources sémantiques sous format informatique et leur accessibilité suscitent de plus en plus de travaux ayant recours à la sémantique. On peut d'ailleurs s'attendre à obtenir, comme ce fut le cas pour la traduction automatique, des résultats intéressants dans des domaines restreints. Il faudra cependant attendre encore de nombreuses années pour envisager le traitement automatique de la langue indépendamment des domaines. De par les limites des descriptions sémantiques disponibles, le recours à la sémantique demeure donc encore problématique.

Même si des ressources sémantiques sont mises à la disposition des ordinateurs, on peut s'interroger sur l'apport de ces derniers au découpage au terme. Il est en effet encore difficile d'envisager qu'un ordinateur puisse statuer, même avec l'accès à un thésaurus, sur le statut terminologique d'une unité nominale complexe. Comme nous l'avons mentionné, l'évaluation de la stabilité de la référence n'est pas toujours une tâche facile pour un humain qui possède une vaste connaissance du monde dans lequel il fonctionne. On peut donc envisager les problèmes auxquels serait confrontée la machine.

Le travail nécessaire à l'élaboration de ressources sémantiques va aussi à l'encontre de l'objectif premier du recours à l'informatique dans le cadre du travail terminologique, qui est d'accélérer le travail de dépouillement du terminologue. Si la conception de dictionnaires sémantiques devient une étape nécessaire au processus d'acquisition des termes, le terminologue devra alors y consacrer une bonne partie de son temps et le gain sera alors négligeable, sinon nul. Par contre, lorsque des ressources sémantiques très élaborées seront disponibles[15], elles pourront probablement permettre au terminologue d'obtenir de meilleurs résultats et de réduire le temps nécessaire à l'analyse de la documentation.

2.1.3.2.3 Le critère quantitatif

Contrairement aux critères qui précèdent, l'observation des phénomènes reliés à la fréquence et à la répartition d'un terme complexe dans un corpus ne constitue pas un obstacle pour l'ordinateur. Il s'agit plutôt d'une tâche qui se prête bien à l'informatisation. L'ordinateur possède un avantage indéniable sur l'humain lorsqu'il s'agit d'étudier les variations de fréquence ou de répartition des termes complexes dans un vaste corpus.

En plus de faciliter la manipulation d'une énorme quantité de données, le recours à l'outil informatique permet aussi de mettre en lumière des variations de fréquence qui sont très fines et qui ne sauraient être observées par le terminologue. De plus, le recours à une analyse probabiliste permet de discriminer entre les variations

---

[15] Certains recherches s'effectuent présentement à l'aide de WordNet (Fellbaum 1998) mais l'exploitation des informations sémantiques dans le cadre du travail quotidien du terminologue ne saurait être envisagée.

significatives et non significatives d'un point de vue statistique. La gamme des variations observables peut ainsi être divisée en bandes plus étroites et on peut donc déterminer avec précision l'ampleur de la variation de fréquence jugée intéressante. Combiné avec le critère formel, le critère quantitatif peut permettre à l'ordinateur d'isoler, en discours, des unités nominales complexes dont le potentiel terminologique est fort.

2.1.3.2.4 Le critère pragmatique

L'identification par un ordinateur, de façon entièrement automatique, de caractéristiques extérieures aux corpus comme les interlocuteurs (auteur et public cible) ou encore le domaine d'un texte, relève de l'utopie. Il est cependant envisageable que l'ordinateur puisse exploiter ce type d'information si elle a été mise à sa disposition par l'humain.

Comme le suggère Pearson (1998 : 60-61), des informations telles que l'auteur, le public cible, le degré de technicité, le domaine, etc. doivent être prises en considération lors de l'élaboration du corpus. Une adéquation entre le corpus et les objectifs de recherche permettra de constituer un corpus riche en termes. Ainsi, les critères pragmatiques ne rendent pas possible l'évaluation du statut terminologique d'une unité lexicale complexe, mais ils permettent de favoriser la présence d'unités terminologiques au sein d'un corpus.

2.1.3.3 Candidats-termes

La question qui se pose à la suite de la discussion qui précède est, sans contredit : peut-on vraiment isoler automatiquement un terme en contexte sans craindre de se tromper? Les multiples travaux

en acquisition automatique de termes ont démontré que ce n'est pas possible et que les listes de termes retenus dans un corpus contiennent des unités qui ne possèdent pas de statut terminologique. La présente section aborde le concept de *candidat-terme* mis en place pour pallier cette impossibilité qu'a la machine de déterminer le statut terminologique d'une unité lexicale.

L'objectif des recherches en acquisition automatique des termes est d'établir des listes exhaustives de termes contenus dans des corpus textuels spécialisés. Cependant, l'identification automatique des termes pose, de par la nature référentielle particulière des termes et leur réalisation morphosyntaxique tout à fait comparable à celle des syntagmes de discours, des problèmes majeurs. Par exemple, les syntagmes nominaux dans les phrases *Il parle à cette* ***fille de programmation*** (non-terme)[16] et *Il utilise un* ***langage de programmation*** (terme) ont une structure syntagmatique identique, mais un statut terminologique très différent.

Cette réalisation en discours identique de groupes nominaux de nature terminologique et des syntagmes nominaux non spécialisés ne fait que compliquer la donne pour les ordinateurs. Étant donné la très grande difficulté qu'ont ces derniers à gérer les informations de type sémantique, il est donc impossible, pour le moment, de penser à distinguer automatiquement les termes des non-termes sans une intervention humaine non négligeable.

Afin d'adopter une terminologie représentative des résultats d'un système d'acquisition automatique, nous devons considérer que le système fournit à l'humain une liste de CT et non une liste de

---

[16] Cet exemple est emprunté à David et Plante (1990 :144).

termes. L'adoption d'une telle terminologie peut sembler trop prudente à certains, mais nous croyons que les systèmes d'acquisition automatique des termes s'inscrivent dans une chaîne de travail interactive[17] où le terminologue est celui qui porte un jugement sur la valeur terminologique d'une proposition faite par l'outil. Ainsi, il incombera au terminologue de distinguer, dans l'ensemble des résultats proposés par le système, les termes, ou les CT valides, des propositions sans intérêt terminologique.

L'adoption de la notion de CT ne remet cependant pas en cause l'objectif des systèmes, qui est de repérer les termes. Pour le moment, nous désirons, comme la majorité des auteurs s'intéressant à la reconnaissance automatique des termes, nuancer nos propos. La terminologie que nous adoptons rejoint celle mise de l'avant par Bourigault (1994a) et qui semble reprise par la majorité des chercheurs oeuvrant dans le domaine : Habert *et al.* (1997), Jacquemin (1997), Bouveret (1998), et Daille (1999).

2.1.3.3.1 Mesure d'efficacité

Comme nous l'avons décrit plus tôt, une liste de CT est composée à la fois de termes et d'unités qui ne sont pas des termes. Ces dernières, malgré qu'elles n'ont pas de statut terminologique bien défini, sont tout de même intéressantes pour le terminologue dans la

---

[17] L'étape d'acquisition des termes est entièrement automatique, mais la chaîne de travail est qualifiée d'*interactive* dans la mesure où le terminologue valide les résultats obtenus automatiquement par le logiciel.

majorité des cas[18]. Nous ne pouvons cependant pas en tenir compte dans un cadre d'acquisition automatique de termes. Ces formes sont cependant généralement conservées pour l'évaluation et la validation des résultats.

C'est pour cette raison que certains indices ont été mis en place afin d'évaluer la performance des systèmes et leur capacité à isoler le *pertinent* du *non-pertinent*. Afin de permettre de comparer les résultats obtenus par les divers systèmes d'acquisition automatique de termes au travail d'un terminologue, on peut envisager le scénario présenté par la figure 1.

L'ensemble TR représente une liste de termes de référence compilée par un terminologue ou un spécialiste; c'est cette liste qui sera utilisée pour l'évaluation des performances du logiciel. La liste des termes de référence contient la section TR- qui correspond aux termes de la liste de référence qui n'ont pas été identifiés par l'outil.

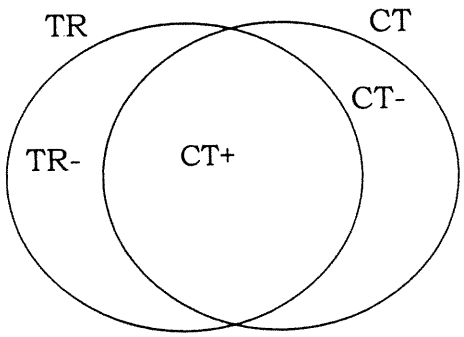

Figure 1. Mesure d'efficacité des logiciels

[18] Par exemple, on y retrouve souvent l'expression du phénomène de collocation entre un verbe et un substantif d'un domaine. Alors que la tendance de ces deux unités peut passer inaperçue aux yeux d'un terminologue qui n'a pas recours à un outil informatique, le logiciel ne manquera pas de le mettre en évidence.

L'ensemble des CT identifiés par le logiciel se nomme CT. Les CT dont le statut terminologique est confirmé par un terminologue ou un spécialiste sont regroupés dans le sous-ensemble CT+ alors que les CT ayant été retenus par erreur par le logiciel se trouvent dans l'aire étiquetée CT-.

Parmi les indices les plus fréquents retenus pour évaluer les performances à l'aide de divers sous-ensembles illustrés par la figure 1, on a souvent recours aux notions de *rappel,* de *précision*, de *silence* et de *bruit.* Ces indices nous viennent du domaine de l'indexation et de la recherche documentaire où ils sont utilisés depuis de nombreuses années[19].

Le *rappel (R)* correspond à la proportion des réponses pertinentes extraites par un système donné par rapport à l'ensemble des réponses pertinentes possibles. Ainsi, si un corpus contient 100 termes attestés par un terminologue (TR) et qu'un système identifie 150 CT, dont 95 qui apparaissent dans la liste validée par le terminologue (CT+), on dira que le taux de rappel est de 0,95 (95/100). Cet indice nous permet donc de nous concentrer uniquement sur l'intersection entre la liste qui nous sert de référence et la liste générée par le logiciel.

$$R = \frac{CT+}{TR}$$

---

[19] Dans le cadre du projet ATTRAIT (Atelier de travail informatisé du terminologue), le RINT (2001) a recours au silence et au bruit pour l'évaluation de logiciels.

À notre avis, la majorité des logiciels obtiennent une bonne performance lorsque l'on examine leurs résultats selon l'angle du rappel. La majorité des outils présentement disponibles sur le marché ont un taux de rappel supérieur à 90 %. On peut donc considérer que les systèmes sont à même de repérer les termes dans les textes et d'en dresser une liste.

Par contre, cette liste devient moins intéressante dès que les termes sont dilués dans une liste de CT erronés. Le terminologue se doit alors de procéder au dépouillement terminologique de la liste des CT extraits et l'attrait pour un système automatique en est grandement diminué. Le gain de productivité que le système automatisé laisse miroiter est ainsi mis en veilleuse par le temps nécessaire au nettoyage des données.

La *précision (P)* correspond au nombre de réponses pertinentes (CT+) identifiées par un système donné par rapport à l'ensemble des réponses (pertinentes ou non) identifiées par le même système (CT). Reprenons l'exemple cité plus haut où un corpus contient 100 termes. Le logiciel a, à la suite de son analyse, recensé 150 CT, dont 95 CT+. On dira alors que la précision est de 0,63, ce qui correspond au nombre de termes identifiés (95) sur l'ensemble des CT identifiés (150).

$$P = \frac{CT+}{CT}$$

Pour sa part, le *silence (S)* compare le nombre de termes d'une liste de référence établie par un terminologue ou un spécialiste qui n'ont pas été identifiés (TR-) par un logiciel avec le nombre total de termes dans cette même liste de référence (TR). Ainsi, en reprenant

l'exemple cité plus haut, si 5 termes qui apparaissent dans la liste de référence n'ont pas été relevés par le logiciel, on peut parler d'un silence de 5/100 (0,05).

$$S = \frac{Tr -}{Tr}$$

Malheureusement, le silence n'est généralement pas pris en considération dans le cadre des recherches en acquisition automatique des termes. Il s'agit cependant d'un indice très pertinent et il serait intéressant d'entreprendre des travaux visant à évaluer les impacts de la recherche d'une précision accrue sur le silence. Une telle démarche permettrait de vérifier combien de bons CT sont mis de côté au profit d'une augmentation de la qualité des résultats.

Enfin, le *bruit (B)* évalue le nombre de CT extraits par le logiciel qui sont absents de la liste de référence (CT-) par rapport au nombre total de CT extraits par le logiciel (CT). En reprenant notre exemple, on obtient une valeur de 55 CT erronés (CT-) alors que le total de CT proposés est fixe à 150; le bruit est donc de 55/150 (0,37).

$$B = \frac{CT -}{CT}$$

Cet indice recouvre donc la portion négative des CT et peut aussi être exprimé sous la forme $B = 1\text{-}P$ à partir du moment où la précision est une valeur connue.

2.2 Terminologie et informatique

Afin de bien comprendre d'où tire son origine le domaine de l'acquisition automatique des termes, il est important de passer en revue les faits saillants ayant conduit à l'informatisation graduelle du travail du terminologue. De l'informatique lourde à la micro-informatique, les sections suivantes illustrent l'importance de plus en plus grande qu'a su occuper l'informatique en terminologie.

2.2.1 Historique

2.2.1.1 La création des banques de terminologie (1930-1980)

Il est intéressant que noter que Wüster (1981 : 57), lors de l'élaboration de la théorie de la terminologie, décrit l'informatique comme un domaine qui partage des intérêts communs avec la terminologie. Il souligne que la terminologie fournit à l'informatique des moyens pour la classification des connaissances et des concepts. Cette intersection dans les activités de la terminologie et de l'informatique est aussi reprise par Sager (1990 : 5-7) et Cabré (1998 : 100-106). Selon Cabré (1998 : 101), entre les années 1930 et 1960, la terminologie fournit donc à l'informatique une réflexion théorique sur les concepts et leur gestion.

C'est au cours de cette période, qui correspond à celle de la mise en place de la théorie terminologique et de l'élaboration de la méthodologie terminologique, que s'affirmera le caractère systématique de la terminologie. Le rôle joué par l'informatique au cours de cette période est mineur, mais ce domaine saura tirer profit, dans les années qui suivront, de l'aspect systématique dont a hérité la terminologie.

Du début des années 1960 au milieu des années 1970, les terminologues procèdent à l'élaboration de fichiers terminologiques imposants qui sont le fruit de travaux terminologiques sur une multitude de domaines. C'est au cours de cette période que l'informatique passe de discipline connexe à la terminologie à une discipline pouvant permettre de résoudre certains problèmes qui se posent aux terminologues. Les besoins de gestion de la masse de fiches terminologiques accumulées, de concert avec le progrès accompli par l'informatique lourde et l'évolution des techniques documentaires (Cabré 1998 : 28), donnent naissance aux premières banques de terminologie[20].

Les premières banques ont fait leur apparition en Europe avec la mise sur pied en 1963 de DICAUTOM par la Communauté européenne du charbon et de l'acier; ce projet de dictionnaire électronique sera par la suite abandonné pendant quelques années. On a ensuite assisté à la création de LEXIS, en 1966, par l'Office fédéral des langues de la RFA (Bundessprachenamt) et de la banque de terminologie de Siemens (TEAM) en 1967.

Peu de temps après, en 1972, NORMATERM est née de l'effort de l'Association française de normalisation de répertorier et de permettre l'exploitation de la terminologie technique contenue dans les normes françaises et les règlements publiés au *Journal officiel* de France ainsi que dans les normes et les recommandations de l'International Standard Organization (ISO). Les besoins traductionnels de la Commission des communautés européennes raniment le projet DICAUTOM et, dès 1973, EURODICAUTOM est lancé. Il s'agit avant tout d'une banque descriptive n'ayant pas de

---

[20] Voir notamment Auger (1994) et de Schaetzen (1994).

visée normative. Il est intéressant de noter que les banques de terminologie ne se sont pas créées uniquement dans des cadres gouvernementaux, mais aussi dans de grandes sociétés privées comme Siemens, qui perçoivent l'importance de la terminologie pour la gestion de l'information.

Parallèlement, on observe en Amérique du Nord un intérêt pour les banques de terminologie, qui peuvent répondre aux besoins de normalisation et de traduction des organismes gouvernementaux. En 1970, le système BTUM est créé à l'Université de Montréal; ce système deviendra par la suite TERMIUM au sein du Bureau de la traduction du gouvernement canadien en 1975. Au Québec, l'Office de la langue française met sur pied la Banque de terminologie du Québec en 1973.

Au cours de cette période du développement de la terminologie, l'informatique vient aider le terminologue dans la gestion quotidienne des données recueillies et se métamorphose en un outil au service de la terminologie. De science partageant des principes avec la terminologie, l'informatique devient alors un instrument facilitant le travail du langagier.

2.2.1.2 Naissance de la terminotique (1980-1989)

On observe, au début des années 1980, une modification du support utilisé pour la diffusion du contenu des banques de terminologie. À l'origine, ces dernières étaient accessibles par modem et leur accès était restreint à un nombre peu élevé d'usagers. Elles se transforment progressivement en applications de micro-informatique distribuées sur cédérom. Les données terminologiques sont donc directement accessibles à qui possède un micro-ordinateur et tout traducteur ou terminologue est libre de construire son environnement

de bureautique autour de cet accès rapide et convivial à une banque de terminologie. On peut donc dire que l'accès aux banques de terminologie se démocratise et se généralise.

À partir de 1985, on assiste ainsi à la mise en place de réseaux de collaboration entre les divers organismes de normalisation désireux de partager leur travail. Les besoins de diffusion qu'entraîne ce mouvement de collaboration, de concert avec les progrès de la micro-informatique, ont une influence majeure sur la pratique de la terminologie et sur la forme que prendront les banques de terminologie.

Grâce aux nouveaux systèmes d'exploitation tirant profit d'une interface multifenêtres, il est désormais possible d'interroger dictionnaires et banques de terminologie directement à partir d'un traitement de texte ou d'un système de gestion de bases de données utilisé pour consigner les fiches terminologiques. Le travail terminologique s'intégrant dans un environnement de bureautique ayant pour but de faciliter le travail du terminologue, on peut, à partir de ce moment, parler de terminologie assistée par ordinateur. Fruit de cette rencontre entre les outils de bureautique et de la terminologie, le terme *terminotique* commence alors à circuler. Selon André Clas (1987 : 96), la création du terme remonterait à 1983.

Mais qu'est-ce exactement que la terminotique? Selon *Le grand dictionnaire terminologique*, elle se définit de la façon suivante :

> « Ensemble des techniques visant l'utilisation de l'ordinateur comme aide au travail terminologique. »
> OLF (2001)

Dans un ouvrage paru en 1995, Pierre Lerat (1995 : 61) fait l'affirmation suivante : « Quant à la terminotique, nom donné à la bureautique terminologique, en quoi consiste-t-elle, si ce n'est en opérations de lecture d'écrans et d'écriture ? » Cette position est on ne peut plus déroutante si on considère que des travaux visant l'informatisation des tâches terminologiques sont en cours depuis plus de vingt ans et que leur objectif constitue beaucoup plus qu'un simple mécanisme permettant de recopier l'information d'un logiciel à un autre.

Déjà, dans le numéro de la revue *Meta* cité précédemment, on envisage que la terminotique sera plus que la simple mise en place de dictionnaires électroniques et que l'ordinateur prendra en charge certaines tâches répétitives et fastidieuses du terminologue (Paradis et Auger 1987 : 105). À l'instar de la définition proposée précédemment, nous croyons que la notion de terminotique recouvre toute intervention de l'ordinateur à caractère terminologique dans le travail du terminologue. Par intervention à caractère terminologique, nous entendons toute activité dont l'objectif est un produit terminologique : la consultation d'une banque de terminologie, la mise en place d'un fichier terminologique personnel, le dépouillement manuel d'un corpus sur support informatique, le dépouillement assisté par ordinateur, la recherche d'attestations dans Internet, etc.

Les paragraphes qui suivent illustrent l'impact de la terminotique sur le travail quotidien des terminologues. Nous parlons ici volontairement d'*impact* puisque la terminotique, en plus de fournir des outils aux langagiers, modifie leurs habitudes de travail (Schaetzen 1997 : 81).

2.2.1.3 L'explosion de la micro-informatique (1989-2001)

2.2.1.3.1 Le poste de travail

Dans un numéro spécial de la revue *Meta*, André Clas (1987 : 96) évoque le concept de *poste du traducteur*. Ce poste, alors utopique, inclut un logiciel de traitement de texte, un système de messagerie permettant d'échanger des textes, un compteur de mots qui prend en charge la facturation ainsi qu'un logiciel qui compare la terminologie à celle des banques de terminologie personnelles du traducteur et à celle des grandes banques internationales.

Cette idée d'intégration des ressources nécessaires au travail du traducteur sera poussée plus loin par d'autres chercheurs qui mettront en place de tels systèmes[21]. Fait à noter, ces systèmes seront d'ailleurs très proches de celui décrit quelques années plut tôt par André Clas (1987 : 26). Le poste de travail du traducteur sera élaboré en tant que concept; on ne cherchera pas à décrire *le* poste de travail du traducteur, mais plutôt l'ensemble des outils susceptibles d'aider ce dernier à accomplir ses tâches (Macklovitch 1991 : 14). On se concentre donc sur la description des outils et leur intégration. Le traducteur est ainsi libre de choisir son logiciel de traitement de texte préféré et de l'intégrer lui-même à ce poste de travail.

L'apparition des travaux sur le poste de travail coïncide avec l'arrivée sur le marché de la micro-informatique d'environnements d'intégration comme les logiciels *DESQview* et *Windows*[22] (Macklovitch

---

[21] Voir notamment la description du poste de travail du traducteur ou PTT du gouvernement canadien dans Macklovitch (1991 : 10-21).

[22] *Windows* n'était, en 1991, qu'un logiciel d'intégration permettant d'utiliser plus d'un logiciel à la fois et non un système d'exploitation.

1991 : 14, 1993 : 283). Ces logiciels permettaient aux utilisateurs de micro-ordinateurs d'avoir accès en simultané à leur traitement de texte, à une banque de terminologie, à un logiciel de facturation, etc. Il s'agit là, d'un point de vue informatique, d'une grande révolution, qui a suscité un grand intérêt pour les postes de travail.

Au début des années 1990, des travaux ayant pour but la prise en charge de certaines tâches du terminologue se mettent en branle. On note, en autres, les recherches entreprises par le groupe de Pierre Auger (Auger *et al.* 1991 : 121-127) à l'Université Laval. L'objectif principal de ces travaux n'est pas de mettre en place un logiciel qui peut prendre en charge le travail du terminologue, mais de démontrer qu'il est possible, à l'aide de logiciels existants, d'automatiser l'ensemble des étapes de la chaîne de travail terminologique. Ces logiciels pouvant ensuite être intégrés dans un poste de travail du terminologue.

Dans la foulée des travaux sur le *PTT* entrepris au gouvernement canadien, des démarches débutent dès 1991 pour l'élaboration de Latter (Leonhardt 1991a : 257-275, 1991b : 11-12) : l'ATelier du TERminologue. Le Bureau de la traduction du gouvernement du Canada a su reconnaître que la gestion de l'information terminologique est désormais devenue une tâche ardue. Comme le faisait remarquer Sager une année plus tôt (Sager 1990 : 129), le recours à l'ordinateur est dorénavant la seule avenue envisageable pour la gestion de l'information lexicale. Il est donc tout à fait normal qu'on veuille fournir au terminologue un poste de travail adapté à ses besoins et qui l'assiste dans ses tâches quotidiennes.

Latter offre au terminologue les fonctions suivantes (Leonhardt 1991a : 257-275) :

- création de fiches terminologiques unilingues, bilingues ou multilingues,
- gestion des fiches terminologiques (impression, copie, modification et effacement),
- interrogation des fiches personnelles,
- impression de lexiques, vocabulaires, etc.,
- exportation des fiches personnelles vers Termium,
- exportation dans un format propre à Latter pour favoriser l'échange de fiches entre terminologues,
- importation de fiches en format Latter,
- importation de fiches en format Termium.

Avec le temps, l'usage de Latter se généralise au gouvernement du Canada (Leonhardt 1994 : 1-7). La vocation initiale de Latter, qui était de permettre aux terminologues de gérer leurs propres fiches avant de les verser dans Termium, s'est considérablement élargie. Certains terminologues l'utilisent pour modifier leurs fiches déjà stockées dans Termium. Grâce aux modules d'import et d'export entre Latter et Termium, il est désormais possible pour les terminologues d'utiliser le module de gestion de fiches terminologiques de Latter pour procéder à une mise à jour systématique des fiches de Termium.

Une version subséquente de Latter inclut des logiciels qui assistent le terminologue dans sa démarche, depuis le dépouillement terminologique assisté, jusqu'à la publication de lexiques ou de vocabulaires, en passant par la saisie et la gestion des fiches terminologiques (Claude 1996 : 77-80). La publication est prise en charge par le logiciel Publiciel alors que le logiciel Ivanhoé est utilisé pour le dépouillement terminologique (Claude 1996 : 78).

Cette étape de dépouillement est entièrement prise en charge par le terminologue qui insère des balises dans le texte afin de délimiter les termes et les informations qu'il désire conserver sur une fiche terminologique. Latter peut ensuite utiliser ces balises afin d'importer les informations dans une fiche qui peut par la suite être versée dans Termium. On voit donc que l'atelier prend de l'ampleur et devient un élément de plus en plus important de la démarche terminologique.

2.2.1.3.2 Le passage d'outil à acteur

Le concept du poste de travail du terminologue, tout comme celui du traducteur, s'est donc principalement construit autour de logiciels de bureautique que l'on adapte au travail terminologique. Il ne s'agit donc pas de logiciels à vocation terminologique élaborés uniquement pour aider le terminologue dans sa tâche. Les progrès effectués dans le domaine du traitement automatique de la langue entraîneront un changement de paradigme qui contribuera à l'apparition d'outils informatiques conçus spécifiquement pour des applications terminologiques.

La dernière décennie a vu de nombreux chercheurs s'attaquer à la problématique du dépouillement automatique des corpus, l'acquisition automatique des termes[23]. Ces recherches ont pour objet

---

[23] Certains chercheurs (Duchesne 1979 : 485-503; Vinay 1979 : 97) se sont intéressés à l'utilisation de l'ordinateur pour le dépouillement des documents ou ont fait allusion à une telle possibilité dans les années 1970, mais il a fallu attendre au début des années 1990 pour que des recherches plus poussées soient entreprises.

de repousser les frontières de l'automatisation de certaines étapes du travail du terminologue[24]. La section 2.2.2 de la thèse décrit ces travaux. D'instrument, l'informatique devient acteur à part entière dans le travail terminologique, prenant en charge les tâches les plus fastidieuses et répétitives du terminologue. Le terminologue peut ainsi concentrer ses efforts sur le travail linguistique à accomplir.

On peut donc envisager, dans un proche avenir, voir les logiciels d'acquisition automatique de termes faire leur apparition au sein du poste de travail du terminologue. L'ordinateur est donc appelé à jouer un rôle principal dans la démarche terminologique et la prochaine étape de l'informatisation des tâches du terminologue passera par la mise au point des logiciels d'acquisition automatique des termes.

2.2.1.3.3 L'apport d'Internet

Un autre phénomène intéressant observé depuis quelques années est digne de mention. Comme nous l'avons mentionné précédemment, le mode de distribution du contenu des banques de terminologie s'est transformé radicalement vers la fin des années 1980. Ainsi, de banques accessibles en réseau, les banques sont devenues accessibles au langagier par le biais des cédéroms. La croissance fulgurante d'Internet et l'accessibilité accrue des micro-ordinateurs ont à nouveau modifié le mode privilégié de diffusion des banques de terminologie.

---

[24] Voir notamment David et Plante (1990); Enguehard *et al.* (1992); Bourigault (1992a); David (1993); Justeson et Katz (1993); Daille (1994a); Perron (1996) et Jacquemin (1996).

On assiste ainsi à un retour en force des banques de terminologie qui sont désormais accessibles par Internet. La tenue d'une table ronde sur les banques de terminologie en 1996 (RINT 1996) a mis en lumière ce regain d'intérêt pour ces outils ainsi que les nouvelles problématiques[25] qui résultent de leur mise en réseau à l'échelle mondiale. L'accès aux banques à partir d'Internet offre des avantages indéniables puisque, contrairement aux banques sur cédéroms, les données sont mises à jour continuellement. L'utilisateur a donc accès à des données vivantes et n'a pas à attendre la publication d'un nouveau cédérom.

De nombreux utilisateurs peuvent aussi consulter la banque de terminologie en simultané et le logiciel est tout à fait indépendant du type d'appareil utilisé par le langagier. Cette flexibilité accrue ne peut que simplifier la vie des utilisateurs et celle des concepteurs. De nombreux logiciels sont aussi disponibles pour les langagiers qui cherchent une solution à leurs besoins de gestion de fiches terminologiques. Les principaux logiciels ont d'ailleurs été décrits dans le cadre du projet européen POINTER (POINTER 1996)[26].

---

[25] Une séance de la table ronde a été consacrée aux enjeux juridiques que constitue la publication de contextes à l'aide d'Internet, par exemple (RINT 1996 : 10-19).

[26] Le lecteur pourra trouver de l'information supplémentaire sur le projet POINTER dans le site Internet suivant : http://www.computing.surrey.ac.uk/ai/pointer/. Dans la foulée de ce projet, un nouveau projet européen visant la validation des données terminologiques a été lancé : INTERVAL. Le rapport final de ce projet (INTERVAL : 1998) est disponible à l'adresse suivante : http://www.computing.surrey.ac.uk/research/ai/new_interval/final_report/final_report.frames.html

2.2.2 Acquisition automatique des termes

Jusqu'à tout récemment, deux grandes avenues ont été empruntées pour recenser automatiquement les termes : les *modèles linguistiques* et les *modèles statistiques*. De nouvelles recherches entreprises au cours de la dernière décennie tendent à tirer profit de ces deux grandes approches pour proposer des méthodologies qui ne sont ni purement linguistiques, ni purement statistiques (*modèles hybrides*).

Nous avons cru bon, avant de procéder à la présentation des travaux les plus récents, aborder des travaux qui se situent un peu à l'écart de la recherche en acquisition automatique des termes. Nous les regroupons sous l'appellation de *modèles mécaniques*. À l'origine, ces travaux ne portaient pas directement sur l'acquisition automatique des termes, mais plutôt sur l'identification des collocations. Puisqu'ils ont été repris comme point de départ pour d'autres travaux dans ce domaine, nous jugeons essentiel de les décrire.

Les travaux qui font l'objet de descriptions dans les pages qui suivent sont de deux natures : des travaux fondamentaux ne portant pas directement sur l'acquisition automatique des termes, mais qui ont par la suite été utiles dans ce même domaine, et des travaux ayant pour objectif l'élaboration de logiciels d'acquisition des termes. Les systèmes de dépouillement assisté par ordinateur, comme Ivanhoé (Claude 1995 : 78), à l'aide desquels le terminologue prend en charge la tâche d'identification des termes, n'ont pas été retenus. Les postes de travail voués au dépouillement terminologique, comme Adepte (Perron 1996 : 37-46), composante souvent associée au logiciel Nomino, ne font pas non plus l'objet d'une description.

2.2.2.1 Modèles mécaniques

Les travaux décrits ici, que nous qualifions de *mécaniques*, reposent entièrement sur des algorithmes utilisant la force brute des ordinateurs. L'utilisation du qualificatif *mécanique* a pour but de représenter l'aspect très systématique de cette approche entièrement dépourvue de connaissances linguistiques ou statistiques. Leur but est essentiellement de relever des segments de texte qui se répètent à l'intérieur d'un corpus. Ces segments peuvent donc contenir des termes ou une manifestation de tout autre phénomène linguistique répétitif[27], observable en discours.

2.2.2.1.1 Choueka, Klein et Neuwtiz (1983); Choueka (1988)

2.2.2.1.1.1Présentation

Les techniques décrites ici sont tirées des articles de Choueka, Klein et Neuwtiz (1983) et Choueka (1988) portant sur la langue anglaise. L'objectif de ces travaux est de fournir au lexicographe un outil qui rend possible le repérage de ce que l'auteur nomme des *collocations*. Par ce terme, il désigne les séquences de deux ou plusieurs mots consécutifs qui possèdent les caractéristiques d'une unité syntaxique ou sémantique dont la signification ne peut être déduite directement à partir de ses constituants (Choueka 1988 : 612).

---

[27] Les enchaînements privilégiés de verbes et de prépositions, par exemple.

2.2.2.1.1.2Description

Le corpus utilisé dans le cadre de cette étude est d'une taille importante pour l'époque (1983) où elle a été menée. Il est composé de 10 millions de mots tirés de 17 000 articles du quotidien *New York Times* portant sur un éventail très large de sujets. Les en-têtes et les titres des articles ont été exclus du corpus afin d'obtenir un texte le plus homogène possible.

L'approche adoptée est entièrement mécanique et se contente d'identifier les chaînes de caractères qui se produisent côte à côte plus d'une fois à l'intérieur d'un corpus. Pour y arriver, il s'agit de vérifier toutes les occurrences d'une forme f_n[28] avec la forme voisine f_{n+1} afin de vérifier si la séquence $f_n f_{n+1}$ est supérieure à 1 dans le texte. Si c'est le cas, il s'agit de vérifier si la séquence $f_n f_{n+1} f_{n+2}$ se produit également plus d'une fois. Ce processus se poursuit jusqu'à la séquence $f_n f_{n+1} \ldots f_{n+seuil}$. Il s'agit donc d'un algorithme qui balaie un texte mot à mot en examinant l'ensemble de ses voisins à la recherche d'une séquence intéressante. Le processus s'interrompt lorsqu'un nombre déterminé de mots a été atteint (le *seuil*).

2.2.2.1.1.3Commentaires

La méthodologie mise de l'avant par l'équipe de Choueka exclut l'utilisation de techniques linguistiques faisant appel à la syntaxe ou à la sémantique. En effet, l'auteur juge que les techniques à sa disposition au moment de rédiger cet article conduisent à l'obtention de résultats décevants.

---

[28] Où n est le rang ou le numéro d'ordre d'une forme *f* dans le texte; le premier mot recevant le numéro d'ordre 1.

L'auteur ne prévoit pas une utilisation directe des résultats bruts. Ceux-ci doivent être revus par un lexicographe possédant la connaissance nécessaire à l'utilisation des résultats et à leur extrapolation. L'outil proposé s'insère donc dans le cadre d'une station de travail lexicographique et n'en constitue qu'un des maillons.

Les résultats obtenus à l'aide des algorithmes ont été ramenés à 5 listes selon le nombre de mots qui composent les chaînes détectées. On a ainsi obtenu 16 000 chaînes de longueur de 2, 4 800 chaînes de longueur 3, 1 000 chaînes de longueur 4, 200 chaînes de longueur 5 et 10 chaînes de longueur 6.

Étant donné l'approche adoptée, qui repose sur la répétition de segments textuels, le corpus se doit d'être de taille suffisante pour que les résultats soient concluants. L'auteur s'étonne de la précision des résultats obtenus, mais il ne fait malheureusement pas mention d'une évaluation quantitative de cette dernière. Nous ne sommes donc pas en mesure de vérifier à quel point les résultats sont intéressants sur l'ensemble de la liste fournie en fonction de la fréquence des chaînes retrouvées. L'auteur mentionne toutefois que les résultats pour les chaînes de longueur 5 et 6 sont moins intéressants que les autres.

L'utilisation d'un algorithme tel que celui proposé par Choueka *et al.* (1983) conduit à l'identification d'enchaînements répétitifs comme *home run, fried chicken, Magic Johnson, Magic Johnson was, Magic Johnson is, take place, put up, to put up, once upon a time,* etc. Une telle approche entraîne donc, d'un point de vue terminologique,

beaucoup de bruit[29]. L'algorithme est aussi très lent, mais il a cependant l'avantage d'être indépendant des langues, d'être systématique et de repérer l'ensemble des formes répétées dans la mesure où celles-ci ne connaissent pas de variation orthographique.

2.2.2.1.2 Salem (1987); Lebart et Salem (1988, 1994)

2.2.2.1.2.1Présentation

Les travaux sur les segments répétés en français sont issus du *Laboratoire lexicologie et textes politiques* de l'École normale supérieure de Saint-Cloud. Ils ont principalement été décrits dans Salem (1987) ainsi que dans Lebart et Salem (1988, 1994).

L'objectif des auteurs est de représenter un texte, qui est habituellement vu comme un enchaînement de formes simples, comme une succession de formes simples et de *segments répétés* (SR). Ces derniers sont définis comme des « suites de formes graphiques non séparées par un caractère délimiteur de séquence, qui apparaissent plus d'une fois dans ce corpus de textes » (Salem 1987 : 21).

Le recensement des SR dans un corpus résulte en une liste de SR nommée l'*inventaire des segments répétés* (ISR). Cette liste peut être triée selon l'ordre alphabétique, la fréquence, la longueur ou une combinaison de ces critères. Les SR ont principalement été utilisés dans le cadre de l'analyse du discours afin de mettre en évidence des phénomènes reliés à la *langue de bois* politique observée dans divers corpus.

---

[29] Voir la section 2.1.3.3.1 au sujet de la notion de *bruit*.

2.2.2.1.2.2Description

La technique utilisée consiste à repérer, comme pour les travaux de Choueka *et al.* (1983) et de Choueka (1988), les enchaînements de formes qui se répètent plus d'une fois côte à côte dans un texte. Cependant, contrairement aux travaux abordés au point 2.2.2.1.1, les chercheurs du laboratoire de Saint-Cloud utilisent des textes ayant été préalablement lemmatisés. Cette technique favorise l'augmentation de la performance des algorithmes d'extraction puisque des segments qui diffèrent d'un point de vue purement graphique sont regroupés (*le problème financier, les problèmes financiers*).

2.2.2.1.2.3Commentaires

Comme l'indique Salem (1987 : 24), les inventaires des segments répétés constituent un réservoir de renseignements précieux. Nous croyons que, comme dans le cas des travaux de Choueka, les travaux sur les SR sont dignes de mention, car ils ont été repris au sein d'autres études visant l'acquisition automatique des termes.

L'avantage principal des SR est de mettre en lumière des redondances observées en discours. Ces dernières peuvent ensuite être utilisées pour observer les manifestations en surface de constructions syntaxiques sous-jacentes. L'ISR dressé est très flexible et peut servir de matière brute pour des travaux portant autant sur la phraséologie, la cooccurrence ainsi que la terminologie.

le petit, petit chat, chat est, est mort [longueur 2]

le petit chat, petit chat est, chat est mort [longueur 3]
le petit chat est, petit chat est mort [longueur 4]

Comme l'illustrent les exemples de SR qui précèdent obtenus à partir de la séquence *le petit chat est mort* (Lebart et Salem 1994 : 61), la technique utilisée conduit à l'obtention d'une liste d'enchaînements qui n'ont pas toujours un statut linguistique bien défini. L'hypothèse sous-jacente aux travaux sur les SR est que la fréquence des segments permet de mettre en évidence les unités les plus intéressantes au sein d'un corpus.

2.2.2.1.3 Drouin et Ladouceur (1994)

2.2.2.1.3.1Présentation

Les travaux de Drouin et Ladouceur (1994 : 18-28) s'inspirent des techniques proposées par Choueka *et al.* (1983), Choueka (1988), Salem (1987) et Lebart et Salem (1988, 1994) pour le dépistage des enchaînements fréquents. Ces travaux portent sur la langue française et visent l'extraction automatique des unités nominales permettant de cerner le contenu d'un texte : les descripteurs.

2.2.2.1.3.2Description

Les chercheurs utilisent comme point de départ une analyse des segments répétés. L'ISR obtenu à partir d'un corpus technique est ensuite filtré selon divers critères afin d'isoler les unités nominales permettant de cerner le contenu du corpus.

De nombreux indices sont utilisés par les chercheurs afin de réduire le nombre d'entrées retenues par le recensement des segments répétés : la fréquence, la morphologie et une analyse de

similarité entre les SR retenus. Une analyse des contextes d'occurrence des SR est ensuite utilisée afin de trier la liste des SR.

Le premier critère utilisé pour filtrer la liste des SR est la fréquence. Un seuil de fréquence minimal est utilisé afin de ne conserver que les CT qui risquent de représenter de façon significative le contenu du corpus. Le postulat de base de ces travaux est que les unités nominales les plus fréquentes sont représentatives de la thématique d'un corpus.

Les SR qui possèdent une fréquence suffisante sont par la suite soumis à une étape de filtrage morphologique qui a pour but d'éliminer de l'ISR les segments dont la structure de surface ne correspond pas à celle d'une unité nominale complexe potentielle. Une description de ce qui, d'un point de vue morphologique, ne peut pas constituer une unité terminologique complexe est mise en place. Par exemple, il est impossible que les SR qui débutent ou se terminent par un adverbe ou une préposition soient des termes. Les SR qui répondent à ce critère seront donc éliminés.

Les similarités entre les SR de la liste sont ensuite identifiées et les fréquences des SR qui se recoupent sont comparées. Il s'agit d'un deuxième recours à la fréquence dans un objectif de filtrage. Ainsi, si un segment (SR_1) en inclut un plus court (SR_2) et que les deux possèdent la même fréquence, seul le segment le plus long sera conservé. L'identification de l'intersection entre certains SR et la comparaison de leurs fréquences respectives permettent encore une fois de limiter le nombre d'entrées dans la liste des SR.

Quelques indices tirés des contextes d'occurrence des SR sont par la suite exploités afin d'opposer les SR entre eux et de distinguer

les plus aptes à cerner le contenu du corpus. Les algorithmes élaborés vont, par exemple, interpréter la présence de guillemets autour d'un SR comme un indice positif de sa pertinence. Un SR suivi en contexte d'un verbe conjugué sera aussi jugé comme étant potentiellement une unité nominale.

2.2.2.1.3.3Commentaires

Les résultats de ces travaux de recherche sont intéressants d'un point de vue de l'indexation des textes et de l'identification automatique des descripteurs. Les SR retenus peuvent faire l'objet d'une utilisation dans le cadre d'une démarche visant l'acquisition automatique des termes, mais la pertinence des résultats présentés est essentiellement validée par rapport à une démarche documentaire et non terminologique. Ces travaux ont cependant pu démontrer que le système automatisé offre un avantage sur l'indexeur humain : la systématicité. En effet, l'analyseur automatique recense des formes qui n'ont pas été relevées par l'humain (Drouin et Ladouceur 1994 : 25-26).

2.2.2.1.4 Oueslati (1999)

2.2.2.1.4.1Présentation

Dans sa thèse, Rochdi Oueslati (1999) propose une méthode d'aide à l'acquisition des connaissances à partir d'un corpus en langue française. Afin d'y parvenir, l'auteur propose une technique qui fait appel aux travaux sur les segments répétés présentés dans les paragraphes précédents.

L'auteur préconise une approche interactive ayant pour but de mettre en lumière les relations sémantiques entre les termes propres

à un domaine. Il cherche aussi à constituer des classes de termes et à démontrer le rôle de ces dernières au sein des relations identifiées (structures prédicat-arguments).

2.2.2.1.4.2Description

Afin d'identifier les relations prédicat-arguments, l'auteur procède à une étape d'acquisition des termes qui repose sur une analyse des segments répétés. Les étapes adoptées par l'auteur pour l'acquisition des termes sont (1999 : 84) :

- le prétraitement du corpus,
- l'extraction et le filtrage des segments répétés,
- la structuration des segments répétés sous forme d'arbres de termes,
- le filtrage de l'ISR et la constitution d'une liste de termes.

Nous passerons sous silence la première étape puisqu'elle n'offre que très peu d'intérêt pour l'acquisition automatique des termes. Le lecteur remarquera que l'approche globale est semblable à celle adoptée dans les travaux présentés dans la section précédente. Les données font cependant l'objet de traitements plus nombreux qui ont pour but de structurer l'ISR. Une des étapes de traitement a pour but d'identifier les termes au sein des SR recensés. Cette dernière n'est pas entièrement automatique et les résultats seront revus par un linguiste ou un terminologue.

L'étape de structuration exploite la similarité qui existe entre les termes afin de regrouper les SR qui partagent une tête syntaxique commune (*artère brachiale, artère circonflexe, artère coronaire, artère*

coronaire droite, artère coronaire gauche, etc.). Ce processus permet de regrouper les termes au sein d'une structure arborescente dont les nœuds sont les têtes et les branches sont les expansions.

Le système proposé par l'auteur procède ensuite à un repérage de nouveaux termes à partir des termes ayant été validés. Ce processus d'identification de nouveaux termes utilise des formalismes permettant de repérer les têtes et les expansions déjà relevées et de les appliquer à de nouvelles structures.

Les autres étapes de traitement proposées par l'auteur dépassent les limites du domaine de l'acquisition automatique des termes et abordent l'acquisition des connaissances (relations sémantiques). Afin d'y parvenir, l'auteur examine les relations qui existent entre les termes dans les contextes. Les cooccurrences des termes sont examinées afin d'isoler des schémas morphosyntaxiques potentiels.

Une telle démarche permet d'identifier des regroupements du type *terme – verbe - terme* représentatifs du corpus. Par exemple, un schéma <T1 V T2> pour des contextes comme *coronographie montre des lésions* ou *ventriculographie confirme l'existence de problèmes cardiaques.* Ces deux exemples permettent d'envisager la mise en place de structures de type *prédicat-arguments.*

2.2.2.1.4.3Commentaires

L'étape de structuration des termes en fonction de la tête syntaxique ne va pas sans rappeler celle adoptée dans les travaux de Didier Bourigault (1994b). Cette décomposition met en évidence des redondances tant du point de vue des têtes que des expansions; ce

sont ces éléments redondants qui sont filtrés avant la constitution de la liste finale des termes.

L'intervention de l'humain au cours du processus d'acquisition des termes place ces travaux à l'écart de notre approche qui recherche une automatisation complète de cette étape. Les travaux de Oueslati (1999) offrent cependant l'avantage de bien mettre en évidence l'apport non négligeable que l'analyse des segments répétés peut avoir pour le travail terminologique.

L'étape de structuration des termes, préalablement validés par un terminologue, est intéressante dans l'optique d'un processus d'acquisition de termes reposant sur une liste de termes déjà connus. Le système proposé par l'auteur exploite la structure de termes connus afin d'extraire du corpus des termes qui auraient été ignorés au cours du recensement initial des segments répétés.

2.2.2.1.5 Conclusion

Même si les approches mécaniques sont, d'un point de vue informatique, plutôt lentes, elles offrent un avantage indéniable sur le travail manuel de dépouillement d'un corpus. Les algorithmes sont plus systématiques que l'humain peut l'être. Cet avantage est nettement mis en évidence lorsque le corpus analysé est imposant.

La flexibilité de l'ISR est sans doute l'aspect le plus intéressant des travaux sur les SR. La liste des SR obtenue peut en effet être utilisée dans le cadre d'études en politique, en linguistique, en traduction, etc. Même à l'intérieur de chacun de ces domaines, on peut envisager de nombreuses utilisations. Dans le cas de la linguistique, par exemple, on peut autant l'utiliser pour des études

portant sur la thématique de certaines portions du corpus que pour le dépistage des structures prédicat-arguments.

