Université de Montréal

Élaboration d'un corpus étalon pour l'évaluation d'extracteurs de termes

par
Gabriel Bernier-Colborne

Département de linguistique et de traduction
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès arts (M.A.)
en traduction
option recherche

mai, 2012

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

Élaboration d'un corpus étalon pour l'évaluation d'extracteurs de termes

Présenté par :
Gabriel Bernier-Colborne

a été évalué par un jury composé des personnes suivantes :

Mireille Tremblay, président-rapporteur
Patrick Drouin, directeur de recherche
Marie-Claude L'Homme, co-directrice
Philippe Langlais, membre du jury

Résumé

Ce travail porte sur la construction d'un corpus étalon pour l'évaluation automatisée des extracteurs de termes. Ces programmes informatiques, conçus pour extraire automatiquement les termes contenus dans un corpus, sont utilisés dans différentes applications, telles que la terminographie, la traduction, la recherche d'information, l'indexation, etc. Ainsi, leur évaluation doit être faite en fonction d'une application précise.

Une façon d'évaluer les extracteurs consiste à annoter toutes les occurrences des termes dans un corpus, ce qui nécessite un protocole de repérage et de découpage des unités terminologiques. À notre connaissance, il n'existe pas de corpus annoté bien documenté pour l'évaluation des extracteurs. Ce travail vise à construire un tel corpus et à décrire les problèmes qui doivent être abordés pour y parvenir.

Le corpus étalon que nous proposons est un corpus entièrement annoté, construit en fonction d'une application précise, à savoir la compilation d'un dictionnaire spécialisé de la mécanique automobile. Ce corpus rend compte de la variété des réalisations des termes en contexte. Les termes sont sélectionnés en fonction de critères précis liés à l'application, ainsi qu'à certaines propriétés formelles, linguistiques et conceptuelles des termes et des variantes terminologiques.

Pour évaluer un extracteur au moyen de ce corpus, il suffit d'extraire toutes les unités terminologiques du corpus et de comparer, au moyen de métriques, cette liste à la sortie de l'extracteur. On peut aussi créer une liste de référence sur mesure en extrayant des sous-ensembles de termes en fonction de différents critères. Ce travail permet une évaluation automatique des extracteurs qui tient compte du rôle de l'application. Cette évaluation étant reproductible, elle peut servir non seulement à mesurer la qualité d'un extracteur, mais à comparer différents extracteurs et à améliorer les techniques d'extraction.

Mots-clés : Terminologie, terminologie computationnelle, extraction de termes, évaluation, annotation de corpus, variation terminologique.

Abstract

We describe a methodology for constructing a gold standard for the automatic evaluation of term extractors. These programs, designed to automatically extract specialized terms from a corpus, are used in various settings, including terminology work, translation, information retrieval, indexing, etc. Thus, the evaluation of term extractors must be carried out in accordance with a specific application.

One way of evaluating term extractors is to construct a corpus in which all term occurrences have been annotated. This involves establishing a protocol for term selection and term boundary identification. To our knowledge, no well-documented annotated corpus is available for the evaluation of term extractors. This contribution aims to build such a corpus and describe what issues must be dealt with in the process.

The gold standard we propose is a fully annotated corpus, constructed in accordance with a specific terminological setting, namely the compilation of a specialized dictionary of automotive mechanics. This annotated corpus accounts for the wide variety of realizations of terms in context. Terms are selected in accordance with specific criteria pertaining to the terminological setting as well as formal, linguistic and conceptual properties of terms and term variations.

To evaluate a term extractor, a list of all the terminological units in the corpus is extracted and compared to the output of the term extractor, using a set of metrics to assess its performance. Subsets of terminological units may also be extracted, providing a level of customization. This allows an automatic and application-driven evaluation of term extractors. Due to its reusability, it can serve not only to assess the performance of a particular extractor, but also to compare different extractors and fine-tune extraction techniques.

Keywords : Terminology, computational terminology, term acquisition, evaluation, annotated corpora, term variation.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Je souhaite tout d'abord remercier Patrick Drouin de m'avoir encouragé à poursuivre mes études aux cycles supérieurs; je tiens à lui exprimer toute ma reconnaissance. Je remercie également Marie-Claude L'Homme, qui a gentiment accepté de codiriger ce travail. Sans leur constante disponibilité et les commentaires judicieux qu'ils ont toujours su apporter, l'épreuve aurait été considérablement plus difficile. Patrick et Marie-Claude, je vous témoigne toute ma gratitude.