Dans certains cas, cette flexibilité des résultats peut aussi constituer un désavantage. En effet, l'approche des SR exige la mise en place d'une stratégie de filtrage afin d'isoler avec succès le phénomène faisant l'objet de la recherche. Par exemple, des travaux ayant pour but d'isoler les combinaisons *verbe-préposition* devront comporter une étape qui aura pour but d'isoler ces enchaînements dans l'ensemble de l'ISR. On peut cependant envisager d'élaborer un algorithme de reconnaissance des SR qui utilise des contraintes lors de la constitution de l'ISR afin de ne conserver que les éléments désirés.

2.2.2.2 Modèles linguistiques

Les systèmes présentés ici sont qualifiés de *linguistiques* puisqu'ils font appel à des techniques d'analyse reposant sur les connaissances actuelles de la langue et de sa structure. On distingue principalement les systèmes utilisant des informations syntaxiques et ceux qui utilisent des informations lexicales ou morphologiques.

Les premiers reposent sur une analyse complète de la phrase en ses constituants afin d'en dégager les syntagmes intéressants selon les objectifs de la recherche[30]. Dans le second cas, des grammaires locales procèdent à une analyse de surface de la phrase à la recherche de syntagmes potentiels. Ces derniers sont décrits, à leur tour, à

---

[30] Des travaux effectués dans le cadre de la terminologie ne conduiront pas nécessairement au recensement des mêmes formes que des travaux entrepris dans une optique lexicographique.

l'aide de grammaires qui permettent de circonscrire l'ensemble des réalisations potentielles. On peut aussi, à l'inverse, utiliser ces grammaires et un lexique acquis en cours d'analyse ou par le biais d'une collaboration avec des spécialistes pour générer l'ensemble des termes potentiels d'un domaine.

2.2.2.2.1 David et Plante (1990); Plante, Dumas et Plante (2000)

2.2.2.2.1.1Présentation

Nomino compte parmi les systèmes d'acquisition automatique de termes. Il a été élaboré dans le cadre d'une collaboration entre l'Office de la langue française du Québec et une équipe du Centre d'ATO de l'Université du Québec à Montréal. La première version de ce logiciel se nommait Termino (voir David 1990); il a depuis été remplacé par un nouveau système nommé Nomino (voir Perron 1996).

Les auteurs de Termino le présentent comme un système ayant pour but l'identification d'unités nominales syntaxiques susceptibles de se lexicaliser (David 1993 : 224), de *synapsies* (David et Plante 1990 : 145). Pour sa part, Nomino est présenté comme un système de dépouillement terminologique (Perron 1996 : 32). Les deux versions du logiciel sont destinées au traitement de corpus en langue française.

2.2.2.2.1.2Description

La présente description ne porte que sur la plus récente version du logiciel : Nomino. Ce dernier procède à l'acquisition des termes en quatre grandes étapes (Perron 1996 : 34-36). La première consiste en un découpage du document à dépouiller. Le logiciel découpe le texte

en lexèmes et en phrases et identifie au passage certaines unités comme les noms propres, les abréviations, etc.

Lors de la deuxième étape de traitement, chaque lexème identifié est soumis à une analyse morphosyntaxique qui a pour but de lemmatiser la forme et de lui attribuer une catégorie grammaticale. À cette étape du traitement, les lexèmes ne font pas l'objet d'une désambiguïsation et les formes ambiguës peuvent se voir attribuer plusieurs catégories grammaticales. Ainsi, un lexème comme *écrit* sera catégorisé comme *substantif, adjectif* et *verbe*. L'attribution des catégories grammaticales ne s'effectue pas à l'aide de dictionnaires, mais plutôt à partir de l'application de règles qui prennent en considération la morphologie des unités lexicales.

Le troisième traitement effectué par Nomino est une analyse syntaxique en vue de désambiguïser, en contexte, les formes qui ont reçu plus d'une catégorie grammaticale à l'étape précédente. À la fin de cette étape, toutes les unités de la phrase ne possèdent qu'une seule catégorie grammaticale. L'étape de désambiguïsation terminée, le logiciel peut procéder à l'identification des unités nominales complexes.

Cette étape de traitement fait place à une liste de CT qui contient deux types d'unités nominales : les *unités complexes nominales* (*ucn*) et les *unités complexes nominales additionnelles* (*ucna*). La première liste contient les formes dont la qualité est jugée bonne. Ce sont les unités nominales complexes qui ne se construisent pas autour des prépositions *avec, sur, pour* et *sans* ou de verbes à l'infinitif (*machine à laver*). Les autres unités, dont la qualité est jugée moindre, sont placées dans la seconde liste. On y trouve, entre

autres, les CT qui contiennent un déterminant comme *traitement de la parole.*

Nomino possède des fonctions qui permettent à l'utilisateur de décomposer les *ucn* et les *ucna* les plus complexes. Ainsi, on peut obtenir *carte à piste, piste magnétique* et *carte magnétique* à partir de l'ucn *carte à piste magnétique* (Perron 1996 : 35-36). Le système prend aussi en charge la décomposition des unités issues de la coordination de deux CT comme *lecteur et encodeur de carte* ou *système administratif et financier.* Nomino proposera alors les unités nominales *lecteur de carte, encodeur de carte* ainsi que *système administratif* et *système financier* (Perron 1996 : 36). Ce traitement de la coordination se rapproche de celui proposé par Jacquemin (1997 : 126-129) dans le cadre de ses recherches sur les variantes des termes.

2.2.2.2.1.3Commentaires

Nomino est le doyen des logiciels d'acquisition automatique de termes. Ses performances initiales et l'intérêt qu'il a su susciter ont lancé les recherches dans le domaine. Elles ont aussi démontré l'importance de l'ordinateur dans le travail terminologique.

Les auteurs ont recours à une analyse syntaxique puisqu'ils considèrent que les *synapsies* sont des unités qui relèvent directement de la syntaxe. Selon eux, l'utilisation de méthodes

statistiques[31] ou de matrices de formation ne permet pas de cerner la complexité de ces unités.

Le recours à des analyses morphosyntaxique et syntaxique lors des deuxième et troisième étapes de traitement rend le système très dépendant de sources externes d'informations linguistiques. Afin de procéder à de telles analyses, le logiciel doit posséder un ensemble très vaste de règles décrivant la morphologie et la syntaxe d'une langue. Le logiciel est donc étroitement lié à la langue à laquelle il est destiné. L'adaptation de l'approche proposée à une autre langue constitue donc un projet d'envergure non négligeable.

2.2.2.2.2 Bourigault (1992a)

2.2.2.2.2.1 Présentation

Le logiciel LEXTER a été élaboré par Didier Bourigault dans le cadre de son travail à la Direction des Études et Recherches d'Électricité de France et de sa thèse de doctorat (1992a, 1994a). Il a pour but d'enrichir les thésaurus d'un système d'indexation automatique de textes de la société.

En plus de sa vocation initiale d'enrichissement des thésaurus, le logiciel LEXTER a, par la suite, été utilisé dans le cadre du travail proprement terminologique pour l'acquisition et la modélisation des connaissances à partir de textes (voir Bourigault 1994a et 1994b) en langue française. La manipulation de ces connaissances, au sein de LEXTER, passe par l'acquisition des termes.

---

[31] Par méthodes *statistiques*, les auteurs font référence aux travaux sur les segments répétés (David et Plante 1990 : 141-142) que nous préférons ranger sous l'étiquette *mécaniques*.

2.2.2.2.2.2Description

Dans le cadre des approches linguistiques appliquées au repérage automatique des termes, une nouvelle piste de recherche a été explorée par Bourigault (1992a, 1992b, 1993, 1994a, 1994b) : l'exploitation du concept de *frontière de terme*. Bourigault adopte une approche s'articulant autour d'une analyse syntaxique locale ayant pour but non pas de recenser les CT à partir de matrices de formation syntagmatique des termes, mais plutôt à partir des frontières de termes. Il définit cette notion de la façon suivante :

> « ... des unités lexicales constituant des limites d'expressions terminologiques; par exemple, des mots appartenant à certaines catégories grammaticales (*verbe, conjonction, pronom, adverbe, ...*), ainsi que certaines suites ayant des structures particulières comme *préposition + déterminant.* »
> (Bourigault 1992b : 2)

L'auteur justifie le choix d'une approche par analyse locale par sa simplicité d'un point de vue informatique. Cette dernière est particulièrement intéressante dans le cadre d'applications industrielles qui doivent rendre le logiciel capable de traiter tous les types de textes. Toujours selon l'auteur, ces contraintes d'*universalité textuelle* rendent illusoire la rédaction d'un analyseur syntaxique suffisamment performant.

Avec LEXTER, l'acquisition automatique des termes s'effectue en trois grandes étapes principales : l'étiquetage des formes, le découpage des textes en CT par repérage de frontières et la décomposition des groupes nominaux obtenus (voir Bourigault

1994b). Chacune de ces étapes est prise en charge par un module logiciel différent : *catégorisation*, *découpage* et *décomposition*.

Une version antérieure (voir Bourigault 1992) du logiciel comportait 6 modules différents : *catégorisation*, *frontière*, *coordination*, *décomposition*, *filtre* et *navigation*. Les modules *frontière* et *coordination* ont depuis été placés sous le module *découpage*, qui regroupe leurs fonctions respectives. Le module *navigation* a depuis été écarté du logiciel puisqu'il s'agissait en fait d'un module externe au repérage de termes. Ce module avait pour fonction de permettre la consultation des résultats de LEXTER. Les formes relevant de la coordination sont maintenant traitées par le module *décomposition* (voir Bourigault 1994b). Quant au module *filtre*, qui incorporait des éléments de statistique à LEXTER, il n'est plus présent dans les dernières versions du logiciel. Le fonctionnement de LEXTER est donc purement linguistique dans son état actuel. Nous croyons qu'il est tout de même intéressant de décrire l'ensemble des modules présents dans les premières versions de LEXTER, car ils permettent de mieux saisir la complexité des travaux de l'auteur.

Le module *catégorisation* n'est pas un module à part entière de LEXTER, il s'agit plutôt d'une étape de traitement prise en charge par un logiciel externe à LEXTER[32]. Ce dernier découpe les textes en phrases et les phrases en mots afin de les soumettre, dans un second temps, à une analyse morphologique qui leur attribue une catégorie grammaticale (*nom, adjectif, verbe, adverbe,* etc.). Le système résout les ambiguïtés catégorielles et n'attribue qu'une seule valeur à chaque forme. Le module produit donc une sortie qui est constituée du mot,

---

[32] Didier Bourigault utilise présentement le logiciel commercial CORDIAL pour procéder à ce traitement.

de sa catégorie grammaticale et de son lemme. Cette sortie est ensuite reprise en charge par LEXTER qui met de l'avant ses algorithmes d'acquisition des termes.

La première étape du repérage des formes terminologiques complexes est prise en charge par le module *frontière*, qui fonctionne à l'aide de règles de découpage. Comme tous les autres systèmes de repérage automatique de terminologie, LEXTER a pour but la production d'une liste de CT. Le module *frontière* exploite une technique utilisant les connaissances sur les formes terminologiques complexes de façon négative.

Le module *Frontière* balaie ainsi le texte (ou le résultat du module *catégorisation*) et identifie des suites pouvant correspondre à des termes complexes potentiels à partir de l'identification des mots ne pouvant pas faire partie d'un terme (les frontières de terme). Les travaux de l'auteur reposent sur l'hypothèse que les éléments entre les frontières sont des CT. Une phrase comme *Le système verbo-moteur fonctionne indépendamment du système de réfrigération et d'alimentation du réacteur nucléaire et du générateur de vapeur* sera découpée de la façon suivante :

- système verbo-moteur
- système de réfrigération et d'alimentation du réacteur nucléaire
- générateur de vapeur

Les formes complexes proposées par les modules précédents sont parfois hypercomplexes et elles peuvent donc faire l'objet d'un traitement supplémentaire par le module *décomposition*. Ce dernier est constitué d'un ensemble de règles de décomposition qui sont

appliquées sur les CT afin d'identifier les sous-expressions qui peuvent constituer des CT. Cette étape de traitement permet d'identifier des CT qui seraient ignorés par la majorité des systèmes de repérage. On obtiendra ainsi les CT qui suivent :

- système verbo-moteur
- système
- système de réfrigération
- réfrigération
- système d'alimentation du réacteur nucléaire
- système d'alimentation
- alimentation
- réacteur nucléaire
- réacteur
- générateur
- vapeur

Le dernier module de LEXTER conduit à un réseau de termes regroupés autour des notions de tête et d'expansion. Ce réseau a pour but de faciliter la consultation des résultats par un terminologue ou encore d'enrichir un système d'acquisition des connaissances.

2.2.2.2.2.3Commentaires

Les travaux de Bourigault reposent sur une approche qui, bien que linguistique, est totalement différente de celle utilisée pour le logiciel Nomino. Ils adoptent comme point de départ l'hypothèse double selon laquelle une analyse syntaxique complète des phrases n'est pas indispensable et qu'un simple balayage des corpus à l'aide de matrices est insuffisant (voir Bourigault 1992b : 2). On s'écarte donc considérablement des travaux réalisés pour Nomino puisque

David et Plante (1990) prônent comme indispensable une analyse morphosyntaxique de la phrase pour le dépistage des termes.

L'approche à l'origine de LEXTER, qui utilise une analyse syntaxique locale, permet d'obtenir de bons résultats. Ces derniers peuvent être obtenus plus rapidement qu'avec une approche qui repose sur une analyse syntaxique complète de la phrase. En effet, cette dernière ne peut être fiable que si le système possède une grammaire et des dictionnaires exhaustifs qui rendent possible une analyse sans faille. Pour sa part, l'analyse locale permet l'introduction du concept de frontières de termes, point fort des travaux de Bourigault (1992a, 1992b, 1993, 1994a, 1994b). Ces dernières rendent possible un recensement des termes avec un minimum de connaissances linguistiques.

Le module *frontière* constitue l'aspect le plus innovateur du travail de Bourigault qui, plutôt que de procéder à une analyse traditionnelle par matrices syntagmatiques, met de l'avant une technique utilisant les connaissances sur les formes terminologiques complexes de façon négative. Le module *décomposition* constitue aussi un aspect novateur en ce qu'il permet d'identifier (ou de recomposer) des CT qui seraient ignorés par la majorité des systèmes d'acquisition automatique de termes.

L'absence de recours à des connaissances sur les termes (matrices syntagmatiques, liste de termes, banque de terminologie, dictionnaire électronique, etc.) rend l'approche indépendante des domaines de spécialité. On peut ainsi procéder à l'acquisition de termes dans un corpus traitant de chimie en utilisant exactement les mêmes ressources que celles utilisées pour un corpus de botanique.

Il est important de garder en mémoire le fait que le logiciel LEXTER a été développé pour répondre à des besoins pratiques et non pour des raisons de recherches fondamentales en terminotique. Il s'agit, selon nous, d'un avantage dans la réalisation d'un tel système qui doit prendre en considération l'utilité des résultats pour les terminologues confrontés à des besoins de production.

2.2.2.2.3 Voutilainen (1993)

2.2.2.2.3.1Présentation

Les travaux de Voutilainen (1993) sur l'anglais ont donné naissance au logiciel nommé NPtool. Contrairement aux travaux présentés dans les pages précédentes, la raison d'être de NPtool n'est pas l'identification des termes, mais l'identification des syntagmes nominaux. La visée n'est donc pas strictement terminologique, mais beaucoup plus large.

2.2.2.2.3.2Description

Les travaux de Voutilainen (1993) reposent sur une analyse syntaxique complète qui utilise une grammaire à contraintes afin d'identifier les syntagmes nominaux (SN). Le logiciel NPtool procède à une identification des SN en deux grandes étapes.

Le logiciel procède d'abord à une attribution de toutes les catégories grammaticales possibles pour chacun des éléments de la phrase. Les ambiguïtés sont ensuite levées automatiquement afin que tous les mots possèdent une étiquette unique.

La seconde étape d'analyse consiste en une identification des syntagmes nominaux. Pour ce faire, NPtool a recours à deux grammaires qui possèdent des niveaux de contraintes très différents.

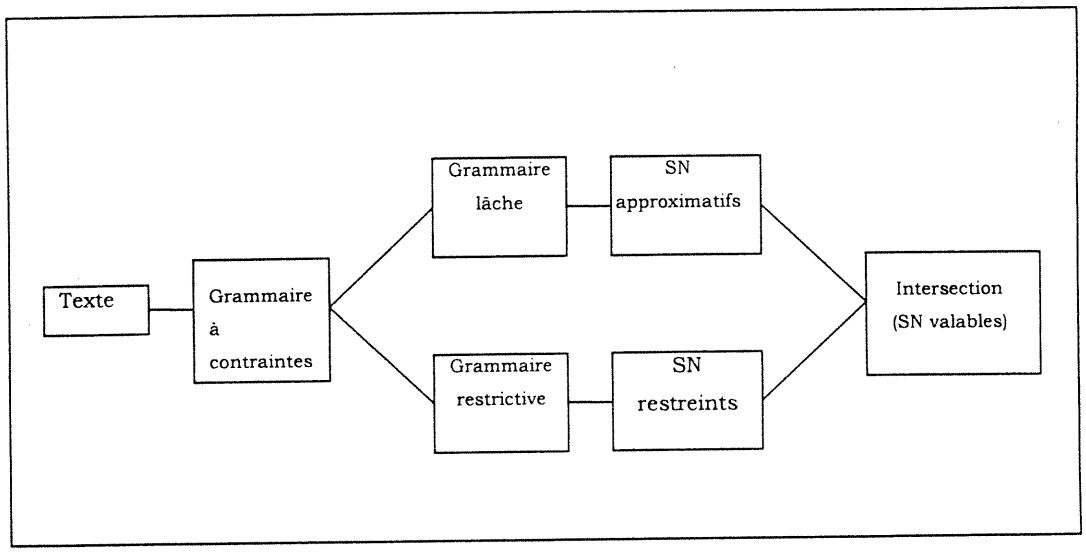

Figure 2. Grammaire utilisée par NPTool

Ces grammaires sont dites *lâches* ou *restrictives* (voir figure 2)[33] selon le nombre de SN qu'elles permettent de recenser. Le premier type de grammaire permet de détecter un grand nombre de SN (SN approximatifs) alors que le second n'accepte que des SN qui répondent à des contraintes beaucoup plus fortes (SN restreints).

Les résultats obtenus à partir des deux grammaires sont ensuite comparés et les CT retenus par les deux grammaires sont considérés comme des SN valables. La logique derrière cette décision repose sur l'intuition de l'auteur qui postule que si un SN est suffisamment flexible pour être détecté par les deux grammaires, il est

---

[33] L'auteur qualifie les grammaires de *NP-hostile* (restrictive) et de *NP-friendly* (lâche).

un CT intéressant. Les autres CT sont étiquetés comme douteux, mais ils ne sont cependant pas éliminés de la liste des résultats.

L'analyse de la séquence *the (DET) oil (N) can (N/V)* à l'aide de la grammaire lâche conduirait à la détection de *oil can* alors que la grammaire restrictive l'écarterait à cause de l'ambiguïté de la forme *can*, qui peut être soit un substantif, soit un verbe. Par contre, les deux grammaires conservent la forme *exhaust manifold* dans *the (DET) exhaust (N) manifold (N)* puisqu'il n'y a aucune forme ambiguë dans cette séquence.

2.2.2.2.3.3Commentaires

L'originalité des travaux de Voutilainen (1993) réside dans son utilisation de grammaires possédant des niveaux de contraintes différents. Les résultats obtenus à l'aide de cette technique sont impressionnants. NPtool permet d'atteindre un taux de rappel variant de 98,5 % à 100 % avec une précision allant de 95 % à 98 %. Parmi les SN erronés proposés par le système, 95 % ont été identifiés comme douteux. Étant donné que les termes constituent un sous-ensemble des syntagmes nominaux, il est possible d'envisager l'exploitation de cette approche pour l'acquisition automatique des termes.

2.2.2.2.4 Jacquemin (1997)

2.2.2.2.4.1Présentation

Les travaux de Jacquemin (1997) ont débouché sur la création d'un logiciel nommé FASTER. Ce dernier peut être utilisé pour traiter des corpus en langue anglaise ou en langue française. L'auteur cherche à décrire les transformations possibles des groupes nominaux terminologiques. Ces descriptions permettent de générer

les diverses variations potentielles des termes en contexte et de les identifier dans un corpus.

2.2.2.2.4.2Description

On peut regrouper les phénomènes de variation des termes sous trois grandes catégories (Jacquemin 1997, 2001) : syntaxique, morphosyntaxique et sémantique. Dans le premier cas, la structure syntaxique de la réalisation textuelle du terme est totalement différente. De sont côté, la variation morphosyntaxique s'opère grâce aux règles de dérivation morphologique et à une modification de la structure syntaxique. Pour leur part, les cas de variations sémantiques sont moins courants, mais il s'agit principalement de remplacement d'un des éléments du terme par un hyperonyme ou un hyponyme. Les principales catégories identifiées par Jacquemin (1997, 2001) sont les suivantes :

Coordination [syntaxique]
Ce type de transformation associe deux termes dans une structure syntaxique qui coordonne deux syntagmes nominaux. Afin de pouvoir s'insérer dans cette structure, les syntagmes doivent avoir recours à la même tête lexicale ou à la même expansion. De plus, le mot de tête ou les arguments doivent posséder la même fonction syntaxique.

Ex. : *abdominal wall* ⇒ *abdominal wall and chest*
fruits tropicaux ⇒ *fruits et agrumes tropicaux*

Modification et substitution [syntaxique]
La modification est l'ajout d'un modificateur à un des arguments d'un terme ou à la tête. Jacquemin considère que la modification devient une substitution à partir du moment où l'élément inséré dans le

terme forme un terme attesté et lexicalisé avec la tête. La distinction repose donc sur le degré de lexicalisation du sous-élément.

Ex. : *activité de l'eau ⇒ activité thermodynamique de l'eau*[34]

Composition et décomposition [syntaxique]
La décomposition consiste en l'utilisation d'une paraphrase pour faire allusion à la forme lexicalisée d'un composé.

Ex. : *consommation d'oxygène ⇒ consommation de l'oxygène*

Dérivation Nom - Nom [morphosyntaxique]
Dans ce cas précis, un des éléments substantifs est remplacé par un autre substantif.

Ex. : *fixation de l'azote ⇒ fixateurs de l'azote*

Dérivation Nom - Verbe [morphosyntaxique]
On observe dans ce type de transformation un remplacement d'un nom par un verbe.

Ex. : *fixation de l'azote ⇒ fixer l'azote*

Dérivation Nom - Adjectif [morphosyntaxique]
Cette transformation tire profit de l'équivalence qui existe entre le syntagme adjectival et le syntagme prépositionnel dans la modification d'un substantif ou d'un syntagme nominal.

Ex. : *variation du climat ⇒ variation climatique*

---

[34] *Activité thermodynamique de l'eau* est une substitution d'*activité de l'eau* dans le cas où *activité thermodynamique* est un terme.

Permutation [syntaxique]
La permutation est une transformation qui déplace un argument de la gauche de la tête à la droite de la tête d'un syntagme. Les permutations sont plus fréquentes en anglais qu'en français.

Ex. : *birth date* ⇒ *date of birth*

Référence anaphorique elliptique [sémantique]
Une anaphore elliptique est une transformation d'un terme ayant le même mot tête et ne reprenant qu'une partie non vide de ses arguments. Cette définition ne permet donc pas à la transformation de s'appliquer aux termes qui contiennent moins de 2 arguments se rattachant à la tête.

Ex. : *bone marrow transplant* ⇒ *transplant*

Le formalisme descriptif de base de FASTER est triple (Jacquemin 1999). La première composante est un lexique de mots simples où sont fournies des informations sur la catégorie et sur la morphologie flexionnelle ou dérivationnelle, des liens morphologiques entre les mots et leur racine et des liens sémantiques entre les mots simples et leurs voisins sémantiques. Ce premier niveau du formalisme repose sur un étiquetage du corpus effectué à l'aide d'un étiqueteur probabiliste nommé TreeTagger (voir Schmid 1994) qui appose des catégories grammaticales aux mots simples contenus dans le texte. Des liens de parenté sémantique et morphologique sont par la suite ajoutés aux mots à l'aide de WordNet (Fellbaum 1998).

Le deuxième niveau est constitué d'une liste d'entrées lexicales complexes, les termes, qui s'appuie sur les mots simples du niveau précédent et qui, en outre, donne les informations relatives au terme dans son ensemble. Finalement, une grammaire des variations

locales des termes génère, à partir des termes donnés au niveau précédent, leurs variantes licites et les identifie en corpus.

2.2.2.2.4.3Commentaires

Le point de départ des travaux de Jacquemin est une observation selon laquelle les variantes de termes attestés[35] représentent environ 15 % de tous les termes contenus dans un texte. Il considère donc, et nous partageons cet avis, qu'il est important de pouvoir procéder à l'acquisition automatique de ces variantes qui ne sont habituellement pas recensées par les logiciels d'acquisition automatique des termes.

FASTER permet de recenser ces variantes. En effet, les métarègles qui décrivent les transformations potentielles rendent possible l'identification des occurrences d'un terme en discours même s'il a fait l'objet d'une élision, d'une coordination, etc. Afin d'y parvenir, le logiciel nécessite cependant le recours à une liste de termes attestés. Une telle liste n'est par contre pas toujours disponible, particulièrement pour les domaines en émergence.

On doit aussi prendre en considération le fait que FASTER repose sur un ensemble de métarègles qui décrivent les réalisations potentielles des termes en discours et que cette dernière particularité

---

[35] L'auteur utilise le terme *variante* [de terme] afin de désigner des enchaînements textuels qui, à notre avis, ne sont pas toujours des termes. Ainsi, selon nous, la chaîne *stations synoptique et climatique* ne consiste pas en une variante des termes *station synoptique* et *station climatique*, mais plutôt en la réalisation textuelle de deux termes distincts. Il s'agit ici d'une distinction terminologique mineure que nous avons cru bon de signaler.

le rend vulnérable aux transformations non décrites par l'utilisateur. Il est effectivement possible que certaines nouvelles réalisations textuelles apparaissent dans un texte (sous l'influence du style de l'auteur par exemple). Les systèmes qui se fondent sur une analyse syntaxique complète et sur le regroupement en constituants, de même que ceux qui utilisent une analyse syntaxique de surface, ont le potentiel d'identifier ces variations non décrites préalablement.

2.2.2.2.5 Conclusion

Même si les approches linguistiques permettent l'obtention de bons résultats, l'intérêt de ces dernières est pondéré par la dilution de l'information causée par un fort taux de bruit. Sur le plan de la reconnaissance des unités complexes, les principaux problèmes proviennent du fait que l'analyse syntaxique fait très rarement appel à la sémantique et, bien souvent, de manière très partielle. Une analyse de la structure de surface, pour sa part, ne permet pas non plus de distinguer le terme du syntagme de discours lorsqu'il possède une structure syntaxique identique comme ceux qu'on retrouve dans les phrases *Il utilise un* ***langage de programmation*** (terme) et *Il parle à cette* ***fille de programmation*** (syntagme de discours).

Le couplage d'un module sémantique afin d'augmenter les performances de cette approche ne peut être une solution économiquement envisageable pour l'élaboration d'un système de reconnaissance des unités terminologiques complexes qui serait appelé à traiter des textes provenant de domaines très différents. En effet, la confection de dictionnaires électroniques ou la réutilisation de

l'information sémantique «cachée»[36] dans les grandes banques de terminologie sont encore trop coûteuses, en temps et en efforts, pour être facilement intégrées dans le cadre d'une démarche flexible. De plus, le processus de mise en place de dictionnaires pour un outil ayant justement pour but l'identification des termes pour l'élaboration de dictionnaires nous semble relever de la circularité.

2.2.2.3 Modèles statistiques

Depuis quelques temps, nous assistons à une recrudescence des techniques statistiques dans le domaine du traitement automatique de la langue naturelle. Le domaine de l'acquisition automatique des termes ne fait pas exception.

L'approche statistique offre, dans certaines circonstances, des avantages indéniables puisqu'elle permet de s'attaquer à des ensembles de données d'une taille imposante qu'il serait tout à fait impensable de traiter manuellement. Elle permet aussi de traiter des ensembles textuels pour lesquels des dictionnaires électroniques n'ont pas été élaborés en vue d'un traitement linguistique.

---

[36] Les banques de terminologie renferment des informations à caractère sémantique sous forme de définitions, notes linguistiques, etc. Étant donné que ces données ne se présentent pas sous forme structurée, elles ne peuvent, pour le moment, être utilisées dans le cadre de l'acquisition automatique des termes.

2.2.2.3.1 Church et Hanks (1989)

2.2.2.3.1.1Présentation

Church et Hanks (1989) font office de pionniers dans le domaine du traitement statistique des données linguistiques, plus particulièrement des données textuelles. Leurs travaux ont influencé de nombreux chercheurs et ont ouvert la voie à de nouvelles approches en acquisition automatique des termes.

Les travaux de Church et Hanks (1989) ont pour but de repérer automatiquement l'ensemble des collocations contenues dans un ensemble de données textuelles. Leurs visées ne sont donc pas terminologiques.

2.2.2.3.1.2Description

Church et Hanks (1989) présentent une mesure théorique, l'*information mutuelle*, qui rend possible l'évaluation du ratio d'association entre deux formes contenues dans un corpus. Si ces dernières, x et y, ont des probabilités d'occurrence P(x) et P(y), alors leur information mutuelle (*IM*) est la suivante :

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

L'information mutuelle tente de comparer la probabilité d'observer x et y ensemble par rapport à leur probabilité d'occurrence indépendante. La probabilité d'occurrence d'un mot x est calculée à partir de sa fréquence totale dans un corpus donné f(x) divisée par le nombre total de mots dans le corpus. Pour sa part, la probabilité

conjointe de x et de y, P(x,y), est calculée à partir du nombre de fois où les mots x et y apparaissent ensemble dans une fenêtre de w mots.

La taille variable de cette fenêtre permet de mettre en évidence des phénomènes différents depuis la cooccurrence simple (fenêtre étroite), jusqu'aux relations sémantiques (fenêtre large). Par exemple, cherchant à identifier une relation entre verbe et préposition, les auteurs ont choisi d'utiliser une fenêtre de taille fixe de 5 mots (w=5).

Une deuxième approche proposée par Church et Hanks (1989) consiste à imposer des contraintes additionnelles à l'aide d'étiquettes grammaticales assignées aux mots grâce à un étiqueteur grammatical. L'idée générale derrière cet ajout d'étiquettes consiste à repérer certains enchaînements qui possèdent un niveau d'information mutuelle supérieur. On pense, par exemple, à la relation qui existe en anglais entre certains verbes et les prépositions *to* et *in*.

La dernière technique utilisée par les auteurs pour contraindre les résultats statistiques consiste à remplacer l'étiqueteur grammatical par un analyseur syntaxique qui attribue à chaque forme une fonction syntaxique. Ce nouvel étiquetage permet d'avoir accès à un niveau supérieur d'information et d'identifier des relations syntaxiques à partir des diverses formes dans le texte. Ainsi, il est possible de mettre en évidence les liens qui existent entre le sujet (S), le verbe (V) et son objet (O). Les auteurs citent en exemple des corrélations observées entre *drink/V martini/O, drink/V beer/O, drink/V milk/O, drink/V cup_water/O, drink/V cup_coffee/O* etc.

On peut ainsi identifier des structures récurrentes plus ou moins figées, selon la spécialisation du texte, utiles au lexicographe

ou au terminologue, selon le cas. Church et Hanks (1989) soulignent l'intérêt d'un marquage supplémentaire qu'ils n'ont pas effectué, mais qui se fonderait sur des catégories sémantiques. On pourrait ainsi identifier les relations SVO en ajoutant une contrainte sur S ou sur O afin de restreindre l'ensemble des CT proposés.

2.2.2.3.1.3Commentaires

Bien que les travaux de Church et Hanks (1989) se situent en marge des recherches en acquisition automatique des termes, nous avons jugé nécessaire de les présenter puisque les techniques qu'ils ont proposées sont devenues le point de départ de nombreuses recherches en acquisition automatique de la terminologie. Même si leurs travaux ont été effectués sur la langue anglaise, l'information mutuelle est purement statistique et elle est indépendante des langues.

À l'analyse de leurs résultats, les auteurs constatent que les formes dont les valeurs de l'IM sont plus élevées sont très intéressantes, alors que les couples dont les valeurs de l'IM tendent vers 0 sont moins intéressants pour le lexicographe. L'application de ces techniques statistiques au dépouillement terminologique donne des résultats intéressants.

Ce n'est dont pas le fruit du hasard ou d'une certaine mode en acquisition automatique des termes qui fait que le test de l'information mutuelle est très populaire. En effet, ce dernier sert de base à de nombreux travaux. Les travaux de Church et Hanks (1989) ont le mérite d'être solides, bien documentés et facilement adaptables. Il est ainsi possible de les reprendre et de les insérer dans le cadre de travaux plus complexes et mieux adaptés à la terminologie.

2.2.2.3.2 Enguehard *et al.* (1992)

2.2.2.3.2.1Présentation

Enguehard *et al.* (1992) proposent le logiciel ANA, élaboré dans le cadre de travaux sur l'indexation automatique de textes en langue française. Les auteurs le décrivent comme un système dédié à l'apprentissage de concepts (Enguehard *et al.* 1992 : 1197-1202).

2.2.2.3.2.2Description

Le système procède au choix des concepts sans recours à l'analyse syntaxique, sémantique ou morphologique et en l'absence de dictionnaires (Enguehard *et al.* 1992 : 1197-1202). L'apprentissage des concepts par le système se fonde entièrement sur le contenu du texte et sur la description des objets familiers qui s'y trouvent. ANA contient les éléments suivants : les connaissances procédurales, le boostrap et les connaissances déclaratives.

Les connaissances procédurales

Les événements fréquents au sein d'un corpus sont utilisés comme point de départ pour l'apprentissage des concepts. On se fonde aussi sur certains enchaînements textuels à partir des connaissances déjà acquises sur les concepts. Ainsi, une configuration textuelle (un enchaînement) qui a la forme [mot inconnu - mot spécifiant un schéma - concept connu] permettra au système de conclure que le mot inconnu est susceptible de devenir un concept. Les mots spécifiant un concept sont obtenus par apprentissage sur les corpus de taille suffisamment grande pour le permettre ou donnés sous forme déclarative dans les autres cas.

Le bootstrap

Le réseau initial de concepts est validé par la suppression progressive des concepts et leur validation individuelle dans les cas où le système est à même de les découvrir à nouveau en cours d'analyse du texte.

Les connaissances déclaratives

Les connaissances déclaratives sont des listes de mots qui seront utilisées par le système pour ses inférences. Les auteurs signalent trois listes : la liste des mots vides, la liste des mots fortement liés et la liste des schémas. Les premiers correspondent aux mots outils de la langue qui ralentissent généralement les analyses et qui doivent être ignorés en cours de traitement. Les auteurs excluent quelques prépositions, conjonctions, adverbes, etc. qu'ils considèrent comme non significatifs. En gros, ce sont les mots qui ne peuvent aspirer au « statut de concept à titre individuel, ni figurer au début ou à la fin d'un concept » (Enguehard *et al.* 1992 : 1199).

Les CT sont validés à l'aide de la liste des schémas ou des configurations textuelles intéressantes selon leur appartenance à une configuration syntagmatique qui les rend aptes à représenter des concepts. Il s'agit essentiellement d'une approche qui recense les CT selon certaines constructions valides ou possibles à partir de matrices syntagmatiques.

Le modèle

Le modèle se construit autour de trois grandes classes d'objets : la classe des concepts, la classe des expressions et la classe des CT. La première recouvre les concepts connus du domaine auxquels s'ajoutent ceux qui sont découverts par le système en cours d'analyse. Les deux dernières recouvrent les mécanismes de découverte de

nouveaux concepts. C'est à ce niveau que les occurrences intéressantes sont emmagasinées de concert avec la fréquence.

2.2.2.3.2.3Commentaires

L'affirmation selon laquelle des mots « ne pourront ni obtenir le statut de concept à titre individuel, ni figurer au début ou à la fin d'un concept » (Enguehard *et al.* 1992 : 1199) nous laisse croire qu'ANA est un système d'acquisition automatique des termes plutôt que de concepts. En effet, il semble que ce soit entre autres par le statut de la structure syntagmatique du terme que transite la validité d'un concept. ANA assure cependant le passage des termes à un réseau de concepts à une étape subséquente.

Malgré l'affirmation des auteurs sur l'absence de nécessité du système ANA d'avoir recours à des ressources extérieures au corpus analysé, la description du logiciel faite par les auteurs (Enguehard *et al.* 1992) souligne l'importance de l'utilisation de listes de mots (les *connaissances déclaratives*). Malgré l'absence de définition, d'information morphologique, etc., ces listes correspondent à des dictionnaires électroniques. Il s'agit tout de même d'information linguistique, tout comme les schémas utilisés pour la validation syntagmatique des chaînes susceptibles de représenter des concepts. Le système n'est donc pas indépendant des langues et des domaines; il repose sur des connaissances préalablement mises à sa disposition.

2.2.2.3.3 Ahmad (1996)

2.2.2.3.3.1Présentation

Les travaux d'Ahmad (1996), effectués sur la langue anglaise, ne visent pas l'identification automatique des termes, mais plutôt de

mettre à la disposition des terminologues des listes de mots ainsi que des concordances utiles pour l'identification de la terminologie. Il nous semble cependant important d'inclure les travaux de cet auteur dans le présent chapitre puisqu'il est le premier à présenter une approche tirant profit de corpus non spécialisés afin d'isoler des particularités lexicales dans des corpus spécialisés.

2.2.2.3.3.2Description

Khurshid Ahmad (1996) adopte une approche qui repose sur une opposition des fréquences observées dans des corpus techniques et non techniques. Le corpus non technique utilisé est le LOB (Lancaster-Olso-Bergen). Ce dernier est un corpus étiqueté qui a été élaboré en 1961 et il compte 1 million d'occurrences.

L'auteur propose un indice qu'il nomme *coefficient d'étrangeté* (*co-efficient of weirdness*) et qui consiste à évaluer le rapport entre la fréquence relative d'une forme dans un corpus non spécialisé et la fréquence relative de la même au sein d'un corpus technique. Les formes qui apparaissent dans le corpus technique, mais qui ne sont pas représentées dans le corpus non technique se voient attribuer une valeur infinie. Il sont donc considérés comme «étranges».

Ahmad (1996) fait la constatation que les valeurs du coefficient d'étrangeté pour le vocabulaire usuel (*in, the, about*) atteignent des valeurs relativement basses en comparaison des termes spécifiques au domaine. Un tri effectué en fonction de ce coefficient permet de placer en tête de liste des formes directement reliées à la thématique du corpus technique.

2.2.2.3.3.3Commentaires

Les recherches d'Ahmad (1996) prennent comme point de départ un corpus qui date de plus de 40 ans. Certains pourraient s'objecter à l'utilisation d'un corpus si ancien. Cependant, nous croyons que le recours à un tel corpus permet non seulement de faire ressortir les variations du lexique entre le corpus technique et un corpus non technique, mais aussi de mettre en évidence des néologismes. Il serait intéressant de vérifier la part de néologismes qui ont été identifiés dans les travaux d'Ahmad (1996). Les formes dont le coefficient d'étrangeté est particulièrement élevé sont spécifiques au corpus technique. Nous croyons qu'elles pourraient aussi faire l'objet d'une analyse de collocation afin de procéder à un dépistage des termes en contexte. L'auteur ne présente cependant pas de suggestion en ce sens.

Les travaux décrits ici fournissent des pistes de recherche intéressantes qui méritent d'être reprises dans le cadre de travaux sur la terminologie et la néologie. Le recours à une simple comparaison entre les fréquences relatives des formes permet de mettre en lumière des phénomènes liés à la prépondérance des termes dans les textes techniques. Nous croyons qu'une démarche poussant plus loin les idées présentées par Ahmad (1996) et qui mettrait en place une stratégie élaborée autour d'un calcul probabiliste permettrait de vérifier que les valeurs obtenues à l'aide du coefficient d'étrangeté ne sont pas dues au hasard.

2.2.2.3.4 Conclusion

Si on les oppose aux techniques mécaniques et linguistiques, les techniques statistiques sont plus rapides, car elles permettent de cibler les sous-ensembles de données intéressants. Les ressources logicielles et matérielles nécessaires sont aussi moins imposantes puisque de telles approches ne requièrent pas de recours à des données linguistiques extérieures au corpus. En effet, elles peuvent fort bien effectuer leur travail en l'absence de dictionnaires et de grammaires. Il s'agit d'un avantage indéniable, car ces dernières ressources sont bien souvent les plus coûteuses à élaborer puisqu'elles sont habituellement le fruit d'un travail manuel. La capacité des approches statistiques de travailler sans avoir recours à des connaissances linguistiques les rend indépendantes des domaines abordés dans les corpus. En effet, les techniques statistiques ne reposent que sur les corpus eux-mêmes.

Malgré tous ces avantages, on note aussi des désavantages. Les résultats obtenus par les méthodes statistiques sont intimement reliés aux corpus utilisés et ne peuvent être interprétés en dehors de ce contexte. On doit aussi s'assurer que les corpus analysés possèdent une taille suffisamment grande pour que les résultats soient significatifs. On considère généralement que l'application de techniques statistiques à des corpus de taille inférieure à 100 000 occurrences ne conduit pas à l'obtention de résultats fiables et justifiables.

Les résultats obtenus par l'application de techniques statistiques conduisent aussi à des résultats dont la validité, d'un point de vue linguistique, est aléatoire. Il s'agit bien souvent de techniques qui sont utilisées pour identifier, dans un ensemble de

phénomènes linguistiques (observés sous un angle purement statistique), les phénomènes qui s'écartent du comportement normal, ceux qui se distinguent. Les résultats obtenus, à moins d'être filtrés à l'aide de techniques linguistiques ou par intervention de l'humain, ne peuvent donc que difficilement être interprétés dans le cadre d'une théorie linguistique.

2.2.2.4 Modèles hybrides

Les modèles hybrides sont, comme leur nom l'indique, à mi-chemin entre les modèles linguistiques et les modèles statistiques. Les études présentées dans les sections qui suivent adoptent un ordre de traitement qui varie. En effet, certains auteurs préfèrent commencer le traitement des corpus par une analyse linguistique dont les résultats sont filtrés à l'aide de techniques statistiques alors que d'autres procèdent à l'inverse.

2.2.2.4.1 Daille (1993)

2.2.2.4.1.1Présentation

Daille (1993) s'intéresse uniquement à l'acquisition automatique des termes complexes et elle n'aborde pas la problématique de l'acquisition des termes simples. Ses travaux portent avant tout sur la langue française, mais l'auteure a aussi appliqué sa méthode dans une optique d'appariement de la terminologie sur des corpus bilingues (anglais-français) dans Daille (1994a) et (1994b). Les recherches de l'auteure ont débouché sur la création du logiciel ACABIT.

2.2.2.4.1.2Description

Le premier maillon de la méthodologie de Daille (1993, 1994a et 1994b) repose sur une technique linguistique qui rejoint celles décrites précédemment pour Nomino et LEXTER : le corpus[37] analysé est préalablement étiqueté grammaticalement. La seconde étape de traitement fait appel à une description des groupes nominaux à l'aide de matrices syntagmatiques. Ces matrices sont identifiées dans les corpus à l'aide d'une grammaire.

L'auteure procède à l'acquisition des CT correspondant aux matrices syntagmatiques suivantes : N ADJ (*dissipation thermique*), N_1 de (DET) N_2 (*signal de raccrochage*), N_1 à (DET) N_2 (*tube à ondes*), N_1 PREP (DET) N_2 (*multiplexage par répétition*) et N_1 N_2 (*voie support*). Les séquences ayant une fréquence égale ou supérieure à deux et ayant été reconnues par un automate sont par la suite lemmatisées et transformées en occurrences d'un couple. Par exemple, le couple [*satellite, orbite*] recouvre les réalisations textuelles suivantes : *satellite sur orbite, satellites sur orbite, satellites en orbite, satellite mis en orbite.*

Des groupes nominaux recensés, seules les extrémités (couples lexicaux) seront prises en considération et soumises à une batterie de tests statistiques afin d'affiner les résultats de l'analyse linguistique. C'est à partir des résultats de ces tests que seront éliminés ou conservés les CT. L'auteure considère, et c'est le postulat de sa thèse, qu'il est nécessaire de procéder à une analyse statistique pour compléter l'analyse linguistique. À son avis, une simple analyse

---

[37] Le corpus utilisé traite du domaine des télécommunications et est constitué de 1 million de mots.

syntaxique locale (ou de surface) ne suffit pas à assurer le caractère terminologique des séquences retrouvées. Elle cherche, à l'aide de mesures statistiques, à décrire le plus fidèlement possible ce qui caractérise les termes en vue de leur identification automatique.

Afin d'évaluer la performance des indices, l'auteure compare la liste des CT isolés à une liste de termes validés par des terminologues. La comparaison est effectuée sans recours à la sémantique et se limite à une comparaison de chaînes de caractères. L'étape de comparaison des données avec le contenu d'une banque de terminologie a pour objet de valider les mesures statistiques élaborées et d'isoler la mesure qui conduit à des résultats qui se rapprochent le plus possible du contenu de la banque. Afin d'améliorer la précision des résultats, l'auteure retient quatre types de caractérisation numérique : la fréquence, les critères d'association, la diversité et les mesures de distance.

La fréquence

Selon les premières constatations de l'auteure faites sur le corpus à l'étude, « la fréquence se révèle comme l'une des mesures les plus performantes pour détecter les termes d'un domaine » (Daille 1994a : 133). Malgré l'apparence favorable des résultats obtenus par la fréquence, l'auteure affirme qu'il ne s'agit pas d'un indice valable pour distinguer les termes des non-termes, puisqu'elle laisse aussi échapper beaucoup trop de termes pour être directement utilisable (Daille 1994a : 133-134). Elle constate que le classement offert par la fréquence présente très rapidement des CT qui ne sont pas valides.

Les critères d'association (Daille 1994a : 136-141)

Daille constate que les performances du *score d'association*[38] pour le dépistage des termes ne sont pas aussi intéressantes que celles obtenues par Church et Hanks (1989) pour le dépistage des collocations. On constate que les couples les plus fréquents, majoritairement des termes complexes, possèdent un score d'association faible qui ne leur permet pas de se distinguer au sein des CT. Ainsi, des CT intéressants sont laissés de côté.

C'est pour cette raison que l'auteure procède à une adaptation du score d'association et propose un nouvel indice, le *score d'association avec le numérateur au cube*. Cette modification de l'indice offre l'avantage de retenir les mêmes CT en tête de liste que le score d'association, mais de ne pas écarter systématiquement les couples les plus fréquents.

Pour sa part, le *coefficient de vraisemblance* présente en tête de liste les même CT que le score d'association au cube. Il se démarque cependant lorsqu'une forme est absente de tous les autres couples recensés dans le texte et il prend une valeur nulle. Ce comportement ne permet donc pas de prendre en considération les associations occasionnelles.

Les résultats obtenus à l'aide du *critère de Fager et MacGowan* sont sensiblement les mêmes que ceux obtenus à partir du score d'association. La principale différence est que cet indice n'écarte pas systématiquement les CT dont la fréquence est élevée. À l'opposé, les

---

[38] Le score d'association correspond à l'information mutuelle (IM) de Church et Hanks (1989).

CT les plus favorisés sont ceux dont les éléments apparaissent souvent ensemble.

Après examen des résultats de chacun des critères d'association proposés, Daille ne retient que le coefficient de vraisemblance (Daille 1994a : 148). Elle justifie sa décision par le fait que les autres critères accordent trop d'importance aux couples d'éléments qui apparaissent toujours ensemble. Elle refuse de retenir le score d'association au cube, qu'elle élabore elle-même et qui conduit à des résulats intéressants, étant donné son caractère empirique qui ne relève pas de lois purement statistiques.

La diversité (Daille 1994a : 141-143)

Cet indice a été élaboré par Shannon (1948), dans le cadre de ses travaux en biologie, afin de positionner les individus dans la hiérarchie genre-espèce. L'auteure l'utilise pour obtenir des renseignements sur la distribution d'une forme au sein de l'ensemble des couples ayant été repérés. La diversité est appliquée à chacun des éléments des couples. Elle laisse ainsi savoir si les éléments sont fréquents en tant que couple, ou en tant qu'élément isolé.

Comme pour les autres mesures, les fréquences utilisées sont celles des éléments lexicaux pleins des CT repérés. Ainsi, l'auteure procède à l'analyse de la diversité des pôles lexicaux N_1 et N_2 pour les éléments qui entrent dans la matrice syntagmatique N_1 (Prep (Det)) N_2 [*liaison sémaphore, mise en œuvre, reconnaissance des signaux,* etc.] ainsi que N et Adj pour la matrice N Adj [*circuit numérique, ligne téléphonique*, etc.].

La diversité ne permet pas d'identifier les termes directement parmi les CT. Elle rend cependant possible la caractérisation du

comportement d'un des éléments dans un couple. On peut donc déterminer si un des éléments d'un couple est fréquent ou non au sein des autres couples recensés par le système. Il s'agit d'un avantage par rapport au coefficient de vraisemblance qui ne permet pas d'identifier l'élément d'un couple qui n'apparaît que dans ce dernier (Daille 1994a : 137).

La variance et l'écart-type (Daille 1994a : 143-147)

La moyenne et la variance servent à calculer la distance, en termes d'éléments lexicaux, qui sépare les membres d'un couple. Ces mesures prennent aussi en considération le nombre d'*éléments pleins* faisant partie du CT, les éléments pleins pouvant *grosso modo* être définis comme étant les mots non grammaticaux.

La variance et l'écart-type, qui ont été calculés pour chacun des couples recensés, ne permettent pas de tirer de conclusion sur le statut terminologique d'un CT. Daille (1994 : 145) en arrive à la conclusion que les termes d'un domaine ne sont pas des structures morphosyntaxiques figées. Cette idée sera ultérieurement reprise dans les travaux de Jacquemin (1997).

2.2.2.4.1.3Commentaires

L'engouement observé pour l'IM au cours de la dernière décennie n'est pas négligeable et l'observation des couples relevés dans les travaux de Daille (1994a : 136) à l'aide de cet indice nous fournit une explication potentielle. Les exemples cités par l'auteure sont *aiguille d'une montre, béton armé* et *dos à dos*. Ces formes sont intéressantes pour un traitement général des formes complexes dans un corpus, mais elles ne le sont pas pour une démarche visant l'acquisition automatique des termes. La performance de ce critère

n'est donc pas sans intérêt, mais elle n'est pas suffisante pour permettre de distinguer les CT qui sont des termes de ceux qui n'en sont pas. Ainsi, on peut conclure que le recours à l'IM se justifie mieux dans le cadre du travail lexicographique que du travail terminologique.

Le *coefficient de vraisemblance* se démarque du score d'association avec le numérateur au cube lorsqu'une forme est absente de tous les autres couples recensés dans le texte. Cet aspect nous semble d'une importance fondamentale puisque ces associations « marginales » nous paraissent primordiales dans une optique terminologique. Ces couples, qui possèdent un élément nouveau, peuvent permettre de déceler des néologismes potentiels. Cette possibilité est soulignée par la valeur de la diversité qui permet même d'identifier lequel des membres d'un couple est l'élément exclusif (ex. : [réseau, *maillé*]) en plus d'indiquer lorsque les éléments apparaissent uniquement ensemble (ex. : [océan, indien]). Il s'agit donc de critères importants dont les résultats adaptés ou combinés à d'autres indices statistiques pourraient faciliter l'identification des hapax ou des néologismes.

Les CT qui se sont vu attribuer une valeur négative pour le critère de Fager et MacGowan sont aussi particulièrement intéressants. En effet, ces derniers correspondent aux couples dont un des éléments apparaît exclusivement avec l'autre membre. Les couples identifiés ont cependant un intérêt terminologique variable. En effet, on remarque [moteur, apogée], qui correspond au terme *moteur d'apogée*, [accusé, réception], qui correspond au mot composé *accusé de réception*, et le couple [bout, bout], qui correspond à l'adverbe *bout à bout*. Le critère ne permet donc pas de discerner les

CT de ceux qui ne possèdent pas un statut terminologique intéressant.

L'utilisation de la variance et de l'écart-type permettent d'identifier les couples qui ont un niveau de figement ou de lexicalisation très élevé. Elle ne permet cependant pas de distinguer entre les formes recensées dont la distance ne varie que très peu et qui ne sont pas des termes (ex. : *organigramme de la figure*) et celles qui sont des termes (ex. : *canal support*).

Les travaux de Daille (1993) sur l'acquisition de termes constituent un premier pas vers une intégration des statistiques aux techniques linguistiques. La simplicité des filtres linguistiques est compensée par la mise en place de filtres statistiques ayant pour but de représenter le statut terminologique des CT. Le recours à la statistique pour accomplir cette tâche constitue une première. Le recensement de CT (*satellite sur orbite, satellites sur orbite, satellites en orbite, satellite mis en orbite*) sous forme de couples (*satellite, orbite*) ouvre aussi la voie aux travaux sur l'acquisition automatique des variantes de termes.

2.2.2.4.2 Justeson et Katz (1993)

2.2.2.4.2.1 Présentation

Justeson et Katz (1993) ont élaboré le logiciel TERMS qui a pour objet l'acquisition automatique des termes en langue anglaise. Les auteurs concentrent leurs efforts sur l'acquisition des termes complexes et laissent entièrement de côté la problématique des unités simples.

2.2.2.4.2.2Description

L'approche de Justeson et Katz (1993) pour l'acquisition automatique des termes repose sur deux grandes contraintes. La première contrainte que doivent satisfaire les CT est un seuil minimal de fréquence supérieur ou égal à 2. Les auteurs justifient en ces termes cette décision :

> « Groups of candidate terms of lower frequency have lower quality than groups of candidate terms of higher frequency; and the most frequent strings recovered from technical text are almost always valid technical terms. »
> Justeson et Katz (1993 : 8)

La seconde contrainte imposée lors de l'identification automatique des termes en est une de conformité à des matrices syntagmatiques. Les auteurs utilisent une grammaire qui décrit les structures possibles pour les termes afin d'éliminer certains CT de la liste retenue par le logiciel. Ils considèrent que les termes sont des chaînes composées de plus d'un élément lexical dont le premier élément (situé à droite) est un substantif suivi, soit d'un substantif, soit d'un adjectif, et qui se termine par un substantif. Ces chaînes peuvent par la suite se combiner entres elles à l'aide d'une préposition. Afin de décrire les réalisations potentielles, les auteurs proposent la grammaire suivante :

$((A \mid N)^{+} \mid (A \mid N)^{*} (NP) (A \mid N)^{*}) N$

où :

- A est un adjectif,
- N est un nom,
- P est un préposition,
- X | Y correspond à X ou Y,
- + indique qu'un élément peut apparaître une ou plusieurs fois,
- * indique qu'un élément peut être absent ou apparaître plusieurs fois.

Le processus d'acquisition des termes est effectué en deux grandes étapes. La première consiste en l'attribution de catégories grammaticales aux formes d'un corpus à l'aide de dictionnaires électroniques. Lorsqu'une forme est soit un substantif, soit un adjectif, soit une préposition, elle est alors conservée pour l'étape suivante. Le logiciel relève ensuite les enchaînements qui sont conformes à la grammaire présentée précédemment, dans la mesure où ils satisfont la contrainte de fréquence.

2.2.2.4.2.3Commentaires

L'analyse des résultats par les auteurs s'effectue sur deux plans. Premièrement, ils évaluent l'adéquation de la grammaire à décrire la structure de termes réels. Ils vérifient ensuite si les contraintes proposées (fréquence et matrices) fonctionnent pour l'acquisition des termes et s'ils permettent de distinguer les termes des non-termes dans la liste des CT proposés par le logiciel.

Le premier volet de l'évaluation est très positif et Justeson et Katz (1993) démontrent que leur grammaire réussit à décrire la structure de 99 % des termes tirés de la nomenclature de quelques dictionnaires techniques. Pour ce qui est de la deuxième portion de l'évaluation, les corpus soumis au logiciel TERMS ont d'abord été soumis à des humains qui avaient pour fonction d'identifier manuellement l'ensemble des termes[39] contenus dans les documents. Les performances du logiciel sont ensuite comparées à celles de l'humain. Comme on peut le constater dans le tableau I, le logiciel est plus performant lors de l'analyse de petits corpus que lors du traitement de gros ensembles textuels.

Texte	**Nombre de mots**	**% des termes identifiés**
1	2 300	90 %
2	6 300	77 %
3	14 900	66 %

Tableau I. Résultats de travaux de Justeson et Katz (1993)

La décision des auteurs d'écarter systématiquement les CT qui n'apparaissent qu'une seule fois dans les corpus peut être remise en question. En effet, certains CT peuvent avoir une fréquence très basse et en même temps constituer des termes importants un sein d'un corpus. L'élimination de ces formes constitue, à notre avis, un point faible de la méthodologie de Justeson et Katz (1993), car elle ne permet pas de recenser l'ensemble des termes d'un corpus.

---

[39] Dans le cas précis de cette étude, les auteurs ont eux-mêmes procédé au dépouillement manuel des textes.

Comme nous l'avons mentionné plus haut, les performances du logiciel diminuent avec l'augmentation de la taille du corpus analysé. Une explication potentielle de cette contre-performance réside dans le fait que le nombre de CT dont la fréquence est égale à 1 augmente avec la taille du corpus et que ces CT posent des difficultés à l'algorithme présenté par les auteurs.

Malgré ces points négatifs, les travaux de Justeson et Katz (1993) ont permis de démontrer que les enchaînements récurrents les plus fréquents dans les textes techniques, dans la mesure où ils correspondent à des matrices syntagmatiques précises, sont en majorité des termes.

2.2.2.4.3 Smadja (1993)

2.2.2.4.3.1Présentation

Conçu par Frank Smadja (1993), le système XTRACT est né de la recherche dans le domaine du repérage d'information et de l'indexation automatique en langue anglaise. L'objectif du logiciel XTRACT n'est pas le repérage de la terminologie, mais celui des collocations.

2.2.2.4.3.2Description

Le système XTRACT exploite une approche hybride qui combine des techniques statistiques et linguistiques. L'application des techniques statistiques précède celle des techniques linguistiques. XTRACT fonctionne en trois grandes étapes : extraction de couples de mots (bigrammes) présentant une information mutuelle importante selon la technique de Church et Hanks (1989); analyse contextuelle des bigrams pour le repérage d'enchaînements plus longs (*n*-grams)

et, finalement, filtrage des collocations obtenues aux étapes précédentes à l'aide d'information syntaxique.

La première étape de repérage du prototype XTRACT, l'extraction des bigrams, se fonde sur l'affirmation de Cruse (1986) selon laquelle une relation lexicale syntagmatique s'exprime par une corrélation directe entre deux mots. L'objectif de cette étape est donc d'identifier cette corrélation dans un contexte limité. L'auteur utilise une fenêtre de 5 mots avant et après le mot sélectionné comme pivot (le mot faisant l'objet d'une concordance).

Les concordances obtenues au cours de la première étape sont filtrées selon trois critères qui cherchent à cerner la force d'association des deux formes, leur répartition, ainsi que leur distance moyenne au sein de constructions syntaxiques récurrentes. Le résultat de cette première étape est une liste de bigrams qui sert de point de départ pour la seconde étape.

Cette dernière étape a pour but de recenser les constructions fréquentes au sein desquelles apparaissent les bigrams. On procède donc à l'analyse des concordances repérées afin de vérifier si des constructions syntaxiques très fréquentes ne s'en dégagent pas. Afin d'effectuer le filtrage des collocations, on fait appel à un analyseur syntaxique pour catégoriser grammaticalement et syntaxiquement les éléments du bigram original au sein des collocations. La catégorisation ne s'effectue donc pas sur l'ensemble des formes de la collocation, mais uniquement sur les pôles ayant été utilisés pour l'identifier. Cette façon de faire permet de mettre en évidence la relation qui unit les deux membres du bigram (*nom-nom, sujet-verbe, verbe-objet, adjectif-nom*, etc.).

2.2.2.4.3.3Commentaires

Selon l'auteur, l'étude des collocations peut amener à dépister divers types d'information, dont les syntagmes nominaux forts. Smadja (1993 : 148) définit ces derniers comme étant des enchaînements non interrompus de substantifs et d'adjectifs comme *stock market* et *foreign exchange*. Cette description fondée sur la linéarité textuelle le rapproche des travaux de Choueka (1988), mais le différencie de ceux de Daille (1993, 1994a et 1994b) et de Jacquemin (1997), qui cherchent à repérer des unités qui sont parfois interrompues.

Nous tenons à mentionner un passage important de l'article où l'auteur discute de l'importance du corpus. Selon lui, deux facteurs reliés au corpus influencent les résultats obtenus : la taille du corpus et son contenu (Smadja 1993 : 168-171). Selon Smadja (1993 : 168-169), la taille du corpus a une grande influence sur les tests statistiques. En effet, ces derniers rejettent des formes potentiellement intéressantes étant donné que leur fréquence est trop basse pour être prise en compte. Dans le cas précis des expérimentations décrites par l'auteur, les formes ayant une fréquence inférieure à 100 (dans un corpus de 10 millions de mots) ont été écartées. L'auteur souligne l'importance de l'utilisation d'un corpus de grande taille, surtout pour la mise en place de lexiques portant sur un domaine précis du savoir.

En second lieu, l'auteur commente sur le lien qui existe aussi entre les résultats et le contenu (le sujet plus particulièrement) du corpus (Smadja 1993 : 169). À titre d'exemple, dans le cas du corpus utilisé par l'auteur, le *Wall Street Journal*, la nourriture n'est pas *consommée* mais *transigée*. Le contenu des documents qui constituent le corpus a donc une grande influence sur les collocations

qui sont repérées. Il faut ainsi faire en sorte que le corpus puisse assurer une certaine représentativité du domaine à traiter quand on cherche à isoler des phénomènes reliés à ce domaine (collocations, termes, etc.).

2.2.2.4.4 Lauer (1994)

2.2.2.4.4.1Présentation

Lauer (1994) procède à une analyse qui a pour but de repérer les syntagmes nominaux de type N_1 N_2 N_3 en langue anglaise. L'auteur propose aussi une approche pour la désambiguïsation automatique de ces structures. Il ne s'agit donc pas à proprement parler de travaux dont le but est terminologique. Par contre, la technique proposée, qui recense l'ensemble des formes nominales d'un corpus, identifie aussi les termes.

2.2.2.4.4.2Description

La première étape de traitement consiste en un étiquetage grammatical du corpus effectué à l'aide de dictionnaires. L'algorithme examine chacune des formes délimitées par des blancs typographiques. Lorsque les dictionnaires permettent de l'identifier de façon catégorique comme étant une unité nominale, elle est étiquetée et retenue pour la prochaine étape du traitement.

Un calcul statistique évalue ensuite le degré d'association conceptuelle[40] entre les éléments du segment de texte retenu. Afin d'évaluer cette association, le logiciel a recours à un thésaurus. Une

---

[40] Cette technique s'oppose à celle de Daille (1993, 1994a, 1994b), qui s'intéresse à l'attirance entre les lexèmes d'un point de vue purement lexical.

analyse syntaxique prend ensuite en considération le poids de l'association conceptuelle entre les divers éléments afin de construire un arbre syntaxique et de trancher en cas d'ambiguïté. Ainsi, si l'association des deux premiers éléments reçoit une valeur plus grande que l'association des deux derniers, l'arbre de ce segment composé de trois éléments sera [[N_1 N_2] N_3]. Dans le cas contraire, l'algorithme propose l'arbre [N_1 [N_2 N_3]].

2.2.2.4.4.3Commentaires

Cette technique, utilisée pour lever l'ambiguïté des syntagmes nominaux, rejoint l'idée à l'origine du module décomposition de LEXTER. Lauer (1994) pousse l'idée plus loin en ajoutant une dimension conceptuelle qui n'est pas présente dans les travaux de Bourigault (1992a). Cependant, contrairement à l'approche de Bourigault (1992a), qui n'utilise que le corpus comme source d'information, l'algorithme de Lauer (1994) nécessite le recours à des connaissances linguistiques pour être à même de procéder à la désambiguïsation des CT.

Les travaux de Lauer (1994) sont intéressants, car ils font appel à la sémantique. Il s'agit d'un type d'information qui n'est pas fréquemment exploité. Par contre, le recours à un thésaurus rend difficile l'utilisation d'une telle technique dans le cadre d'un travail portant sur des domaines de pointe. En effet, les dictionnaires nécessaires à la prise en charge de ces domaines risquent de ne pas être disponibles. De plus, nous nous trouvons encore une fois devant un problème de circularité qui veut que le logiciel ait accès à des dictionnaires alors que l'objectif même du logiciel est l'élaboration de dictionnaires.

2.2.2.4.5 Frantzi et Ananiadou (1997), Frantzi *et al.* (1999)

2.2.2.4.5.1Présentation

Frantzi et Ananiadou (1997) présentent une technique visant l'acquisition automatique des termes en anglais. La démarche de ces auteures ne vise cependant pas uniquement l'acquisition des termes, mais aussi l'élaboration d'un indice permettant de cerner le caractère terminologique d'un CT.

2.2.2.4.5.2Description

L'acquisition automatique des termes s'effectue à l'aide d'une grammaire. Cette dernière identifie les CT qui correspondent à des matrices de formation syntagmatiques dans un corpus qui a préalablement été étiqueté. Pour ce faire, les auteures ont recours à l'étiqueteur d'Éric Brill (1994, 1995). Les séquences retenues par l'algorithme de Frantzi et Ananiadou (1997 : 1) correspondent à la grammaire suivante :

$$(Nom \mid Adjectif)^{+}\ Nom$$

Cette grammaire permet de repérer des enchaînements constitués de noms et d'adjectifs dont la tête potentielle (située à l'extrême droite) est nécessairement un nom. Les CT peuvent donc être composés d'un simple nom ou d'un enchaînement de noms et d'adjectifs de longueur indéterminée.

L'évaluation du statut terminologique des CT se fait par le biais d'un indice, la C-value. Cette dernière prend en considération la fréquence du CT, sa longueur, ainsi que les recoupements entre les

CT recensés par le système. L'indice est calculé de la façon suivante (Frantzi et Ananiadou 1997 : 5) :

$$C-value(c_i) = \begin{cases} \log_2|c_i| \cdot f(c_i) \\ ou \\ \log_2|c_i| \cdot \left(f(c_i) - \frac{1}{P(Tc_i)} \sum c_y \in Tc_i f(c_y) \right) \end{cases}$$

où

c_i correspond à l'entrée i de la liste des CT,

c_y est une entrée de la liste qui inclut le CT i,

$f(c_i)$ correspond à la fréquence absolue du CT i,

$f(c_y)$ correspond à la fréquence absolue du CT y,

$|c_i|$ correspond à la longueur en nombre de mots du CT i,

Tc_i correspond à l'ensemble des termes qui incluent le CT i,

$P(Tc_i)$ correspond au nombre de CT qui composent l'ensemble précédent.

L'indice prend une forme différente selon que le CT fait l'objet d'inclusion[41] ou non dans d'autres CT. Lorsque le CT n'est jamais inclus, le premier cas de figure est utilisé. Dans tous les autres cas, un indice plus complexe prend en considération le nombre de CT dans lesquels le CT plus court apparaît.

Les travaux originaux de ces auteures ont été poursuivis dans (Frantzi *et al.* 1999). Les auteures présentent un indice plus complexe qui prend en considération le contexte d'occurrence des CT. La

---

[41] Comme c'est le cas pour le CT *fiber* dans le CT *optical fiber* par exemple.

NC-value (Frantzi *et al.* 1999 : 149-150) utilise la liste produite lors du tri selon la C-value et combine cette information à de l'information tirée des contextes d'occurrence des CT. La liste finale des CT est par la suite triée selon la NC-value. Afin de valider les résultats, les auteures consultent soit une liste de termes, soit un terminologue qui détermine le statut terminologique des CT[42].

L'approche présentée dans Maynard et Ananiadou (2001 : 261-277) fait suite aux travaux sur la NC-value. On y incorpore un aspect sémantique grâce à un thésaurus. Les auteures exploitent la distance entre deux unités lexicales à l'intérieur d'un réseau sémantique. Cette information est ensuite incorporée à la NC-value pour former un nouvel indice. Dans cet article (Maynard et Ananiadou 2001 : 273-276), les auteures évaluent la précision de leur approche à 76 % pour le premier tiers de la liste des CT.

2.2.2.4.5.3Commentaires

Il est difficile d'évaluer les performances de la C-value (Frantzi et Ananiadou 1997 : 8). En effet, les auteures ne présentent que les premières entrées de la liste des CT sans préciser le nombre total de CT recensés. Il nous est donc impossible de spéculer sur la qualité de la liste complète des CT retenus.

Les auteures ne procèdent d'ailleurs pas à une évaluation de leurs résultats en termes de précision. On peut cependant conclure qu'un tri selon la C-value offre l'avantage de proposer des CT valides

---

[42] Frantzi *et al.* (1999 : 154-155) présentent une liste de CT qui comporte une indication de la validité des entrées, mais les auteures omettent de mentionner comment cette information a été obtenue.

en tête de liste. Quant au second indice proposé par les auteures (Frantzi et Ananiadou 1997 : 11-12), la NC-value, il ne fait pas non plus l'objet d'une évaluation.

Les résultats obtenus grâce au recours à la sémantique (Maynard et Ananiadou 2001 : 261-277) sont intéressants et prometteurs. Cependant, l'utilisation d'une source externe d'information limite la portée de ces algorithmes puisqu'il faut que l'information sémantique relative au domaine dont traite le corpus soit disponible sous forme de thésaurus électronique. Pour le moment, dans la majorité des domaines, cette information n'est pas disponible.

2.2.2.4.6 Conclusion

Les approches hybrides constituent un compromis entre les deux grandes tendances de base et s'en approprient donc les avantages et les inconvénients. En effet, leur puissance de traitement, reposant principalement sur l'adoption de modèles traitant de l'information sous forme numérique plutôt que linguistique, permet de s'attaquer plus facilement à des corpus de taille imposante. Cette caractéristique les sert bien puisque, de façon à obtenir des résultats de qualité et à minimiser le niveau de bruit obtenu, ces algorithmes doivent avoir accès à un volume de données important.

L'approche hybride exploite aussi la systématicité, la rapidité et l'indépendance par rapport au domaine des algorithmes statistiques. Cette indépendance se manifeste aussi par l'absence de besoin *sine qua non* de dictionnaires et de grammaires spécialisées. Il s'agit là d'un avantage indéniable étant donné que ces techniques ont

généralement pour but d'assister l'humain dans l'élaboration de dictionnaires.

Les méthodes hybrides permettent aussi l'injection de connaissances par le linguiste. Cette dernière intervention humaine permet de modeler les résultats sur les intuitions du linguiste et de mettre de côté les aspects plus «froids» des résultats obtenus par les approches statistiques. De résultats purement statistiques, l'intervention des connaissances linguistiques permet l'obtention de résultats plus satisfaisants pour le linguiste en fonction des phénomènes qu'il cherche à observer et à décrire.

2.3 Linguistique quantitative

Comme nous l'avons mentionné dans l'introduction de la thèse, nous désirons mettre de l'avant une méthodologie d'acquisition automatique des termes qui se fonde sur une analyse statistique des corpus. Les travaux qui précèdent doivent donc s'inspirer des travaux effectués dans le cadre de la linguistique quantitative et utiliser leurs résultats.

Les paragraphes qui suivent présentent les travaux entrepris en linguistique quantitative ayant pour but de mettre en évidence les phénomènes liés à la variation de la fréquence des unités lexicales dans un corpus ou un sous-ensemble d'un corpus. Bien que ces travaux n'aient pas eu comme objectif d'observer le comportement des unités lexicales dans des corpus techniques, nous croyons qu'il est très intéressant de les adapter à cet objectif et qu'ils s'y prêtent bien. L'avantage principal des travaux en linguistique quantitative est leur indépendance des langues en présence dans les corpus.

2.3.1 Notation utilisée

Afin de simplifier la lecture des paragraphes suivants, nous proposons un système de notation. Dans tous les cas, nous considérons que deux corpus sont à l'étude : un corpus de référence (CR) et un corpus d'analyse (CA). Ces derniers correspondent à des sous-ensembles du corpus global (CG). C'est dans le CA que les fréquences des formes sont observées afin de voir si elles se démarquent sur le plan statistique.

Un corpus est constitué de *formes* uniques qui se manifestent à une ou plusieurs reprises dans un même corpus. Les diverses apparitions d'une forme sont nommées *occurrences*. Les formes des textes sont numérotées par ordre d'apparition de la première occurrence à l'aide de l'indice i où l'indice prend des valeurs entre 1 et n, où n correspond au nombre de formes observées dans le corpus.

La fréquence des formes est dénotée à l'aide de la variable k. Ainsi, la fréquence de la forme 3 dans le corpus *CA* sera notée $k_{3\,a}$. La taille totale du corpus *CR* sera notée k_r; elle est équivalente à la somme des valeurs de $k_{1\,r}$ à $k_{i\,r}$. La même notation est utilisée pour faire référence aux fréquences du corpus *CA*. La fréquence totale de la forme i dans les corpus *CR* et *CA* (*CG*) sera notée k_i. . La taille du corpus *CG* est notée k ($k_r + k_a$).

2.3.2 Muller (1979, 1992a)

Les travaux de cet auteur ont avant tout pour objet de déterminer le vocabulaire spécifique à des ouvrages tels que des pièces de théâtre, des romans, etc. Leur objectif est avant tout

stylistique. L'utilisation faite par Muller (1979, 1992a) des mesures statistiques permet de mette en lumière les divers thèmes chers à un auteur au fil des années et dans ses diverses réalisations. On peut ainsi procéder à des tests ayant pour but d'analyser le vocabulaire du corpus d'un point de vue synchronique, thématique, selon le personnage, etc.

Selon Muller (1992a : 77), le calcul statistique permet de donner un contenu à la notion subjective de *vocabulaire caractéristique* d'un texte. Les observations sur la fréquence ne sont justifiées que si elles sont comparées à une fréquence théorique calculée sur un ensemble textuel plus grand. Ce n'est qu'à partir de ce type de comparaison qu'on peut identifier le vocabulaire qui se démarque significativement à l'intérieur d'un corpus textuel.

Afin d'identifier le vocabulaire caractéristique d'un corpus, Muller (1979) propose l'utilisation d'un modèle théorique, le modèle hypergéométrique. Ce dernier permet d'établir si un corpus d'analyse consiste en un sous-ensemble aléatoire du corpus global. Ainsi, si X dénote la fréquence théorique de la forme i dans le corpus d'analyse, on peut alors écrire :

Loi hypergéométrique

$$P(X = k_{ia}) = \frac{\binom{k_{i\cdot}}{k_{ia}}\binom{k - k_{i\cdot}}{k_a - k_{ia}}}{\binom{k}{k_a}}$$

Dans le même ouvrage, Muller démontre que le modèle binomial permet une bonne approximation du résultat que donnerait un tirage

exhaustif dans les cas où k (taille de CR) est de grande taille par rapport à la taille de CA, soit k_a (Muller 1979 :170). Hubert et Labbé (1988 : 77-91) procèdent à la même démonstration à la fois sur le plan théorique et sur le plan pratique. La loi binomiale s'exprime de la façon suivante :

Loi binomiale

$$P(X = k_{ia}) = \binom{k_a}{k_{ia}} \left(\frac{k_{i\cdot}}{k}\right)^{k_{ia}} \left(1 - \left(\frac{k_{i\cdot}}{k}\right)\right)^{k_a - k_{ia}}$$

Pour Muller (1992a : 77), le vocabulaire n'est considéré comme caractéristique que dans les cas où la fréquence observée dans un échantillon se démarque de 5 % de la fréquence théorique attendue. Les formes dont la fréquence observée est de 5 % inférieure à la fréquence théorique sont qualifiées de *vocabulaire caractéristique négatif* alors que celles dont la fréquence observée est de 5 % supérieure à la fréquence théorique sont nommées *vocabulaire caractéristique positif.*

2.3.3 Lafon (1980); Lebart et Salem (1988, 1994)

Les travaux de Lafon (1980) ainsi que de Lebart et Salem (1988, 1994) portant sur les particularités lexicales d'un sous-corpus se situent principalement dans le cadre de l'analyse du discours. Les auteurs cherchent à identifier les thèmes récurrents dans des réponses à des sondages, à procéder à des regroupements de thèmes selon l'âge, le sexe, la provenance géographique des répondants. Ces travaux s'appliquent donc très bien à des recherches sociolinguistiques, sociologiques, politiques, etc. On peut donc s'intéresser à un ensemble de données textuelles sous divers angles.

L'idée d'isoler les *spécificités* dans une tranche de corpus proposée par Muller (1979) a été reprise dans les travaux de Lafon (1980) et de Lebart et Salem (1988 et 1994). Ces derniers ont eux-aussi recours à un modèle probabiliste afin d'isoler les formes dont la fréquence observée dévie de façon significative de la fréquence théorique.

Le modèle hypergéométrique utilisé rejoint celui de Muller (1979) et la terminologie utilisée est aussi très semblable. Les résultats sont aussi divisés de façon polaire entre des formes dont le comportement est significativement positif (*spécificités positives*) et des formes dont le comportement est significativement négatif (*spécificités négatives*). Par contre, on tient aussi compte des formes qui ne se démarquent pas d'un point de vue statistique et qui sont qualifiées de *banales*. On dit de ces formes qu'elles constituent le *vocabulaire de base* d'un corpus.

2.3.4 Camlong (1996)

Camlong (1996) propose une étude poussée des variations des fréquences au sein d'un corpus. Il s'intéresse principalement à la structuration du lexique ou à ce qu'il nomme l'*analyse spectrale des lexiques* (1996 : 127-129). Il identifie 8 ensembles de vocabulaire en se fondant sur une échelle qui découle de tests statistiques fondés, comme dans le cas des études de Muller (1979 et 1992a) et de Lebart et Salem (1988, 1994), sur la loi hypergéométrique.

Le *vocabulaire préférentiel* est un vocabulaire de prédilection ayant une portée thématique prépondérante. C'est, selon l'auteur, le vocabulaire le plus caractéristique d'un corpus. Dans le cas de textes techniques ou scientifiques, on peut envisager qu'il s'agisse de la

terminologie. Ce vocabulaire est donc omniprésent dans l'ensemble du corpus ou sous-corpus étudié et sa répartition est uniforme.

Le *vocabulaire différentiel*, pour sa part, se situe à l'autre extrémité du spectre; il s'agit d'un vocabulaire de rejet qui est volontairement mis à l'écart du texte à l'étude par l'auteur. Ce vocabulaire de rejet ne peut être identifié qu'à partir d'analyses laissant présager sa présence, il n'est accessible que par des comparaisons de corpus. Selon Camlong (1996 : 128), il s'agit d'un vocabulaire indispensable, bien maîtrisé par l'auteur, mais qui est volontairement mis en retrait.

Quant au *vocabulaire de base (ou de masse),* décrit comme un vocabulaire qui appartient à la fois à la langue, au genre et au style, nous croyons qu'il correspond au *français fondamental* de Gougenheim *et al.* (1964). Ce vocabulaire fournit à l'auteur l'ensemble des ressources lexicales de base nécessaires à l'élaboration de sa pensée. Sans ce vocabulaire, le discours ne peut exister.

Le quatrième ensemble décrit est nommé *vocabulaire de base à tendance positive* et il est présenté comme un vocabulaire de base qui tend à étayer le vocabulaire préférentiel. C'est un vocabulaire qui semble rejoindre le *vocabulaire général d'orientation scientifique (VGOS)* de Phal (1971) et qui complète les ressources lexicales mises à la disposition de l'auteur pour lui permettre de construire son discours dans un cadre scientifique et technique. Ce spectre lexical, de concert avec le vocabulaire préférentiel, correspond de près au concept de *spécificité positive* évoqué précédemment (voir 2.3.3).

La dernière couche qui nous intéresse plus particulièrement et qui est décrite par Camlong (1996 :129) se nomme *vocabulaire de*

base à tendance négative. La description de ce vocabulaire est plutôt nébuleuse et l'auteur insiste sur le fait qu'il n'existe que pour contrebalancer le vocabulaire marqué positivement. Selon les mots même de l'auteur, il s'agit d'*un vocabulaire grammatical où la thématique vient mourir* (1996 : 129). Son positionnement par rapport au vocabulaire de base à tendance positive nous laisse croire que ce concept correspond à celui de *spécificité négative* évoqué dans la section précédente.

3. CORPUS

La démarche entreprise dans la présente thèse repose sur l'opposition de deux corpus : un corpus d'analyse et un corpus de référence. Le corpus d'analyse est le corpus sur lequel le processus d'acquisition automatique des termes est effectué; c'est à l'intérieur de ce document que les spécificités lexicale sont isolées. Ce corpus est composé de trois documents afin de valider les algorithmes que nous mettons en place et de voir si nous sommes à même de faire des observations comparables sur les trois documents. Cette approche offre l'avantage d'assurer que les observations sont indépendantes d'un document particulier.

Pour être à même d'isoler les formes qui ont un comportement spécifique dans le corpus d'analyse, le logiciel a besoin d'un point de comparaison, d'un corpus qui est utilisé comme point de référence. C'est cette dernière fonction que remplit le corpus de référence. Le contenu de ce corpus ne fait pas l'objet d'une analyse approfondie mais, sans sa présence, l'extraction des PLS est impossible à mettre en œuvre puisque le comportement des unités lexicales dans le CA doit être évalué par rapport à celui qu'elles adoptent dans le corpus de référence. On peut ainsi affirmer que ce dernier corpus constitue, dans le cadre de la présente thèse, une norme.

Les deux corpus font l'objet d'une description dans les sections qui suivent. Les points 3.1 et 3.2 contiennent des descriptions plus approfondies des corpus et des documents qui les composent. Pour sa part, le paragraphe 3.3 décrit les étapes nécessaires à la préparation des documents contenus dans le corpus de référence. Les étapes de préparation du corpus d'analyse, qui sont prises en charge par le logiciel TermoStat, ne font pas l'objet d'une description dans la

présente section de la thèse; elles sont cependant décrites en détail dans le chapitre 4.

3.1 Corpus de référence

La nature informatique de l'approche mise en œuvre dans le cadre de la présente thèse nous amène à constituer le corpus de référence à partir de documents écrits. Si l'on respecte la terminologie établie dans le domaine, l'utilisation très spécifique du corpus fait du corpus de référence un *corpus spécialisé* au sens où l'entendent Habert *et al.* (1997 : 143-145) étant donné qu'il a été constitué spécifiquement pour l'acquisition automatique des termes.

Les documents qui composent le corpus de référence ont été mis à notre disposition par le *Groupe de recherche en sémantique, lexicologie et terminologie (GRESLET)* de l'Université de Montréal. Ces documents font partie d'un corpus plus important, TEXTUM, utilisé pour la recherche en lexicographie par le GRESLET. Les éléments de TEXTUM ayant été retenus pour le corpus de référence sont tirés du journal *The Gazette* publié à Montréal entre mars 1989 et mai 1989.

Le corpus est composé de 13 746 articles de journaux distincts portant sur des sujets variés. Cette grande variété de sujets est importante et nécessaire à notre démarche puisqu'elle vient minimiser l'uniformité thématique de notre corpus et que c'est cet aspect hétérogène qui permet au corpus de référence de se distinguer du corpus d'analyse qui est plus uniforme du point de la thématique abordée.

La taille totale du corpus est d'environ 7 400 000 occurrences, qui correspondent à environ 82 700 formes différentes. Bien que la

taille du corpus de référence soit importante, il est difficile d'affirmer qu'il s'agit d'un corpus de *grande taille*. En effet, l'utilisation d'un corpus de 7,4 millions d'occurrences en 2002 peut nous sembler adéquate, mais des corpus de cette taille paraîtront probablement ridicules d'ici quelques années.

3.2 Corpus d'analyse

Le corpus d'analyse est un corpus spécialisé (Habert *et al.* 1997 : 143-145) puisqu'il a été élaboré en vue de répondre à un besoin très précis. Il est composé de trois documents écrits de nature technique. Il s'agit ici d'un choix essentiellement dicté par les objectifs du travail à accomplir. En effet, les documents utilisés se doivent d'être représentatifs de ceux dépouillés par les terminologues au sein de la société Nortel Networks. Dans le cadre de leur travail, ces derniers ne procèdent qu'au dépouillement de documents techniques écrits.

Le recours à des documents écrits rejoint aussi les principes de la terminologie textuelle qui identifie le texte comme source principale d'information sur les termes. Comme l'indique cette citation de Pierre Auger, cette prédominance de l'écrit en terminologie n'est cependant pas nouvelle (1989 : 411) : « L'écrit demeure et demeurera, sans doute, le véhicule privilégié de la transmission et du développement du savoir. »

Les documents qui font l'objet de notre dépouillement terminologique ont été mis à notre disposition par la société Nortel Networks et ils sont l'œuvre du service de documentation du groupe *Optical Networks*. Les documents ont été rédigés au cours de l'année

2000 et sont au nombre de trois. Afin de les désigner, la notation suivante est utilisée : *CA1*, *CA2* et *CA3*.

Bien qu'il soit difficile de classifier catégoriquement un document comme relevant d'un seul domaine de l'activité humaine, nous considérons que le corpus d'analyse traite du domaine des télécommunications. La nature multidisciplinaire de ce domaine conduit cependant à l'inclusion de concepts venus du domaine de la physique optique, de l'informatique, etc. Par contre, pour simplifier la discussion, nous adoptons l'attitude de Kageura (1999 : 29-30), qui considère que la classification en domaines, bien que naïve, est extrêmement utile aux terminologues.

Bien que le domaine des télécommunications serve de dénominateur commun au corpus d'analyse, le document CA_1 traite de considérations informatiques décrivant l'interface de programmation des composantes décrites dans les deux autres documents. Pour leur part, les documents *CA2* et *CA3* traitent d'un sujet plus étroit au sein du domaine des télécommunications, celui de la structure physique des réseaux de fibres optiques et de leurs composantes.

Le document *CA1* s'adresse à des informaticiens qui conçoivent des applications destinées aux composantes décrites dans les documents *CA2* et *CA3*. Ces derniers sont rédigés pour des intervenants du domaine des télécommunications ayant une bonne connaissance de la structure physique des réseaux de fibres optiques. Leur public cible est principalement composé d'architectes de réseaux, d'installateurs, de réparateurs, d'ingénieurs, de testeurs, d'administrateurs de réseaux, etc. Les documents décrivent les

possibilités, les caractéristiques, l'entretien, l'utilisation et l'installation des éléments d'un tel réseau.

Corpus	**Nombre d'occurrences**	**Nombre de mots**
CA_1	11 947	1 207
CA_2	28 583	2 066
CA_3	8 676	1 053

Tableau II. Taille des documents du corpus d'analyse

La taille des documents qui composent le corpus d'analyse varie considérablement. Cette variation nous permet de tester la stabilité de nos hypothèses dans des conditions diverses. Les observations effectuées sur les documents du corpus d'analyse peuvent aussi être comparées entre elles afin de vérifier si la taille du document joue un rôle quelconque.

Comme on peut le constater dans le tableau III, la taille moyenne des documents précédents est relativement petite. Les chercheurs s'entendent habituellement sur le fait que 1 million de mots forment une taille minimale pour constituer un corpus spécialisé (Pearson 1998 : 56-57). On considère que cette taille permet de bien représenter un domaine du savoir ou un type de textes. Comme le fait remarquer Pearson (1998 : 59), si un consensus se dégage quant à la taille étalon que devrait posséder un corpus spécialisé, il n'existe pas de critères objectifs permettant d'évaluer la représentativité d'un corpus de cette taille.

Pearson est d'avis que la taille d'un corpus peut être influencée par la disponibilité des documents sous forme électronique

(1998 : 59). Elle remet aussi en question l'idée selon laquelle un certain seuil de fiabilité peut être déterminé pour un corpus donné.

La taille des documents composant le corpus a, bien sûr, une certaine influence sur l'homogénéité globale. Ainsi, un corpus d'un million de mots composé d'un document de 500 000 mots et de 100 documents de 5 000 mots risque de ne pas être aussi diversifié qu'un corpus composé de 500 documents de 1 000 mots. Doit-on cependant rechercher une telle homogénéité? À notre avis, cette quête peut ou non se justifier en fonction des phénomènes à observer et décrire.

Nous croyons que la taille d'un corpus peut être déterminée en fonction des objectifs de travail. Ainsi, dans le cas des documents qui composent le corpus d'analyse, leur taille doit correspondre à un échantillon représentatif traité par les terminologues en situation de travail. L'objectif que nous avons adopté en début de thèse impose donc une restriction significative sur le corpus. Le corpus se doit d'être identique à ceux traités en entreprise; la taille des corpus d'analyse est donc avant tout dictée par des critères externes à la linguistique et à la terminologie. Ce sont les contraintes de production en entreprise qui ont priorité.

Jusqu'où pouvons nous pousser cette logique et où devons-nous placer le seuil quant à la taille des corpus? Nous ne tentons pas d'apporter une réponse définitive à cette question, mais nous suggérons que l'objectif de la recherche entreprise possède une importance significative sur le seuil adopté. Le type d'expérimentations aura aussi une importance primordiale sur la taille des corpus. Ainsi, un corpus devant être soumis à des tests statistiques devra respecter un certain seuil dicté par la précision recherchée. Un corpus visant la description de phénomènes

langagiers qui serait trop petit ne saurait que mener à des conclusions erronées. La constitution d'un corpus est avant tout un exercice d'équilibre qui s'insère dans un ensemble de contraintes qui doivent être soupesées afin de déterminer une taille qui répond aux objectifs visés.

3.3 Préparation du corpus de référence

3.3.1 Nettoyage

Le corpus, avant de pouvoir être utilisé pour des analyses linguistiques, doit faire l'objet d'un certain nettoyage. À l'état brut, il contient des caractères de contrôle qui sont indésirables et qui peuvent nuire aux analyses et même les fausser. Dans le cas qui nous préoccupe, le nettoyage des codes insérés dans le texte s'avère relativement facile puisque ces derniers sont délimités par les caractères suivants : <>.

Voici un exemple du bloc de descripteurs qui précède les extraits qui composent le corpus :

<99><0>
<14>890301</14>
<15>Wed</15>
<19>NEWS</19>
<23>A3</23>
<29>Police think 'teenager' leads sob-and-rob ring</29>
<31>By ALBERT NOEL of The Gazette</31>
<35>GAZETTE</35>
<41>ROBBERIES</41>
<60> .....
...LOCAL KEYWORDS: ROBBERIES</60>

De ce bloc, nous devons retenir celui qui est délimité par les caractères de contrôle *<29> </29>* puisqu'ils servent à délimiter les titres. Les autres lignes sont tout simplement éliminées du corpus de référence. L'extrait lui-même est délimité par le bloc *<60> </60>* et il contient, en plus du texte qui nous intéresse, une série de descripteurs pour la manchette. Ces derniers ont été éliminés afin de ne pas gonfler artificiellement les fréquences de certaines formes. Les routines ayant pour but d'éliminer les codes de contrôle du corpus sont rédigées à l'aide du langage de programmation *Perl* et de l'utilitaire UNIX *sed*.

Le texte résultant de cette première étape du nettoyage du texte nous donne accès à un texte suivi qui contient l'ensemble des extraits de *The Gazette* sans interférences. Il s'agit donc de ramener le plus possible le corpus à son contenu tel qu'il apparaît au lecteur du quotidien (sans les images et les illustrations). Une fois ce résultat obtenu, une deuxième série d'algorithmes de nettoyage prend la relève afin de procéder à une nouvelle série de mises en forme. Cette dernière s'effectue au niveau des mots et non au niveau du texte lui-même et de sa présentation; il s'agit de l'étape de segmentation.

3.3.2 Segmentation

La notion de *segmentation* repose sur les notions d'*occurrence* et de *séparateur*. Les occurrences sont souvent associées aux *mots*, mais nous tenons à nous éloigner de cette terminologie floue qui, comme il a été souligné par Silberztein (1993 : 111-136), pose de sérieux problèmes. Les séparateurs sont des caractères typographiques qui ne font pas partie intégrante des occurrences et qui ne possèdent pas de statut d'un point de vue linguistique. Les occurrences correspondent aux unités lexicales délimitées par les séparateurs.

L'algorithme de segmentation est fondé sur celui versé dans le domaine public par Robert MacIntyre de la University of Pennsylvania. Ce script en langage interprété par l'utilitaire UNIX *sed* a été rédigé en 1995 dans le cadre du projet Penn Treebank (voir Marcus *et al.* 1993). Nous avons légèrement adapté cet algorithme de segmentation au corpus et à notre objectif de travail. Dans le cas qui nous intéresse, tous les caractères typographiques ont été considérés comme des délimiteurs : le point, la virgule, le point-virgule, l'apostrophe, les guillemets, etc. Ainsi tout segment de texte délimité à gauche et à droite par un délimiteur constitue une occurrence. Ces dernières font par la suite l'objet de deux autres étapes de traitement : l'étiquetage grammatical et la lemmatisation.

3.3.3 Étiquetage morphosyntaxique

La procédure d'étiquetage consiste à attribuer à chacune des occurrences une partie du discours (substantif, verbe, adverbe, etc). Cette étape peut s'effectuer de bien des façons, la solution la plus simple étant celle qui se limite à la consultation d'un dictionnaire afin de trouver l'étiquette appropriée. Par contre, une telle approche possède deux lacunes importantes : la couverture du dictionnaire par rapport au texte et l'attribution d'étiquettes multiples à une même occurrence. Ainsi, on ne pourra distinguer entre deux formes identiques, d'un point de vue de la graphie, comme dans le cas du substantif et du verbe *porte*.

Premièrement, on peut facilement envisager des cas où une occurrence n'est pas connue d'un dictionnaire électronique de référence. Puisque les dictionnaires électroniques disponibles ne sauraient être exhaustifs, surtout dans un contexte terminologique, il

vaut mieux utiliser une stratégie différente afin d'attribuer les parties du discours aux occurrences.