De plus, je remercie le Fonds québécois de recherche (société et culture) et le Conseil de recherches en sciences humaines du Canada pour leur appui financier, ainsi que le Département de linguistique et de traduction et la Faculté des études supérieures et postdoctorales de l'Université de Montréal pour les bourses qu'ils m'ont octroyées. Je remercie également le gouvernement du Canada pour les bourses qui m'ont été offertes dans le cadre du Programme de renforcement du secteur langagier au Canada, ainsi que les donateurs de la bourse Gabriel-Kucharski. Enfin, j'aimerais remercier mes collègues à l'Observatoire de linguistique Sens-Texte; je me compte chanceux de pouvoir travailler dans un environnement aussi stimulant.

Introduction

Les extracteurs de termes sont des outils informatiques conçus pour recenser les termes dans un corpus (un ensemble de textes assemblé à des fins spécifiques). L'ensemble des termes que l'on cherche à repérer dépend de l'application qui motive l'utilisateur à faire appel à un extracteur; il peut s'agir de la terminographie, de la traduction, de l'indexation, etc. Par exemple, dans le texte présenté dans la Figure 1, on pourrait souhaiter extraire seulement les termes de nature nominale de la mécanique automobile apparaissant dans leur forme habituelle (formes en caractères gras). En revanche, on pourrait aussi s'intéresser aux variantes de ces termes, qu'il s'agisse de réductions anaphoriques (formes soulignées) ou de réductions par coordination (forme en italiques).

Compression rings are designed to prevent leakage between the **piston** and the **cylinder**, Figure 2-31.

The idea is to create an internal stress within the ring. This stress will tend to cause the ring to twist in such a fashion that the lower edge of the ring is pressed against the **cylinder wall** on the **intake stroke**. This will cause the ring to act as a mild scraper. The scraping effect will assist in the removal of surplus **oil** that may have escaped the **oil control rings**, Figure 2-32.

On *compression* and **exhaust strokes**, the rings will tend to slip lightly over the **oil** film. This will prolong the life of the ring, Figure 2-33. On the **power stroke**, pressure of the burning gases will force the top edge of the ring downward. This causes the ring to rub the wall with full face contact and provides a good seal for the enormous pressure generated by the **power stroke**.

Figure 1 : Exemples de termes à extraire

L'extracteur de termes prend en entrée un corpus, et produit une liste de candidats termes, généralement triée en fonction de la probabilité qu'un candidat soit effectivement un terme (voir section 4.4, Figure 21).

Si les techniques d'extraction de termes ont atteint une certaine maturité, leur évaluation pose toujours problème, car il n'existe pas de protocole d'évaluation standard comme il y en a dans d'autres branches du traitement automatique de la langue.

Si on souhaite comparer différentes techniques d'extraction, ou voir comment des modifications faites à un système influent sur sa performance, cela présuppose l'existence d'une référence à laquelle on peut comparer la sortie des extracteurs, comme le soulignent Nazarenko et al. (2009). Or, à l'état actuel, nous savons peu de choses sur la construction d'une liste de référence pour l'évaluation des extracteurs. Doit-on utiliser une ressource terminologique existante (p. ex. une banque de terminologie)? ou doit-on la construire nous-même à partir du corpus qui servira pour l'évaluation? En l'occurrence, comment choisit-on les termes dans le corpus? Qui les choisit, et pour quelle application? Quelles sortes de termes devraient être inclus dans la liste de référence? Doit-on inclure les variantes terminologiques?

Ces questions sont au coeur de ce travail, qui vise la construction d'un corpus étalon permettant d'évaluer automatiquement les extracteurs de termes d'une manière quantitative et qualitative, objective et reproductible. Le corpus rendra compte de la variation terminologique et sera assorti d'annotations riches, permettant un paramétrage de la liste de référence.