Deuxièmement, la résolution d'ambiguïtés est une étape nécessaire lors de l'attribution automatique de parties du discours à des occurrences. Par exemple, une occurrence comme *droite* peut aisément se voir attribuer deux étiquettes différentes : *adjectif*, *substantif*. Il faut donc ensuite procéder à une étape supplémentaire d'analyse du contexte immédiat de la forme afin d'être à même de trancher[43].

Les étiqueteurs probabilistes sont des logiciels qui nous permettent de capitaliser sur la simplicité de l'approche par dictionnaires et de prendre en charge l'attribution d'étiquettes pour les occurrences inconnues. Dans le cas qui nous intéresse plus particulièrement, l'étiqueteur utilisé prend en charge la désambiguïsation des formes. Notre choix s'est arrêté sur l'étiqueteur par règles conçu par Éric Brill (voir Brill 1994), qui offre l'avantage d'être distribué gratuitement. Son utilisation est aussi très répandue et il est reconnu pour sa grande fiabilité (Brill 1994 : 5 ; Habert *et al.* 1997 : 170).

L'étiquetage s'effectue en deux grandes étapes. Dans la première, le logiciel consulte son dictionnaire et attribue à chaque occurrence, à la suite d'une analyse probabiliste, une partie du discours. Si la forme ne se trouve pas dans le dictionnaire, l'étiqueteur attribue automatiquement l'étiquette NNP (nom propre) si le mot commence par une lettre majuscule ou NN (nom commun) si le

---

[43] Dans certains cas, il faut nécessairement procéder à une analyse complète de la phrase afin de lever l'ambiguïté.

mot débute par une minuscule. L'attribution des valeurs par défaut déclenche une analyse du contexte à l'aide des règles lexicales afin de vérifier la valeur de l'étiquette par défaut. Si le système rencontre un contexte semblable dans l'ensemble de ses règles lexicales, il procède à un ajustement des étiquettes. Si la recherche est infructueuse, le logiciel laisse en place les valeurs par défaut.

Au cours de la deuxième étape d'analyse, le système revient sur l'étiquetage précédemment effectué et applique systématiquement des règles de transformations contextuelles, dans le but d'affiner les résultats. Ces règles prennent en compte les catégories affectées aux occurrences du texte, telles qu'elles sont au moment de l'appel de la règle dans un contexte « local » assez réduit des quelques occurrences à droite et à gauche. Voici quelques exemples :

> « From common noun to plural common noun if the word has suffix -s. From common noun to adjective if the word has suffix -ly.
> To adverb if the word has suffix -al. »
> Brill (1994 :4)

Ces règles sont simples, mais elles sont cependant très puissantes. Étant donné l'application séquentielle et systématique de toutes les règles disponibles, il est important de placer les règles en ordre décroissant de puissance. Il faut donc les manipuler avec soin et adopter un processus d'essais et d'erreurs afin de trouver les bonnes combinaisons. Les règles distribuées avec l'étiqueteur sont au nombre de 284 et elles peuvent être modifiées par l'utilisateur. Nous avons choisi d'utiliser les règles par défaut.

L'étiqueteur possède une fonction lui permettant de procéder à un apprentissage lorsque l'utilisateur le désire. Un texte étiqueté par le logiciel doit donc être revu manuellement par l'utilisateur qui corrige les erreurs commises par le logiciel. Le corpus est ensuite soumis à nouveau à l'étiqueteur qui assimile les corrections faites par l'humain. Ce processus répétitif d'apprentissage a été laissé de côté dans le cadre de notre recherche. Cette décision peut conduire à une perte de qualité lors de l'étiquetage des données, mais les résultats bruts de l'étiqueteur sont de très haute qualité (voir Brill 1994 : 5).

Lorsque les deux étapes de l'étiquetage sont terminées, le texte a la forme suivante :

> But/CC they/PRP refuse/VBP to/TO believe/VB the/DT woman/NN in/IN custody/NN is/VBZ only/RB 14/CD years/NNS old/JJ ,/, as/IN her/PRP$ passport/NN indicates/VBZ ,/, and/CC she/PRP is/VBZ being/VBG held/VBN in/IN a/DT juvenile/JJ detention/NN centre/NN until/IN further/JJ checks/NNS on/IN her/PRP$ nationality/NN and/CC the/DT authenticity/NN of/IN her/PRP$ passport/NN have/VBP been/VBN made/VBN ./.

Santorini (1999) propose une description détaillée du jeu d'étiquettes attribuées par le système. Ces étiquettes, bien qu'uniquement morphosyntaxiques, contiennent une mine d'informations qui peut être réutilisée pour différents traitements linguistiques.

3.3.4 Stockage des données

Les données traitées par l'étiqueteur sont ensuite stockées dans une base de données sous la forme de couples. Les éléments de ce

couple sont la forme et la partie du discours. On obtient, par exemple, le couple (*interface*, substantif) qui s'oppose au couple (*interface*, verbe)[44]. Il s'agit là de l'information primordiale conservée au sujet des formes recensées dans le corpus de référence. Afin de procéder à des analyses statistiques à des fins de comparaison du comportement du lexique, nous ajoutons à ce couple sa fréquence absolue dans l'ensemble du texte.

Information	**Description**
Numéro de forme	Numéro unique associé à chaque forme
Forme	Forme telle que recensée dans le texte
Partie du discours	Partie du discours attribuée par l'étiqueteur
Fréquence absolue	Fréquence de la forme dans le document

Tableau III. Matrice générée pour chaque forme du corpus de référence

Chaque nouveau couple reçoit aussi un numéro de forme qui est déterminé de façon séquentielle. Les couples ne sont donc stockés qu'une seule fois dans la matrice suivante et leurs occurrences sont stockées dans une autre matrice (voir Tableau IV).

Information	**Description**
Numéro de forme	Numéro unique associé à chaque forme
Position	Position en nombre de mots

Tableau IV. Matrice générée pour chaque occurrence d'une forme

La matrice décrite dans le tableau III peut être reliée à celle du tableau IV par le champ *Numéro de forme*. Le champ *position*

---

[44] L'absence de marquage sémantique du corpus ne peut permettre de gérer les divers sens d'une forme ambiguë comme le substantif *interface*.

correspond à l'ordre d'apparition du mot dans le corpus original, le premier mot ayant la position 1. Les numéros de position sont attribués de façon séquentielle au fur et à mesure que le corpus est analysé. Cette information peut ensuite être utilisée pour reconstruire le corpus original ou encore les contextes d'occurrence d'une forme dans le corpus.

3.3.5 Lemmatisation

Au cours de l'étape de lemmatisation, les ressources du corpus sont exploitées au maximum puisque l'information qu'il contient sera utilisée pour prendre des décisions relatives au statut de formes tirées elles aussi du corpus. On rejoint ici la position de Brill (1994 et 1995) ainsi que de Bourigault et Gonzalez (1994) et leur approche par apprentissage endogène dans une optique d'acquisition de la terminologie.

> « Corpus-based methods are often able to succeed while ignoring the true complexities of language, banking on the fact that complex linguistic phenomena can often be indirectly be observed through simple epiphenomena. »
> Brill (1995 : 544)

Comme le souligne Brill, il est parfois possible d'envisager de décrire indirectement des phénomènes linguistiques relativement complexes. Nous croyons qu'il est tout à fait possible d'utiliser une approche empirique qui se fonde sur les connaissances linguistiques injectées dans le corpus par l'étiqueteur stochastique. L'analyse de la liste des formes identifiées laisse prévoir qu'à l'aide de quelques règles simples, il est possible de procéder à une lemmatisation automatique. Voici un extrait de la liste des formes triées par ordre alphabétique;

les formes en italique sont celles qui seront lemmatisées par notre algorithme alors que les formes en gras seront ignorées.

Forme	**Partie du discours**
against	IN
agency	*NN*
agencies	*NN*
alarm	NN
ambulance	*NN*
ambulances	*NN*
annoying	JJ
anonymous	JJ
another	DT
any	DT
apartment	*NN*
apartments	*NN*
apparently	RB
april	NN
are	VB
area	NN
arrested	VB
artists	NN
buses	NN
business	*NN*
businesses	*NN*
but	CC
by	IN
called	VB
calls	VB
checks	NN
cities	NN
citizens	NN

city	NN
close	VB
clothing	NN
coach	*VB*
coaches	*NN*
coaches	*VB*

Tableau V. Liste des formes avant lemmatisation

Lorsque l'on observe la liste des unités lexicales du tableau V, on constate que de simples règles permettraient de procéder à une lemmatisation. Par exemple, l'utilisation d'une règle qui retranche le suffixe *-s* des formes nominales et qui recherche la forme nominale qui résulte de la troncation permet d'identifier le lemme des formes *ambulances* et *apartments*.

On peut envisager une règle plus complexe qui tronque le suffixe *-ies*, ajoute le suffixe *-y* et qui recherche la forme nominale correspondante. Cette règle identifie le lemme de la forme *agencies* dans cette liste. Il est possible d'identifier, pour l'anglais, un ensemble restreint de règles de lemmatisation qui affinent rapidement les résultats de l'analyse faite par l'étiqueteur. Nous avons établi 8 règles qui couvrent les cas les plus fréquents.

Numéro	Suffixe	Suffixe retranché	Suffixe ajouté	Partie du discours	Longueur minimale	Exemple
1	*-ices*	*-ces*	*-x*	NN	5	*matrices / matrix*
2	*-ives*	*-ves*	*-fe*	NN	5	*knives / knife*
3	*-sses*	*-es*		NN	5	*accesses / access*
4	*-ches*	*-es*		NN	5	*switches / switch*
5	*-eet*	*-eet*	*-oot*	NN	4	*feet / foot*
6	*-ies*	*-ies*	*-y*	NN	4	*possibilities / possibility*
7	*-i*	*-i*	*-us*	NN	4	*stimuli / stimulus*
8	*-s*	*-s*		NN	4	*cars / car*

Tableau VI. Règles de lemmatisation

L'algorithme de lemmatisation consiste donc à identifier une forme nominale, à tenter d'appliquer les règles présentées dans le tableau précédent, à ajouter le suffixe correspondant à la règle lorsque applicable et à rechercher un couple correspondant dans la liste des couples identifiés dans le corpus. Si un couple correspondant est identifié, on considère que le processus de lemmatisation a été effectué avec succès.

La règle la plus fréquemment utilisée est de toute évidence la règle 8, qui retranche le *-s* final dans les cas où la longueur de la forme nominale est supérieure à 4. Les règles sont appliquées dans l'ordre dans lequel elles apparaissent dans le tableau précédent afin d'éviter les conflits entre les règles et de réduire la puissance de certaines règles comme la règle 8, qui recouvre presque toutes les

règles qui précèdent. Elles sont donc classées en ordre décroissant de puissance.

Nous avons décidé d'imposer une contrainte sur la longueur des formes après avoir effectué des tests qui nous ont permis de constater que la majorité des tentatives de lemmatisation des formes dont la longueur est inférieure à 4 caractères conduit à un taux d'erreur élevé. En effet, des 13 formes de notre échantillon qui font partie de cette catégorie, nous n'observons que 3 bonnes applications des règles (23 % des cas).

Cet ensemble de règles de lemmatisation simples nous permet d'obtenir des résultats satisfaisants. Afin de le démontrer, nous avons procédé à une analyse des résultats sur un échantillon de 1 000 formes nominales prélevées au hasard. L'algorithme proposé adopte la bonne solution dans 98,7 %. À titre de comparaison, l'inclusion des formes dont la longueur est limitée à moins de 4 caractères conduit à une précision légèrement moins élevée soit, 97,8 %.

4. ACQUISITION AUTOMATIQUE DES TERMES FONDÉE SUR LES PLS

4.1 Approche retenue

Cette première partie du chapitre 4 a pour objet de présenter la méthodologie élaborée afin de procéder à l'acquisition automatique des termes en anglais. Les sections qui suivent décrivent l'approche retenue à la lumière des travaux présentés dans le chapitre 2 et des objectifs identifiés dans l'introduction de la présente thèse. La méthodologie qui en découle et les résultats obtenus font l'objet d'une description détaillée au point 4.2.

4.1.1 Pivots lexicaux spécialisés

Les travaux sur les spécificités lexicales reposent sur l'opposition de deux corpus (ou sous-corpus) en vue d'en comparer le lexique et de faire ressortir les divergences d'un des deux éléments par rapport à l'autre. Nous postulons qu'une comparaison des fréquences d'occurrences des mots entre un corpus de référence (de type journalistique) et un corpus d'analyse (de type technique) nous permettra de faire ressortir les spécificités lexicales (Lafon 1980, Lebart et Salem 1988 et 1994) du corpus d'analyse et que ces dernières sont étroitement liées à la terminologie du corpus technique.

Comme le font remarquer Sager *et al.* (1980 : 2), le style que l'on retrouve habituellement au sein de corpus techniques se démarque tant par son utilisation d'unités lexicales non spécialisées (prépositions, déterminant, etc.) et que par son utilisation d'unités spécialisées (termes). Les spécificités peuvent appartenir à ces deux sous-ensembles d'unités lexicales.

Pour fins de description des cas de figure possibles, on considère que le corpus de référence sert de point de départ pour les observations. Une comparaison des fréquences dans les deux corpus peut conduire aux observations suivantes :

- la fréquence observée dans le corpus d'analyse correspond à la fréquence théorique projetée à partir du corpus de référence;
- la fréquence observée dans le corpus d'analyse est inférieure à la fréquence théorique projetée à partir du corpus de référence;
- la fréquence observée dans le corpus d'analyse est supérieure à la fréquence théorique projetée à partir du corpus de référence.

Une unité lexicale dans un corpus d'analyse peut ainsi avoir une fréquence inférieure, équivalente ou supérieure à sa fréquence établie à partir d'observations faites dans le corpus de référence. La méthode proposée dans le cadre de la présente thèse nous permet d'élaborer une méthodologie et un outil qui peuvent être utilisés pour vérifier si les variations de fréquence observées sont dues ou non au hasard. Afin d'y parvenir, nous utilisons la technique de Lebart et Salem (1994) pour le repérage des spécificités dans un corpus technique.

Le comportement particulier qui nous intéresse est envisagé d'un point de vue de la fréquence des unités et non d'un point de vue de la sémantique ou de la syntaxe. Le présent travail se limite à l'étude des manifestations textuelles, sous forme de variations de

fréquence des utilisations propres à un corpus. La couche linguistique qui nous intéresse est donc purement lexicale.

Jusqu'à maintenant, les travaux sur les spécificités ont été effectués sur des corpus homogènes. L'identification des spécificités a donc uniquement été faite sur des divisions (sous-corpus) de corpus homogènes. Nos recherches sont effectuées sur un *corpus global* (CG) que nous qualifions d'hétérogène puisqu'il est artificiellement composé de documents qui opposent deux types de discours : un corpus journalistique (le corpus de référence) et un corpus technique (le corpus d'analyse).

Cette dichotomie dans le type de discours est utilisée pour mettre en lumière les particularités lexicales du corpus technique par rapport à l'ensemble du corpus global. Cette approche se distingue aussi des recherches visant à identifier les points de variation de thématique au sein de corpus (voir Ferret et Gruau 2001). En effet, ces dernières cherchent à identifier les frontières thématiques alors que, dans le cas du corpus élaboré ici, nous savons où cette dernière se situe et nous avons pour but d'identifier son effet sur le plan lexical.

Nous proposons une nouvelle notion, le *pivot lexical spécialisé*. Ce dernier chapeaute à la fois les notions de *vocabulaire préférentiel* et de *vocabulaire de base à tendance positive* énoncées au paragraphe 2.3. Elle recouvre aussi les notions de *spécificité* et de *vocabulaire caractéristique* présentées dans la même section. Nous ne nous intéressons donc qu'aux éléments du vocabulaire qui sont sur-représentés dans un corpus spécialisé et non à ceux qui sont sous-représentés.

Afin d'isoler les PLS, nous utilisons une technique qui s'inspire de celle décrite par Muller (1979, 1992a) ainsi que par Lebart et Salem (1988, 1994), qui ont recours à la loi hypergéométrique. Lorsque le corpus d'analyse est beaucoup plus petit que le corpus de référence, on peut alors substituer à cette dernière loi la loi binomiale (voir paragraphe 2.3.2). La littérature nous indique que cette loi peut, à son tour, s'approximer par la loi normale dans les cas où k_a>25[45].

En effet, lorsque que la taille de k_a croît, la loi binomiale tend vers une distribution limite qui est la distribution normale (voir Baillargeon 1989 : 200). Dans le contexte de la linguistique de corpus et de son orientation de plus en plus marquée vers l'utilisation de gros corpus, il est plutôt rare de travailler avec des échantillons qui ont une taille inférieure à 25 occurrences. L'adoption de la loi normale se fait donc dans la majorité des cas lorsqu'on traite des corpus volumineux.

Loi normale

$$P\left(k_{ia}-0{,}5 \le X \le k_{ia}+0{,}5\right)= P\left(\frac{\left(k_{ia}-0{,}5\right)-\left(\frac{k_a k_{i\cdot}}{k}\right)}{\sqrt{k_a\left(\frac{k_{i\cdot}}{k}\right)\left(1-\frac{k_{i\cdot}}{k}\right)}} \le N(0,1) \le \frac{\left(k_{ia}+0{,}5\right)-\left(\frac{k_a k_{i\cdot}}{k}\right)}{\sqrt{k_a\left(\frac{k_{i\cdot}}{k}\right)\left(1-\frac{k_{i\cdot}}{k}\right)}}\right)$$

La loi normale nous permet d'identifier à la fois les formes qui sont significativement fréquentes et celles qui sont significativement peu fréquentes dans le corpus d'analyse par rapport au corpus de

---

[45] La notation utilisée respecte les conventions présentées au paragraphe 2.3.1.

référence. Nous sommes donc face à deux cas de figure où soit la fréquence observée dans le corpus d'analyse est supérieure ou égale à la valeur théorique, soit la fréquence observée est inférieure à sa valeur théorique.

Le calcul suivant nous permet de recenser les formes dont la fréquence observée dans *CA* ($k_{i\,a}$) est inférieure à celle attendue (cas de figure P_1); ces formes correspondent aux spécificités négatives de Lebart et Salem (1994).

$$P_1 = P\left(X \leq k_{ia}\right) = P\left(N(0,1) \leq \frac{\left(k_{ia} + 0{,}5\right) - \left(\frac{k_a k_{i\cdot}}{k}\right)}{\sqrt{k_a \left(\frac{k_{i\cdot}}{k}\right)\left(1 - \frac{k_{i\cdot}}{k}\right)}} \right)$$

L'équation suivante (cas de figure P_2) nous permet de recenser les formes qui se démarquent de par leur fréquence élevée dans le corpus d'analyse, soit les *spécificités positives* de Lebart et Salem (1994).

$$P_2 = P\left(X \geq k_{ia}\right) = P\left(N(0,1) \geq \frac{\left(k_{ia} - 0{,}5\right) - \left(\frac{k_a k_{i\cdot}}{k}\right)}{\sqrt{k_a \left(\frac{k_{i\cdot}}{k}\right)\left(1 - \frac{k_{i\cdot}}{k}\right)}} \right)$$

Il est important de souligner un troisième ensemble lexical obtenu à l'aide de l'application de la loi normale. Il s'agit des éléments lexicaux dont le comportement correspond à la répartition constatée

dans le corpus de référence et que Habert *et al.* (1997 :196) qualifient de *formes banales*.

Les PLS sont issus de l'analyse des spécificités, ils correspondent à un sous-ensemble des spécificités positives (voir paragraphe 2.3.3). Les PLS sont donc des (1) spécificités (2) nominales ou adjectivales positives (3) qui possèdent une fréquence observée qui a moins de 1 chance sur 1 000 d'être le fruit du hasard (4) au sein d'un corpus technique. Les PLS sont donc fortement représentatifs du lexique d'un corpus; nous désirons tenter d'établir un lien entre ces formes et les termes qui, eux aussi, sont directement liés au contenu d'un corpus.

4.1.2 Acquisition automatique des termes fondée sur les PLS

Comme nous l'avons évoqué dans la section 2.2.2, tous les modèles proposés pour l'acquisition automatique de la terminologie possèdent des points faibles et des points forts. Des modèles mécaniques, nous retiendrons leur aspect systématique et leur exploitation du phénomène de répétition de certaines chaînes lexicales.

Les modèles linguistiques offrent l'avantage d'utiliser diverses sources d'information linguistique afin de trancher sur le statut linguistique des éléments étudiés. Nous croyons que cet aspect est essentiel lorsqu'un système cherche à distinguer les chaînes d'éléments lexicaux qui possèdent un statut linguistique de celles qui n'en possèdent pas. La lourdeur et la lenteur de traitement apportées par les modèles purement linguistiques seront réduites au maximum par l'adoption d'une analyse syntaxique locale reposant sur le concept de frontière de termes. Dans le cadre de la présente recherche, le

concept de frontière est simplifié et nous adoptons une définition lâche qui peut se résumer ainsi : toute forme ne pouvant apparaître à l'intérieur d'un CT est considérée comme une frontière, indépendamment du contexte d'occurrence.

Les modèles statistiques permettent d'isoler un certain nombre de données de façon à concentrer les efforts des analyseurs linguistiques. C'est de ce dernier aspect dont nous tentons de tirer profit. Les unités lexicales dont le comportement est propre au corpus d'analyse sont d'abord identifiées à l'aide de tests statistiques pour ensuite être traitées à l'aide de techniques linguistiques. Ainsi, le processus d'acquisition ne porte que sur les formes qui se situent dans le contexte immédiat des PLS. Nos travaux se situent dans le cadre des modèles hybrides et font donc appel à la fois à la statistique et à la linguistique. Le travail d'acquisition des termes se divise en trois grandes étapes; les paragraphes suivants donnent un aperçu de chacune de ces étapes.

4.1.2.1 Pré-traitement des données

Le pré-traitement des données a pour but de préparer les données pour les étapes d'acquisition automatique des PLS et des termes. La première étape, purement mécanique, segmente le texte en unités lexicales simples. Cette étape ne fait appel à aucune connaissance extratextuelle et se limite à diviser le texte en fonction d'une série de délimiteurs.

Les unités recensées sont par la suite soumises à un processus de lemmatisation. Des heuristiques de lemmatisation sont appliquées à chacune des formes du corpus afin de vérifier si elles peuvent faire l'objet d'une lemmatisation. Cette décision repose entièrement sur

une comparaison du lemme potentiel d'une forme avec les autres formes du corpus. Les algorithmes de lemmatisation n'ont donc recours qu'à la connaissance tirée du corpus pour prendre leur décision.

4.1.2.2 Acquisition des pivots lexicaux spécialisés

L'identification des PLS constitue la dernière étape avant la production par le logiciel TermoStat d'une liste de CT. Nous nous intéressons au corpus d'analyse dans une optique strictement lexicale, en particulier au comportement des formes[46] afin de les comparer, d'un point de vue statistique, à leur comportement dans le corpus de référence. Nous laissons ainsi de côté toute information syntaxique, sémantique, etc. pour nous concentrer uniquement sur le lexique du corpus à l'étude.

Nous reprenons à notre compte les éléments d'analyse identifiés par Lebart et Salem (1994) : les *spécificités positives*, les *spécificités négatives* et les *formes banales*. Les spécificités sont des formes qui possèdent une fréquence dans le CA qui se démarque significativement de la fréquence observée dans le CG. Pour leur part, les formes banales sont des formes qui adoptent un comportement qui n'est ni spécifiquement positif, ni spécifiquement négatif dans le CA. Leur comportement est donc conforme à celui attendu à la suite des observations faites sur le corpus de référence. Dans le cadre de la

---

[46] Nous adoptons le terme *forme* afin de faire référence aux éléments des sous-ensembles sur lesquels portent nos observations puisqu'ils sont constitués à la fois des unités lexicales et des unités qui n'ont pas de statut linguistique particulier (nombre, ponctuation, etc.).

présente thèse, seules les spécificités positives feront l'objet d'une étude.

Le recours à la loi normale met à notre disposition deux indices qui peuvent être utilisés pour l'analyse des résultats : la *valeur-test* et la *probabilité*. Nous reprenons ici la terminologie de Lebart et Salem (1994 : 317), qui décrivent la valeur-test comme une quantité permettant d'apprécier la signification de la position d'un élément illustratif sur un axe factoriel. En d'autres termes, il s'agit de la position d'un point sur la courbe normale. Pour sa part, la probabilité correspond à l'aire sous la même courbe (obtenue à partir d'une table) exprimée relativement au point correspondant à la valeur-test (voir Muller 1992b : 174-177).

Nous avons décidé d'adopter les valeurs-tests pour regrouper les formes plutôt que les probabilités puisque les valeurs calculées pour ces dernières sont beaucoup moins nuancées et notre analyse perdrait ainsi beaucoup de sa subtilité. En effet, l'examen d'une table des probabilités reliées aux valeurs-tests permet de constater que, pour des valeurs supérieures à 3,09, les probabilités sont inférieures à 0,001 (1/1 000) et que les écarts sont de plus en plus petits et difficiles à cerner.

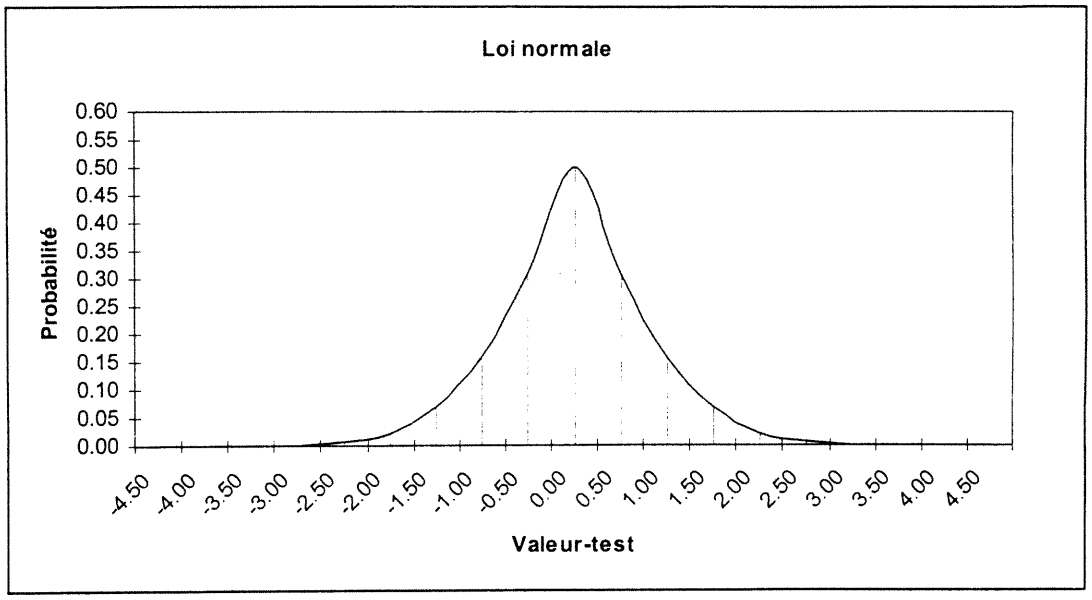

Figure 3. Loi normale

L'utilisation des valeurs-tests permet aussi d'avoir une meilleure idée des écarts de comportement des diverses formes et d'apporter des nuances là où les probabilités ne permettent pas de le faire. La figure 3 illustre bien la diminution extrêmement rapide de la valeur de la probabilité lorsque les valeurs-tests s'éloignent de zéro.

Afin de ne retenir que les formes intéressantes et très significatives, nous nous attardons sur celles dont les valeurs-tests sont supérieures à 3,09, comme le suggèrent Lebart et Salem (1994 : 183). L'adoption de ce seuil nous permet d'assurer qu'il n'y a que 1 chance sur 1 000 que la fréquence observée dans le texte CA_i soit due au hasard. À titre de comparaison, mentionnons qu'un seuil de 1,96 correspond à une probabilité de 5 %[47] alors qu'un seuil de 2,33 correspond à une probabilité de 1 %. La figure 3 permet de visualiser

[47] C'est le seuil adopté par Muller (1992a :77) pour l'identification de son vocabulaire caractéristique et par Lebart et Salem (1994) pour l'identification des spécificités.

la distribution normale et la variation de la probabilité en fonction de la valeur-test.

La technique d'acquisition des PLS retenue repose donc sur une approche statistique proposée par Lebart et Salem (1988 et 1994) pour représenter les spécificités lexicales d'un corpus par rapport à un autre corpus. Afin de pouvoir être considérées comme PLS, les spécificités doivent répondre aux critères suivants :

- être une spécificité positive,
- être une forme nominale ou adjectivale,
- posséder une fréquence qui a moins de 1 chance sur 1 000 d'être le fruit du hasard.

Les PLS identifiés sont ensuite mis à profit au cours du processus d'acquisition automatique des termes.

4.1.2.3 Acquisition des termes

La technique d'acquisition automatique de la terminologie que nous mettons de l'avant se fonde sur les travaux effectués sur l'acquisition par frontières de termes. Cette approche repose sur une réflexion linguistique et sur une approche empirique de tests effectués sur gros corpus (Bourigault 1994b). L'expérimentation sur corpus offre l'avantage de valider très rapidement les intuitions du linguiste et de corriger le tir dans les cas où les algorithmes mènent à des résultats peu intéressants.

L'approche par frontières proposée par Bourigault (1992a, 1992b, 1993, 1994a et 1994b) se rapproche de celle documentée dans Drouin et Ladouceur (1994). Dans le premier cas, l'algorithme balaie

le texte et recense les formes au sein d'un CT jusqu'à ce que la forme analysée corresponde à une partie du discours ayant été identifiée comme une frontière potentielle. La forme sera alors considérée ou non comme une frontière à partir d'une analyse du contexte syntaxique immédiat.

Dans le second cas, les algorithmes élaborés utilisent aussi les connaissances au sujet des termes de façon négative. En d'autres termes, les auteurs mettent en place une description informatique de ce qui ne peut pas constituer un terme d'un point de vue morphosyntaxique. Ainsi, les CT qui contiennent certaines prépositions ou déterminants (*système d'information à, système d'information sur les*) (Drouin et Ladouceur 1994 : 22) ne seront pas retenus car ils ne correspondent pas, d'un point de vue morphosyntaxique, à la structure potentielle d'un terme. Le processus d'acquisition se fait donc en deux étapes : le recensement des CT et le filtrage des CT en fonction de leur structure syntagmatique.

Nous reprenons donc l'idée selon laquelle certaines parties du discours sont désignées comme frontières entre les termes potentiels puisqu'elles ne peuvent apparaître à l'intérieur des termes (Dubuc 1979 : 55; Kocourek 1991 : 139). Étant données les matrices de formation syntagmatique des termes en langue anglaise, les décisions concernant les frontières sont simples et ne nécessitent pas le recours à une analyse syntaxique locale; les coupures sont donc nettes et indépendantes du contexte.

Un ensemble de parties du discours est considéré comme des frontières dans l'ensemble des contextes. Cette conception de la frontière de terme est beaucoup plus lâche que celle proposée par

Bourigault (1992a, 1992b, 1993, 1994a et 1994b) qui a recours à une analyse locale du contexte de la frontière potentielle afin de confirmer son statut. L'approche est aussi différente de celle utilisée par Drouin et Ladouceur (1994) puisque cette dernière consiste à procéder à l'acquisition de l'ensemble des segments répétés contenus dans un document pour ensuite éliminer ceux contenant des parties du discours qui ne peuvent se retrouver au sein de termes.

La contrainte la plus importante qui sera imposée au processus d'acquisition de la terminologie est que tous les éléments composant un CT doivent être des PLS. L'hypothèse derrière ce choix est que si une forme est suffisamment spécifique au corpus, elle possède un intérêt terminologique. Du même souffle, on peut postuler que certaines suites de formes spécifiques correspondent aux termes. Cette contrainte a pour but d'assurer que les CT recensés par l'outil TermoStat se construisent autour d'éléments lexicaux importants pour le texte spécialisé analysé. Nous ajoutons à la contrainte initiale une autre exigence qui veut que la tête du syntagme soit une unité nominale.

L'algorithme d'acquisition des termes procède de la droite vers la gauche, la limite à droite est constituée par un PLS nominal. On isole ce dernier et il sert de point de départ pour la construction d'un terme d'une longueur maximale de six unités simples. Cette contrainte sur la longueur des CT est imposée à la lumière des travaux de Justeson et Katz (1993), Nkwenti-Azeh (1994) et de Jacquemin (1996). Le positionnement de la tête à l'extrémité droite du terme constitue le mode de construction le plus courant en langue anglaise (Sager *et al.* 1980 : 268); c'est pour cette raison que nous adoptons cette règle de base pour l'acquisition de nos CT. L'adoption

d'une telle technique ne peut que mettre à l'écart certains termes[48]; nous devons cependant rappeler que nous ne cherchons pas à procéder à un dépouillement intégral des corpus.

4.1.2.4 Validation des CT

Les CT issus du processus d'acquisition sont automatiquement validés à l'aide d'une banque de terminologie spécialisée en télécommunications. Cette banque multilingue (allemand, anglais, chinois, espagnol, français, japonais, portugais) compte un peu plus de 100 000 termes au total. Au cours du processus de validation des données, seule la partie anglaise de la banque est sollicitée (61 000 termes).

L'étape de validation des résultats ne fait pas partie intégrante du logiciel TermoStat et elle a été intégrée afin d'évaluer la qualité des résultats obtenus à l'aide des algorithmes proposés. Dans le cadre du travail en entreprise, les résultats ne seraient pas validés à l'aide d'une banque de terminologie et ils seraient directement présentés aux terminologues.

La présente étape de traitement des données n'élimine pas de CT de la liste retenue par le logiciel, mais vient plutôt confirmer le statut terminologique de certains d'entre eux. Le logiciel TermoStat consulte le contenu de la banque de terminologie afin de comparer sa liste de CT avec la liste contenue dans la banque. Les CT qui sont

---

[48] Les constructions dont la tête n'est pas située à droite seront ignorées. Les termes qui ont recours à la préposition *of* présentent souvent cette particularité : ***degree*** *of freedom*. Ces cas sont cependant peu fréquents.

identifiés dans la banque de terminologie sont considérés comme valides.

La liste de termes mise à la disposition de TermoStat par la banque de terminologie est une liste dite *à plat*, c'est-à-dire que cette liste ne comporte pas d'information sémantique permettant de confirmer hors de tout doute que le CT relevé correspond au terme de la banque de terminologie. Les CT sont donc comparés, sur le plan de la graphie, avec la liste des termes contenus dans la banque de terminologie. Cette démarche s'inspire de celle décrite dans Daille (1994a : 123-124). Étant donné l'étroitesse du domaine et la corrélation entre les documents du corpus d'analyse et la banque de terminologie consultée[49], nous croyons qu'il s'agit d'une source fiable pour assurer la qualité des décisions prises par TermoStat.

Afin d'assurer la validation des CT recensés dans le corpus qui n'apparaissent pas dans la banque de terminologie, nous avons eu recours à trois terminologues experts du domaine des télécommunications. Les experts avaient à leur disposition la liste des CT ainsi que tous les contextes pour chacune des occurrences des CT afin de déterminer la validité des CT apparaissant dans la liste. Cette seconde étape de validation a pour objet de s'assurer que l'ensemble des CT recensés par TermoStat est soumis au processus de validation. Les données validées sont ensuite utilisées pour évaluer les performances du logiciel.

---

[49] La banque de terminologie et les documents du corpus d'analyse ont été mis à notre disposition par la société Nortel Networks.

4.1.3 Indice terminogénique

La recherche d'un indice terminogénique, capable de représenter l'intérêt terminologique d'un CT, repose sur notre conviction que l'on se doit de présenter au terminologue une liste dont l'information est la plus pertinente possible. Elle n'a cependant pas pour objectif de retrancher de la liste des CT, mais de trier la liste des CT en ordre de pertinence. Des travaux semblables ont été effectués par Daille (1993, 1994a et 1994b) ainsi que Frantzi et Ananiadou (1997).

À cette étape, nous choisissons de procéder à un tri plutôt qu'à une élimination des CT puisque les informations auxquelles nous avons recours ne permettent pas de déterminer de façon catégorique si un CT possède un statut terminologique ou non. En cherchant à atteindre une liste qui serait triée en ordre de pertinence, nous soulevons le problème de l'évaluation du statut terminologique du CT et de son approximation par le terminologue ou par le logiciel. Dans le cadre de la présente thèse, nous cherchons à déterminer une façon de qualifier ou de quantifier ce statut, une façon de représenter le caractère terminogénique d'un CT. C'est autour de ce caractère terminogénique d'un CT que nous articulons la notion d'indice terminogénique.

4.1.4 TermoStat

Les concepts décrits dans les sections précédentes ont été intégrés au sein d'un prototype de logiciel d'acquisition automatique des termes : TermoStat. La création du logiciel a pour but de tester nos hypothèses et de permettre des expérimentations futures dans le domaine de la linguistique quantitative et de l'acquisition automatique des termes fondée sur les PLS.

4.2 TermoStat : Logiciel d'acquisition automatique des termes fondée sur les PLS

Le logiciel TermoStat a été conçu afin de tester la méthodologie d'acquisition des PLS et des termes. La présentation de la méthodologie élaborée débute avec une description des traitements effectués sur les données du corpus d'analyse afin de procéder à l'acquisition automatique des PLS et des termes. Ces deux étapes sont ensuite décrites en détail et les résultats sont analysés. Nous proposons enfin un indice terminogénique visant à cerner l'intérêt terminologique d'un CT.

4.2.1 Pré-traitement des données

L'étape de pré-traitement des données sert à préparer le terrain pour l'acquisition automatique des PLS et des termes. Elle consiste en diverses étapes de manipulation du corpus brut étiqueté, de la segmentation du texte en unités simples à la lemmatisation.

4.2.1.1 Segmentation et étiquetage

Les procédures de segmentation et d'étiquetage utilisées à cette étape correspondent à celles qui ont été décrites pour la préparation du corpus de référence. TermoStat fait appel aux mêmes outils et utilise directement leurs sorties au cours de l'étape suivante. Nous ne reprenons donc pas la description de ces outils en détail.

Le corpus, une fois enrichi par l'étiqueteur de Brill, contient une séquence de paires constituées d'une forme et de sa partie du discours. Le premier module de TermoStat isole donc cette paire et

porte une attention toute particulière à l'étiquette associée au mot par le logiciel de Brill afin de déterminer si nous sommes en présence d'une frontière de terme[50].

En effet, c'est dès l'étape de segmentation que certaines formes sont identifiées comme étant des frontières de termes. Cette décision est fondée sur la partie du discours, mais aussi sur le contexte linguistique et typographique entourant la forme. D'un point de vue linguistique, le logiciel considère que les seules unités qui ne constituent pas des frontières sont les formes nominales et les adjectivales. Le concept de frontière de terme est décrit plus en détail au paragraphe 4.2.3.2.2.

En plus de la partie du discours, certains éléments non linguistiques du texte sont considérés comme des frontières de terme :

- un changement de paragraphe,
- un retour à la ligne forcé,
- un changement de cellule dans un tableau,
- un signe de ponctuation (à l'exception du trait d'union).

---

[50] Nous tenons à rappeler au lecteur la définition de ce concept proposée par Didier Bourigault : « ... des unités lexicales constituant des limites d'expressions terminologiques; par exemple, des mots appartenant à certaines catégories grammaticales (verbe, conjonction, pronom, adverbe, ...), ainsi que certaines suites ayant des structures particulières comme préposition + déterminant. » (Bourigault 1992b : 2).

Les frontières que nous utilisons sont donc à la fois linguistiques et extralinguistiques, en ce qu'elles dépendent aussi de la mise en page et de la structure du texte. Étant donné l'absence de marquage dans le corpus utilisé, le système ne peut pas toujours prendre une décision sans équivoque, mais il arrive tout de même, dans la majorité des cas, à gérer les contextes que nous venons de décrire. Il s'agit donc d'un cas où un niveau de marquage donnant de l'information complémentaire sur la structure du texte permettrait d'obtenir de meilleurs résultats. Le recours aux langages SGML ou XML pourrait se révéler, dans de telles situations, très bénéfique.

4.2.1.2 Stockage des données

L'étape de segmentation permet de transformer l'information linéaire du corpus sur les formes et leur partie du discours en des matrices plus complexes qui servent de point de départ au repérage des PLS et des CT. Ces matrices sont gérées et utilisées directement par TermoStat tout au long du processus d'acquisition automatique des termes.

Une forme est considérée comme nouvelle lorsqu'elle remplit les conditions suivantes : le couple composé de l'occurrence et de la partie du discours constitue une combinaison unique pour le document analysé. Le tableau VII correspond à la matrice mise en place lors du processus de segmentation. Le contenu des cases ombragées est ajouté au cours d'étapes subséquentes du processus de pré-traitement.

Information	Description
Numéro de document	Numéro unique associé à un document
Numéro de forme	Numéro unique associé à chaque forme
Forme	Forme telle que recensée dans le texte
Partie du discours	Partie du discours attribuée par l'étiqueteur
Fréquence absolue	Fréquence de la forme dans le document
Valeur-test	Valeur-test obtenue à partir du calcul des spécificités.
Probabilité	Probabilité associée à la valeur-test.
PLS	Vrai/Faux

Tableau VII. Matrice générée par TermoStat pour chaque forme

Le numéro de document est un numéro unique séquentiel attribué à chaque document analysé; ce numéro est utilisé à l'interne par le logiciel TermoStat, qui offre la possibilité de traiter plusieurs documents et de conserver les résultats en mémoire. Les données reliées à un texte sont donc regroupées autour d'un numéro de document.

Quant au numéro de forme, il s'agit d'un nombre unique attribué à chacune des formes d'un même texte. On retrouvera donc une forme numéro 1 pour le document 1, une pour le document 2, etc. La graphie de la forme relevée par le logiciel est banalisée, c'est-à-dire qu'elle est systématiquement convertie en caractères minuscules. Les occurrences ayant la même graphie banalisée et la même partie du discours sont regroupées et la fréquence absolue de la forme est ajustée en conséquence.

Les cellules ombragées du tableau VII ne sont pas utilisées au cours de la segmentation du document. L'information contenue dans

ces cases sera ajoutée à la matrice lors de l'étape d'acquisition des PLS (voir le point 4.2.3). La matrice précédente (Tableau VII) ne permet que de connaître la liste des formes qui composent le texte et ne permet pas de reconstituer le texte comme un enchaînement d'occurrences. Afin de faciliter cette représentation à deux dimensions, une autre matrice contient les détails de chacune des occurrences d'une forme.

Information	**Description**
Numéro de document	Numéro unique associé à un document
Numéro de forme	Numéro unique associé à chaque forme
Position	Position en nombre de mots
Graphie originale	Graphie originale de la forme en contexte

Tableau VIII. Matrice générée pour chaque occurrence d'une forme

Les deux premières rangées de la matrice relient l'information à la matrice décrite précédemment (Tableau VII). La position dans le texte est décrite en nombre de mots depuis le premier mot du texte qui possède la position 1. Étant donné que la graphie des occurrences est normalisée dans la matrice précédente, la seconde matrice est utilisée pour conserver la graphie recensée dans le texte analysé et pour la rétablir au besoin. Chaque contexte pourra ainsi être fidèlement recomposé.

4.2.1.3 Lemmatisation

La technique de lemmatisation des données utilisée pour le traitement du corpus d'analyse est conforme à celle décrite au point 3.3.5. Nous ne la décrivons donc pas dans la présente section de la thèse. Le logiciel balaie la liste des entrées de la matrice et applique les règles de lemmatisation décrites précédemment. Lorsqu'une forme

plurielle est repérée et lemmatisée, sa fréquence est automatiquement additionnée à la fréquence absolue de la forme singulière correspondante.

L'étape de lemmatisation a une influence non négligeable sur l'étape d'acquisition automatique des termes et nous désirons en décrire les impacts des règles sur la liste de CT obtenus. Le tableau qui suit présente divers cas relevés dans les documents analysés. Pour les besoins de la présente thèse, nous avons modifié TermoStat de façon à rendre l'étape de lemmatisation optionnelle et à observer l'impact sur les résultats.

La première colonne présente des exemples obtenus sans lemmatisation alors que la deuxième contient des exemples obtenus alors que le logiciel procédait à une lemmatisation des CT. On trouve, dans la dernière colonne, notre évaluation de l'impact des règles de lemmatisation sur la qualité des CT obtenus. L'impact est qualifié de *positif* lorsqu'il facilite l'acquisition des termes et de *négatif* lorsqu'il résulte en l'inclusion de bruit dans la liste des CT.

Sans lemmatisation	Avec lemmatisation	Impact
adapter *adapter kit table* *adapter kits description* *adapterless circuit pack* ***adapterless circuit packs*** ***adapters***	*adapter* *adapter kit description* *adapter kit table* *adapterless circuit* *adapterless circuit pack*	Positif
additional steps	*additional steps*	Nul
amplifier group number *amplifier groups*	*amplifier group number*	Négatif
amplifier connector losses	*amplifier connector los*	Négatif négligeable
application description topologies	*application description topology*	Positif
network element *network element configurations* *network element connections* *network element equipment* ***network elements***	*network element* *network element configuration* *network element connection* *network element equipment*	Positif
wavelength translator *wavelength translator application* *wavelength translator circuit* *wavelength translator interfaces* ***wavelength translators***	*wavelength translator* *wavelength translator application* *wavelength translator circuit* *wavelength translator interface*	Positif

Tableau IX. Effet de la lemmatisation sur l'acquisition des CT

Comme l'illustrent les données du tableau IX, la majorité des décisions prises par l'algorithme de lemmatisation ont un effet positif sur la liste des CT retenus. Les CT retenus lors de la première analyse et qui apparaissent en caractères gras dans la première colonne ont été reconnus comme des variations de CT déjà repérés au singulier et éliminés dans la liste finale.

On note aussi une lemmatisation erronée de la forme *losses* (*amplifier connector losses*) en *los* (*amplifier connector los*), qui conduit à la présentation d'un CT plus difficile à analyser par le terminologue consultant la liste produite par le système. Cependant, l'information est tout de même conservée. Nous considérons que l'impact sur la qualité des résultats, bien que négatif, est négligeable puisque la situation peut être corrigée par le terminologue.

Dans le cas de la perte du CT *amplifier groups*, nous considérons qu'elle n'est pas négligeable parce que les deux CT correspondent à des concepts différents et que le générique ne peut être identifié. Cette règle a donc un effet négatif sur les résultats obtenus à l'aide du logiciel. En effet, le CT *amplifier group* devrait avoir été retenu, mais a tout simplement été écarté.

4.2.2 Identification des PLS

L'identification des PLS constitue la première étape de traitement des corpus par le logiciel TermoStat. Lorsque cette étape est terminée et que la liste des PLS a été dressée, le logiciel peut concentrer ses efforts aux contextes immédiats de ces quelques formes spécifiques au corpus d'analyse afin de procéder à l'acquisition automatique des termes.

Les résultats de l'acquisition des PLS peuvent être analysés sous divers angles. Les paragraphes qui suivent s'attardent au rapport entre les spécificités et les PLS et aux liens qui existent entre les PLS et les substantifs les plus fréquents du corpus d'analyse.

Afin de vérifier la stabilité de la liste des PLS obtenus et l'adéquation du corpus de référence utilisé, nous avons effectué des

tests au cours desquels nous avons fait varier la composition du corpus de référence utilisé. Les PLS sont aussi validés afin d'évaluer la pertinence des formes en fonction du corpus d'analyse puisqu'elles serviront ensuite à l'acquisition automatique des termes.

4.2.2.1 Seuil sélectionné pour les valeurs-tests

Les PLS sont des spécificités positives qui possèdent une fréquence observée qui a moins de 1 chance sur 1 000 d'être le fruit du hasard au sein d'un corpus technique. Cette probabilité correspond à une valeur-test égale ou supérieure à 3,09. La sélection d'un tel seuil conduit à la mise à l'écart de certaines formes qui ont un comportement spécifique au corpus d'analyse.

La présente section s'intéresse plus particulièrement aux formes qui possèdent une valeur-test qui se situe sous le seuil que nous avons adopté, mais dont la fréquence observée dans le corpus d'analyse a moins de 1 chance sur 100 d'être due au hasard. Il s'agit donc des unités dont la valeur-test se situe entre 1,96 et 3,09.

	CA_1	**CA_2**	**CA_3**
Nombre PLS +	784	1454	655
Nombre PLS	528	941	374
Ratio	67,4 %	64,7 %	57,1 %

Tableau X. Ratio PLS versus PLS +

Le tableau X représente le ratio observé entre le nombre de formes dont la valeur-test se situe au-delà de 1,96 (PLS +) et le

nombre de PLS observés dans chacun des corpus d'analyse[51]. En moyenne, sur l'ensemble des documents, on observe que 35 % des PLS+ ne sont pas pris en compte dans le cadre de l'acquisition des termes. L'annexe A dresse la liste complète des formes qui font partie des PLS+, mais qui n'ont pas été retenues à titre de PLS pour chacun des documents du CA.

Les listes sont triées en ordre décroissant de valeur-test. Les formes en tête de liste sont donc les plus caractéristiques des corpus parmi celles qui n'obtiennent pas la note de passage pour être considérées comme des PLS. Dans cette même liste, nous avons mis en caractères gras les substantifs[52] puisqu'ils constituent les formes les plus intéressantes.

Les substantifs sont en effet l'élément déclencheur de l'acquisition des termes pour le logiciel TermoStat. L'élimination de ces derniers de la liste peut donc conduire à une perte du point de vue du rappel puisque certains CT valides ne sont pas recensés.

	CA_1	**CA_2**	**CA_3**
Nombre PLS+ >=1,96 <= 3,09	48	94	68
Nombre de substantifs	22	34	19

Tableau XI. Nombre de substantifs au sein des PLS+ écartées

[51] Les PLS (valeur-test >=3,09) constituent donc un sous-ensemble des PLS+ (valeur-test >= 1,96).

[52] Certaines de ces unités ne sont pas réellement des substantifs (*change, none*, etc.), mais elles ont été identifiées comme des substantifs au cours du traitement. Étant donné que la catégorisation faite par l'étiqueteur n'a pas fait l'objet d'une révision manuelle, certaines erreurs du genre ont été introduites dans les corpus.

Le tableau XI présente le nombre de substantifs rencontrés au sein du sous-ensemble des PLS+ écartés. La prise en charge de ces formes par le logiciel au cours de l'acquisition automatique des termes conduit à l'apparition de CT supplémentaires. Le tableau XII présente la productivité des formes écartées.

	CA_1	**CA_2**	**CA_3**
CT valides	10	13	9
CT non valides	7	12	3

Tableau XII. Répercussion de l'exclusion de certains substantifs des PLS

Une analyse de ces résultats conduit à la constatation que, d'un point de vue du rappel, l'inclusion de ces formes serait intéressante puisque le logiciel pourrait faire l'acquisition de 10 CT valides dans le corpus CA_1, de 13 dans le deuxième corpus et de 9 dans le dernier corpus. La couverture de l'algorithme en termes de rappel bénéficierait donc légèrement de l'inclusion de ces formes au sein des PLS.

D'un point de vue de la précision, le gain varie de façon importante selon le corpus et il est nettement plus important dans le corpus CA_3. Il est cependant intéressant de noter que, dans tous les cas, l'inclusion de ces substantifs au sein des PLS pourrait conduire à une légère augmentation de la précision. À la lumière des résultats obtenus dans les documents qui composent le corpus d'analyse, on peut donc conclure que l'utilisation d'un seuil moins élevé, soit une valeur-test égale ou supérieure à 1,96, conduirait à une légère augmentation de la performance des algorithmes, tant du point de vue du rappel que de la précision.

4.2.2.2 Fréquence versus PLS

La méthode documentée au point 4.1.2.2 accorde une place de choix à la fréquence des formes. On pourrait donc croire qu'il existe une corrélation exacte entre les substantifs les plus fréquents d'un corpus et les PLS. La présente section a donc pour but de vérifier si l'importance accordée à la fréquence est démesurée et si cette corrélation existe.

Les figures qui suivent (4, 5 et 6) illustrent la répartition des substantifs et des PLS en fonction de la fréquence dans les trois documents du corpus d'analyse[53]. On remarque que la distribution des PLS et des substantifs le long de l'éventail des fréquences est la même pour les trois documents. On note aussi que, dans tous les cas, les fréquences les plus basses conduisent à un plus grand nombre de substantifs éliminés de la liste des PLS.

---

[53] Les données correspondantes sont présentées à l'annexe C.

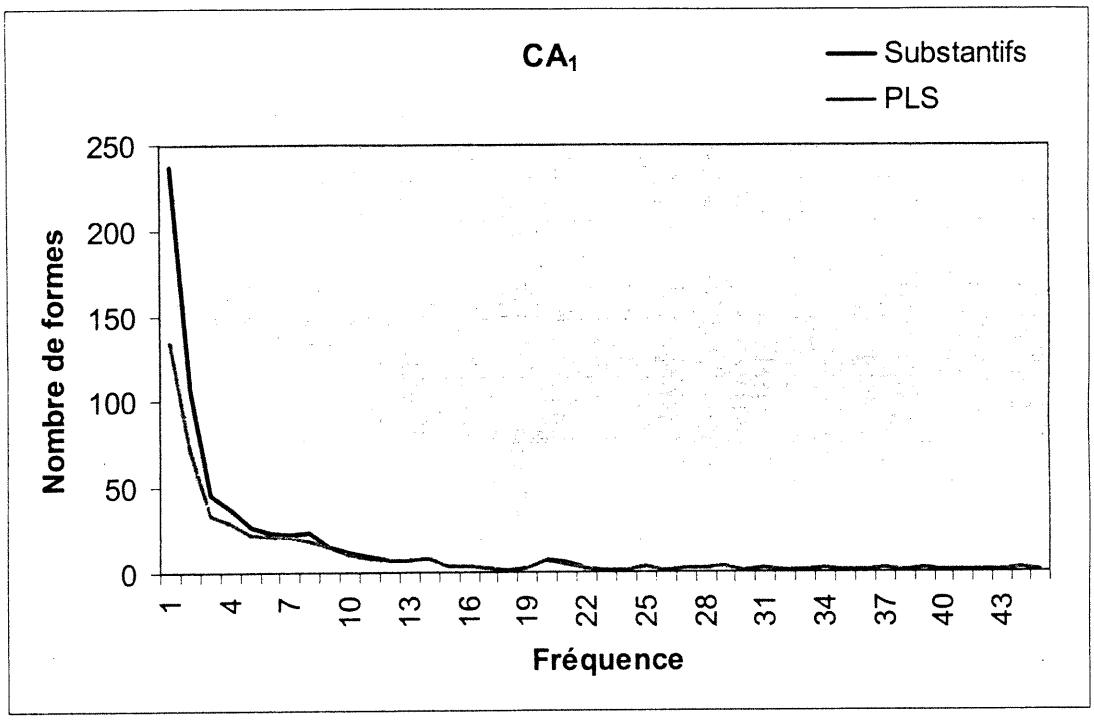

Figure 4. Répartition des PLS en fonction de la fréquence dans CA1

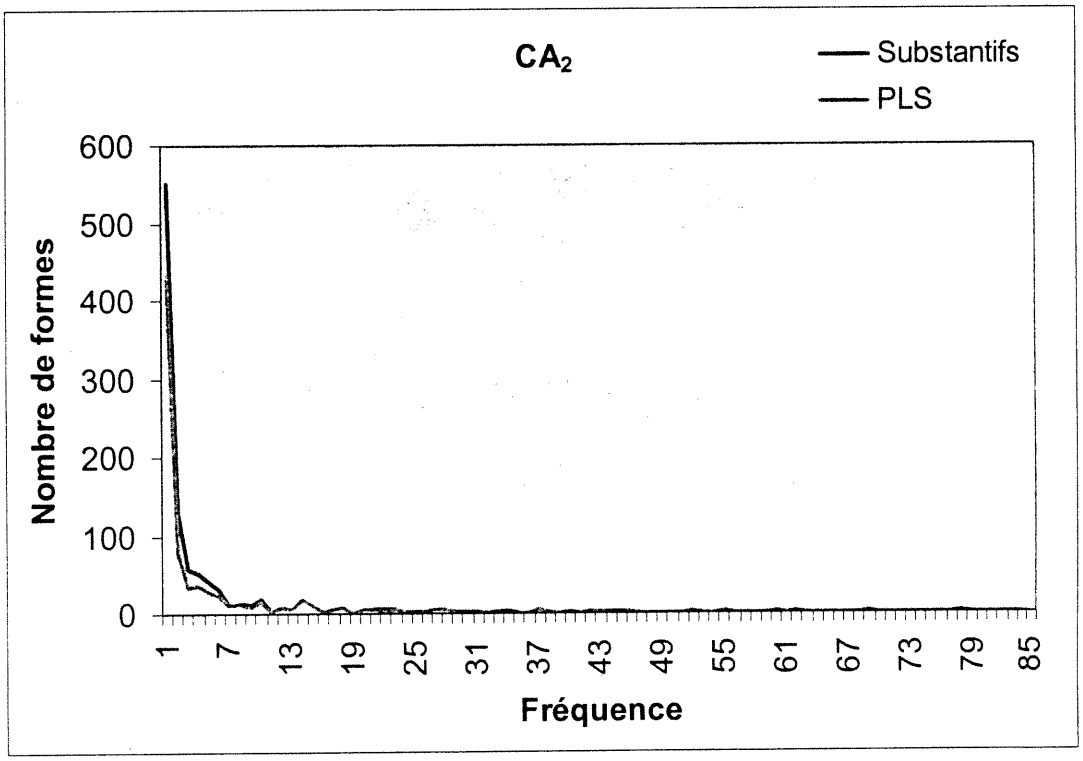

Figure 5. Répartition des PLS en fonction de la fréquence dans CA_2

Malgré la différence importante de taille entre les documents CA_1 et CA_2, on remarque des similitudes entre ces derniers. Les formes ayant une fréquence inférieure à 5 sont les plus touchées par l'algorithme de sélection des PLS.

Ce type de comportement de la part de l'algorithme d'acquisition des PLS n'est cependant pas surprenant puisque le ratio de formes éliminées correspond à la répartition des formes selon la fréquence dans le corpus. En effet, comme c'est le cas dans la majorité des corpus, les formes les plus fréquentes sont aussi les moins nombreuses au sein de tous les corpus. Il est donc tout à fait normal que l'effet de la sélection se fasse sentir de façon plus importante au sein des basses fréquences.

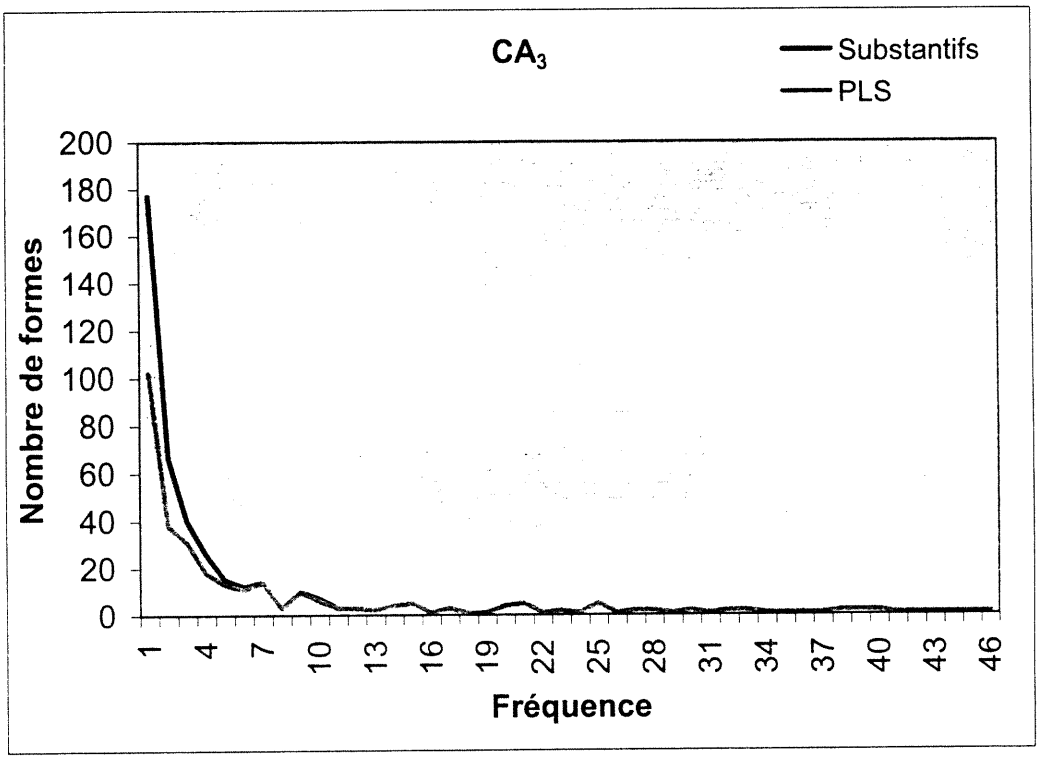

Figure 6. Répartition des PLS en fonction de la fréquence dans CA_3

Malgré que les graphiques précédents soient utiles et qu'ils apportent une bonne description de la répartition des PLS en fonction de la fréquence, ils ne permettent pas de bien saisir le travail de sélection de l'algorithme pour les fréquences les plus élevées. Le tableau présenté dans l'annexe C contient la liste des substantifs qui ont été éliminés par l'algorithme d'identification des PLS pour chaque corpus d'analyse.

Comme les figures précédentes, la consultation de la liste permet de constater que peu de formes fréquentes sont mises à l'écart. Par contre, des spécificités ayant une fréquence absolue élevée ont été écartées de la liste : *time* (CA_1 - 22), *end* (CA_1 - 11), *m* (CA_2 - 35), *group* (CA_2 - 22), *point* (CA_3 - 10), *building* (CA_3 - 7). Le document CA_1 ne contient que 246 spécificités dont la fréquence est supérieure à 8 sur un total de 1 237 et la sélection des PLS procède à l'élimination de 7 d'entre elles. Le même phénomène est observé pour les autres documents du CA; en effet seules 17,6 % des spécificités ont une fréquence supérieure à 10 dans CA_2 et 30,6 % dépassent le seuil de 4 dans CA_3.

En conclusion, nous pouvons dire que les formes les plus fréquentes sont donc plus souvent considérées comme des PLS que les substantifs de fréquence basse. Par contre, il n'y a pas de correspondance exacte entre les PLS et les substantifs les plus fréquents du corpus d'analyse.

4.2.2.3 Stabilité des PLS

Les tests effectués dans les paragraphes qui suivent ont pour but de vérifier l'influence du corpus de référence sur la stabilité des PLS. Afin de tester les impacts de la variation du CR sur la liste

obtenue, ce dernier a été divisé en quatre sous-corpus nommés CR1, CR2, CR3 et CR4. Ces quatre corpus ont une taille identique.

Afin de procéder à des tests sur des corpus ayant des tailles diverses et des contenus plus ou moins aléatoires, les corpus CR1 et CR3 ont été concaténés en un corpus plus important nommé CR1_CR3. L'acquisition des PLS a été effectuée sur le document CA_1 à l'aide des corpus de référence CR, CR2, CR1_3 et CR4. les tests ont donc pour but d'opposer le lexique du corpus d'analyse à différents corpus de référence afin d'identifier les divergences au sein de la liste des PLS obtenus.

La première vague de tests avait pour but d'opposer la liste des PLS obtenus à l'aide du corpus CR2 à celle obtenue à partir de CR4. Ces corpus ont une taille identique. Une deuxième comparaison a été faite entre le corpus de référence original (CR) et le corpus de CR1_3 dont la taille correspond à la moitié de celle du premier. Finalement, afin de maximiser les écarts de taille entre les corpus, la liste obtenue à partir du CR a été comparée à celle de CR4.

Une comparaison manuelle des listes a été effectuée pour identifier les différences entre ces dernières. La figure 7 recense ces différences en fonction de la fréquence des formes relevées. Par exemple, nous avons identifié 27 différences entre les PLS obtenus à l'aide du corpus de référence CR et ceux obtenus à l'aide de CR1_3. Les différences relevées sont observées sur les deux listes. Ainsi un PLS qui se trouve dans la liste obtenue à partir de CR et qui n'est pas recensé à partir de CR1_3 constitue une différence; la situation inverse est aussi vraie.

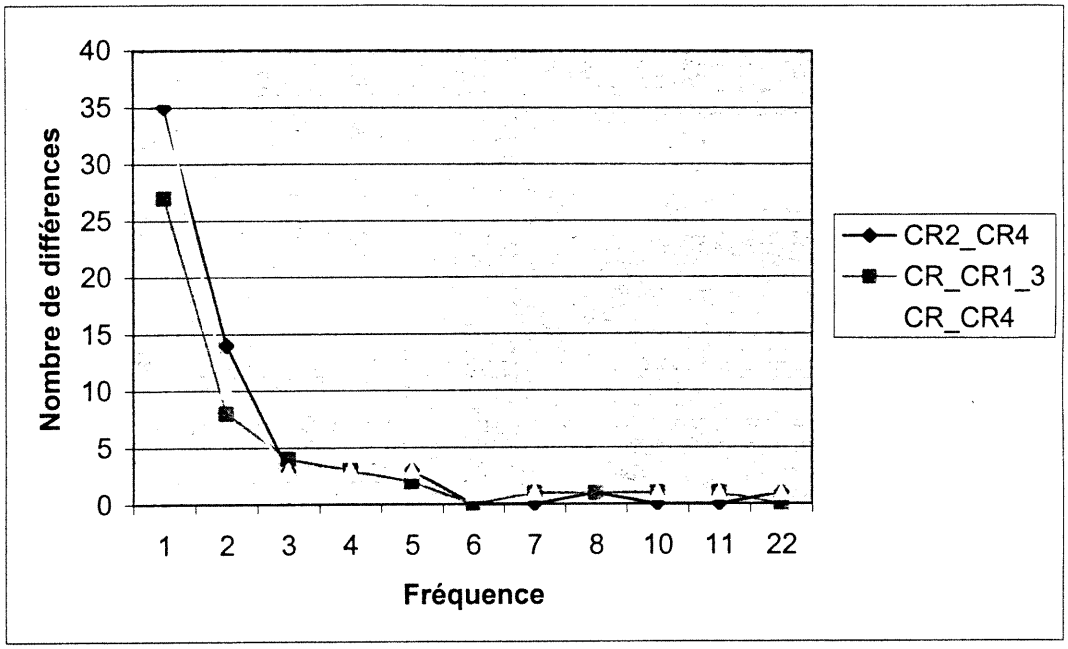

Figure 7. Variation du CR et stabilité des PLS

On constate rapidement que les variations observées sont sensiblement les mêmes sur les diverses tranches du corpus de référence. C'est sur le plan des formes dont la fréquence est basse (inférieure à trois) que les listes divergent le plus. Les divergences s'atténuent cependant très rapidement lorsque la fréquence est supérieure à ce seuil. Le nombre total de divergences entre les listes n'est cependant pas important et il est de l'ordre de 20 % pour la comparaison de CR et de CR4, de 11,5 % pour CR2 et CR4 et de 9,2 % pour CR et CR1_3.

On peut donc conclure, à la lumière de ces expérimentations, que le corpus de référence a une influence sur les PLS qui sont recensés. Étant donné l'écart moins important observé entre les corpus dont les tailles se rapprochent le plus l'une de l'autre, soit CR et CR1_3, on peut postuler que plus la taille du corpus de référence

utilisé est importante, plus la liste des PLS est fiable. Ainsi, des corpus de référence de taille semblable risquent de conduire à des résultats comparables. La vérification de cette hypothèse se situe cependant en marge de la présente thèse.

Pour les besoins de notre approche, il suffit de constater que les PLS obtenus à partir de divers corpus de référence sont relativement stables et indépendants des corpus utilisés. Cette affirmation est particulièrement vraie pour les PLS dont la fréquence est supérieure à deux. L'interprétation des résultats obtenus lors de l'acquisition automatique des termes à l'aide des PLS de fréquence inférieure à ce seuil devra donc se faire avec prudence.

4.2.2.4 Validation des PLS

Afin de valider les données issues de l'acquisition des PLS, nous avons recours à une banque de terminologie et à des terminologues spécialistes du domaine des télécommunications. La banque de terminologie a été mise à notre disposition par la société Nortel Networks et comporte essentiellement de la terminologie reliée au domaine des télécommunications.

La validation à l'aide de la banque de terminologie consiste en une comparaison à plat des listes de PLS construites à partir des documents qui composent le CA et de la liste extraite de la banque. Les PLS qui sont présents au sein de la banque de terminologie sont considérés comme pertinents, les autres sont ensuite soumis à une validation humaine afin de juger de leur pertinence. Les consignes données aux terminologues pour la validation des PLS sont simples et ces derniers doivent se limiter à évaluer les deux aspects :

- la pertinence de la forme pour le document,
- la pertinence de la forme pour le domaine des télécommunications.

Ainsi, si la forme est utilisée dans le domaine des télécommunications ou est représentative du contenu du document, elle est alors considérée comme valide. En effet, l'analyse des PLS ne peut être utilisée que pour déterminer la pertinence par rapport à un corpus particulier[54] tiré d'un domaine d'activité plus ou moins spécifique. Ces consignes étendent ainsi le champ de validité d'un PLS à l'ensemble d'un domaine et non uniquement à un corpus. Certaines unités lexicales pourraient en effet paraître banales au sein d'un corpus, mais elles n'en demeurent pas moins essentielles. Dans le cas des documents du corpus d'analyse, le substantif *network* en est un bon exemple puisque son apparition au sein des documents n'est pas étonnante bien qu'elle soit en relation directe avec le domaine d'activité.

L'étape de validation, bien que fiable, est subjective. Le recours à une validation humaine, bien qu'il soit souhaitable, rend la démarche non systématique et dépendante de l'humain qui procède à la validation des résultats. Une approche visant à évaluer la constance de la validation faite par les terminologues et pondérant les

---

[54] On pourrait objecter à cette affirmation qu'un corpus représentatif d'un domaine dans son ensemble pourrait permettre d'obtenir un tel résultat d'un algorithme d'acquisition des spécificités. Par contre, une telle représentativité est difficile à décrire et, par le fait même, à démontrer. Nous préférons donc parler de représentativité des résultats par rapport à un corpus tiré d'un domaine d'activité.

résultats serait peut-être souhaitable, mais elle se situe au-delà des objectifs de la présente thèse.

	CA_1	**CA_2**	**CA_3**
PLS pertinents	444	810	273
PLS non pertinents	84	131	101
Précision	84,1 %	86,1 %	73,0 %

Tableau XIII. Évaluation de la pertinence des PLS pour les CA

Le tableau ci-dessus donne la précision du processus d'acquisition des PLS du logiciel TermoStat pour les trois documents qui composent le corpus d'analyse. Cette bonne performance doit être interprétée en contexte et en fonction des consignes données lors de l'étape de validation. Ces formes sont donc, dans la majorité des cas, représentatives du corpus ou du domaine des télécommunications.

Les spécificités qui n'ont pas été retenues comme PLS n'ont pas fait l'objet d'une validation par les terminologues. Il est donc difficile de savoir si des formes pertinentes ont été écartées du processus d'acquisition des termes. La lecture de l'annexe C (Substantifs non retenus à titre de PLS) apporte cependant quelques renseignements à ce sujet.

Des formes comme *time, rate, process* dans le CA_1, *house, loss, exchange* dans le CA_2 ou encore *point, building, state* et *manager* dans le CA_3 sont typiques des documents liés au domaine des télécommunications, mais leur fréquence ne leur permet pas de se distinguer dans le corpus d'analyse. Dans tous les cas, il s'agit de formes polysémiques ayant un sens non technique (mot) et un sens relevant du domaine des télécommunications (terme). La spécificité de

ces formes est donc sémantique et non purement lexicale. Le processus d'acquisition des PLS ne permet malheureusement pas d'identifier cette particularité.

4.2.2.5 Conclusion

Nous avons proposé, au point 4.2.2, des techniques statistiques ayant pour but d'isoler des phénomènes linguistiques, plus particulièrement, des phénomènes lexicaux. La méthodologie élaborée permet d'opposer le comportement du lexique d'un corpus d'analyse à celui d'un corpus de référence. Ce dernier corpus est considéré comme norme pour les besoins de l'expérimentation et les déviations significatives du comportement du lexique dans le corpus d'analyse sont recensées automatiquement par un outil informatisé.

Nos algorithmes, qui font appel à la loi normale, mettent à notre disposition deux indices pour la classification des résultats : la probabilité et la valeur-test. Étant donné le manque de discernement qu'entraîne une description des résultats à l'aide de la probabilité, nous avons décidé d'utiliser la valeur-test pour recenser les spécificités positives. Des contraintes propres à notre approche sont ensuite ajoutées pour isoler, au sein des spécificités, le sous-ensemble lexical des PLS. Ces derniers sont des adjectifs ou des substantifs anormalement fréquents identifiés au sein d'un corpus technique et dont la fréquence observée a moins de 1 chance sur 1 000 d'être le fruit du hasard.

À la lumière des résultats obtenus, on peut conclure que, malgré une intersection importante au niveau des fréquences les plus élevées, il n'y a pas de correspondance exacte entre les PLS et les substantifs les plus fréquents d'un corpus d'analyse. On observe

même une tendance de l'algorithme d'identification des PLS à éliminer une plus grande quantité de substantifs lorsque l'étendue du corpus d'analyse augmente.

Malgré l'absence de marquage sémantique dans les corpus de référence et d'analyse, le prototype TermoStat est à même de dresser une liste de PLS qui sont, en majorité, pertinents par rapport au contenu du corpus d'analyse et au domaine d'activité auquel appartiennent les corpus.

4.2.3 Acquisition des termes

L'acquisition automatique des termes est une étape cruciale dans le traitement des documents, car c'est à ce moment que sont recensés tous les CT par le logiciel TermoStat. La présente section se divise en trois parties. La première présente la structure de données utilisée par le logiciel TermoStat afin de procéder à l'acquisition des termes. La deuxième partie est consacrée à une description des contraintes imposées au processus d'acquisition. C'est par le biais de ces contraintes que se définissent les principes de base de notre méthodologie. Une analyse des résultats est enfin présentée pour vérifier dans quelle mesure l'approche à l'aide des PLS conduit à une augmentation de la précision.

4.2.3.1 Structure des données de TermoStat

La présente section de la thèse décrit les structures de données utilisées par TermoStat pour entreposer les CT qu'il recense au sein des corpus d'analyse. Les CT sont, comme dans le cas des formes, gérés à l'aide de deux matrices séparées, une pour les CT et une pour les occurrences des CT.

Information	Description
Numéro de document	Numéro unique associé à un document.
Numéro de CT	Numéro unique associé à chaque CT.
Candidat	Candidat-terme.
Longueur	Nombre d'unités lexicales composant le terme.
Fréquence absolue	Fréquence du CT dans le document.
Tête	Tête du terme.
Matrice	Structure syntagmatique du CT

Tableau XIV. Matrice générée par TermoStat pour chaque CT

Le tableau XIV correspond à la matrice que TermoStat remplit pour chacun des CT. Le numéro de document est un numéro unique séquentiel attribué à chaque document analysé; ce numéro est utilisé à l'interne par le logiciel TermoStat, qui offre la possibilité de traiter plusieurs documents et de conserver les résultats en mémoire. Les CT reliés à un document sont donc regroupés autour d'un numéro de document.

Quant au numéro de CT, il s'agit d'un nombre unique attribué à chacun des CT d'un même document. Ce numéro est aussi attribué de façon séquentielle à l'intérieur d'un même document. Le CT est ajouté à la matrice sous un aspect banalisée; c'est-à-dire qu'il est systématiquement converti en caractères minuscules. La longueur du CT est exprimée en nombre de mots. Étant donné qu'une seule entrée est créée par CT, un champ permet de conserver sa fréquence d'occurrence. Les deux derniers éléments d'information sont de nature plus terminologique; au sein de chaque CT, la forme située à

l'extrême droite est identifiée comme la tête potentielle du terme[55]. Le dernier élément d'information contient la matrice syntagmatique du CT.

Numéro Document	Numéro CT	CT	Longueur	Fréquence Absolue	Tête	Matrice
1	1	*Feature*	1	45	*feature*	*N*
1	2	*object class*	2	4	*class*	*N N*
1	3	*attribute list*	2	5	*list*	*N N*
1	4	*termination point*	2	8	*point*	*N N*
1	5	*office terminal*	2	56	*terminal*	*N N*
1	6	*graphical network browser*	3	34	*browser*	*A N N*

Tableau XV. Extrait de la matrice générée par TermoStat

Le tableau XV illustre une tranche de la matrice élaborée pour un corpus d'analyse. Nous ne présentons que les premières entrées dans la liste. Afin de permettre un accès rapide et facile au logiciel, les données sont stockées par TermoStat dans une base de données relationnelle.

Le logiciel permet de trier la matrice selon divers critères et d'en consulter les résultats. On peut ainsi trier les CT par ordre alphabétique, de fréquence, etc. La présence d'une colonne spécifiant

---

[55] TermoStat ne possédant pas d'information sur la structure syntaxique du CT, le logiciel identifie la forme la plus à droite comme étant la tête du syntagme. Dans la majorité des cas, en anglais, cette information se révèle juste; voir à ce sujet, Sager *et al.* (1980 : 268).

la tête de chaque CT permet aussi de regrouper les CT en fonction du mot constituant la tête des CT.

Information	**Description**
Numéro de document	Numéro unique associé à un document.
Numéro de CT	Numéro unique associé à chaque CT.
Position	Position dans le texte en nombre de mots.

Tableau XVI. Matrice générée pour chaque occurrence d'un CT

La matrice qui précède ne permet pas d'établir de lien entre la liste des CT et le texte d'où ils sont extraits. Une seconde matrice vient ajouter une autre dimension à l'information emmagasinée par TermoStat sur les CT qui sont prélevés en cours d'analyse. Pour chaque occurrence du CT dans le texte, une autre matrice est alimentée afin de permettre au logiciel de pouvoir revenir sur les occurrences du CT.

Les deux premiers champs de la matrice décrite par le tableau XVI correspondent à ceux du tableau précédent. La dernière colonne du tableau exprime la position du CT dans le document analysé. Le tableau XVII correspond à la partie de la matrice comportant l'information relative au CT *object class*.

Numéro Document	**Numéro CT**	**Position**
1	3	2
1	3	45
1	3	351
1	3	784

Tableau XVII. Extrait de la matrice générée pour chaque occurrence d'un CT

C'est cette seconde matrice qui permet au logiciel de préserver chacun des contextes pour tous les CT. Le terminologue utilisant le logiciel peut ainsi naviguer dans le texte d'une occurrence d'un CT à une autre et être à même de juger de la validité des entrées retenues par TermoStat. Les deux matrices sont donc complémentaires puisque la première représente l'ensemble des CT recensés alors que la seconde permet d'obtenir de l'information sur les occurrences des CT en contexte dans le corpus d'analyse.

4.2.3.2 Contraintes imposées aux algorithmes

Ce travail de repérage des termes potentiels s'effectue dans le cadre de contraintes précises imposées au logiciel. On peut associer ces dernières à une vision prototypique[56] des réalisations potentielles du terme en discours. Il est en effet très difficile de décrire exactement ce qui constitue un terme[57] et les logiciels ne peuvent avoir recours qu'à des descriptions approximatives, ou encore incomplètes, de ce qui constitue un terme. Ces descriptions permettent d'encadrer le travail du logiciel et de prendre des décisions sur le fait de recenser ou d'ignorer une forme ou une suite de formes dans un contexte précis.

L'élaboration d'algorithmes d'acquisition automatique de termes réside ainsi dans l'art du compromis. Il faut savoir cerner l'ensemble des observations théoriques et empiriques faites au sujet des termes et les transformer en contraintes. La souplesse ou la sévérité de ces contraintes vient déterminer la couverture du premier repérage, la

---

[56] Voir Kleiber (1999) au sujet de la notion de *prototype*.

[57] Si la description des réalisations textuelles du terme était chose facile, le défi de l'acquisition automatique des termes serait sans intérêt.

performance du logiciel en termes de rappel. Des restrictions plus lâches imposées au logiciel font augmenter le rappel alors que l'inverse entraîne sa diminution rapide.

Les limites que nous avons choisies d'imposer à nos algorithmes sont de natures diverses et tiennent à la fois de la linguistique, de la statistique et de l'informatique. Le contexte suivant nous permettra d'illustrer la façon dont le logiciel perçoit le texte à l'étude au cours de son cheminement.

> *For Dual MSA sites (line sites with high OADM counts) shown in Figure 4-12, the signal flow is the same except that a second MSA (DSCM or OADM filter) is placed between the Booster18 and Booster21 circuit packs.*

Le texte est d'abord segmenté et se présente donc comme une liste de formes simples; il est ensuite soumis à l'étiqueteur de Brill. L'extrait suivant consiste en une sortie brute telle qu'elle est mise à la disposition de TermoStat par l'étiqueteur. Comme pour le corpus de référence, nous avons décidé de ne pas procéder à une étape supplémentaire d'apprentissage lors de l'étiquetage du corpus d'analyse.

> For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

Dans les paragraphes qui suivent, ce contexte sera repris à titre d'exemple afin d'illustrer le travail de TermoStat et la progression de l'algorithme d'acquisition automatique des termes.

4.2.3.2.1 Tête des CT

La principale contrainte imposée comme point de départ pour l'algorithme est l'utilisation d'une forme nominale tirée de la liste des PLS comme tête des CT. La notion de tête repose sur le découpage binaire des unités syntagmatiques en une tête et une expansion pouvant toutes les deux être des unités complexes ([E](*IBM System/360 operating system*) [T](*assembler programme*)). Dans cet exemple, tiré de Sager *et al.* (1980 : 273), le terme complexe *assembler programme* est déterminé par l'unité terminologique complexe *IBM System/360 operating system.* Les têtes prises en charge par TermoStat ne sont cependant que des unités simples et, dans ce cas précis, la forme *programme* serait considérée comme la tête du CT. Cette vision du CT n'est pas erronée puisque *programme* correspond effectivement à la tête d'*assembler programme.*

Comme l'ont démontré les travaux de Bourigault (1994b) et de Assadi et Bourigault (1996), la productivité[58] de la tête est un bon indice de la qualité des CT. Elle n'est cependant disponible qu'à la fin du processus d'acquisition automatique des termes et ne peut donc pas être utilisée comme point de départ par nos algorithmes. Le critère de fréquence utilisé dans les travaux sur la productivité n'a vraiment de sens que lorsqu'une liste de termes potentiels a été élaborée par le logiciel. Cet indice ne peut donc pas être utilisé en

---

[58] La productivité correspond au nombre de CT différents engendrés par une même tête.

cours de repérage. L'utilisation de la fréquence de la tête comme critère décisif pour l'acquisition des termes ne saurait être suffisante en elle-même; c'est pourquoi nous proposons d'utiliser les résultats des tests statistiques qui permettent de recenser les PLS.

La sélection des têtes à partir d'un test statistique nous permet d'éliminer dès le départ des CT potentiels qui se construiraient à partir de têtes moins pertinentes pour le domaine à l'étude. Nous croyons que cette sélection permet d'écarter initialement certains syntagmes nominaux récurrents qui n'auraient pas de valeur terminologique.

> For/IN Dual/JJ **MSA/NNP sites/NNS** (/(**line/NN sites/NNS** with/IN high/JJ **OADM/NNP counts/NNS**)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN **flow/NN** is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ **MSA/NNP** (/(**DSCM/NNP** or/CC **OADM/NNP filter/NN**)/SYM is/VBZ placed/VBN between/IN the/DT **Booster18/NNP** and/CC **Booster21/NNP circuit/NN packs/NNS** ./.

Le contexte précédent permet d'illustrer la vision de l'algorithme en fonction des têtes potentielles de CT identifiées à partir d'une liste de PLS. Dans l'extrait, les PLS nominaux apparaissent en caractères gras. Le logiciel TermoStat identifie donc des points chauds terminologiques autour desquels le travail d'acquisition sera effectué.

4.2.3.2.2 Frontières de termes

Notre conception des frontières de termes diffère quelque peu de celle qui a été exposée au point traitant des travaux de Bourigault (1992a, 1992b, 1993, 1994a et 1994b). En effet, le logiciel LEXTER détermine si une forme constitue une frontière en fonction de sa

partie du discours, mais aussi à partir du contexte immédiat de cette dernière. Notre conception des frontières de termes est plus lâche et les règles qui sont appliquées afin de déterminer ce qui constitue une frontière sont plus simples, car elles n'impliquent pas d'analyse syntaxique locale.

Nous qualifions nos règles de plus lâches parce qu'elles n'impliquent que deux catégories grammaticales qui ne constituent pas des frontières : l'adjectif et le substantif. Ainsi, toutes les autres catégories sont qualifiées de frontières sans égard au contexte. Par exemple, LEXTER procède à une analyse locale du contexte autour de la préposition afin de déterminer si elle constitue une frontière. TermoStat considère la préposition comme une frontière dans tous les contextes.

Les règles utilisées dans le cadre de notre démarche sont donc moins contraignantes linguistiquement que celles utilisées dans les travaux précédents sur les frontières de termes. Par contre, une contrainte extralinguistique vient s'ajouter : toutes les formes au sein d'un CT doivent faire partie de la liste des PLS. Ainsi, une dimension statistique s'ajoute à la définition de la frontière. Des adjectifs et des substantifs peuvent donc être considérés comme des frontières dans les cas où ils n'obtiennent pas le statut de PLS.

Notre décision de n'admettre au sein des CT que des adjectifs ou des substantifs se fonde sur les travaux précédents (Justeson et Katz 1993; Frantzi et Ananiadou 1997) effectués sur l'anglais. Les résultats obtenus par ces chercheurs indiquent qu'il s'agit de la combinaison optimale lorsqu'on cherche à limiter le bruit au cours du processus d'acquisition des termes. Les premiers auteurs considèrent que la préposition peut faire partie des termes, mais ils proposent

tout de même un prototype flexible qui permet de l'exclure des formes pouvant faire partie des CT. En effet, leurs résultats sont assez éloquents et prouvent que 97 % des CT sont composés uniquement d'adjectifs et de substantifs (Justeson et Katz 1993 : 3-4). Sur un total de 800 termes, ils recensent 2 termes qui contiennent une conjonction et 17 qui contiennent une préposition[59].

Frantzi et Ananiadou (1997 : 4) adoptent une approche plus restrictive qui exclut les prépositions et qu'elles opposent à celle de Justeson et Katz (1993), qu'elles qualifient d'ouverte. La justification de leur décision repose essentiellement sur le fait qu'elles cherchent à réduire le bruit que l'on retrouve dans les résultats.

Nos travaux rejoignent donc les précédents sur cet aspect puisque, nous tenons à le répéter, l'objectif de notre recherche n'est pas de maximiser la couverture des algorithmes d'acquisition automatique des termes, mais de dresser une liste de CT pertinents au détriment d'un dépouillement intégral des corpus. Cette décision d'exclure les prépositions des formes possibles a donc un impact direct sur les CT qui seront rassemblés par notre algorithme et certains termes sont laissés de côté.

En plus d'une modification du concept initial de frontière, nous procédons d'une certaine façon à son extension en prenant en compte de l'information telle que la ponctuation, les éléments de mise en page, etc. Ainsi, à l'exception du tiret, les signes de ponctuation sont considérés comme des frontières de terme. La conception de frontière exploitée dans la présente thèse repose donc à la fois sur une vision linguistique, statistique et textuelle du fonctionnement des termes en

---

[59] Dans 15 des 17 cas il s'agit de la préposition *of*.

corpus. En résumé, les éléments suivants sont considérés comme des frontières de termes :

- les unités qui n'obtiennent pas le statut de PLS,
- les signes de ponctuation à l'exception du tiret,
- les retours de chariot,
- les changements de cellule dans un tableau,
- les tabulations.

Reprenons le contexte que nous avons utilisé au point précédent afin d'illustrer les formes qui sont considérées comme des frontières soit par leur catégorie grammaticale, soit par l'information paratextuelle qui les entoure, soit par leur absence de catégorisation à titre de PLS :

> For/IN Dual/JJ **MSA/NNP sites/NNS** (/(**line/NN sites/NNS** with/IN high/JJ **OADM/NNP counts/NNS**)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN **flow/NN** is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ **MSA/NNP** (/(**DSCM/NNP** or/CC **OADM/NNP filter/NN**)/SYM is/VBZ placed/VBN between/IN the/DT **Booster18/NNP** and/CC **Booster21/NNP circuit/NN packs/NNS** ./.

Les unités lexicales soulignées constituent les frontières. L'identification de ces dernières constitue la dernière étape du processus d'acquisition automatique des CT. Les enchaînements lexicaux situés entre les frontières forment la première liste brute de termes potentiels. Avant d'obtenir officiellement le statut de CT, ils doivent satisfaire aux contraintes qui sont présentées dans les deux sections qui suivent.

En anglais, l'adoption d'une approche par frontière n'exclut pas nécessairement la description des résultats à l'aide d'une grammaire. L'équivalence de ces approches dans cette langue est rendue possible grâce à la morphologie des unités terminologiques plus simple qu'en français. Étant donné les contraintes énoncées précédemment, nous pouvons prévoir les structures valides qui seront recensées comme des termes potentiels. L'identification de toutes les catégories grammaticales, à l'exclusion des substantifs et des adjectifs, comme frontières de termes facilite la description des résultats en termes formels. Voici la grammaire qui permet de décrire les CT recensés dans le corpus d'analyse :

$$(\underline{A} \mid \underline{N})^{?}\ (\underline{A} \mid \underline{N})^{?}\ (\underline{A} \mid \underline{N})^{?}\ (\underline{A} \mid \underline{N})^{?}\ (\underline{A} \mid \underline{N})^{?}\ \underline{N}$$

où :

- A est un adjectif,
- N est un nom,
- (A | N) correspond à A ou à N,
- ? indique un élément qui peut être absent ou apparaître une seule fois,
- _ fait partie de l'ensemble des PLS.

Cette grammaire se rapproche de celle décrite dans Frantzi et Ananiadou (1997 : 4), mais elle impose une longueur maximale aux CT recensés qui ne fait pas partie des travaux des auteurs précédents. L'ajout d'une contrainte sur le statut de PLS des formes qui apparaissent au sein des CT est aussi un aspect nouveau. Les termes potentiels identifiés par l'algorithme vont donc du simple substantif à un enchaînement d'adjectifs et de substantifs d'une longueur maximale de six éléments.

Notre grammaire se distingue aussi de celle de Justeson et Katz (1993) par l'impossibilité d'insérer des prépositions dans les termes et par la contrainte sur la longueur maximale qui ne figure pas dans leurs travaux. Leur algorithme permet donc de repérer une chaîne de longueur indéterminée composée d'adjectifs, de substantifs et de prépositions. L'inclusion de ces dernières est cependant facultative et déterminée par l'utilisateur du logiciel.

À la lecture des paragraphes qui précèdent, on peut s'interroger sur la pertinence d'une approche par frontières lorsqu'une grammaire permet d'obtenir les mêmes résultats. Notre décision d'adopter une telle approche repose avant tout sur le postulat qu'une description méthodologique par frontières de termes facilitera l'adaptation de nos travaux à des langues qui ne sont pas aussi facilement décrites à l'aide de grammaires syntagmatiques.

4.2.3.2.3 Fenêtre de repérage

La fenêtre de repérage correspond au nombre de mots autour de la tête potentielle que le logiciel peut explorer à la recherche d'un CT. Comme le fait remarquer Goffin (1992 : 434) au sujet des unités terminologiques, les limites de l'expansion du terme ne sont pas déterminées à partir de critères logiques. On ne peut donc pas se fier à des critères purement syntaxiques afin de déterminer où se termine l'expansion d'un terme. Afin de faciliter l'identification des termes potentiels par un logiciel, nous devons tout de même contraindre les réalisations potentielles et mettre en place une fenêtre maximale qui sera utilisée par nos algorithmes.

Les travaux effectués sur l'acquisition automatique des termes nous ont amené à limiter la longueur des termes potentiels identifiés

par TermoStat. Le tableau suivant illustre les résultats obtenus par Justeson et Katz (1993 : 4) ainsi que par Nkwenti-Azeh (1994) sur l'anglais, et par Jacquemin (1996 : 456) sur le français.

	1 mot	2 mots	3 mots	4 mots	5+ mots
Justeson et Katz	29,5 %	54,5 %	12,4 %	3,9 %	0 %
Nkwenti-Azeh 1	9,15 %	71,86%	16,93 %	2,06 %	0 %
Nkwenti-Azeh 2	7,30 %	49,02 %	32,83 %	8,88 %	1,97 %
Nkwenti-Azeh 3	30,73 %	49,84 %	15,13 %	3,5 %	0,8 %
Jacquemin		74,5 %	9,2 %	0,9 %	0,1 %

Tableau XVIII. Pourcentage des CT en fonction de la longueur.

Les données du tableau XVII relatives aux travaux de Justeson et Katz (1993) sont tirées d'observations faites sur des nomenclatures de dictionnaires. Les observations qu'ils ont effectuées sur des corpus confirment cependant celles faites sur les nomenclatures (Justeson et Katz 1993 : 4-5). Ces auteurs constatent que lorsque la longueur du terme augmente, le nombre de termes repérés par leur algorithme diminue. Ils notent aussi que, règle générale, le nombre de termes composés de deux mots est plus élevé que le total de termes composés d'un seul mot.

La seule exception à cette dernière règle découle d'une observation sur des textes médicaux dans lesquels les formes simples sont plus nombreuses que les formes binaires. Justeson et Katz (1993) attribuent cette dernière observation au recours intensif à la composition à l'aide de racines latines et grecques, caractéristique de la langue médicale, mais moins fréquent dans les autres domaines analysés.

Les observations de Blaise Nkwenti-Azeh (1994) ont été effectuées sur trois corpus différents. Le premier corpus est une liste de termes extraits manuellement d'un corpus portant sur le domaine des communications par satellite (1994 : 67). La deuxième liste de termes examinée est tirée d'une banque de terminologie[60] (1994 : 67) alors que la troisième liste a été établie à partir d'un dictionnaire traitant de la théorie des antennes (1994 : 67).