Le mémoire est organisé de la façon suivante. Le premier chapitre portera sur certaines notions fondamentales liées à la terminologie et à l'extraction de termes. Dans le deuxième chapitre, nous recenserons les travaux qui ont touché à l'évaluation des extracteurs de termes. Le troisième chapitre portera sur la méthodologie d'annotation qui nous a permis de construire le corpus étalon. Enfin, nous validerons et analyserons nos résultats dans le quatrième chapitre, et montrerons comment exploiter le corpus étalon.

Dans la suite de ce travail, en ce qui concerne les indices typographiques, nous utilisons :

- les italiques pour souligner une *forme linguistique*;
- les guillemets pour souligner le « concept » dénoté par une forme linguistique;
- les caractères gras pour souligner quelques **termes importants** à la compréhension de ce mémoire.

1. L'extraction de termes

Ce chapitre porte sur des notions fondamentales liées à l'extraction de termes. Dans un premier temps, nous ferons un survol de quelques perspectives théoriques de la terminologie, ce qui nous amènera à adopter une optique principalement lexico-sémantique du travail terminologique. Puis, nous établirons une définition de la notion de « terme » adaptée à notre travail. Ensuite, nous présenterons l'extraction de termes, ainsi que les principales techniques mises en place pour réaliser cette tâche. Enfin, nous décrirons le concept de « variation terminologique », qui figure dans de nombreux travaux sur l'extraction de termes et qui occupera une place prépondérante dans notre méthodologie d'évaluation des extracteurs de termes.

1.1 Termes et terminologie

Il importe pour commencer de se faire une idée claire de ce qui constitue la terminologie et des principes théoriques qui sous-tendent la terminologie en tant que discipline. Une définition souvent citée de la terminologie nous est fournie par Sager : « Terminology is the study and the field of activity concerned with the collection, description and presentation of terms, i.e. lexical items belonging to specialised areas of usage of one or more languages » (Sager 1990 : 2). Cette définition nous donne une idée de ce qui constitue l'activité terminologique, mais il existe par ailleurs d'autres notions qui sont rattachées au terme *terminologie*. En effet, Rondeau nous offre trois définitions de la terminologie :

> 1. L'ensemble des termes d'un domaine, comme par exemple la biochimie, ou d'une discipline, comme par exemple la linguistique. On dira alors : la terminologie de la biochimie, de la linguistique, etc.
>
> 2. Les méthodes de collecte et de classement des termes, de création néologique, de normalisation des termes, de diffusion des termes : c'est ce à quoi s'emploient terminologues et terminographes.

> 3. Une science dont l'objet est d'ordre linguistique, mais qui est essentiellement pluri-disciplinaire et participe à la fois de la linguistique, de la logique, de l'ontologie, de la classologie et de l'informatique (Rondeau 1984 : 18).

En somme, le terme *terminologie* désigne à la fois une activité et le résultat de cette activité.

1.1.1 Deux optiques terminologiques

La théorie de la terminologie a été traversée de nombreux courants. Le premier modèle théorique, qu'on appelle la *théorie générale de la terminologie* (TGT), est associé à Eugen Wüster, ingénieur autrichien membre du Cercle de Vienne dont les travaux ont, dès le début des années 1930, contribué largement à façonner une approche que l'on peut appeler **optique conceptuelle**, qui a longtemps dominé le champ de la terminologie.

Le concept, dans cette optique, est défini de la façon suivante : « a unit of thought constituted through abstraction on the basis of properties common to a set of objects » (norme ISO 1087, cité dans Pearson 1998 : 15). L'existence de cette entité abstraite est posée a priori :

> The primary objects of terminology, the terms, are perceived as symbols which represent concepts. Concepts must therefore be created and come to existence before terms can be formed to represent them. In fact, the naming of a concept may be considered the first step in the consolidation of a concept as a socially useful or usable entity (Sager 1990 : 22).

Le concept est dénoté par un terme, dans un lien de référence qu'il convient de fixer, selon cette optique, par un travail de normalisation afin d'obtenir un vocabulaire spécialisé qui présente un minimum d'ambiguïté. La TGT est effectivement caractérisée par l'importance qu'elle accorde à la **biunivocité** du lien entre terme et concept : chaque terme ne doit dénoter qu'un concept, et chaque concept ne peut être désigné que par un seul terme.