Contrairement aux travaux précédents, ceux de Jacquemin (1996), qui portent sur le français, n'incluent pas de total pour les formes simples. Cette décision de l'auteur s'explique par le fait que ses travaux portent sur les diverses variations que subissent les termes complexes en contexte; les termes simples se situent donc en marge de ses intérêts de recherche.

On remarque une importante différence entre le pourcentage d'unités composées de deux mots repérées dans le cadre des trois recherches. Pour les travaux portant sur la langue anglaise, l'écart s'explique probablement par l'influence qu'a eue l'inclusion de textes médicaux sur les résultats de Justeson et Katz (1993), plus précisément sur le nombre de termes simples identifiés, qui fait diminuer significativement le total des termes composés de deux mots. Par contre, on constate dans tous les cas que le nombre de termes identifiés lorsque la longueur est supérieure à trois est plutôt négligeable et qu'il chute très rapidement.

En plus d'avoir une influence directe sur le nombre total de termes recensés, la longueur maximale utilisée par les algorithmes

---

[60] Il s'agit de la banque de terminologie EURODICAUTOM mentionnée dans la section 2.2.1.1.

agit directement sur la qualité des résultats. En effet, comme le démontre Jacquemin (1996 : 458), une utilisation d'une fenêtre de repérage de plus en plus large conduit à des résultats de moins en moins intéressants. Selon cet auteur, la précision et le rappel atteignent leur valeur maximale lorsqu'une fenêtre de trois mots est utilisée.

Les données que nous avons présentées dans le tableau XVII, ainsi que les arguments avancés par les auteurs précédents, nous amènent à sélectionner une longueur maximale de six mots pour nos CT. Une telle contrainte fait en sorte que TermoStat laisse de côté quelques termes, mais elle offre l'avantage de concentrer les efforts du logiciel sur les unités les plus fréquentes, de minimiser le bruit et d'offrir une bonne couverture de l'éventail des termes potentiels.

Reprenons le contexte utilisé précédemment pour illustrer l'influence de l'imposition d'une fenêtre de repérage d'une longueur de six mots. Les unités en caractères gras sont les PLS nonimaux, celles soulignées sont les frontières de termes et les crochets servent à délimiter la fenêtre de travail de TermoStat[61].

For/IN [[Dual/JJ **MSA/NNP**] **sites/NNS**] (/([[**line/NN**] **sites/NNS**] with/IN [[high/JJ **OADM/NNP**] **counts/NNS**])/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN [**flow/NN**] is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ [**MSA/NNP**] (/([**DSCM/NNP**] or/CC

---

[61] Il est à noter que les signes de ponctuation comptent pour un mot au même titre que les autres formes.

[[**OADM/NNP**] **filter/NN**]])/SYM is/VBZ placed/VBN between/IN the/DT [**Booster18/NNP**] and/CC [[[**Booster21/NNP**] **circuit/NN**] **packs/NNS**] ./.

L'algorithme de repérage des termes procède de façon séquentielle à partir d'un PLS qui sert de point de départ. Notre algorithme n'exploite que le contexte situé à gauche des unités nominales dont la fréquence a été jugée significative par les tests statistiques. Le recours à un repérage qui s'effectue de la droite vers la gauche respecte le mode de composition syntagmatique le plus fréquent pour les unités terminologiques complexes en anglais. En effet, dans la grande majorité des cas, la tête des termes se situe à l'extrême droite alors que l'expansion est à la gauche du terme.

- *msa*
- *dual msa*
- *sites*
- *msa sites*
- *dual msa sites*
- *line*
- *sites*
- *line sites*
- *oadm*
- *high oadm*
- *counts*
- *oadm counts*
- *high oadm counts*
- *flow*
- *msa*
- *dscm*
- *oadm*
- *filter*
- *oadm filter*
- *booster18*
- *booster21*
- *circuit*
- *booster21 circuit*

- *packs*
- *circuit packs*
- *booster21 circuit packs*

L'acquisition des CT situés entre les frontières dans l'exemple décrit précédemment donne naissance à la liste qui précède. On remarque que le logiciel procède à partir de la tête vers la gauche en ajoutant une occurrence à la fois au CT et en ajoutant ce CT à la liste. Ainsi, l'analyse du segment de texte [*dual msa sites*] à partir de la tête potentielle *sites* donne naissance, dans l'ordre, au CT *sites*, *msa sites* et *dual msa sites*. Les CT potentiels issus de cette étape de traitement doivent cependant satisfaire à un critère supplémentaire afin d'obtenir le statut de CT : ils doivent faire preuve d'une certaine autonomie linguistique.

4.2.3.2.4 Degré d'autonomie

La contrainte d'autonomie consiste à vérifier que les CT atteignent un certain degré de fonctionnement linguistique autonome dans le corpus d'analyse. En comparant les CT recensés, les recoupements entre ces CT et leur fréquence respective, le logiciel TermoStat procède à un élagage de CT. La technique consiste à examiner chacun des CT et à vérifier s'il est utilisé à titre de tête ou d'expansion d'un CT plus long. Si c'est le cas, la fréquence des deux CT sont comparées. Les CT qui sont inclus dans ces CT plus longs sont considérés comme des fragments potentiels de CT plus longs. À titre d'exemple, le CT *slot* est considéré comme un fragment des CT plus longs *slot filler* et *slot filler circuit*, tout comme *slot filler* est un fragment du CT plus long. La comparaison des fréquences des CT peut conduire à deux cas de figure :

- la fréquence du fragment et du CT qui l'inclut est identique,
- la fréquence du fragment est supérieure à celle du CT plus long.

Lorsque la fréquence d'un fragment est égale à celle d'un CT qui l'inclut, le logiciel ne conserve que la chaîne la plus longue. Étant donné que ces chaînes ont une fréquence identique, nous pouvons affirmer que le fragment n'a pas d'autonomie à l'extérieur de la chaîne plus longue au sein du corpus analysé. Il s'agit donc d'un fragment et non d'un CT valide. Si la fréquence d'un fragment potentiel est supérieure à une chaîne qui l'inclut, nous pouvons déduire que le premier a une autonomie linguistique qui lui est propre et qu'il doit ainsi être traité séparément et conservé puisqu'il s'agit d'un CT autonome.

Dans le tableau qui suit, composé des CT issus du contexte utilisé dans les paragraphes précédents, les CT rayés ont été éliminés puisqu'ils correspondaient à des fragments de termes ayant la même fréquence absolue que des CT plus longs.

CT	**Fréquence**
booster18	1
~~*booster21*~~	1
~~*booster21 circuit*~~	1
booster21 circuit packs	1
~~*circuit*~~	1
~~*circuit packs*~~	1
~~*counts*~~	1
dscm	1

~~*dual msa*~~	1
dual msa sites	1
~~*filter*~~	1
flow	1
~~*high oadm*~~	1
high oadm counts	1
~~*line*~~	1
line sites	1
msa	2
msa sites	1
oadm	2
oadm counts	1
oadm filter	1
~~*packs*~~	1
sites	2

Tableau XIX. Exemples d'application de la contrainte d'autonomie

On remarque la réaction en chaîne dans le cas de *booster21*, *booster21 circuit* qui sont éliminés pour laisser place au CT *booster21 circuit packs*. Certaines fragments éliminés correspondent parfois à des termes en usage dans le domaine (*circuit, dual msa*), mais leur élimination est sans grande conséquence puisque, si le besoin se fait sentir, un terminologue pourra les identifier lors de la consultation de la liste finale des CT. Cette contrainte simple, qui ne procède qu'à des observations sur le corpus, permet d'éliminer deux CT de la liste qui sera présentée au terminologue.

4.2.3.3 Résultats de l'acquisition

Nous passons ici en revue les résultats de la phase d'acquisition automatique des termes. Afin de déterminer la qualité des CT recensés par TermoStat, une validation des résultats obtenus est nécessaire. Le paragraphe 4.2.3.3.1 décrit cette étape de validation. La section suivante s'intéresse plus particulièrement à la longueur des CT valides retenus par rapport aux résultats obtenus par d'autres chercheurs (Justeson et Katz 1993, Nkwenti-Azeh 1994 et Jacquemin 1996). Les matrices syntagmatiques des termes identifiés sont aussi examinées afin d'identifier les plus productives. La discussion sur les résultats de l'acquisition se termine par une évaluation des performances du logiciel TermoStat d'un point de vue de la précision et du rappel.

4.2.3.3.1 Validation des résultats

La validation des résultats a essentiellement pour but de nous permettre de vérifier si l'objectif d'augmentation de la précision du processus d'acquisition à l'aide de PLS, tel que fixé au début de la présente recherche, a été atteint. Afin d'assurer une validation de tous les CT recensés, la méthodologie adoptée comprend deux étapes : une validation automatique suivie d'une validation manuelle.

4.2.3.3.1.1 Validation automatique

Afin de valider automatiquement les données, TermoStat consulte une banque de terminologie. Cette étape vient confirmer le statut terminologique des CT. La liste de termes mise à la disposition du logiciel par la banque de terminologie est une liste dite *à plat*, c'est-à-dire que cette liste ne comporte pas d'information sémantique. La validation automatique ne permet donc pas de confirmer hors de

tout doute que le CT relevé correspond au terme de la banque de terminologie.

- *access network element*
- *application*
- *emergency technical support*
- *entry point*
- *extension*
- *interface specification*
- *main transport shelf*
- *mbit*
- *network management protocol*
- *oc*
- *operation interface*
- *shelf layout*
- *standard*
- *transmission control protocol*

Les termes précédents représentent quelques exemples de CT qui ont été validés à l'aide de la consultation de la banque de terminologie. On y note des formes simples comme des termes simples ou des abréviations ainsi que des formes terminologiques complexes. La consultation d'une liste de termes peut cependant entraîner des problèmes et une comparaison strictement graphique du CT identifié dans le texte et d'un terme inclus dans la banque de terminologie ne permet pas de résoudre les cas d'ambiguïtés. C'est d'ailleurs le cas du terme *standard*, qui apparaît dans la liste précédente et qui peut être soit un substantif, soit un adjectif.

Ainsi, nous ne pouvons pas considérer que toutes les formes trouvées à la fois dans le corpus et dans la banque sont des termes et que leur validité terminologique ne peut être remise en question. On peut cependant considérer qu'il s'agit d'un bon indice de la validité d'un CT. Étant donné l'étroitesse du domaine et la corrélation entre les textes et la banque de terminologie utilisée, nous croyons qu'il s'agit d'une source suffisante pour assurer la qualité des décisions prises par TermoStat.

4.2.3.3.1.2Validation manuelle

Le processus de validation automatique ne permet pas de valider l'ensemble des CT puisque tous les CT ne sont pas contenus dans la banque de terminologie. Afin de pallier cette lacune, des terminologues, spécialistes du domaine des télécommunications et possédant une bonne connaissance des documents qui composent le corpus d'analyse, ont consulté la liste des CT n'ayant pu être validés automatiquement. Les trois documents du CA ont été répartis entre trois terminologues qui ont effectué une validation de la liste des CT pour un document. Le terminologue le plus expérimenté du groupe a ensuite révisé les décisions des autres terminologues.

La validation des résultats par les terminologues a été effectuée à l'aide d'une interface qui présente, pour chaque document du corpus d'analyse, l'ensemble des CT n'ayant pas été validés à la suite de la consultation de la banque de terminologie. Afin d'être en mesure d'évaluer la validité des CT, les terminologues ont accès à l'ensemble des occurrences d'un CT au sein d'un corpus. La validation d'un CT dans un document entraîne la validation de ce même CT dans l'ensemble des documents du corpus d'analyse. Ce comportement est semblable à celui obtenu à l'aide de la banque de

terminologie et d'une consultation à plat puisqu'une forme graphique est jugée valide pour l'ensemble des itérations effectuées sur le CA.

Afin d'encadrer leur travail et de s'assurer d'obtenir des résultats comparables, les consignes suivantes ont été distribuées aux terminologues :

- Vous êtes terminologue au sein d'une société spécialisée en télécommunications et vous devez dépouiller un document en vue de sa réécriture en anglais contrôlé.

- Vous devez relever les CT pertinents pour la conversion en anglais contrôlé. Les CT sont jugés pertinents s'ils doivent faire partie du vocabulaire à inclure ou à ne pas utiliser dans le document en langue contrôlée.

- Les fragments de termes sont jugés pertinents dans la mesure où ils peuvent adopter un comportement autonome dans le domaine des télécommunications ou être utiles pour le dépouillement de corpus reliés aux documents dépouillés.

La première consigne fixe l'objectif du dépouillement et permet au terminologue d'avoir une meilleure idée du produit terminologique visé. La deuxième consigne fait référence au vocabulaire permis au sein d'un document en langue contrôlée et au vocabulaire d'exclusion. Par exemple, un document dépouillé peut contenir les variantes *fibre management* et *fiber management* alors qu'une seule forme serait conservée dans le document en anglais contrôlé. La tâche des terminologues consiste à recenser les deux formes puisque la connaissance des formes à ne pas utiliser est aussi une information utile pour la construction de glossaires.

Lors de la consultation des listes produites par TermoStat, les terminologues peuvent être confrontés à ce qu'ils jugent être un fragment de terme. Par exemple, TermoStat peut ne pas avoir été capable de recenser un CT complet comme *common object request broker architecture*, mais avoir retenu *common object request broker*. Les terminologues doivent alors faire appel à leur jugement et à leur connaissance du domaine des télécommunications pour déterminer si le CT doit être considéré comme valide. Lors de la consultation des occurrences du CT dans le document, les terminologues pourront constater que le logiciel n'a retenu qu'un fragment du terme à relever. S'ils jugent que le fragment peut constituer un terme valide, ils se doivent de l'approuver.

La validation humaine, de par sa nature subjective, peut entraîner des variations qui ne sont pas observées dans un processus de validation automatique. Dans ce dernier cas, un CT sera jugé valide ou non de façon systématique. L'intervention humaine fait intervenir le jugement du terminologue et sa décision peut varier en fonction du temps ou de la situation. Ainsi, deux terminologues recevant les mêmes consignes ne prendront pas toujours la même décision au sujet d'un CT. Cette constatation est aussi valable pour le même terminologue à deux moments différents.

Il serait intéressant de procéder à des études plus poussées sur un tel processus de validation en soumettant à plusieurs reprises au même terminologue la liste de CT obtenue pour un CA et en comparant les divergences dans la liste des formes validées ou rejetées. Un tel processus itératif pourrait peut-être mettre en lumière une approche permettant de pondérer les résultats obtenus afin de tenir compte du facteur humain. Cette préoccupation se situe

cependant au-delà de nos objectifs de recherche. Les résultats de la validation des CT retenus par le système sont présentés dans les paragraphes 4.2.3.3.4 et 4.2.3.3.5.

4.2.3.3.2 Fenêtre de repérage

Il est intéressant de vérifier dans quelle mesure les termes identifiés par TermoStat occupent la fenêtre de repérage imposée à l'algorithme d'acquisition. L'adoption d'une fenêtre de repérage de 6 mots permet d'obtenir une bonne idée de la répartition des CT en fonction de la longueur des CT. Afin de ne pas inclure des données qui viendraient fausser ou influencer vainement les observations sur la longueur des unités, nous ne considérons dans la description que les CT valides : les termes. La figure qui suit présente la répartition observée dans les trois corpus qui composent le corpus d'analyse.

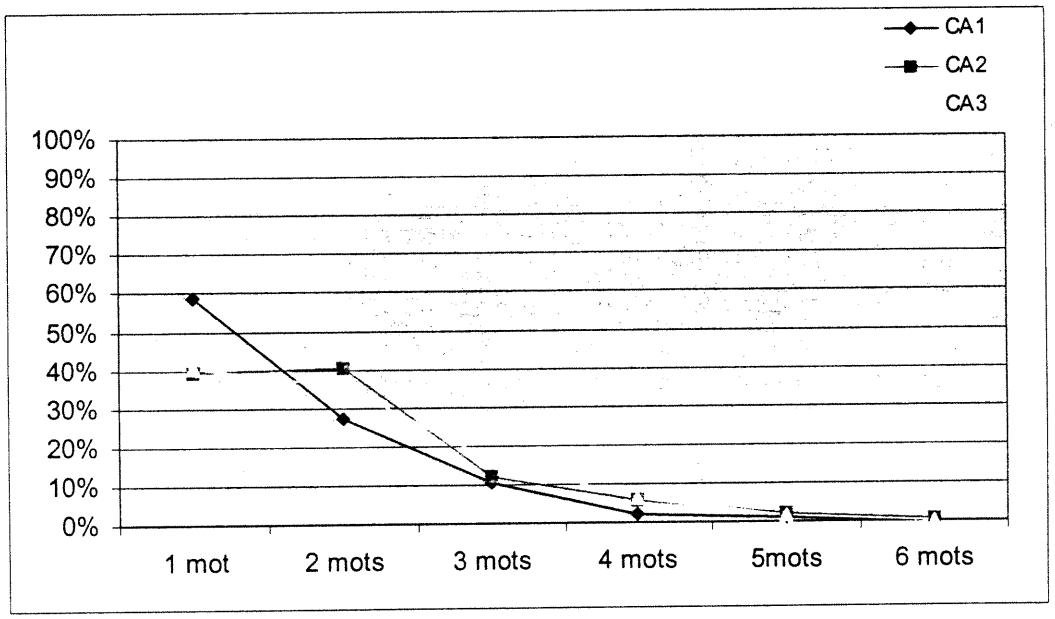

Figure 8. Pourcentage des termes en fonction de la longueur

On constate que le comportement des CT de longueur 1 est très différent entre le document CA_1 et les autres documents. Ces CT occupent une place beaucoup plus importante au sein du premier document. La consultation des termes recensés dans CA_1 ne permet malheureusement d'identifier ce qui cause la prépondérance de termes simples dans ce document. La répartition des CT de longueur supérieure à 1 est cependant relativement uniforme au sein de tous les documents. Comme on pouvait le prévoir, plus les CT sont longs, moins ils occupent une place de choix au sein des résultats; cette proportion tend rapidement vers 0 %.

À titre de comparaison, la figure qui suit présente le pourcentage de CT trouvés en fonction de leur longueur en nombre de mots dans les études présentées au paragraphe 4.2.3.2.2.

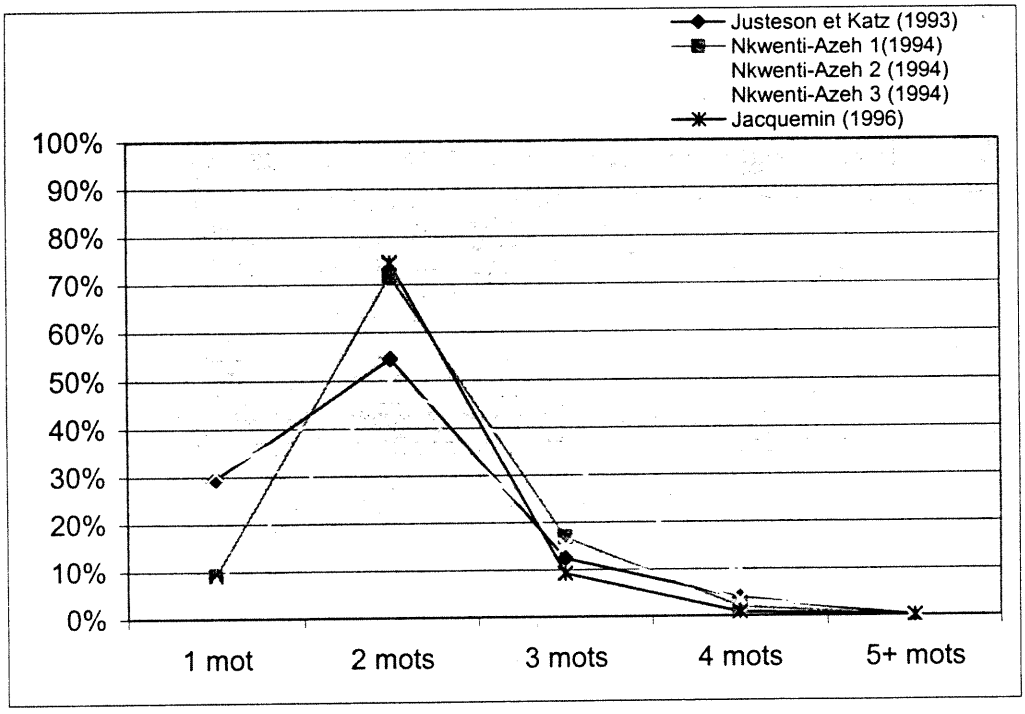

Figure 9. Pourcentage des CT en fonction de la longueur – autres études

La comparaison de ces résultats avec ceux obtenus par TermoStat est surprenante et intéressante puisque les résultats divergent fortement. Cette divergence se situe principalement au niveau des formes composées d'une seule unité. Dans tous les cas, à l'exception de l'étude sur le français de Jacquemin (1996), qui n'évalue pas la proportion de CT de longueur 1, la proportion de termes de longueur 1 est plus basse que celle de longueur 2. Ces résultats s'opposent à ceux observés dans le cadre de la présente étude où la proportion des CT valides de longueur 1 est égale ou supérieure à ceux composés d'une unité de plus.

N'ayant pas un accès exhaustif aux résultats des chercheurs précédents, il est difficile d'expliquer ce qui cause cette disproportion. Cette dernière pourrait être attribuée à notre corpus d'analyse, au domaine d'activité à l'étude, au style propre aux documents de la société Nortel Networks, etc. Pour ce qui est des CT de les plus longs, la répartition au sein des résultats est sensiblement la même que celle observée au sein des listes générées par TermoStat. En effet, la très grande majorité des termes ont une longueur qui est inférieure à 4 mots.

4.2.3.3.3 Structure des CT

Le tableau qui suit détaille les observations faites sur les trois textes spécialisés qui composent le corpus d'analyse. Seules les matrices des CT validés ont été conservées et classées en ordre décroissant de fréquence. Le tiret a été conservé à titre de composant de la matrice terminologique. Ainsi, les matrices *N N* (*software upgrade*) et *N - N* (*l - band*) sont considérées comme différentes.

À la consultation du tableau XX, on remarquera que les proportions sont sensiblement les mêmes d'un document à un autre et que la fréquence relative des matrices les unes par rapport aux autres est respectée dans l'ensemble des documents. Cette uniformité pourrait être due au fait que les documents traitent du même domaine et utilisent une terminologie commune.

Matrice	**CA_1**	**CA_2**	**CA_3**	**Moyenne**
N	58,78 %	39,27 %	40,39 %	46,15 %
N N	23,65 %	30,24 %	25,92 %	26,60 %
N N N	9,29 %	8,28 %	10,80 %	9,46 %
A N	3,72 %	9,91 %	9,72 %	7,78 %
N N N N	2,03 %	2,13 %	4,54 %	2,90 %
A N N	0,34 %	2,26 %	4,32 %	2,31 %
N N N N N	1,18 %	0,38 %	1,08 %	0,88 %
N A N	0,51 %	0,88 %	0,43 %	0,61 %
N - N N	0 %	1,76 %	0 %	0,59 %
A N N N	0,17 %	0,88 %	0,43 %	0,49 %
N - N	0,17 %	0,50 %	0,22 %	0,30 %
A A N	0,17 %	0,25 %	0,22 %	0,21 %
N - N N N	0 %	0,63 %	0 %	0,21 %
N - N A N	0 %	0,50 %	0 %	0,17 %
N A N N	0 %	0,25 %	0,22 %	0,16 %
N - A N N	0 %	0 %	0,43 %	0,14 %
N N A N	0 %	0 %	0,43 %	0,14 %
N N A N N	0 %	0,13 %	0,22 %	0,11 %
A N - N	0 %	0,25 %	0 %	0,08 %
A N A N	0 %	0,25 %	0 %	0,08 %
A N N - N	0 %	0,25 %	0 %	0,08 %
A N N N N	0 %	0,25 %	0 %	0,08 %
N - N N A N	0 %	0,25 %	0 %	0,08 %
N N - N	0 %	0,25 %	0 %	0,08 %
A A N N	0 %	0 %	0,22 %	0,07 %

N - A N	0 %	0 %	0,22 %	0,07 %
N N N N N N	0 %	0 %	0,22 %	0,07 %
A - N N N N	0 %	0,13 %	0 %	0,04 %
N - N A N N	0 %	0,13 %	0 %	0,04 %

Tableau XX. Matrices identifiées au sein des CT valides

Les exemples suivants illustrent certaines des matrices identifiées :

amplifier [N]
network element [N N]
line amplifier site [N N N]
object request broker architecture [N N N N]
l - band input signal [N - N N N]
line trail termination point id [N N N N N]
pmbb common object request broker architecture [N N N N N N]
dwdm optical link [N A N]
vt100 - compatible terminal [N - A N]
unidirectional osc [A N]
transparent wavelength translator [A N N]
dual amplifier circuit pack [A N N N]
bidirectional optical amplifier [A A N]

Les CT uniquement formés de substantifs sont de loin les plus fréquents et recouvrent entre 83 % et 95 % de l'ensemble des CT. Ainsi, dans le cas du corpus qui fait l'objet de notre analyse, une procédure d'étiquetage des corpus possédant des règles permettant d'isoler avec une excellente fiabilité les substantifs réussirait à isoler une majorité des CT. On peut aussi envisager d'obtenir des résultats probants avec l'approche inverse; c'est-à-dire l'exploitation des

connaissances par la négative. Il suffirait alors d'identifier avec certitude ce qui ne peut être un substantif.

Il serait intéressant de procéder à des échantillonnages semblables sur des textes spécialisés portant sur un sujet différent afin de vérifier si l'importance attribuée à chaque matrice de formation demeure la même ou si on peut observer un changement marqué d'un domaine à un autre, semblable à la différence de répartition des CT en fonction de la longueur observée entre nos résultats et ceux des autres chercheurs.

4.2.3.3.4 Précision

Nous le rappelons, la précision vise à mesurer la capacité d'un système d'identifier uniquement les CT pertinents (voir 2.1.3.3.1). On ne peut vraiment apprécier les performances d'un logiciel que dans la mesure où un point de comparaison des résultats est disponible.

Afin de mettre en lumière le rôle joué par les PLS dans l'acquisition des termes, nous avons modifié le logiciel TermoStat de façon à obtenir une liste de CT qui ne repose pas sur les PLS. Les algorithmes utilisés sont les mêmes pour les deux prototypes mais, dans un cas, l'acquisition est effectuée à l'aide des PLS. Ainsi, au lieu de procéder à une acquisition à partir de têtes nominales qui se qualifient à titre de PLS, l'algorithme modifié utilise toutes les unités nominales.

Les tableaux des sections suivantes comportent deux colonnes intitulées *précision*; celle de gauche présente la performance de TermoStat sans recours aux PLS alors que celle de droite comporte les résultats de l'acquisition à l'aide des PLS. La colonne intitulée *Fréq.*

permet de constater l'influence de l'imposition d'un seuil de fréquence minimal pour les CT sur les résultats de l'acquisition.

4.2.3.3.4.1Précision brute

La précision brute correspond à une évaluation de la précision des résultats sur l'ensemble des CT retenus par le prototype. Les tableaux XXI, XXII et XXIII détaillent les variations de la précision (exprimée en pourcentage) pour les divers seuils de fréquence[62]. Ainsi, pour le document CA_1, si on considère les CT dont la fréquence est égale ou supérieure à 1, le prototype procédant à l'acquisition sans la contrainte des PLS obtient une précision de 74,84 % alors que l'approche par PLS conduit à une précision de 77,28 %. L'utilisation des PLS entraîne donc un gain en précision de 2,44 % tel qu'indiqué dans la colonne qui porte le titre *Delta.*

CA_1 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_1 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	954	714	240	74,84		1	766	592	174	77,28	2,44
	2	517	422	95	81,62		2	454	377	77	83,04	1,41
	3	339	294	45	86,73		3	311	274	37	88,10	1,38
	4	257	230	27	89,49		4	240	216	24	90,00	0,51
	5	206	184	22	89,32		5	196	175	21	89,29	-0,03
	6	171	156	15	91,23		6	164	149	15	90,85	-0,37
	7	142	132	10	92,96		7	137	127	10	92,70	-0,26
	8	125	117	8	93,60		8	123	115	8	93,50	-0,10
	9	95	92	3	96,84		9	94	91	3	96,81	-0,03
	10	82	79	3	96,34		10	81	78	3	96,30	-0,05

Tableau XXI. Précision brute – CA_1

[62] La première colonne (Fréq.) correspond à une fréquence minimale d'occurrence (Fréquence>=1) et non à une fréquence particulière (Fréquence=1).

CA_2 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_2 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	1524	984	540	64,57		1	1236	867	369	70,15	5,58
	2	770	603	167	78,31		2	655	522	133	79,69	1,38
	3	537	443	94	82,50		3	468	399	69	85,26	2,76
	4	429	359	70	83,68		4	384	331	53	86,20	2,51
	5	347	292	55	84,15		5	316	273	43	86,39	2,24
	6	299	254	45	84,95		6	280	243	37	86,79	1,84
	7	261	222	39	85,06		7	248	213	35	85,89	0,83
	8	236	202	34	85,59		8	226	194	32	85,84	0,25
	9	211	181	30	85,78		9	203	174	29	85,71	-0,07
	10	191	164	27	85,86		10	185	159	26	85,95	0,08

Tableau XXII. Précision brute – CA_2

CA_3 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_3 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	752	540	212	71,81		1	621	463	158	74,56	2,75
	2	374	298	76	79,68		2	327	272	55	83,18	3,50
	3	233	199	34	85,41		3	218	191	27	87,61	2,21
	4	165	144	21	87,27		4	157	139	18	88,54	1,26
	5	134	116	18	86,57		5	130	114	16	87,69	1,13
	6	101	90	11	89,11		6	100	89	11	89,00	-0,11
	7	87	77	10	88,51		7	87	77	10	88,51	0,00
	8	72	65	7	90,28		8	72	65	7	90,28	0,00
	9	58	55	3	94,83		9	58	55	3	94,83	0,00
	10	50	47	3	94,00		10	50	47	3	94,00	0,00

Tableau XXIII. Précision brute – CA_3

On constate que, pour les deux approches, la précision augmente rapidement lorsque le seuil de fréquence utilisé augmente. Les CT dont la fréquence est plus élevée sont donc généralement de bons CT. Cette constatation n'est pas surprenante puisque les termes ont habituellement tendance à être réutilisés au sein des documents techniques. C'est ce qui explique aussi que les logiciels d'acquisition

automatique de termes, afin d'augmenter la qualité des résultats obtenus, vont imposer au CT un seuil minimal de fréquence.

La dernière colonne des tableaux précédents met cependant en lumière un phénomène intéressant relié à la performance de l'acquisition à l'aide des PLS : son apport diminue rapidement lorsque le seuil de fréquence augmente. Ainsi, l'utilisation des PLS pour l'acquisition des termes est plus intéressante lorsqu'on s'intéresse à l'ensemble des CT de la liste, indépendamment de leur fréquence, ou aux CT dont la fréquence est très basse.

Le gain en précision observé pour tous les documents, lorsque qu'aucun seuil de fréquence n'est imposé, confirme notre hypothèse selon laquelle les PLS permettent de cibler les CT les plus intéressants au sein d'un corpus. La performance moins satisfaisante de cette approche pour des seuils de fréquence plus élevés est un peu décevante. Par contre, le fait de pouvoir recenser l'ensemble des termes avec une plus grande précision est non négligeable puisqu'il est, à notre avis, plus intéressant de pouvoir procéder à l'acquisition des termes sans l'imposition d'un seuil de fréquence élevé et de pouvoir atteindre une précision accrue.

L'instabilité des PLS de basse fréquence observée lors des tests décrits au point 4.2.2.4.3 ne semble pas avoir eu d'influence sur l'acquisition des termes. Malgré les appréhensions soulevées à la lumière de ces expérimentations, il semble que les PLS peu fréquents conduisent à l'acquisition des CT de qualité.

Le peu de gain qu'entraîne le recours au PLS par rapport à l'utilisation d'un simple seuil de fréquence est tout de même légèrement décevant. En effet, on pourrait croire que le fait de cibler

les mots dont le comportement se démarque des autres pourrait conduire à une meilleure performance. C'est cette attente qui nous amène à regarder de plus près les résultats et à tenter de voir si le gain en précision est plus important sur certains sous-ensembles des CT.

4.2.3.3.4.2Termes simples

Les terminologues opposent souvent les termes simples (une seule unité) aux termes complexes (plus d'une unité), tout en considérant que les deux types de termes sont aussi importants lors du dépouillement d'un corpus. La problématique des termes simples est cependant souvent laissée de côté par les chercheurs qui s'intéressent à l'acquisition automatique de termes.

D'un point de vue de leur délimitation ou de leur découpage en discours, ces derniers sont faciles à isoler puisqu'un simple découpage du texte en unité permet de les recenser. Par contre, ils sont beaucoup plus difficiles à distinguer des mots que les termes complexes. En effet, les outils informatiques, tout comme les terminologues, ne possèdent pas pour ces unités les mêmes indices textuels liés à la structure syntaxique, à la répétition, etc., qui sont mis à leur disposition pour les termes complexes.

Les tableaux XXIV, XXV et XXVI illustrent la précision de l'acquisition sur les CT constitués d'une seule forme. La précision de l'approche par PLS est encore une fois comparée à celle n'exploitant pas cette contrainte.

CA_1 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_1 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	515	434	81	84,27		1	382	348	34	91,10	6,83
	2	311	275	36	88,42		2	260	242	18	93,08	4,65
	3	218	202	16	92,66		3	193	184	9	95,34	2,68
	4	181	170	11	93,92		4	164	156	8	95,12	1,20
	5	152	143	9	94,08		5	142	134	8	94,37	0,29
	6	137	129	8	94,16		6	130	122	8	93,85	-0,31
	7	115	110	5	95,65		7	110	105	5	95,45	-0,20
	8	104	100	4	96,15		8	102	98	4	96,08	-0,08
	9	83	81	2	97,59		9	82	80	2	97,56	-0,03
	10	72	70	2	97,22		10	71	69	2	97,18	-0,04

Tableau XXIV. Précision CT simples – CA_1

CA_2 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_2 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	656	452	204	68,90		1	450	383	67	85,11	16,21
	2	406	313	93	77,09		2	315	266	49	84,44	7,35
	3	309	250	59	80,91		3	253	217	36	85,77	4,86
	4	260	215	45	82,69		4	222	194	28	87,39	4,70
	5	214	177	37	82,71		5	189	165	24	87,30	4,59
	6	186	156	30	83,87		6	172	150	22	87,21	3,34
	7	162	138	24	85,19		7	153	133	20	86,93	1,74
	8	149	130	19	87,25		8	141	124	17	87,94	0,69
	9	139	122	17	87,77		9	132	116	16	87,88	0,11
	10	126	110	16	87,30		10	121	106	15	86,60	0,30

Tableau XXV. Précision CT simples – CA_2

CA_3 Fréquence						CA_3 PLS						
	Fréq.	Total	Valides	Rejetés	Précision		Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	312	239	73	76,60		1	227	187	40	82,38	5,78
	2	194	158	36	81,44		2	155	136	19	87,74	6,30
	3	123	109	14	88,62		3	111	102	9	91,89	3,27
	4	95	85	10	89,47		4	89	81	8	91,01	1,54
	5	76	69	7	90,79		5	74	68	6	91,89	1,10
	6	62	57	5	91,94		6	61	56	5	91,80	-0,13
	7	58	53	5	91,38		7	58	53	5	91,38	0,00
	8	48	45	3	93,75		8	48	45	3	93,75	0,00
	9	40	38	2	95,00		9	40	38	2	95,00	0,00
	10	33	31	2	93,94		10	33	31	2	93,94	0,00

Tableau XXVI. Précision CT simples – CA_3

Comme le démontrent les tableaux qui précèdent, l'approche par PLS conduit ici à une augmentation importante de la précision pour les seuils de fréquence égaux ou inférieurs à 5. Le meilleur gain en précision est atteint lors de l'analyse du document CA_2. Cette dernière est en partie due au fait que le document comporte une série de formes simples semblables à des acronymes, qui servent à désigner des caractéristiques techniques et qui ne se retrouvent pas dans les autres documents. Sans l'apparition de ces formes, les performances du prototype seraient comparables sur l'ensemble des documents du corpus d'analyse.

L'utilisation d'une technique qui recense les spécificités lexicales d'un document et isole les plus pertinents d'un point de vue terminologique conduit donc à de bons résultats. Il semble exister une corrélation étroite entre les formes nominales et adjectivales et les caractéristiques d'un corpus et sa terminologie. La fréquence absolue a souvent été utilisée pour l'acquisition des termes, mais il semble qu'une étude qui prend aussi en considération la probabilité d'occurrence des formes conduise à de meilleurs résultats. En effet,

comme le démontre la section gauche des tableaux précédents, l'adoption d'un seuil de fréquence élevé permet une acquisition des termes de bonne qualité.

Pour l'acquisition des termes simples, l'apport des PLS se situe principalement au niveau des basses fréquences. Ces derniers permettent de recenser des formes qui passent habituellement inaperçues aux yeux des terminologues. Ce sont souvent ces formes plus ou moins fréquentes qui représentent un défi pour le terminologue puisqu'il est difficile de distinguer une forme peu fréquente au sein d'un corpus volumineux. L'approche par PLS facilite donc le recensement de ces formes, souvent laissées de côté par l'humain ou par les techniques d'acquisition automatique de termes exploitant un seuil minimal de fréquence.

4.2.3.3.4.3Termes complexes

Nous observons ici de plus près la seconde tranche des CT, c'est-à-dire les termes complexes. La bonne performance obtenue pour les CT simples n'est pas égalée lors de l'acquisition des termes complexes. Les tableaux XXVII, XXVIII et XIX présentent la précision obtenue lors de l'analyse des documents du CA.

CA_1 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_1 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	439	280	159	63,78		1	384	244	140	63,54	-0,24
	2	206	147	59	71,36		2	194	135	59	69,59	-1,77
	3	121	92	29	76,03		3	118	90	28	76,27	0,24
	4	76	60	16	78,95		4	76	60	16	78,95	0,00
	5	54	41	13	75,93		5	54	41	13	75,93	0,00
	6	34	27	7	79,41		6	34	27	7	79,41	0,00
	7	27	22	5	81,48		7	27	22	5	81,48	0,00
	8	21	17	4	80,95		8	21	17	4	80,95	0,00
	9	12	11	1	91,67		9	12	11	1	91,67	0,00
	10	10	9	1	90,00		10	10	9	1	90,00	0,00

Tableau XXVII. Précision CT complexes – CA_1

CA_2 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_2 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	868	532	336	61,29		1	786	484	302	61,58	0,29
	2	364	290	74	79,67		2	340	269	71	79,12	-0,55
	3	228	193	35	84,65		3	215	184	31	85,58	0,93
	4	169	144	25	85,21		4	162	139	23	85,80	0,60
	5	133	115	18	86,47		5	127	110	17	86,61	0,15
	6	113	98	15	86,73		6	108	94	14	87,04	0,31
	7	99	84	15	84,85		7	95	81	14	85,26	0,41
	8	87	72	15	82,76		8	85	71	14	83,53	0,77
	9	72	59	13	81,94		9	71	59	12	83,10	1,15
	10	65	54	11	83,08		10	64	54	10	84,38	1,30

Tableau XXVIII. Précision CT complexes – CA_2

CA_3 Fréquence	Fréq.	Total	Valides	Rejetés	Précision	CA_3 PLS	Fréq.	Total	Valides	Rejetés	Précision	Delta
	1	440	300	140	68,18		1	394	275	119	69,80	1,62
	2	180	139	41	77,22		2	172	135	37	78,49	1,27
	3	110	89	21	80,91		3	107	88	19	82,24	1,33
	4	70	58	12	82,86		4	68	57	11	83,82	0,97
	5	58	46	12	79,31		5	56	45	11	80,36	1,05
	6	39	32	7	82,05		6	39	32	7	82,05	0,00
	7	29	23	6	79,31		7	29	23	6	79,31	0,00
	8	24	19	5	79,17		8	24	19	5	79,17	0,00
	9	18	16	2	88,89		9	18	16	2	88,89	0,00
	10	17	15	2	88,24		10	17	15	2	88,24	0,00

Tableau XXIX. Précision CT complexes – CA_3

Alors que les performances de l'algorithme fondé sur les PLS sur l'ensemble des formes (seuil de fréquence égal ou supérieur à 1) étaient supérieures dans le cas des CT de longueur 1, on observe dans les tableaux qui précèdent une baisse de la précision lorsque l'on prend en considération les formes les moins fréquentes. L'apport des PLS à ce niveau est donc négatif ou négligeable.

Pour l'ensemble des résultats décrits, le gain en précision, lorsqu'il existe, est minime. Une approche par PLS ne donne donc pas les résultats espérés pour les CT complexes et le seuil de performance n'est en rien comparable à celui obtenu pour les CT simples. En fait, c'est dans cette opposition CT simples versus CT complexes que nous croyons que réside l'explication de la précision décevante de l'acquisition des termes complexes à l'aide des PLS.

Les PLS trouvent leur source dans l'analyse des spécificités telle qu'elle est décrite par Lebart et Salem (1988 et 1994). À l'aide de cette technique, les auteurs tentent de cerner les unités lexicales dont le

comportement se démarque le plus au sein d'un corpus. Dans le cadre de cette analyse, seules les unités simples sont envisagées et les unités complexes sont laissées de côté.

L'approche pour l'acquisition automatique des termes proposée dans le cadre de la présente thèse repose sur l'hypothèse qu'une fois les PLS identifiés à l'aide d'une technique qui s'inspire fortement de celle des spécificités, il est ensuite possible des réutiliser ces formes simples pour avoir un accès plus pertinent aux formes complexes. Les résultats obtenus laissent présager que ce n'est pas le cas et que les enchaînements syntagmatiques ne sont pas bien décrits par une analyse reposant sur les PLS.

Il ne faut cependant pas conclure que leur utilité est nulle; nous croyons qu'il faut au contraire pousser les recherches plus loin pour découvrir des techniques qui permettent d'utiliser les informations fournies par une approche par PLS afin de bien représenter l'aspect combinatoire des CT complexes. Une telle approche pourrait faire entrer en jeu, une fois la phase d'acquisition automatique des PLS terminée pour un corpus d'analyse, des indices comme ceux testés par Daille (voir 2.2.2.4.1) visant à mettre en lumière la tendance de deux ou plusieurs formes à s'associer ou à partager des contextes communs.

D'ici à ce que des recherches sur le sujet soient entreprises, l'acquisition automatique des termes fondée sur les PLS permet tout de même d'obtenir, sur l'ensemble des CT identifiés par le logiciel TermoStat, une augmentation de la performance par rapport à un prototype qui ne fait appel qu'à un seuil de fréquence.

4.2.3.3.5 Rappel

Le rappel évalue la capacité d'un système de recenser l'ensemble des termes dans un document (voir 2.1.3.3.1). Bien que cet objectif se situe en marge de ceux établis pour la présente recherche, il est tout de même intéressant de vérifier les répercussions qu'entraîne la recherche d'une augmentation de la précision sur le rappel. En effet, bien que l'objectif de nombreux systèmes soit l'obtention d'un rappel et d'une précision maximum, l'augmentation d'un des indices passe bien souvent par une diminution de l'autre.

Le problème principal de l'évaluation du rappel obtenu par un logiciel réside dans le fait que l'ensemble des termes d'un document doit avoir été recensé. Dans le cas précis des documents qui composent notre corpus d'analyse, ce dépouillement systématique n'a pas été effectué. Afin de pallier cette difficulté et de mettre en lumière l'effet de l'utilisation des PLS pour l'acquisition automatique de termes, nous avons décidé de choisir comme mesure étalon la liste des termes approuvés par les terminologues produite à l'aide du logiciel TermoStat lorsque ce dernier procède à l'acquisition des termes à l'aide d'un seuil de fréquence. Ces résultats sont directement accessibles et correspondent à ceux utilisés pour l'évaluation de la précision à la section précédente. Étant donné que cette version du prototype utilise à titre de tête potentielle l'ensemble des formes nominales des documents, les résultats de l'acquisition à l'aide des PLS sont nécessairement plus restreints.

Les paragraphes qui suivent reprennent la présentation utilisée pour la précision. Nous présentons d'abord le rappel brut, ensuite le rappel pour les termes simples et finalement, le rappel pour les termes complexes.

4.2.3.3.5.1Rappel brut

Nous comparons, dans les lignes qui suivent, la performance du logiciel TermoStat exploitant la contrainte d'acquisition des termes à l'aide des PLS par rapport à sa performance sans cette dernière contrainte sur l'ensemble des résultats. Les tableaux suivants indiquent le nombre de termes recensés sans les PLS (*Nombre de termes*) et celui recensé à l'aide des PLS (*Termes identifiés*). La colonne intitulée *Rappel* correspond au ratio entre les deux premières colonnes[63].

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	714	592	77,28	82,91
2	422	377	83,04	89,34
3	294	274	88,10	93,20
4	230	216	90,00	93,91
5	184	175	89,29	95,11
6	156	149	90,85	95,51
7	132	127	92,70	96,21
8	117	115	93,50	98,29
9	92	91	96,81	98,91
10	79	78	96,30	98,73

Tableau XXX. Impact de l'utilisation des PLS sur le rappel – CA_1

[63] Comme dans le cas des tableaux précédents, la première colonne (Fréquence) correspond à une fréquence minimale d'occurrence (Fréquence >=1) et non à une fréquence particulière (Fréquence =1).

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	984	867	70,15	88,11
2	603	522	79,69	86,57
3	443	399	85,26	90,07
4	359	331	86,20	92,20
5	292	273	86,39	93,49
6	254	243	86,79	95,67
7	222	213	85,89	95,95
8	202	194	85,84	96,04
9	181	174	85,71	96,13
10	164	159	85,95	96,95

Tableau XXXI. Impact de l'utilisation des PLS sur le rappel – CA_2

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	540	463	74,56	85,74
2	298	272	83,18	91,28
3	199	191	87,61	95,98
4	144	139	88,54	96,53
5	116	114	87,69	98,28
6	90	89	89,00	98,89
7	77	77	88,51	100,00
8	65	65	90,28	100,00
9	55	55	94,83	100,00
10	47	47	94,00	100,00

Tableau XXXII. Impact de l'utilisation des PLS sur le rappel – CA_3

On observe que le rappel augmente rapidement avec le seuil de fréquence utilisé pour la sélection des CT. Cette situation peut sembler contradictoire puisque l'imposition d'une contrainte de fréquence a nécessairement comme conséquence de faire diminuer le nombre de termes retenus par rapport à l'ensemble du bassin de termes. Le rappel devrait donc diminuer et non augmenter.

Ce comportement du rappel est dû à la technique que nous utilisons pour calculer ses valeurs. Notre calcul repose sur le nombre de termes identifiés par le prototype en l'absence de contrainte pour le même seuil de fréquence. Ainsi, même si le document CA_3 contient 540 termes, nous jugeons que le rappel de TermoStat, lorsque la contrainte d'utilisation des PLS est activée, atteint un rappel de 100 % avec un seuil de fréquence égal à 7.

On pourrait aussi considérer que le rappel doit être calculé sur le nombre total de termes contenus dans un document; le logiciel atteindrait alors un rappel de 14,26 % pour le même seuil de fréquence. Nous croyons cependant que le calcul présenté dans les tableaux précédents permet de mieux cerner les conséquences de la contrainte que nous imposons à nos algorithmes.

On constate que les résultats obtenus à l'aide des PLS s'approchent très rapidement de la performance de l'algorithme sans restriction. Ainsi, plus le seuil de fréquence augmente, moins l'écart, calculé en nombre de termes, est important. Cette constatation n'est pas si surprenante étant donné que l'écart entre les PLS et les substantifs diminue lorsque la fréquence augmente. Comme nous l'avons souligné au point 4.2.2.2.2, sans conserver l'ensemble des substantifs fréquents, l'analyse des PLS tend à en conserver une majorité. Il n'est donc pas étonnant que les valeurs de rappel obtenues par les deux approches convergent.

Sans l'imposition d'une contrainte de fréquence, une approche souvent utilisée par les chercheurs du domaine, TermoStat parvient tout de même à identifier au minimum 82,91 % des termes. Le rappel passe rapidement au-dessus de la barre des 90 % dès que le seuil de

fréquence est supérieur ou égal à 3. La recherche de précision qui caractérise notre approche ne se fait donc pas entièrement au détriment d'un dépouillement des documents.

4.2.3.3.5.2Rappel – Termes simples

Le point de départ du processus d'acquisition automatique des termes, pour l'algorithme évoluant sans la contrainte des PLS, est composé de l'ensemble des substantifs des documents. Ce bassin est donc très large et le logiciel ne tente pas de déterminer quelles formes sont plus ou moins représentatives du document. Les seules formes qui sont éliminées au cours du processus d'acquisition sont celles qui ne parviennent pas à satisfaire la contrainte d'autonomie. L'approche par PLS a recours à un ensemble encore plus restreint parce que les unités nominales doivent d'abord se qualifier à titre de PLS pour être retenus comme terme potentiel et ensuite satisfaire la contrainte d'autonomie.

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	434	348	91,10	80,18
2	275	242	93,08	88,00
3	202	184	95,34	91,09
4	170	156	95,12	91,76
5	143	134	94,37	93,71
6	129	122	93,85	94,57
7	110	105	95,45	95,45
8	100	98	96,08	98,00
9	81	80	97,56	98,77
10	70	69	97,18	98,57

Tableau XXXIII. Termes simples : impact de l'utilisation des PLS sur le rappel – CA_1

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	452	383	85,11	84,73
2	313	266	84,44	84,98
3	250	217	85,77	86,80
4	215	194	87,39	90,23
5	177	165	87,30	93,22
6	156	150	87,21	96,15
7	138	133	86,93	96,38
8	130	124	87,94	95,38
9	122	116	87,88	95,08
10	110	106	86,60	96,36

Tableau XXXIV. Termes simples : impact de l'utilisation des PLS sur le rappel – CA_2

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	239	187	82,38	78,24
2	158	136	87,74	86,08
3	109	102	91,89	93,58
4	85	81	91,01	95,29
5	69	68	91,89	98,55
6	57	56	91,80	98,25
7	53	53	91,38	100,00
8	45	45	93,75	100,00
9	38	38	95,00	100,00
10	31	31	93,94	100,00

Tableau XXXV. Termes simples : impact de l'utilisation des PLS sur le rappel – CA_3

Comme le montrent les tableaux qui précèdent, le recours aux PLS entraîne une diminution du nombre de termes simples identifiés. Cette diminution est en relation directe avec la nature des PLS. La sélection de certaines formes au sein du bassin de formes nominales

ne peut conduire qu'à un sous-ensemble de ces mêmes formes une fois les contraintes appliquées. C'est sur le plan des unités simples que les contraintes des PLS se manifestent le plus puisqu'elles entraînent une élimination immédiate de certains CT.

Cette élimination est cependant en accord avec l'objectif initial de la présente thèse qui consiste à favoriser la précision par rapport au rappel. Il est cependant intéressant de mettre en relation la sélection des substantifs faite par les contraintes de notre algorithme et la bonne performance, d'un point de vue de la précision, du logiciel TermoStat pour le même sous-ensemble de termes. Un nombre restreint de CT simples sont retenus par le logiciel, mais ils sont généralement plus pertinents.

Les exemples suivants sont tirés de la listes des termes identifiés par l'approche sans contrainte et qui ne font pas partie de la liste générée par la version régulière du logiciel TermoStat :

building
bus
canada
care
case
division
end
exchange
field
incident
manager
need
page

performance
point
state
store

On retrouve, parmi les termes non retenus, des unités nominales ayant une fréquence relativement élevée au sein du corpus d'analyse. Les exemples qui précèdent nous permettent de rejoindre les commentaires faits précédemment sur la spécificité sémantique de certaines unités (voir 4.2.2.4.) et son impact sur l'approche par PLS. Il est possible que les substantifs qui apparaissent au sein de ces exemples soient, d'un point de vue sémantique, hautement spécifiques des documents analysés. L'approche par PLS ne permet cependant pas de bien cerner cette spécificité sémantique (opposition mot – terme) et les formes sont ainsi retranchées de la liste des termes simples. Le retrait de ces substantifs de la listes des PLS débouche sur une réduction du nombre de CT complexes recensés. La section qui suit aborde le sujet plus en détail.

4.2.3.3.5.3Rappel – Termes complexes

C'est au cours du processus d'acquisition des termes, lorsqu'il s'agit de recenser les termes complexes, que la contrainte des PLS exerce sa pression maximale sur l'algorithme. Ainsi, pour qu'un CT soit recensé par TermoStat, toutes les unités qui le composent doivent appartenir à la liste des PLS.

Les valeurs élevées de rappel affichées dans les tableaux qui suivent sont élevées et varient entre 87,14 % et 91,67 %. L'effet plus ou moins négatif exercé par la contrainte de PLS sur la précision lors de l'acquisition des ternes complexes ne semble pas se répercuter

avec la même force sur le rappel. Cette bonne performance implique donc que les PLS se retrouvent régulièrement au sein des termes complexes.

Ces résultats viennent démontrer la pertinence des PLS pour l'acquisition automatique des termes. Comme nous l'avons souligné précédemment, une approche utilisant les PLS comme point de départ pour l'acquisition des termes complexes, mais qui permettraient de représenter de façon plus adéquate le rôle des PLS au sein des termes, conduirait probablement à de meilleurs résultats tant au niveau de la précision que du rappel.

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	280	244	63,54	87.14
2	147	135	69,59	91.84
3	92	90	76,27	97.83
4	60	60	78,95	100.00
5	41	41	75,93	100.00
6	27	27	79,41	100.00
7	22	22	81,48	100.00
8	17	17	80,95	100.00
9	11	11	91,67	100.00
10	9	9	90,00	100.00

Tableau XXXVI. Termes complexes : impact de l'utilisation des PLS sur le rappel – CA_1

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	532	484	61,58	90,98
2	290	269	79,12	92,76
3	193	184	85,58	95,34
4	144	139	85,80	96,53
5	115	110	86,61	95,65
6	98	94	87,04	95,92
7	84	81	85,26	96,43
8	72	71	83,53	98,61
9	59	59	83,10	100,00
10	54	54	84,38	100,00

Tableau XXXVII. Termes complexes : impact de l'utilisation des PLS sur le rappel – CA_2

Fréquence	Nombre de termes	Termes identifiés	Précision	Rappel
1	300	275	69,80	91,67
2	139	135	78,49	97,12
3	89	88	82,24	98,88
4	58	57	83,82	98,28
5	46	45	80,36	97,83
6	32	32	82,05	100,00
7	23	23	79,31	100,00
8	19	19	79,17	100,00
9	16	16	88,89	100,00
10	15	15	88,24	100,00

Tableau XXXVIII. Termes complexes : impact de l'utilisation des PLS sur le rappel – CA_3

Les exemples suivants sont tirés de la liste des termes retenus par la version du logiciel qui n'exploite pas la contrainte des PLS. Les formes en caractères gras correspondent aux substantifs dont la

fréquence ne se démarquait pas suffisamment au sein du corpus d'analyse pour obtenir le statut de PLS.

amplifier ***group***
amplifier l - band overlay ***migration***
bell ***canada***
bidirectional osc ***loop***
channel *power detection*
contiguous ***path***
control point
controller ***list window***
digital ***equipment corporation***
fiber ***cut***
fiber ***manager***
graphical network ***editor***
interface definition ***language***
lucent technologies ***inc***
network ***manager***
network ***surveillance***
optical fiber ***plant***
optical ***infrastructure***
optical span ***range***
redundant communication ***path***
removable ***media***
system fault ***detection***
transition ***strategy***

Ces unités nominales correspondent aux substantifs éliminés des CT simples parce qu'ils n'obtiennent pas le statut de PLS. Un substantif qui n'est pas un PLS, ne pourra être utilisé à titre de tête de terme (*manager, detection*) pour l'acquisition de CT plus longs. Il

ne pourra pas non plus apparaître au sein d'un terme (*channel*) puisqu'il interrompra le processus d'acquisition des termes complexes et une forme plus courte sera recensée (*power channel*).

Ainsi, la perte de substantifs lors de la détection des PLS a un effet double : elle empêche à la fois l'acquisition de CT simples et de CT complexes. On doit donc s'assurer que l'acquisition des PLS s'effectue correctement et qu'elle n'est ni trop étroite, ni trop large.

4.2.3.3.6 Conclusion

Le présente section de la thèse avait pour objet de décrire la méthode d'acquisition automatique des termes mise en œuvre par le logiciel TermoStat. Ce dernier utilise comme point de départ les PLS nominaux identifiés au cours de l'étape précédente (voir le point 4.2). Les PLS nominaux sont utilisés à titre de tête de CT pouvant atteindre une longueur maximale de 6 mots. Une contrainte fondamentale est imposée au logiciel au cours du processus d'acquisition : les CT doivent être entièrement constitués de PLS.

Afin de pouvoir évaluer les résultats de l'acquisition automatique, les listes de CT ont fait l'objet d'un processus de validation en deux temps : une validation automatique et une validation manuelle. La première étape utilise le contenu d'une banque de terminologie pour valider l'ensemble des CT recensés par le logiciel. Les CT ne pouvant être validés de cette façon sont soumis à une équipe de terminologues. À la fin de ces deux étapes, le statut terminologique de l'ensemble des CT est connu.

Les résultats de l'acquisition automatique sont comparés à une liste de CT établie à l'aide d'une version modifiée du logiciel qui

procède à l'acquisition sans la contrainte des PLS. Cette comparaison des résultats permet de mettre en lumière les avantages et les désavantages de l'approche par PLS. D'un point de vue de la précision, la performance du logiciel est intéressante; elle atteint une moyenne de 74 % sur l'ensemble des documents du corpus d'analyse. La comparaison de la précision par rapport à la liste générée sans contrainte laisse entrevoir que l'approche par PLS est surtout avantageuse pour les CT dont la fréquence est égale ou inférieure à 3. Il s'agit d'une particularité intéressante parce que les CT les moins fréquents sont souvent laissés de côté par les logiciels d'acquisition automatique de termes. De plus, ces CT peu fréquents risquent fort de passer inaperçus lors du dépouillement manuel d'un corpus volumineux.

L'approche par PLS est particulièrement performante lorsqu'on s'intéresse aux CT simples puisqu'elle permet d'atteindre un seuil de précision moyen de 86,2 %. La méthode est donc bien adaptée à l'acquisition de ces formes qui sont, elles aussi, habituellement laissées de côté par la majorité des chercheurs. La précision obtenue lors de l'acquisition des CT complexes est cependant moins intéressante; elle est de 64,97 % en moyenne. De la performance précédente, on peut conclure que la technique PLS ne permet pas de bien cerner le rôle que jouent les PLS au sein des termes complexes et que les recherches méritent d'être poursuivies.

L'évaluation du rappel obtenu par un logiciel sous-entend la disponibilité d'une mesure étalon, de documents dépouillés pouvant servir de point de référence. Étant donné l'absence de ce point de comparaison pour la présente recherche, nous adoptons le dépouillement effectué par le prototype d'acquisition sans contrainte comme point de référence. Les performances générales de TermoStat,

d'un point de vue du rappel, sont satisfaisantes. En moyenne, le logiciel parvient à obtenir un rappel de 85,59 % sur le corpus d'analyse.

Comme dans le cas de la précision, l'opposition des résultats obtenus sur les termes simples et les termes complexes permet de mettre en lumière quelques phénomènes intéressants. La comparaison des listes de CT valides recensés par le logiciel procédant à l'acquisition avec ou sans contrainte permet de constater que les formes qui ont été écartées comme non spécifiques possèdent une spécificité qui n'est pas lexicale mais sémantique. On y rencontre des unités polysémiques qui possèdent un sens général (*manager*, appellation d'emploi) et un sens technique (*manager*, informatique).

Cette impossibilité de distinguer les utilisations variées dans le corpus de référence et le corpus d'analyse a aussi un impact sur la performance pour l'acquisition des formes complexes. En effet, la contrainte qui veut que toutes les formes qui composent un CT complexe fassent partie de l'ensemble des PLS rend impossible la présence de ces unités polysémiques au sein des CT relevés. Le rappel s'en trouve ainsi automatiquement diminué.

Malgré ces répercussions sur le rappel de l'approche par PLS, l'algorithme atteint un seuil de 90,03 % lors de l'acquisition des termes complexes. On se doit donc de conclure que les PLS sont fortement présents au sein des termes complexes. Par contre, comme il a été démontré au paragraphe dans la section traitant de la précision, la même approche ne permet pas de bien cerner le rôle des PLS au sein des termes complexes. Il faudra donc pousser plus loin la recherche sur leur utilisation pour l'acquisition des CT complexes

afin d'identifier une méthode qui réussira à cerner le rôle des PLS au sein de ces unités syntagmatiques.

Les exemples de CT présentés au paragraphe 4.2.3.3.5.3 ayant été écartés parce qu'une des formes qui les composent n'obtenait pas le statut de PLS (*amplifier* ***group,*** *amplifier l - band overlay* ***migration,*** *bell* ***canada,*** *bidirectional osc* ***loop,*** *contiguous* ***path***) laissent croire qu'un relâchement de la contrainte voulant que la tête des CT soit un PLS pourrait permettre d'obtenir de meilleures performances. Une mesure comme celle de la diversité de Shannon, telle qu'utilisée par Daille (1994a : 120), conduirait peut-être à de meilleurs résultats en prenant en compte la productivité de certaines formes[64]. De telles observations nous permettent de penser que l'approche par PLS ouvre la porte à de nombreuses pistes de recherche dans le domaine de l'acquisition automatique des termes.

4.2.4 Indice terminogénique

La recherche d'un indice terminogénique, capable de représenter l'intérêt terminologique d'un CT, repose sur notre conviction que l'on se doit de présenter au terminologue une liste dont l'information est la plus pertinente possible. Elle n'a cependant pas pour objectif de retrancher CT de la liste, mais de trier la liste des CT en ordre de pertinence.

Nous cherchons ainsi à présenter en tête de liste les CT dont le statut terminologique est le plus probable. L'évaluation de ce

---

[64] Voir les travaux sur la productivité de Bourigault (1994b) et d'Assadi et Bourigault (1996) à propos de la faculté que possèdent certaines formes de donner naissance à de nombreux termes.

caractère terminologique des CT passe par l'observation des indices contenus dans le corpus afin de tenter de distinguer les CT qui sont des termes de ceux qui n'en sont pas. Nous cherchons à déterminer une façon de qualifier ou de quantifier le caractère terminologique d'un CT. C'est autour de ce caractère terminologique des CT que nous articulons la notion d'indice terminogénique.

L'approche adoptée au sein de la présente thèse rejoint celle de Daille (1993, 1994a et 1994b) qui utilise divers indices statistiques afin de déterminer le caractère terminologique d'un CT. Cette dernière effectue cependant ses recherches sur des couples de mots, ce qui ne permet pas la validation de ses hypothèses sur nos résultats, qui contiennent à la fois des CT composés, qui atteignent une longueur de six mots, et des CT simples.

Frantzi et Ananiadou (1997) proposent une approche semblable à celle décrite précédemment et élabore un indice, la *C-value*. Elles l'utilisent afin de trier la liste des CT retenus par leurs algorithmes. L'indice proposé par ces auteures sera testé afin de vérifier sa pertinence au sein de notre démarche et sa capacité à mettre en évidence les CT pertinents.

Les indices que nous avons proposés dans le cadre de la présente thèse possèdent un dénominateur commun : ils n'utilisent que de l'information directement disponible à partir de la liste des CT ou du corpus. Nous ne désirons pas avoir recours à une source d'information extérieure de façon à conserver l'autonomie du prototype par rapport au domaine traité.

Cette indépendance est primordiale pour un logiciel qui a comme mission principale de dépouiller des textes dont le contenu

est, dans la majorité des cas, à la fine pointe de la technologie. Un tel logiciel ne peut se permettre d'exiger la création de dictionnaires électroniques spécialisés afin de l'aider dans son travail puisqu'il a justement comme objectif principal l'élaboration de dictionnaires.

L'évaluation de la pertinence des indices repose sur un tri de la liste des CT retenus afin de concentrer dans la première partie de la liste les CT dont le statut terminologique a été confirmé au cours de l'étape de validation. La précision du tri sera donc évaluée sur les premiers 50 % de la liste des CT pour chacun des documents qui composent le corpus d'analyse.

L'évaluation de la pertinence ne peut se faire que dans le cadre d'une démarche précise qui repose sur les consignes données lors de la validation des résultats. En effet, le statut terminologique des CT a été confirmé par les terminologues en fonction des consignes données (voir 4.3.3.1). Ainsi, les résultats obtenus à l'aide des indices élaborés dans les paragraphes qui suivent sont valides dans le cadre de la présente thèse. Des tests supplémentaires devront être effectués afin de les valider dans le cadre d'approches différentes utilisant des corpus de référence et d'analyse différents de ceux utilisés ici.

Les points suivants présentent les indices que nous avons élaborés afin de dresser une liste de CT triée en ordre de pertinence pour le travail du terminologue. Nous testons d'abord la *C-value* (voir le paragraphe 2.2.2.4.5) afin de vérifier sa pertinence dans le cadre de notre approche avant de proposer de nouveaux indices.

4.2.4.1 La C-value

Afin de pouvoir utiliser la *C-value* pour le tri des CT ainsi que pour déterminer leur statut terminologique, nous devons apporter une modification à l'algorithme original. En effet, les travaux de Frantzi et Ananiadou (1997) ne portent que sur les CT composés de longueur 2 ou 3 alors que nous travaillons aussi sur des CT simples.

Une utilisation directe de l'indice tel que proposé conduirait à l'obtention d'une *C-value* de 0 pour tous les CT simples. Le recours au log_2 appliqué à une longueur de 1 conduit à l'annulation du reste de l'équation. Afin d'éviter de tels résultats, nous modifions l'indice original (voir 2.2.2.4.5) en nous assurant que la partie de gauche de l'équation est toujours égale ou supérieure à 1 :

$$C-value(c_i)=\begin{cases}\left(\log_2|c_i|+1\right)\cdot f(c_i)\\ ou\\ \left(\log_2|c_i|+1\right)\cdot\left(f(c_i)-\frac{1}{P(Tc_i)}\sum c_y \in Tc_i f(c_y)\right)\end{cases}$$

La modification est relativement mineure et n'influence pas l'ordre de grandeur obtenu par l'indice original. En effet, en plus de prendre en considération les entrées de la liste qui sont constituées d'un seul mot, l'indice conduit à une échelle de valeurs qui respecte la classification obtenue grâce à l'indice tel que proposé par les auteurs.

Le tableau XXXIX contient les données obtenues à l'aide de la *C-value*. La précision est mesurée en fonction du nombre de termes identifiés par rapport au nombre de CT non valides qui se trouve dans la première moitié de la liste.

	Pertinents	**Non pertinents**	**Total**	**Précision**
CA_1	235	47	282	83,33 %
CA_2	493	125	618	79,77 %
CA_3	250	60	310	80,65 %
Précision moyenne				81,25 %

Tableau XXXIX. Performance de l'indice *C-value*

Ainsi, dans l'ensemble des documents, un tri à l'aide de la *C-value* permet d'obtenir une bonne concentration des termes en tête de la liste de CT. Dans tous les cas, la précision obtenue sur cette portion de la liste est supérieure à la précision observée sur l'ensemble de la liste (voir le point 4.2.3.3.4).

Il est difficile de comparer nos résultats avec ceux de Frantzi et Ananiadou (1997), qui ne calculent pas la précision de leur algorithme et qui n'offrent que les premiers CT de la liste générée par leur logiciel. Nous ne possédons pas l'information nécessaire à l'évaluation de la précision de leurs travaux et il nous est donc impossible de déterminer si la précision de l'indice dans le cadre de leurs travaux était inférieure ou supérieure à celle que nous obtenons.

Les performances obtenues à l'aide de la *C-value* nous conduisent à conclure qu'il s'agit d'un indice permettant de bien cerner le caractère terminologique de certains des CT recensés. Les éléments qui entrent en jeu dans le calcul de l'indice (fréquence, longueur et recoupement entre les CT) sont donc pertinents pour la mise en place d'un indice terminogénique. Les travaux de Frantzi et Ananiadou (1997) accordent une place importante à la fréquence des

CT, au nombre de mots qui les composent ainsi qu'au nombre de fois où les CT sont inclus à l'intérieur de CT plus longs.

Nous partageons l'avis de ces auteures et croyons que ces indices peuvent être utilisés pour déterminer la pertinence des CT. Les paragraphes qui suivent tentent de mettre en lumière la pertinence et le rôle de chacun de ces critères pour l'élaboration d'un indice à caractère terminologique.

4.2.4.2 La fréquence absolue

L'utilité de la fréquence pour le dépistage des termes a souvent été mentionnée et exploitée au sein de logiciels. Les travaux de Daille (1994a : 133) et Justeson et Katz (1993 : 6) en sont de bons exemples. La présente section de la thèse s'intéresse à l'effet d'un tri en ordre décroissant de fréquence absolue sur la précision. Encore une fois, cette dernière est évaluée sur la première partie de la liste des CT.

Ce tri de la liste des CT accorde nécessairement une très grande importance aux CT les plus fréquents. La répartition inégale des fréquences absolues, telles que rencontrées dans les corpus, aura pour effet de concentrer les CT les plus fréquents en tête de liste. On remarque en effet peu de CT dont la fréquence est très élevée et un nombre croissant de CT lorsque la fréquence diminue.

	Pertinents	**Non pertinents**	**Total**	**Précision**
***CA*$_1$**	246	36	282	87,23 %
***CA*$_2$**	506	112	618	81,88 %
***CA*$_3$**	259	51	310	83,55 %
Précision moyenne				84,22 %

Tableau XL. Performance de la fréquence absolue

Le tableau précédent résume les observations faites sur les documents du corpus d'analyse à la suite d'un tri selon la fréquence. L'efficacité d'un tri aussi simple est surprenante si on compare les résultats à ceux obtenus à l'aide de la *C-value*. Cette comparaison est d'autant plus marquante lorsqu'on prend aussi en considération la complexité de la *C-value* alors que la fréquence absolue est directement observable dans le corpus. La précision moyenne obtenue à l'aide de cet indice (84,22 %) est supérieure à celle obtenue par la *C-value*.

Tout comme le soulignent Justeson et Katz (1993 : 8), un filtrage des données ayant recours à la fréquence est efficace, mais il a cependant le désavantage de réduire considérablement le rappel des algorithmes. Cela entraîne une perte des CT les moins fréquents, qui n'est peut-être pas entièrement justifiée. En effet, selon les objectifs visés lors du processus d'acquisition des CT, on peut tout aussi bien s'intéresser au CT les moins fréquents dans la mesure où ils sont pertinents. Ces derniers sont d'ailleurs souvent les plus difficiles à identifier.

Malgré l'attrait d'un tri aussi simple, il est important de mentionner son manque de finesse lorsque l'on compare des tranches de fréquence très productives. En effet, cette approche ne permet pas d'opposer les CT qui partagent la même fréquence et de déterminer lequel de ces CT possède un caractère terminologique plus important. Le point suivant a pour objectif d'apporter un élément supplémentaire afin de nuancer la répartition obtenue à l'aide de la fréquence absolue.

4.2.4.3 La longueur : iLong

À l'exception des aspects linguistiques comme la morphologie et la syntagmatique, la quantité d'information qu'on peut obtenir par l'observation des CT est limitée. Un des éléments intéressants et directement observables exploités dans le calcul de la *C-value* est le nombre de mots qui composent le CT. L'indice élaboré dans la présente section vise à intégrer cette information à celle fournie par la fréquence afin de nuancer les résultats obtenus à partir de cette dernière. L'indice *iLong* combine donc un élément d'information tiré du corpus (la fréquence absolue) à une observation faite sur le CT lui-même (la longueur) :

$$iLong(c_i) = f(c_i) \cdot \frac{1}{|c_i|}$$

où

c_i correspond à l'entrée i de la liste des CT,

$f(c_i)$ correspond à la fréquence absolue du CT i,

$|c_i|$ correspond à la longueur en nombre de mots du CT i.

L'effet de cet indice sur la liste des CT est double : plus la fréquence d'un CT augmente, plus le CT est bonifié et placé en tête de liste. Par contre, plus le CT comporte de mots, plus son importance est diminuée et plus il apparaît tard dans la liste des CT.

Ainsi, grâce à l'influence qu'a la longueur des CT sur l'indice, les CT qui se voyaient réunis dans une même partie de la liste triée selon la fréquence se verront potentiellement répartis dans des sections différentes[65]. Par exemple, le CT *information*, qui a une

[65] Les techniques utilisées pour distinguer les CT les uns des autres relèvent de l'empirisme et non de la théorie.

fréquence absolue de 19, se verra attribuer un indice *iLong* de 19 alors que le CT *performance management*, qui a la même fréquence que le CT précédent, obtiendra 9,5. Cependant, deux CT qui ont en commun la longueur et la fréquence seront classés ensemble au sein de la liste et l'indice ne permet pas de les distinguer.

Notre approche rejoint ici les travaux sur la *C-value* qui prennent aussi en considération la longueur des CT afin de déterminer leur statut terminologique. Cependant, nous l'utilisons à l'inverse de Frantzi et Ananiadou (1997). Nous diminuons l'importance d'un CT s'il est plus long alors que la *C-value* la bonifie[66].

	Pertinents	**Non pertinents**	**Total**	**Précision**
CA_1	256	26	282	90,78 %
CA_2	516	102	618	83,50 %
CA_3	258	52	310	83,23 %
Précision moyenne				85,84 %

Tableau XLI. Performance de l'indice *iLong*

L'utilisation de cet indice pour le tri de la liste des CT conduit à une augmentation (1,62 %) de la précision moyenne des CT situés dans les premiers 50 % de la liste par rapport à un tri reposant sur la fréquence absolue. Pour les deux premiers documents du corpus d'analyse, le gain en précision est important; dans le cas du document CA_3, un CT pertinent est éliminé de la liste et fait légèrement baisser la précision.

---

[66] L'utilisation du $log_2|c_i|$ dans le calcul de la *C-value* conduit à une bonification des CT dont la longueur est supérieure à 2.

Les tableaux suivants décrivent la répartition des CT en fonction de leur longueur pour les listes de CT utilisant la fréquence comme critère de tri et l'indice *iLong*. On remarque, dans le deuxième cas, une forte augmentation de la représentation des CT plus courts. L'indice *iLong* favorise donc l'apparition des CT de longueur 1 dans la première portion de la liste des CT.

Longueur du CT	CA_1	CA_2	CA_3
1	164	278	139
2	94	234	110
3	20	53	38
4	3	35	19
5	1	12	4
6	0	6	0

Tableau XLII. Complexité des termes pour la fréquence absolue

Longueur du CT	CA_1	CA_2	CA_3
1	207	315	155
2	67	234	111
3	6	43	36
4	2	21	7
5	0	5	1
6	0	0	0

Tableau XLIII. Complexité des termes pour l'indice iLong

L'augmentation de la précision observée est en relation directe avec la présence plus important de ces unités simples en tête de liste. Comme nous l'avons vu dans la section portant sur l'évaluation de la précision de l'approche par PLS sur les formes simples (voir 4.2.3.3.4), il s'agit d'un groupe de CT où la précision est très élevée. Les bonnes

performances du logiciel TermoStat pour ce groupe de termes sont donc mises en évidence à l'aide de l'indice *iLong*, qui accorde moins d'importance aux CT les plus longs. Les CT les plus longs ne sont conservés que dans la mesure où leur fréquence relative au sein d'un corpus leur permet de se démarquer.

L'indice *iLong* offre une bonne performance d'un point de vue de la précision et permet de recenser à la fois des termes simples et des termes complexes. L'acquisition des termes simples est très avantageuse puisque la majorité des logiciels créés pour l'acquisition automatique des termes ne s'attardent qu'aux termes complexes. De plus, étant donné la capacité de l'approche par PLS de bien identifier les termes simples, il est avantageux d'avoir recours à un indice permettant de les mettre en évidence au sein du bassin des CT.

L'impossibilité pour l'indice iLong de discriminer entre des CT de longueur et de fréquence égales nous amène à chercher un autre indice. Afin d'y parvenir, nous tentons d'exploiter une source supplémentaire d'information sur le CT : les recoupements qui existent entre les entrées de la liste des CT.

4.2.4.4 Le recoupement à droite : iInc

L'indice *iInc* ne relève pas directement d'observations faites sur les CT ou sur leurs occurrences en corpus, mais plutôt du croisement d'observations faites sur plusieurs CT. On tente ici d'examiner l'ensemble des termes qui partagent des éléments lexicaux communs et les divers recoupements entre les CT.

Les recoupements entre les CT ont déjà fait l'objet d'une discussion au paragraphe 4.2.3.2.4 puisqu'ils ont été exploités à titre

d'indices permettant d'éliminer des CT de la liste des CT. Les fragments de CT possédant la même fréquence que des CT plus longs ont été éliminés. Par contre, les fragments qui possèdent des fréquences supérieures à celles des CT plus longs qui les incluent ont été conservés puisqu'ils sont considérés comme autonomes.

Le logiciel TermoStat s'intéresse plus particulièrement au recoupement à droite dans les termes, soit des CT plus courts qui sont potentiellement utilisés à titre de tête de CT plus longs. On cherche ainsi à quantifier l'autonomie observée lors de l'acquisition (voir 4.2.3.2.4).

Fragment	**CT**
monitoring	*performance monitoring*
interface	*PMBB interface*
	application interface
	operation interface
technical support	*emergency technical support*
transport signal	*synchronous transport signal*

Tableau XLIV. Exemples de fragments de termes

Le tableau XLIV présente quelques exemples de recoupements entre des CT considérés comme des fragments de CT plus longs. Les CT présentés dans la colonne de gauche du tableau précédent jouent le rôle de tête dans les termes plus longs qui apparaissent dans la colonne de droite. Le phénomène, vu d'un point de vue sémantique, consiste en une restriction de l'extension. L'expansion du terme vers la gauche vient préciser la notion désignée par le terme.

Sans prendre en considération cet aspect sémantique, l'indice que nous proposons, nommé *iInc*, se limite à observer et à totaliser le nombre de fois où un CT agit à titre de tête potentielle d'un CT plus long. C'est ce phénomène que nous nommons *recoupement à droite* et qui est décrit par l'indice suivant :

$$iInc(c_i) = f(Tc_i)$$

où

c_i correspond à l'entrée i de la liste des CT,

Tc_i correspond à l'ensemble des termes qui utilisent le CT i à titre de tête potentielle,

$f(Tc_i)$ correspond au nombre de fois où l'inclusion précédente est observée.

Nous nous situons encore une fois en marge des travaux sur la *C-value*. En effet, cette dernière ne s'intéresse pas directement au recoupement à droite entre les CT, mais à l'inclusion pure et simple d'un CT plus court dans un CT plus long. L'indice de Frantzi et Ananiadou (1997) considère donc que l'ensemble des recoupements entre les CT est pertinent pour représenter le caractère terminologique des entrées de la liste des CT. L'indice en question fait aussi intervenir la fréquence absolue de tous les CT qui incluent une forme plus courte alors que l'indice *iInc* ne prend pas cette information en considération. Ce dernier se concentre uniquement sur le nombre de fois où un CT est inclus et non sur la fréquence des CT l'incluant.

La *C-value* entraîne aussi une perte d'intérêt rapide pour un CT lorsqu'il est inclus dans des CT plus longs. Nous croyons, pour notre

part, qu'il s'agit au contraire d'une information très importante, qui caractérise les CT valides.

	Pertinents	**Non pertinents**	**Total**	**Précision**
CA_1	251	32	283	88,69 %
CA_2	521	97	618	84,30 %
CA_3	263	47	310	84,84 %
Précision moyenne				85,94 %

Tableau XLV. Performance de l'indice *iInc*

Les résultats présentés dans le tableau XLV viennent appuyer cette intuition. En effet, la prise en compte du recoupement à droite permet d'obtenir un effet de concentration des CT valides en tête de liste. À l'exception du corpus CA_1, le recours à l'indice *iInc* conduit à une augmentation de la précision par rapport à l'indice *iLong*.

Par définition, les CT les plus courts, qui sont plus susceptibles de se retrouver à l'intérieur de CT plus longs, sont encore une fois avantagés et placés en tête de liste. Les CT les plus sévèrement pénalisés sont ceux qui ne sont jamais inclus et qui ne peuvent espérer être bonifiés puisque leur *iInc* est automatiquement nul. Les CT les plus longs risquent donc d'être pénalisés. Les listes de CT obtenues à partir des trois documents comportent cependant des CT de longueur 4 (1 dans CA_1, 16 dans CA_2 et 4 dans CA_3) ou 5 (4 dans CA_2 et 1 dans CA_3) dont la fréquence d'utilisation à titre de tête potentielle est suffisamment élevée pour les aider à se démarquer.

La performance de l'indice *iInc* est donc intéressante et nous croyons qu'il est important de prendre en considération le phénomène

d'inclusion à droite observé dans la liste des CT. De la même façon, nous croyons que les bonnes performances obtenues à partir des observations faites sur la fréquence et la relation inverse entre le nombre de mots qui sont inclus dans un CT et sa fréquence méritent d'occuper une place importante. Nous proposons donc, dans la section qui suit, un indice qui cherche à fusionner les indices proposés dans les paragraphes précédents.

4.2.4.5 L'indice terminogénique : iTer

Puisque les indices *iLong* et *iInc* permettent d'obtenir une bonne qualité de résultats, nous cherchons ici à tirer profit des informations mises à notre disposition par les deux indices. L'indice *iTer* a pour objectif de cerner à la fois l'importance du recoupement entre les termes et celle du rapport inverse entre la fréquence absolue et la longueur des CT. En combinant les indices *iLong* et *iInc*, nous obtenons l'indice *iTer* qui se lit comme suit :

$$iTer(c_i) = \left(f(c_i) \cdot \frac{1}{(|c_i|+1)} \right) \cdot (f(Tc_i)+1)$$

où

c_i correspond à l'entrée i de la liste des CT

$f(c_i)$ correspond à la fréquence absolue du CT i

$|c_i|$ correspond à la longueur en nombre de mots du CT i

Tc_i correspond à l'ensemble des termes qui utilisent le CT i à titre de tête potentielle,

$f(Tc_i)$ correspond au nombre de fois où l'inclusion précédente est observée.

La portion gauche de l'équation qui nous vient de *iLong* prend en considération à la fois la fréquence absolue des CT et leur longueur. Ainsi, une augmentation de la fréquence conduit à une bonification du CT par rapport à l'ensemble des résultats; la longueur du CT vient cependant contrebalancer la valeur de la fréquence.

La portion de droite de l'équation, qui correspond à *iInc* légèrement modifié, vient moduler la fréquence en fonction de l'inclusion du CT à l'étude à titre de tête potentielle à l'intérieur de CT plus longs. Ainsi, plus le CT est inclus à l'intérieur des CT plus longs à titre de tête potentielle, plus il sera bonifié dans la liste finale.

Au sein de *iTer*, l'indice *iInc* a été modifié afin de ne jamais être nul. Ainsi, même si une forme n'est jamais incluse à l'intérieur de CT plus longs, la portion droite de l'équation ne conduira pas à une valeur de 0 pour *iTer*. Cette approche offre l'avantage de venir moduler l'information fournie par *iLong* sans l'annuler dans les cas où *iInc* est nul; elle se contente alors de multiplier par 1.

	Pertinents	**Non pertinents**	**Total**	**Précision**
CA_1	257	25	282	91,14 %
CA_2	519	99	618	84,00 %
CA_3	264	46	310	85,16 %
Précision moyenne				86,77 %

Tableau XLVI. Performance de l'indice *iTer*

La consultation du tableau précédent nous permet de constater un gain de précision négligeable par rapport aux résultats obtenus avec l'indice précédent (0,06 %) et de 0,83 % par rapport à *iInc*.

Cependant, il est important de retenir que la fusion des indices *iLong* et *iInc* ne conduit qu'à une perte d'information minime. Le prototype réussit à augmenter légèrement la précision des résultats obtenus à partir du premier et du troisième document et, par le fait même, la précision globale.

Malgré que la précision ne soit pas grandement influencée par la fusion des deux indices, nous croyons important de les mettre en commun puisqu'ils véhiculent de l'information d'origine diverse sur les CT. On réussit ainsi à combiner les observations faites sur le corpus (la fréquence), sur les CT eux-mêmes (longueur) et sur l'ensemble de la liste des CT (recoupement à droite). Cette diversification de la provenance de l'information représentée au sein de l'indice ne peut qu'être bénéfique et éviter d'accorder trop d'importance à un phénomène particulier.

Encore une fois, les CT les plus pénalisés sont ceux qui ne font pas l'objet d'inclusion et qui ne peuvent être bonifiés puisque la partie droite de l'équation conduit à une valeur égale à celle de l'iInc. Cependant, dès que le CT fait l'objet d'une seule inclusion, il peut espérer voir sa fréquence absolue faire augmenter son rang dans la liste des termes. Un tel scénario n'est pas envisageable avec l'indice *iInc* puisque le tri est uniquement fondé sur l'inclusion sans tenir compte de la fréquence.

À la lumière de la précision obtenue sur la première moitié de la liste des CT, nous considérons que l'indice *iTer* permet de bien représenter le caractère terminologique des entrées. Il offre la possibilité de trier une liste de CT en ordre de pertinence sans avoir recours à une source extérieure de connaissances. Il est donc à la fois

autonome face aux corpus et au domaine d'activité dont traitent les corpus.

4.2.4.6 Conclusion

Dans cette section de la thèse nous avons élaboré un indice terminogénique permettant de représenter l'intérêt des CT. L'indice proposé est utilisé pour trier la liste des CT afin de les présenter au terminologue en ordre décroissant de pertinence.

Après avoir procédé à des expérimentations à l'aide de la *C-value*, indice proposé par Frantzi et Ananiadou (1997), de nouveaux indices ont été élaborés de façon à augmenter la précision des résultats. Les indices proposés tentent d'exploiter au maximum l'information mise à leur disposition par la liste des CT, les CT eux-mêmes ainsi que le corpus.

Un tri effectué à l'aide de la fréquence permet d'augmenter la précision de la première portion de la liste (84,22 %) par rapport aux résultats obtenus à l'aide de la *C-value* (81,25 %). Un troisième indice testé prend en considération la longueur des CT. Nous avons pu démontrer que la modulation de la fréquence d'un CT par l'inverse de sa longueur permet de préciser le caractère terminologique de deux formes de fréquence identique mais de longueur différente (*iLong*). L'indice *iLong* accorde une place de choix aux CT les plus courts et il atteint une précision de 86,71 %. Cette bonne performance est en relation directe avec la capacité de l'approche par PLS de bien isoler les termes simples d'un document.

Des observations faites sur la liste des CT, plus particulièrement sur les recoupements qui existent entre les CT

recensés par le logiciel, nous ont conduit à l'élaboration d'un quatrième indice : *iInc*. Ce dernier reflète le nombre de fois où un CT est utilisé à titre de tête potentielle de un ou plusieurs CT plus longs. Sur l'ensemble des documents du corpus d'analyse, l'indice *iInc* obtient une valeur moyenne de 85,94 %. Tout comme *iLong*, *iInc* laisse de côté une partie des CT. En effet, les CT qui ne sont jamais inclus à l'intérieur de CT plus longs sont défavorisés et se retrouvent en fin de liste puisqu'ils obtiennent une valeur nulle.

Afin de pallier ce problème, l'indice *iTer* combine les informations utilisées par *iInc* et *iLong*. En moyenne, la précision obtenue à l'aide de l'indice *iTer* dépasse celle obtenue avec les autres indices. *iTer* constitue donc un bon compromis permettant de bien cerner le caractère terminologique des CT et de concentrer les plus pertinents en tête de liste.

5. CONCLUSION

L'objectif de la présente thèse était d'élaborer une approche d'acquisition automatique de termes. Afin d'atteindre cet objectif, une méthodologie de travail a été mise au point pour :

- identifier les particularités lexicales d'un corpus technique : les PLS,
- mettre sur pied une stratégie d'acquisition automatique des termes fondée sur les PLS,
- proposer un indice terminogénique représentant l'intérêt terminologique des CT,
- intégrer les stratégies précédentes au sein d'un logiciel : TermoStat.

Les paragraphes qui suivent passent en revue les éléments principaux de la méthodologie élaborée. Ces derniers ont été intégrés dans un logiciel : TermoStat. Les apports de l'étude, ses limites et les perspectives de travail qui en découlent sont discutés.

5.1 Pivots lexicaux spécialisés

La première étape du processus d'acquisition automatique des termes, telle que nous l'avons envisagée dans le cadre de la présente étude, repose sur l'opposition de deux corpus. Cette dernière a pour objet l'identification d'unités lexicales ayant un comportement différent, dans un corpus d'analyse, de celui qu'elles adoptent dans un corpus de référence. Afin d'y parvenir, nous avons adopté une approche documentée dans le domaine de l'analyse du discours et de la linguistique quantitative ayant pour but l'identification des spécificités lexicales propres à un corpus.

La technique des spécificités avait, jusqu'à maintenant, été principalement utilisée dans le cadre de l'analyse du discours afin d'analyser des réponses à des questions ouvertes et d'en faire ressortir le lexique prédominant en fonction de certains critères démographiques. D'autres expérimentations ont aussi été réalisées sur des textes littéraires afin d'identifier le vocabulaire propre à une section d'un ouvrage ou à un ouvrage dans l'ensemble de l'œuvre d'un auteur.

Notre utilisation de la technique des spécificités se démarque sensiblement de ce qui a été fait par le passé puisque nous cherchons à mettre en opposition deux types de corpus (non technique et technique) afin de faire ressortir les particularités lexicales du corpus d'analyse. L'analyse des spécificités nous permet d'identifier les formes dont la fréquence observée n'est pas due au hasard et de les diviser en trois sous-ensembles : les spécificités positives, les spécificités négatives et les formes banales.

La notion de corpus est aussi envisagée selon un angle différent puisque le corpus utilisé dans le cadre de la présente recherche prend une dimension dynamique qui est, elle aussi, nouvelle. En effet, afin de vérifier la validité de nos hypothèses, nous constituons à chaque fois un corpus global (*CG*), lui-même constitué du corpus de référence (*CR*) et d'un corpus d'analyse (*CA*) qui varie en fonction des documents techniques à analyser. Cette approche a essentiellement pour but de vérifier si, d'un point de vue lexical, le corpus *CA* se comporte de façon identique au corpus *CR*.

Toute cette démarche visant à mettre en lumière les spécificités a un seul objectif : l'introduction de la notion de *pivot lexical*

spécialisé. Pour qu'une spécificité positive soit considérée comme un PLS, elle doit répondre aux conditions suivantes : être une forme nominale ou adjectivale et avoir une fréquence observée supérieure à sa fréquence théorique attendue. Afin de faire partie des PLS, la fréquence absolue d'une forme doit avoir moins de 1 chance sur 1 000 d'être due au hasard.

Les résultats de l'acquisition des PLS ont été validés. Cette étape de validation nous a permis de démontrer que les PLS recensés sont pertinents dans une proportion de 81 %. La pertinence des PLS a été déterminée par des terminologues en fonction de consignes préalablement établies.

L'analyse des formes non retenues à titre de PLS nous a permis d'observer que certaines unités lexicales reliées au domaine d'activité du corpus d'analyse n'ont pas été retenues. Ces formes possèdent une spécificité sémantique et le processus d'acquisition des PLS n'arrive pas à gérer ces oppositions.

La poursuite de nos travaux sur les spécificités nous semble invariablement passer par la prise en charge de documents plus volumineux au sein du corpus d'analyse. Les documents utilisés dans le cadre de la présente thèse sont de petite taille (voir 3.2) et il serait bon de vérifier la validité d'une technique d'acquisition des PLS sur des documents plus grands. Des expérimentations effectués sur un corpus d'analyse formé de documents provenant de domaines différents pourraient aussi conduire à des observations intéressantes sur le comportement des PLS.

Il serait aussi important de procéder à des analyses à l'aide d'un corpus de référence moins homogène. Bien que le CR utilisé dans le

cadre de la présente thèse traite d'une foule de sujets différents, il ne relève qu'un seul type de discours, le style journalistique. On peut envisager qu'un corpus de référence plus varié, d'un point de vue du style, pourrait avoir une certaine influence sur les PLS identifiés dans un corpus technique. Les analyses visant à vérifier la stabilité des PLS effectuées dans la section 4.2.2.2.3 méritent aussi d'être reproduites sur des corpus de référence de même taille que celui utilisé ou sur des corpus plus vastes. Étant donné le rôle fondamental joué par le corpus de référence dans le processus d'acquisition des PLS, des tests exhaustifs visant à identifier la taille et la composition idéals permettraient d'affiner les résultats.

5.2 Acquisition automatique des termes

Les PLS identifiés sont utilisés pour procéder à l'acquisition automatique des termes. La technique d'acquisition élaborée repose sur une approche hybride qui fait appel à la statistique et à la linguistique. La technique d'acquisition utilisée se distingue de par son utilisation des PLS. Cette contrainte imposée à l'algorithme d'acquisition des termes exige que l'ensemble des unités qui apparaissent au sein d'un terme possèdent le statut de PLS.

Le recours aux PLS permet au logiciel de limiter son analyse aux zones lexicales hautement spécifiques du corpus. Une fois les PLS recensés, le logiciel peut se contenter de scruter uniquement le contexte immédiat de ces derniers en quête de frontières de termes.

En plus de la prise en charge des PLS, notre algorithme d'acquisition de termes doit respecter d'autres contraintes dont celle qui veut que la tête du CT (située à l'extrême droite) soit un PLS nominal. Lors de l'analyse du contexte, le logiciel TermoStat exploite

le concept de frontière de termes[67]. Les éléments suivants sont considérés comme des frontières de termes :

- toute unité lexicale qui n'est pas un PLS,
- toute ponctuation à l'exception du tiret,
- certains éléments paratextuels (retour de chariot, tabulation, etc.).