L'optique conceptuelle repose sur un effort cognitif de représentation des concepts, qui débouche sur l'apposition d'étiquettes linguistiques aux concepts. Ce sont des méthodes dites **onomasiologiques** :

> Traditional terminological theory therefore identifies its approach as 'onomasiological', i.e. a 'naming' approach, because in principle it starts from concepts and looks for the names of these concepts. By contrast, the lexicographical approach is called 'semasiological', i.e. a 'meaning' approach, because it starts from words and looks for their meaning (Sager 1990 : 56).

Bien que les postulats de la TGT se conjuguent bien avec la normalisation industrielle, leur mise en pratique est problématique d'un point de vue linguistique. Les travaux de **terminographie** (la compilation de ressources telles que les dictionnaires spécialisés et les banques de terminologie) prennent fréquemment les corpus comme point de départ, et les terminologues praticiens privilégient généralement l'approche contraire, qu'on dit **sémasiologique**, utilisée notamment en lexicographie. Celle-ci part du texte pour appréhender le sens des termes, et prend en compte la dépendance contextuelle des termes. À ce sujet, Gaudin a affirmé que « nous ne nous intéresserons pas à ces entités obscures que sont les représentations mentales, individuelles, mais au langage grâce auquel nous les construisons » (Gaudin 2003 : 61).

De nombreuses difficultés engendrées par la TGT ont été soulevées : le fait que l'on n'observe pas systématiquement une biunivocité totale dans les discours scientifiques et techniques; la difficulté de délimiter les concepts et les domaines, de définir la notion d'expert, de structurer les connaissances, d'assurer la diffusion et l'implantation des terminologies normalisées officielles, de rendre compte de la genèse et de l'évolution des termes. Ces problèmes ont été abordés, à partir des années 1980, par de nouvelles approches théoriques, parmi lesquelles on peut nommer la terminologie textuelle, la théorie communicative de la terminologie, l'approche sociocognitive et la socioterminologie (L'Homme et al. 2003). Ces multiples définitions du cadre théorique de la terminologie ont

permis, entre autres, de rendre compte de phénomènes tels que la synonymie, la polysémie et la variation contextuelle (voir section 1.3).

Aux antipodes de l'optique conceptuelle, on trouve une optique qualifiée de **lexico-sémantique**. Dans cette optique, le terme est d'abord une unité lexicale, c'est-à-dire l'association d'une forme linguistique et d'un sens.

> La particularité du terme, par rapport aux autres unités lexicales d'une langue, est d'avoir un *sens spécialisé*, c'est-à-dire un sens qui peut être mis en rapport avec un domaine de spécialité. La définition du « terme », contrairement à celle qui est donnée pour d'autres unités linguistiques, est donc relative. Elle dépend de la délimitation qu'on a faite d'un domaine spécialisé. En outre, il n'est pas possible d'envisager la notion de *terme* en faisant abstraction des objectifs visés par une description terminographique (L'Homme 2004 : 33).

Dans l'optique lexico-sémantique, les sens sont dégagés en observant les termes en contexte, en portant une attention particulière à leur interaction avec d'autres termes et des unités de la langue générale. Cette approche est donc sémasiologique, car elle part de la forme pour dégager le sens, et s'aligne sur le travail concret des terminologues praticiens. En outre, elle s'intéresse aux relations qui existent entre les sens des termes, ces relations sémantiques pouvant être de type paradigmatique (synonymie, antonymie, méronymie, etc.) ou de type syntagmatique (en particulier, la combinatoire des termes). Ainsi, elle admet l'existence de plusieurs sens pour une même forme (**polysémie**) et l'existence de multiples formes pour dénoter un même concept (synonymie et variation contextuelle).

L'optique conceptuelle et l'optique lexico-sémantique sont fondamentalement différentes, mais les deux ont un rôle à jouer dans le travail terminologique : « La terminographie doit donc s'intéresser aux deux optiques. Elle peut avoir recours aux travaux d'une terminologie plus classique pour aborder le plan conceptuel, mais elle doit régulièrement se tourner vers la lexicologie et, plus particulièrement, vers la sémantique lexicale » (L'Homme 2004 : 38). Ce travail exploitera les deux optiques. Nous utiliserons

des techniques de repérage des termes propres à l'optique lexico-sémantique, et les descriptions terminologiques résultantes seront plus proches de cette optique, car celles-ci porteront, entre autres, sur les liens de variation terminologique et les relations paradigmatiques auxquelles participent les termes. En revanche, l'optique conceptuelle sera reflétée par l'accent qui sera placé sur les concepts concrets du domaine de la mécanique automobile, ce qui se traduira notamment par la sélection d'unités nominales seulement.