L'imposition de ces exigences permet au logiciel TermoStat d'obtenir une précision moyenne de 74 % sur l'ensemble des documents techniques analysés. La précision atteinte est particulièrement intéressante puisqu'elle porte sur l'ensemble des CT. En effet, aucun seuil de fréquence minimale n'a été imposé lors de l'acquisition des CT. Les logiciels d'acquisition de termes ont souvent recours à un seuil de fréquence sous lequel les CT sont systématiquement écartés. L'adoption d'une fréquence minimale de 3 permettrait à la précision d'atteindre une valeur moyenne de 86,99 %.

La performance de TermoStat est particulièrement intéressante lorsqu'on ne considère que les CT simples (composés d'une seule unité lexicale). En effet, la précision atteint alors 86,20 % pour ce sous-ensemble des CT. L'approche des PLS s'applique donc très bien aux unités simples et permet de cibler avec succès les termes simples. Par contre, la performance observée pour les CT complexes est légèrement moins élevée et elle se situe à 64,97 %. On peut en conclure que notre approche d'acquisition automatique de termes à l'aide des PLS ne permet pas de bien comprendre le rôle joué par les

---

[67] Voir le paragraphe 4.2.3.2.1 au sujet de l'équivalence, dans le cas précis des unités terminologiques que nous cherchons à identifier, entre l'approche par frontière et l'utilisation d'une grammaire syntagmatique.

PLS au sein des unités complexes. Par contre, cette performance est équivalente à celle observée à l'aide d'un algorithme qui ne prend pas en charge les PLS. Le recours à une approche par PLS ne conduit donc pas à une diminution de la précision.

La recherche d'une précision maximale n'est pas sans impact négatif sur le rappel des algorithmes d'acquisition des termes. Bien que cette problématique se situe légèrement en marge de la recherche entreprise ici, nous avons tout de même tenu à en déterminer les conséquences. Afin de mesurer le rappel, nous avons utilisé comme point de référence les sorties d'une version modifiée du logiciel TermoStat qui procède à l'acquisition des CT sans la contrainte des PLS.

Le taux de rappel obtenu, évalué de façon différentielle à partir d'une version de TermoStat n'ayant pas recours à la contrainte des PLS, est élevé. Ce dernier atteint une moyenne globale de 85,58 %. Cette bonne performance du prototype laisse présager que les PLS occupent une place de choix dans l'ensemble des termes. Comme dans le cas de la précision, le rappel a été évalué pour les termes simples et pour les termes complexes. Le rappel atteint 81,05 % pour les premiers et 90,03 % pour les unités terminologiques complexes.

Les chiffres précédents démontrent que les PLS occupent une place de choix au sein des termes. Par contre, la précision relativement basse obtenue pour les termes complexes met en lumière la difficulté de l'approche par PLS de bien cerner le rôle de PLS au sein des termes complexes. Il serait donc intéressant de se tourner vers des méthodes probabilistes telles que celles mises de l'avant par Daille (1993, 1994a et 1994b) ou des méthodes plus linguistiques comme celles de Jacquemin (1996 et 2001) afin de vérifier si elles se

combinent mieux à l'utilisation des PLS comme point de départ de l'acquisition des termes.

5.3 Indice terminogénique

L'élaboration d'un indice permettant de représenter le potentiel terminologique d'un terme est une conséquence directe de la recherche de pertinence à l'origine de la présente thèse. Nous désirons présenter les résultats obtenus en positionnant les CT les plus pertinents au sommet de la liste des CT.

En cherchant à atteindre une liste qui serait triée en ordre de pertinence, nous soulevons le problème de l'évaluation du statut terminologique du CT et de son approximation par le logiciel. On cherche donc à qualifier ou à quantifier ce statut, à représenter le caractère terminogénique d'un CT. C'est autour de ce caractère terminologique d'un CT que nous articulons la notion d'indice terminogénique.

Les indices que nous avons élaborés possèdent un dénominateur commun : ils n'utilisent pas d'information provenant d'une source extérieure aux CT, à la liste des CT ou au corpus. Nous jugeons fondamental de conserver cette liberté des indices afin qu'ils puissent être testés et utilisés dans le cadre de travaux portant sur d'autres domaines, d'autres corpus ou ayant des objectifs différents.

Étant donné que nous désirons concentrer les CT pertinents en tête de la liste des CT, les données sont d'abord triées en fonction des divers indices et la précision des résultats est ensuite évaluée sur les premiers 50 % de la liste. Cette opération est répétée pour chacun des documents qui composent le corpus d'analyse.

Tous les indices testés ont permis d'obtenir un niveau de précision satisfaisant. Le point de référence utilisé est la *C-value*, mise au point par Frantzi et Ananiadou (1997), qui conduit à une précision de 81,25 %. Le deuxième indice testé correspond à la fréquence absolue des CT telle qu'observée dans le corpus d'analyse. À l'aide de cet indice très simple, la précision dépasse celle obtenue à l'aide de la *C-value* pour atteindre 84,22 %.

La longueur des CT, en nombre de mots, peut être couplée à la fréquence absolue et être utilisée à titre d'indice. L'indice *iLong* exploite ces deux sources d'information et atteint un précision de 86,71 %. Le troisième indice, *iInc*, cherche à caractériser la productivité de certains CT et leur prédisposition à donner naissance à des CT plus longs. Cet indice recense le nombre de fois où un CT est utilisé à titre de tête potentiel d'un CT plus long. La performance de cet indice est légèrement plus basse que le précédent à 85,94%.

L'inconvénient majeur de l'indice *iInc* est qu'il conduit à des valeurs nulle dès qu'un CT ne fait pas l'objet d'inclusion. Les expérimentations sur le corpus d'analyse ont permis de constater que la fusion des indices *iLong* et *iInc* en un indice plus complet, *iTer*, conduisent à une précision accrue. Les résultats obtenus grâce à l'aide de l'indice terminogénique *iTer* sont satisfaisants; la précision de la première partie de la liste est de 86.77 %.

Il serait très intéressant, sinon nécessaire, de procéder à des tests supplémentaires afin de vérifier si les bonnes performances des indices élaborés dans la présente thèse peuvent être reproduites dans des conditions différentes. Des expérimentations faisant varier le corpus d'analyse, le domaine représenté au sein de ce dernier, la taille

des corpus et leur niveau de technicité pourraient permettre de vérifier nos hypothèses. Ces variations pourraient aussi nous permettre de vérifier si l'indice *iTer* représente un indice terminogénique applicable à l'ensemble des démarches d'acquisition automatique de termes.

5.4 TermoStat

Les techniques présentées dans les paragraphes précédents ont donné naissance au logiciel TermoStat. Ce dernier a été utilisé afin de valider nos intuitions sur les PLS et sur leur utilité dans le cadre d'un processus d'acquisition automatique des termes.

La conception d'un système qui repose sur un ensemble de catégories grammaticales pour déterminer les frontières des unités linguistiques à identifier ouvre aussi la porte à une myriade d'utilisations en dehors de l'application terminologique première. On peut en effet, grâce à la redéfinition des unités qui constituent des frontières, envisager de procéder au repérage des combinaisons *verbe + préposition* ou de tout autre phénomène de cooccurrence et de combinatoire se démarquant de façon statistique dans un corpus d'analyse. Les avenues de recherche avec un outil comme TermoStat sont donc multiples et l'acquisition automatique des termes ne constitue qu'une première réalisation.

6. BIBLIOGRAPHIE

AHMAD, Khurshid (1996). *Language engineering and the processing of specialist terminology*, http://www.computing.surrey.ac.uk/ai/pointer/paris.html, 27 juin 1996.

ASSADI, Houssem et Didier BOURIGAULT (1996). « Acquisition et modélisation des connaissances à partir de textes : outils informatiques et éléments méthodologiques », dans *Actes du 10ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, Rennes, p. 505-514.

AUGER, Pierre (1979). « La syntagmatique terminologique, typologie des syntagmes et limite des modèles en structure complexe », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 9-26.

AUGER, Pierre (1994). « Les outils de la terminotique : typologie des outils d'aide à la terminologie et/ou d'automatisation de la chaîne de travail en terminographie », dans *Terminologies nouvelles*, nº 11, juin, p. 46-52.

AUGER, Pierre et Marie-Claude L'HOMME (1994). « La terminologie selon une approche textuelle : une représentation plus adéquate du lexique dans la langue de spécialité », dans *ALFA. Actes de langue française et linguistique. Terminologie et linguistique de spécialité. Études de vocabulaires et textes spécialisés*, vol. 7-8, p. 17-21.

AUGER, Pierre; Marie-Claude L'HOMME et Patrick DROUIN (1991). « Automatisation des procédures de travail en terminographie », dans *META*, vol. 36, nº 1, Montréal, Presses de l'Université de Montréal, p. 121-127.

BAILLARGEON, Michel (1989). *Probabilités statistiques et techniques de régression*, Trois-Rivières, Les Éditions SMG, 630 p.

BENVENISTE, Émile (1966). « Formes nouvelles de la composition nominale » dans *Bulletin de la société linguistique de Paris*, repris dans *Problèmes de linguistique générale*, tome 2, Paris, Gallimard, 1974, p. 163-176.

BOULANGER, Jean-Claude (1979). « Commentaire de Jean-Claude Boulanger », dans *Table ronde sur les problèmes du découpage du*

terme, 26 août 1978, Montréal, Office de la langue française, p. 169-182.

BOULANGER, Jean-Claude (1995). « Présentation : images et parcours de la socioterminologie », dans *META. Usages sociaux des termes : théories et terrains*, vol. 40, nº 2, Montréal, Presses de l'Université de Montréal, p. 194-205.

BOURIGAULT, Didier (1992a). « Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases », dans *Proceedings of the Fourteenth International Conference on Computational Linguistics-COLING 92*, Nantes, p. 977-981.

BOURIGAULT, Didier (1992b). « LEXTER, un logiciel d'extraction de Terminologie», dans *Actes de TAMA 92 : 2e symposium international de TermNet*, Avignon, mai, p. 229-258.

BOURIGAULT, Didier (1993). « Analyse syntaxique locale pour le repérage de termes complexes dans un texte », T.A.L., vol. 34, nº 2, p. 105-117.

BOURIGAULT, Didier (1994a). *Un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*, thèse de doctorat, Paris, École des Hautes Études en Sciences Sociales, 352 p.

BOURIGAULT, Didier (1994b). « Extraction et structuration automatique de terminologie pour l'aide à l'acquisition des connaissances à partir de textes », dans *Actes du 9ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'94)*, Paris, p. 397-408.

BOURIGAULT, Didier et Isabelle GONZALEZ (1994). « Acquisition automatique des termes complexes en français et en anglais, approche comparative », dans *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, Genève, ISSCO, p. 29-43.

BOURIGAULT, Didier et Monique Slodzian (1999). « Pour une terminologie textuelle», dans *Terminologies nouvelles*, nº 19, décembre et juin, p. 29-32.

BOURIGAULT Didier; JACQUEMIN, Christian et Marie-Claude L'HOMME (éditeurs) (2001). *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia, John Benjamins Publishing Company, xviii + 380 p.

BOUVERET, Myriam (1998). « Approche de la dénomination en langue spécialisée », dans *Meta*, vol. 43, nº 3, Montréal, Presses de l'Université de Montréal, p. 393-410.

BOWKER, Lynne (1996). « Towards a corpus-based approach to terminography », dans *Terminology*, vol. 3, nº 1, p. 27-52.

BRILL, Eric (1994). « Some Advances in Transformation-Based Part-of-Speech Tagging », *dans Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, tiré à part, 6 p.

BRILL, Eric (1995). « Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging », dans *Computational Linguistics*, vol. 21, nº 4, p. 543-565.

CABRE, Maria Teresa (1998). *La terminologie : théorie, méthode et applications*, traduit du catalan et adapté par Monique C. Cormier et John Humbley, Ottawa/Paris, Les Presses de l'Université d'Ottawa/André Colin, 322 p.

CABRÉ, Maria Teresa (1999). « Do we need an autonomous theory of terms? », dans *Terminology*, vol. 5, nº 1, p. 5-19.

CABRÉ, Maria Teresa (2000a). « Elements for a theory of terminology : Towards an alternative paradigm », dans *Terminology*, vol. 6, nº 1, p. 35-37.

CABRE, Maria Teresa (2000b). « Terminologie et linguistique : la théorie des portes », dans *Terminologies nouvelles*, nº 21, juin, Bruxelles, RINT, p. 10-15.

CAMLONG, André (1996). *Méthode d'analyse lexicale textuelle et discursive*, Paris, Orphrys, 199 p.

CHANSOU, Michel (1997). « Étude d'implantation des arrêtés de terminologie. Domaines : audiovisuel et publicité », dans *La mesure des mots : cinq études d'implantation terminologique*, Rouen, Publications de l'Université de Rouen, p. 133-233.

CHOLETTE, Marie (1994). « La problématique de la variation et de l'implantation : pour une socioterminologie », dans *Actes du colloque sur la problématique de l'aménagement linguistique : enjeux théoriques et pratiques*, colloque tenu les 5, 6, et 7 mai 1993 à l'Université du Québec à Chicoutimi, Montréal, Office de la langue française et Université du Québec à Chicoutimi, p. 495-514.

CHOUEKA, Yaacov (1988). « Looking for Needles in a Haystack or Locating Interesting Collocationnal Expressions in a Large Textual Database », dans *Actes de colloque du RIAO 88*, Cambridge, Cambridge University Press, p. 609-623.

CHOUEKA, Yaacov, KLEIN, S. T. et E. NEUWITZ (1983). « Automatic Retrieval of Frequent Idiomatic and Collocationnal Expressions in a Large Corpus », dans *ALLC Journal*, vol. 4, nº 1, p. 34-39.

CHURCH, Kenneth Ward et Patrick HANKS (1989). « Word Association Norms, Mutual Information, and Lexicography », dans *Computational Linguistics*, vol. 16, nº 1, mars, p. 22-29.

CLAS, André (1987). « Éditorial », dans *Meta. Vers l'an 2000 : la terminotique, bilan et prospective*, Montréal, Les Presses de l'Université de Montréal, vol. 32, nº 2, juin, p. 96-97.

CLAUDE, Louise (1996). « L'ATelier du TERminologue (Latter©) », dans *Terminologies nouvelles*, nº 15, juin et décembre, Bruxelles, RINT, p. 77-80.

CONDAMINES, Anne (1995). « Terminology : New needs, new perspectives », dans *Terminology*, vol. 2, nº 2, p. 219-238.

CONDAMINES, Anne et Chantal ENGUEHARD (éditeurs) (1999). *Terminologies nouvelles. Actes du colloque terminologie et intelligence artificielle*, Bruxelles, RINT, 136 p.

CRUSE, Alan (1986). *Lexical Semantics*, Cambridge, Cambridge University Press, xiv + 310 p.

DAILLE, Béatrice (1993). « Extraction automatique de terminologie monolingue», dans *Actes du colloque Informatique et langue naturelle*, décembre 1993, Nantes, 21 p.

DAILLE, Béatrice (1994a). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, thèse de doctorat, Paris, Université de Paris 7, 228 p.

DAILLE, Béatrice (1994b). « Extraction de noms composés terminologiques du domaine des Télécommunications », dans *5ièmes Journées ERLA-GLAT (Études et Recherches Lexicales Appliquées)*, Brest, 13 p.

DAILLE, Béatrice (1999). « Identification des adjectifs relationnels en corpus » dans *TALN '99. 6e Conférence annuelle sur le traitement automatique des langues naturelles*, 12 au 17 juillet 1999, Cargèse, ATATA, p. 105-114.

DAVID, Sophie (1990). *Différentes approches dans la composition nominale*, document du RDLC, Centre d'ATO, UQÀM, novembre, sans pagination.

DAVID, Sophie (1993). *Les unités nominales polylexicales : éléments de description et reconnaissance automatique*, thèse de doctorat, Paris, Université Paris 7, 281 p.

DAVID, Sophie et Pierre PLANTE (1990). « De la nécessité d'une approche morpho-syntaxique en analyse de textes », dans *Intelligence Artificielle et Sciences Cognitives au Québec*, vol. 2, nº 3, septembre, p. 140-155.

DROUIN, Patrick et Jacques LADOUCEUR (1994). « L'identification automatique de descripteurs complexes dans des textes de spécialité », dans *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, Genève, ISSCO, p. 18-28.

DROZD, Lubomir (1981). « Science terminologique : objet et méthode », dans *Textes choisis de terminologie. I. Fondements théoriques de la terminologie*, V. I. Siforov (dir.), Québec, Groupe interdisciplinaire de recherche scientifique et appliquée en terminologie (GIRSTERM), p. 115-131.

DUBUC, Robert (1979). « Découpage de l'unité terminologique », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 53-64.

DUCHESNE, Mariette (1979). « L'analyse morphologique automatique et ses applications en terminologie minière », dans *Actes du 6e colloque international de terminologie*, Pointe-au-Pic, 2 au 6 octobre 1979, Montréal, Office de la langue française, p. 485-503.

ENGUEHARD,Chantal (1994). « Automatic natural acquisition of a terminology », dans *Actes, 2nd International Conference on Quantitative Linguistics (QUALICO 94)*, Moscou, p. 83-88.

ENGUEHARD, Chantal; MALVACHE, Pierre et Philippe TRIGANO (1992). « Indexation de textes : l'apprentissage automatique de concepts »,

dans *Actes du XV^ème colloque international en linguistique informatique*, Nantes, p. 1197-1202.

FELBER, Helmut (1984). *Terminology Manual*, Paris, Infoterm, xii + 144 p.

FELLBAUM, Christiane (dir.) (1998). *WordNet : An Electronic Lexical Database*, Cambridge, The MIT Press, 423 p.

FERRET, Olivier et Brigitte GRUAU (2001). « Utiliser des corpus pour amorcer une analyse thématique », dans *Traitement automatique de la langue*, n° 2, vol. 42, Paris, Hermès, p. 517-545.

FILIPEC, Josef (1994). « Les rapports du lexique spécialisé et courant dans le texte et le système » , dans *ALFA. Actes de langue française et linguistique. Terminologie et linguistique de spécialité. Études de vocabulaires et textes spécialisés*, vol. 7-8, p. 349-359.

FOSSAT, Jean-Louis (1997). « Étude d'implantation des arrêtés de terminologie. Domaine : télédétection et aérospatiale », dans *La mesure des mots : cinq études d'implantation terminologique*, Rouen, Publications de l'Université de Rouen, p. 97-132.

FRANTZI, Katerina T. et Sophia ANANIADOU (1997). « Automatic Term Recognition Using Contextual Cues », dans *Proceedings of the 3rd DELOS Workshop,* Zurich, tiré à part, 8 p.

FRANTZI, Katerina T.; ANANIADOU, Sophia et Junichi TSUJII (1999). « Classifying Technical Terms », dans *Proceedings Third ICCC/IFIP Conference on Electronic Publishing,* Ronneby, p. 144-155.

GAZDAR, Gerald et Chris MELLISH (1989). *Natural language processing in PROLOG: an introduction to computational linguistics*, New York, Addison-Wesley Publishing Company, xv + 504 p.

GAUDIN, François (1993). *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*, Rouen, Université de Rouen, 255 p.

GOFFIN, Roger (1979). « Le découpage du terme à des fins lexicographiques : critères formels, sémantiques, quantitatifs et taxinomiques », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 157-168.

GOFFIN, Roger (1992). « Du synthème au phaséolexème en terminologie différentielle », dans *Terminologie et traduction*, Commission des Communautés européennes, Service de traduction, Unité terminologie, Bruxelles - Luxembourg, nº 2-3, p. 431-438.

GOUADEC, Daniel (1997). « Étude d'implantation des arrêtés de terminologie. Domaine : informatique », dans *La mesure des mots : cinq études d'implantation terminologique*, Paris, Délégation générale à la langue française, p. 235-493.

GOUGENHEIM, Georges; MICHEA, René; Paul RIVENC et Aurélien SAUVAGEOT (1964). *L'élaboration du français fondamental (1er degré : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier, 302 p.

GUILBERT, Louis (1965). *La formation du vocabulaire de l'aviation*, Paris, Librairie Larousse, 712 p.

HABERT, Benoît; NAZARENKO, Adeline et André SALEM (1997). *Les linguistiques de corpus*, Paris, Armand Colin, 240 p.

HUBERT, Pierre et Dominique LABBÉ (1988). « Note sur l'approximation de la loi hypergéométrique par la formule de Muller », dans *Études sur la richesse et la structure lexicales*, Labbé *et al.* éditeurs, Genève, Slatkine, p. 77-91.

INTERVAL (1998). *A European project on the validation of terminology resources. Final Report*, http://www.computing.surrey.ac.uk/research/ai/new_interval/final_report/final_report.frames.html, 3 septembre 1998.

JACQUEMIN, Christian; KLAVANS, Judith L. et Evelyne TZOUKERMANN (1997). « Expansion of Multi-Word Termes for Indexing and Retrieval Using Morphology and Syntax », dans *Proceedings of the Thirty-fifth Annual Meeting of the Association of for Computational Linguistics*, Barcelone, 7-10 juillet, tire à part, 8 p.

JACQUEMIN, Christian (1996). « What is the tree that we see through the window : A linguistic approach to windowing and term variation», dans *Information Processing & Management*, vol. 32, nº 4, p. 445-458.

JACQUEMIN, Christian (1997). *Variation terminologique : reconnaissance et acquisition automatique des termes et de leurs variantes en*

corpus, Habilitation à diriger des thèses, Nantes, Université de Nantes, 323 p.

JACQUEMIN, Christian (1999). *Activités de recherche de Christian Jacquemin*, http://m17.limsi.fr/Individu/jacquemi/Rech.html, 7 avril 1999.

JACQUEMIN, Christian (2001). *Spotting and dicovering temrs through natural language processing*, Cambridge, The MIT Press, 378 p.

JUSTESON, John et Slava KATZ (1993). *Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906*, IBM Research Division, tire à part, 13 p.

KAGEURA, Kyo (1995). « Toward the theoretical study of terms : A sketch from the linguistic viewpoint », dans *Terminology*, vol. 2, nº 2, p. 239-258.

KAGEURA, Kyo (1999). « Theories "of" terminology : A quest for a framework for the study of term formation », dans *Terminology*, vol. 5, nº 1, p. 21-40.

KITTREDGE, Richard (1982). « Variation and Homogeneity of Sublanguages », dans *Sublanguage : Studies of Language in Restricted Domains*, Richard Kittredge et John Lehrberger editors, Berlin-New York, Walter de Gruyter, p. 107-137.

KLEIBER, Georges (1990). *La sémantique du prototype : catégorie et sens lexical*, Paris, Presses universitaires de France, 199 p.

KOCOUREK, Rostislav (1991). *La langue française de la technique et de la science*, 2e édition, Wiesbaden, Brandstetter Verlag, xviii + 327 p.

LAFON, Pierre (1980). « Sur la variabilité de la fréquence des formes dans un corpus », dans *MOTS*, nº 1, p. 128-165.

LAUER, Mark (1994). « Conceptual Association for Compound Noun Analysis », dans *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Student Session*, juin, Las Cruces, tire à part, 6 p.

LEBART, Ludovic et André SALEM (1988). *Analyse statistique des données textuelles : questions ouvertes et lexicométrie*, Paris, Dunod, 210 p.

LEBART, Ludovic et André SALEM (1994). *Statistique textuelle*, Paris, Dunod, viii + 342 p.

LEONHARDT, Christine (1991a). « LATTER, The terminologist's workstation », dans *Actes du symposium international Terminologie et documentation dans la communication spécialisée*, Ottawa, Secrétariat d'État du Canada, p. 257-275.

LEONHARDT, Christine (1991b). « The terminologist's workstation project », dans *L'actualité terminologique*, vol. 24, nº 2, Ottawa, Secrétariat d'État du Canada, p. 11-12.

LEONHARDT, Christine (1994). *Les applications en terminotique à la direction de la terminologie et de la documentation*, Communication présentée à l'Association canadienne-française pour l'avancement des sciences (ACFAS), Montréal, 17 mai, tiré à part, 7 p.

LERAT, Pierre (1995). *Les langues spécialisées*, Paris, Presses universitaires de France, 201 p.

LOTTE, D. S. (1981). « Principes d'établissement d'une terminologie scientifique et technique », dans *Textes choisis de terminologie. I. Fondements théoriques de la terminologie*, V. I. Siforov (dir.), Québec, Groupe interdisciplinaire de recherche scientifique et appliquée en terminologie (GIRSTERM), p. 1-53.

LOUBIER, Christiane (1993). « L'mplantation du français comme langue de travail au Québec : vers un processus de changement linguistique planifié », dans *L'implantation du français : actualisation d'un changement linguistique planifié*, André Martin et Christiane Loubier (éditeurs), Montréal, Office de la langue française, p. 57-133.

LOUBIER, Christiane et Louis-Jean ROUSSEAU (1994). « L'acte de langage, source et fin terminologique », dans *ALFA. Actes de langue française et linguistique. Terminologie et linguistique de spécialité. Études de vocabulaires et textes spécialisés*, vol. 7-8, p. 75-87.

MACKLOVITCH, Elliott (1991). « Le poste de travail du traducteur... En prose claire et simple », dans *Actes du symposium international Terminologie et documentation dans la communication spécialisée*, Ottawa, Secrétariat d'État du Canada, p. 10-21.

MARCUS, Mitch; SANTORINI, Béatrice et Mary Ann MARCINKIEWICZ (1993). « Building a large annotated corpus of English : The Penn

Treebank », dans *Computational Linguistics*, vol. 19, n°2, p. 313-330.

MARTINET, André (1979). « Intervention d'André Martinet », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 183-190.

MAYNARD, Diana et Sophia ANANIADOU (2001). « Term extraction using a similarity-based approach », dans *Recent advances in Computational Terminology*, Bourigault *et al.* (éditeurs), Amsterdam/Philadelphia, John Benjamins Publishing Company, p. 261-278.

MEL'ČUK, Igor (1981). « Meaning-text models : a recent trend in Soviet linguistics » dans *The Annual Review of Anthropology*, n° 10, p. 27-62.

MULLER, Charles (1979). *Langue française et linguistique quantitative : recueils d'articles*, Genève, Slatkine, xiii + 470 p.

MULLER, Charles (1992a). *Principes et méthodes de statistique lexicale*, Paris, Honoré Champion, 207 p.

MULLER, Charles (1992b). *Initiation aux méthodes de la statistique linguistique*, Paris, Honoré Champion, 188 p.

NAKAGAWA, Hiroshi et Tatsunori MORI (1998). « Nested Collocation and Compound Noun for Term Extraction » dans *Computerm '98. First Workshop on Computational Terminology. Proceedings of the Workshop*, 15 août 1998, Université de Montréal, p. 64-70.

NKWENTI-AZEH, Blaise (1994). « Positional and combinational characteristic of terms : consequences for corpus-based terminography », dans *Terminology*, vol. 1, n° 1, p. 61-95.

OLF (1979). *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, 214 p.

OLF (2001). *Le grand dictionnaire terminologique*, http://www.granddictionnaire.com, mai 2001.

OTMAN, Gabriel (dir.) (1995). « Terminologie et intelligence artificielle », *La Banque des mots*, numéro spécial, Paris, Conseil international de la langue française, n° 7, 112 p.

OUESLATI, Rochdi (1999). *Aide à l'acquisition de connaissances à partir de corpus,* thèse de doctorat, Strasbourg, Université Louis Pasteur, 417 p.

PARADIS, Claude et Pierre AUGER (1987). « La terminotique ou la terminologie à l'ère de l'informatique », dans *Meta,* vol. 32, nº 2, Montréal, Presses de l'Université de Montréal, p. 102-110.

PEARSON, Jennifer (1998). *Terms in Context,* Amsterdam/Philadelphie, John Benjamins Publishing, xii + 243 p.

PERRON, Jean (1996). « ADEPTE-NOMINO : un outil de veille terminologique », dans *Terminologies nouvelles,* nº 15, juin et décembre, Bruxelles, RINT, p. 32-47.

PHAL, André (1971). *Vocabulaire général d'orientation scientifique et technique : part du lexique commun dans l'expression scientifique,* Paris, CRÉDIF, 128 p.

PICHT, Heribert (1987). « Terms and their LSP Environment - LSP Phraseology », dans *Meta,* vol. 32, nº 2, Montréal, Presses de l'Université de Montréal, p. 149-155.

PICHT, Heribert et Jennifer DRASKAU (1985). *Terminology: An introduction,* England, The University of Surrey, 265 p.

PLANTE, Pierre; DUMAS, Lucie et André PLANTE (2000). *Nomino. Synopsis,* http://www.ling.uqam.ca/nomino/synopsis.htm, 13 mars 2000.

POINTER (1996). *Pointer Final Report. Proposal for an operational infrastructure for terminology in Europe,* http://www.computing.surrey.ac.uk/ai/pointer/report/index.html, 6 août 1996.

QUIRION, Jean (1996). « L'implantation terminologique : aspects évaluatifs et terminologiques », dans *ALFA. Actes de langue française et de linguistique. Comparaison, contrastes, correspondances : le français et l'anglais en terminologie et en langue de spécialité,* vol. 9, p. 143-152.

QUIRION, Jean (2000). *Aspects évaluatifs de l'implantation terminologique,* thèse de doctorat, Montréal, Université de Montréal, xix + 288 p.

REY, Alain (1999). « Terminology between the experience of reality and the command of sign », dans *Terminology*, vol. 5, nº 1, p. 121-134.

RINT (1996). *Terminologie nouvelles. Banques de terminologie*, nº 15, juin et décembre, 176 p.

RINT (2001). *Atelier de travail informatisé du terminologue (ATTRAIT)*, http://www.rint.org/attrait/index.htm, 15 février 2001.

RONDEAU, Guy (1984). *Introduction à la terminologie*, 2e édition, Chicoutimi, Gaëtan Morin, xlv + 238 p.

ROUSSEAU, Louis-Jean (1979). « Commentaire de Louis-Jean Rousseau », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 27-36.

SAGER, Juan Carlos (1979). « Commentary by Prof. Juan Carlos Sager », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 37-52.

SAGER, Juan Carlos (1990). *A Practical Course in Terminology Processing*, Amsterdam/Philadelphie, John Benjamins Publishing Company, 254 p.

SAGER, Juan Carlos (1999). « In search of a foundation : Towards a theory of the term », dans *Terminology*, vol. 5, nº 1, p. 51-57.

SAGER, Juan Carlos; DUNGWORTH, David et Peter F. MCDONALD. (1980). *English Special Languages. Principles and Practice in Science and Technology*, Wiesbaden, Brandstetter, 368 p.

SALEM, André (1987). *Pratique des segments répétés : essai de statistique textuelle*, Institut national de la langue française - INaLF, URL Lexicométrie et textes politiques, Publications de l'INaLF, Collection Saint-Cloud, Paris, Klincksieck, 333 p.

SANTORINI, Beatrice (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47*, Department of Computer and Information Science, Philadelphie, University of Pennsylvania, tiré à part, 21 p.

SCHAETZEN, Caroline de (1994). « Terminotique: un peu d'histoire », dans *Terminologies nouvelles*, nº 11, juin, p. 60-62.

SCHAETZEN, Caroline de (1997). « Typologie des outils de terminotique », dans *Études de linguistique offertes à Rostislav Kocourek*, Halifax, Les presses d'ALFA, p. 81-87.

SCHMID, Helmut (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees », dans *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 9 p.

SHANNON C. E. (1948). « A mathematical theory of communication », dans *Bell System Tech. Journal*, nº 27, p. 379-423, p. 623-656.

SILBERZTEIN, Max (1993). *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Paris, Masson, xiv + 233 p.

SMADJA, Frank (1993). « Retrieving Collocations from Text: Xtract », dans *Computational Linguistics*, vol. 19, nº 1, Association for Computational Linguistics, p. 143-177.

TEMMERMAN, Rita (1999). « Why traditional terminology theory impedes a realistic description of categories and terms in the life sciences », dans *Terminology*, vol. 5, nº 1, p. 77-92.

TOUSSAINT, Yannick; ROYAUTE, Jean; MULLER, Chantal et Xavier POLANCO (1997). « Analyse linguistique et infométrique pour l'acquisition des connaissances », dans *TIA-97 : actes des deuxièmes rencontres terminologie et intelligence artificielle*, Toulouse-le Mirail, Équipe de Recherche en Syntaxe et Sémantique, p. 27-45.

URI-INIST-CNRS (2001). *TIA-2001 : actes de quatrièmes rencontres Terminologie et Intelligence Artificielle*, 3-4 mai, Nancy, URI-INIST-CNRS, 283 p.

VINAY, Jean-Paul (1979). « Problèmes du découpage du terme », dans *Table ronde sur les problèmes du découpage du terme*, 26 août 1978, Montréal, Office de la langue française, p. 81-100.

VOUTILAINEN, Atro (1993). « Nptool, a detector of English noun phrases », dans *Proceedings of the Workshop on Very Large Corpora*, June, Columbus, Ohio State University, p.48-57.

WÜSTER, Eugen (1968). *Dictionnaire multilingue de la machine-outil*, Londres, Technical Press, 744 p.

WÜSTER, Eugen (1981). « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses », Guy Rondeau et Helmut Felber (réd.), *Textes choisis de terminologie. I. Fondements théoriques de la terminologie*, Québec, GIRSTERM, p. 55-114.

7. ANNEXES

7.1 Annexe A – Formes spécifiques non retenues à titre de PLS

CA1		CA2		CA3	
Forme	**Valeur-test**	**Forme**	**Valeur-test**	**Forme**	**Valeur-test**
library	3,05	reasonable	3,07	allowed	3,05
publication	3,01	this	3,05	**incident**	3,04
editor	2,94	express	3,03	**pre**	3,04
discover	2,93	**wayside**	3,03	**distortion**	3,03
gather	2,88	**lan**	3,03	dashed	3,03
chart	2,87	exiting	3,03	**building**	2,98
bb	2,82	**entry**	3,00	6	2,98
variable	2,82	**logo**	2,97	extend	2,95
none	2,79	**transportation**	2,97	apply	2,89
empty	2,77	meet	2,96	process	2,87
customer	2,76	**frequency**	2,93	**grid**	2,87
process	2,75	internal	2,92	serial	2,87
compare	2,74	**length**	2,92	terminated	2,87
payload	2,73	**leaf**	2,91	222	2,87
activate	2,73	20	2,91	managed	2,86
appear	2,72	**number**	2,90	updated	2,80
end	2,69	**exchange**	2,90	**infrastructure**	2,80
set	2,65	minimal	2,86	fixed	2,80
explicitly	2,64	**path**	2,84	34	2,74
cd	2,60	34	2,84	operating	2,72
enter	2,56	**copyright**	2,82	less	2,72
remaining	2,53	28	2,82	**need**	2,70
change	2,51	permit	2,81	**equal**	2,67
g	2,49	**shipping**	2,78	allocated	2,67
non	2,49	utilized	2,78	downstream	2,67
number	2,48	restart	2,78	maintained	2,66
an	2,45	contrast	2,78	reduced	2,64
target	2,42	**ta**	2,78	reasonable	2,63
centralized	2,42	**airborne**	2,78	**x**	2,63

currently	2,40	**pip**	2,78	enabling	2,62
see	2,38	predetermined	2,78	minimum	2,57
defining	2,35	ui	2,78	locking	2,56
231	2,35	96	2,76	**detection**	2,56
viii	2,35	66	2,74	**route**	2,56
updated	2,29	**los**	2,70	**multi**	2,53
issue	2,20	1993	2,65	masked	2,51
e	2,19	correct	2,61	employing	2,51
periodically	2,18	**compensation**	2,61	horizontal	2,51
allocated	2,18	include	2,60	loaded	2,49
gateway	2,18	120	2,59	travelling	2,48
activated	2,18	1603	2,58	**standard**	2,46
directives	2,18	reconstruct	2,58	decrease	2,46
greenwich	2,13	computing	2,58	**deployment**	2,46
corresponding	2,09	6000	2,58	passive	2,46
maintained	2,06	migrating	2,58	**bus**	2,38
reside	2,04	only	2,57	granted	2,38
alert	2,04	form	2,57	current	2,35
return	1,97	environmental	2,56	technical	2,34
		expansion	2,56	balance	2,28
		tape	2,56	**point**	2,24
		23	2,54	**loop**	2,24
		v	2,54	vertical	2,24
		planned	2,53	routed	2,20
		remote	2,52	reinforced	2,16
		160	2,51	onto	2,14
		0	2,49	comparing	2,13
		comply	2,47	displaying	2,13
		kg	2,44	**l**	2,12
		disclose	2,41	within	2,12
		except	2,41	add	2,11
		optic	2,41	the	2,10
		reactive	2,41	focuses	2,10
		762	2,41	independently	2,06
		slat	2,41	32	2,05
		intended	2,40	both	2,05
		processor	2,39	connecting	2,00
		adjacent	2,39	**difference**	1,99

		assist	2,39	allow	1,97
		fm	2,36		
		bounce	2,36		
		disclosure	2,36		
		compensate	2,31		
		issue	2,28		
		fed	2,27		
		electrically	2,26		
		clamp	2,26		
		1830	2,26		
		bypassed	2,26		
		performance	2,26		
		typically	2,26		
		total	2,25		
		132	2,21		
		r	2,20		
		yellow	2,13		
		representatio n	2,13		
		accumulated	2,10		
		operate	2,10		
		edge	2,07		
		automatic	2,07		
		exceed	2,06		
		affecting	2,04		
		sanity	2,03		
		feeding	1,99		
		pin	1,99		

7.2 Annexe B – Répartition des PLS en fonction de la fréquence

7.2.1 CA_1

Substantifs	**PLS**	**Fréquence**
237	134	1
108	71	2
46	34	3
37	29	4
27	22	5
23	21	6
22	21	7
23	19	8
15	15	9
12	11	10
9	8	11
7	7	12
7	7	13
8	8	14
4	4	16
4	4	17
2	2	18
1	1	19
2	2	20
7	7	21
6	5	22
2	2	23
1	1	24
1	1	26
3	3	27
1	1	29
2	2	31

2	2	33
3	3	34
1	1	35
2	2	36
1	1	39
1	1	41
2	2	42
1	1	46
1	1	47
2	2	48
1	1	54
2	2	57
1	1	58
1	1	61
1	1	63
1	1	72
2	2	74
1	1	88

7.2.2 CA_2

Substantifs	PLS	Fréquence
550	432	1
129	79	2
58	33	3
53	37	4
42	28	5
30	24	6
11	10	7
12	12	8
10	8	9
17	15	10
3	2	11
7	7	12
6	4	13
17	17	14
11	10	15
3	2	16
4	4	17
8	7	18
1	1	19
6	6	20
4	4	21
4	3	22
4	4	23
2	2	24
3	3	25
3	3	26
5	5	28

6	6	29
3	3	30
2	2	31
3	3	32
1	1	33
2	2	34
2	1	35
1	1	36
1	1	37
5	5	38
3	3	39
1	1	40
2	2	41
1	1	42
3	3	43
2	2	44
2	2	45
2	2	47
2	2	48
1	1	49
1	1	51
1	1	54
1	1	55
1	1	57
2	2	60
1	1	61
1	1	63
2	2	68
1	1	69
1	1	71
1	1	73
1	1	74

2	2	75
1	1	78
2	2	83
1	1	84
1	1	86
1	1	103
1	1	110
1	1	113
1	1	114
2	2	116
1	1	117
1	1	120
1	1	129
1	1	133
1	1	136
1	1	142
1	1	156
1	1	157
2	2	218
1	1	224
1	1	228
1	1	237
1	1	267
1	1	280
1	1	337
1	1	338

7.2.3 CA_3

Substantifs	PLS	Fréquence
177	102	1
67	38	2
40	31	3
26	18	4
15	13	5
12	11	6
14	14	7
3	3	8
10	10	9
7	6	10
3	3	11
3	3	12
2	2	13
4	4	14
5	5	15
1	1	16
3	3	18
1	1	19
1	1	20
4	4	21
5	5	22
1	1	23
2	2	24
1	1	25
5	5	26
1	1	29
2	2	32

2	2	33
1	1	37
2	2	40
1	1	42
2	2	44
2	2	47
1	1	50
1	1	52
1	1	53
1	1	54
2	2	64
2	2	66
2	2	68
1	1	72
1	1	77
1	1	79
1	1	86
1	1	92
1	1	151

7.3 *Annexe C – Substantifs non retenus à titre de PLS*

CA1		CA2		CA3	
Forme	**Fréquence**	**Forme**	**Fréquence**	**Forme**	**Fréquence**
time	22	m	35	point	10
end	11	group	22	building	7
number	10	number	20	state	6
group	8	house	18	need	5
day	8	los	16	manager	5
m	8	plan	15	page	4
change	8	end	13	end	4
issue	7	issue	13	incident	4
period	6	part	11	field	4
level	6	exchange	10	bus	4
rate	5	s	10	care	4
process	5	performance	10	division	4
return	5	store	9	l	4
p	5	canada	9	case	4
control	5	case	7	difference	3
e	4	t	6	way	3
hour	4	r	6	route	3
customer	4	p	6	name	3
none	4	view	6	area	3
editor	4	air	6	pre	3
card	4	use	6	company	3
book	4	manager	5	standard	3
environment	4	edge	5	time	3
g	4	addition	5	s	3
library	3	document	5	result	2
difference	3	pair	5	management	2
report	3	value	5	bit	2
way	3	entry	5	rate	2
face	3	change	5	length	2
friday	3	total	5	other	2
monday	3	degree	5	exchange	2
case	3	leaf	5	event	2
effect	3	list	5	size	2
language	3	unit	5	bay	2

november	3	need	5	situation	2
transport	3	transportation	5	environment	2
event	3	floor	5	relationship	2
publication	3	length	5	part	2
side	2	strategy	4	degree	2
generation	2	g	4	multi	2
multi	2	path	4	place	2
code	2	example	4	ste	2
chart	2	expansion	4	x	2
total	2	compensation	4	open	2
index	2	bottom	4	ltd	2
speed	2	tape	4	corporation	2
o	2	contact	4	canada	2
scheme	2	transport	4	united	2
telephone	2	kg	4	cut	2
training	2	rating	4	trouble	2
representative	2	care	4	july	2
union	2	show	4	property	2
category	2	red	4	world	2
reason	2	united	4	entry	1
trading	2	distance	3	chain	1
committee	2	none	3	devices	1
cd	2	d	3	amount	1
menu	2	form	3	interference	1
action	2	plant	3	approach	1
open	2	green	3	book	1
america	2	rate	3	provision	1
part	2	route	3	trend	1
purchase	2	oct	3	error	1
administration	2	range	3	account	1
history	2	test	3	noise	1
june	2	term	3	distortion	1
paper	2	reference	3	sum	1
size	2	budget	3	response	1
october	2	distribution	3	have	1
transfer	2	customer	3	capacity	1
industry	2	position	3	equal	1
equivalent	2	architecture	3	placement	1

s	2	day	3	comparison	1
maintenance	2	v	3	separation	1
international	2	international	3	fact	1
model	1	america	3	services	1
drop	1	north	3	voice	1
connection	1	user	3	t	1
criteria	1	multi	3	plan	1
confirmation	1	europe	2	grid	1
range	1	index	2	use	1
vt	1	new	2	gain	1
payload	1	fm	2	spectrum	1
ui	1	tran	2	plane	1
external	1	provision	2	exception	1
c	1	e	2	peak	1
d	1	carrier	2	compensation	1
l	1	analysis	2	technique	1
migration	1	preparation	2	contact	1
degree	1	format	2	guidance	1
minimum	1	working	2	loop	1
child	1	publication	2	alternative	1
bottom	1	space	2	master	1
staff	1	pin	2	sample	1
problem	1	memory	2	media	1
flag	1	processor	2	view	1
gap	1	property	2	purpose	1
manner	1	other	2	testing	1
greenwich	1	head	2	support	1
fm	1	environment	2	mix	1
place	1	peer	2	database	1
database	1	zone	2	problem	1
contact	1	map	2	out	1
shelves	1	set	2	concept	1
cp	1	planning	2	distinction	1
maximum	1	sub	2	equivalent	1
not	1	size	2	understanding	1
r	1	organizer	2	reach	1
t	1	radio	2	activity	1
analysis	1	frequency	2	c	1

view	1	devices	2	surveillance	1
ap	1	method	2	phoenix	1
year	1	weight	2	plant	1
month	1	condition	2	range	1
_	1	handling	2	km	1
care	1	representation	2	deployment	1
plan	1	water	2	infrastructure	1
viii	1	reduction	2	core	1
mechanism	1	disclosure	2	stage	1
q	1	logo	2	detection	1
is	1	world	2	low	1
violation	1	inc	2	user	1
label	1	package	2	lineup	1
amount	1	shipping	2	test	1
bb	1	safety	2	limit	1
creation	1	growth	2	curve	1
use	1	offering	2	conversion	1
factory	1	copyright	2	counter	1
distribution	1	july	2		
co	1	assist	2		
gateway	1	replacement	1		
water	1	h	1		
mark	1	n	1		
alert	1	o	1		
ring	1	modem	1		
symbol	1	km	1		
angle	1	remainder	1		
explanation	1	exception	1		
click	1	office	1		
keyboard	1	subject	1		
window	1	following	1		
device	1	arrangement	1		
b	1	question	1		
directives	1	color	1		
european	1	shape	1		
south	1	left	1		
caribbean	1	family	1		
north	1	fashion	1		

region	1	greater	1		
phone	1	in	1		
mail	1	suite	1		
consultation	1	business	1		
integration	1	respect	1		
site	1	present	1		
coverage	1	mirror	1		
agreement	1	investment	1		
help	1	device	1		
height	1	november	1		
need	1	dc	1		
j	1	com	1		
audience	1	committee	1		
regular	1	turn	1		
visibility	1	combination	1		
business	1	wayside	1		
personnel	1	out	1		
september	1	look	1		
february	1	library	1		
ware	1	usage	1		
brand	1	eastbound	1		
design	1	classic	1		
layout	1	voice	1		
devices	1	presence	1		
core	1	bus	1		
march	1	media	1		
institute	1	read	1		
national	1	drive	1		
american	1	disk	1		
association	1	administration	1		
exchange	1	engine	1		
		insert	1		
		lan	1		
		star	1		
		status	1		
		inventory	1		
		sanity	1		
		detection	1		

		collection	1		
		gateway	1		
		noise	1		
		supply	1		
		migration	1		
		bear	1		
		location	1		
		principle	1		
		tail	1		
		button	1		
		integration	1		
		mind	1		
		pg	1		
		bridge	1		
		effect	1		
		current	1		
		life	1		
		production	1		
		march	1		
		digital	1		
		ta	1		
		damage	1		
		proximity	1		
		close	1		
		wave	1		
		response	1		
		draft	1		
		bed	1		
		simulator	1		
		truck	1		
		shipment	1		
		airborne	1		
		september	1		
		risk	1		
		construction	1		
		factor	1		
		sea	1		
		environmental	1		

		high	1		
		low	1		
		duration	1		
		year	1		
		re	1		
		speed	1		
		responsibility	1		
		field	1		
		feed	1		
		battery	1		
		pip	1		
		place	1		
		placement	1		
		series	1		
		processing	1		
		blue	1		
		w	1		
		u	1		
		bounce	1		
		security	1		
		slat	1		
		fat	1		
		density	1		
		am	1		
		multiple	1		