1.1.2 Le terme

Que l'on s'intéresse plutôt à sa dimension cognitive ou à sa dimension linguistique, l'objet central de la terminologie est bien sûr le terme. Or, comme il existe plusieurs courants théoriques en terminologie, il n'existe pas de consensus sur la définition de *terme*. On doit à Gaussier (2001 : 168) la citation : « There is no fully operational definition of terms ». Pour leur part, Kageura et al. (1999 : 417) soulignent que, à la différence d'autres domaines du TAL tels que la recherche documentaire, l'objet de la terminologie suscite toujours des débats, ce qui rend l'identification, et en particulier l'identification automatisée de ces termes une tâche difficile. Vivaldi & Rodríguez (2007 : 229) affirment qu'en pratique, il est difficile de définir exactement ce qui constitue un terme, et ajoutent que les concepts utilisés pour déterminer le statut terminologique présupposent un choix binaire : soit c'est un terme, soit ce ne l'est pas. Le repérage des termes pose donc des problèmes liés à cette distinction entre les mots et les termes, et il en résulte nécessairement des incertitudes : « In attempting to distinguish between terms and non-terms and to identify the boudaries of terms, we are dealing with continua rather than dichotomies, and so there are always bound to remain areas of uncertainty, even in the human interpretation of such data » (Ahmad & Rogers 1997 : 755).

Un survol de quelques définitions du terme montrera l'absence de consensus à ce sujet, et le peu d'aide que ces définitions fournissent à qui cherche à repérer les termes, que ce soit manuellement ou automatiquement.

Guilbert (1973) présente une analyse de la distinction entre les termes et les mots de la langue générale en ce qui concerne, entre autres, leur mode de signification :

> Les signes du lexique commun, d'une manière générale, sont porteurs de connotations psychologiques et sociales infiniment complexes exprimant la personnalité du locuteur et la spécificité de la communication. Les signes des vocabulaires scientifiques et techniques au contraire tendraient à être univoques. Mais cette façon particulière de signifier n'est pas inhérente à la forme signifiante elle-même, mais seulement à l'emploi qui en est fait par les locuteurs et à la référence impliquée (Guilbert 1973 : 6).

Le terme aurait ce mode de signification spécialisé lorsqu'il est employé par un spécialiste dans un mode de communication spécialisé. La notion de « terme » serait donc dépendante d'une situation d'usage spécifique (idée apparaissant d'ailleurs dans Cabré (2003)). De plus, les mots de la langue générale prendraient leur sens seulement lors de leur réalisation dans un énoncé; en revanche, « le terme technique se définit par rapport à l'usage qu'on fait de la chose, aux composants de l'objet, aux caractères perçus par les sens (forme, couleur, dimension), à la localisation géographique, ou par la référence à une taxinomie des êtres de la nature [...] » (Guilbert 1973 : 10). Le mode de signification des termes serait donc ancré dans la réalité et dans la perception de celle-ci. De plus, les termes tendraient à être monoréférentiels (voir section 1.1.3.2) dans un domaine donné, ne désignant qu'un concept, pour éviter toute ambigüité dans la communication spécialisée. La spécificité des termes résulterait donc de leur mode de signification axé sur la référence à la réalité, et de leur utilisation par des spécialistes dans un contexte de communication spécialisée.

Ce mode de signification, axé sur la référence, se manifesterait dans les ressources terminologiques par la prédominance des noms. Des noms, on dériverait, selon différents procédés de dérivation morphologique, des termes d'autres catégories grammaticales, tels que les adjectifs et les verbes.

Guilbert décrit également les différences entre termes et mots du point de vue statistique, affirmant que les formes les plus fréquentes, « dans une masse de vocabulaire indifférenciée », appartiennent à la langue générale, et que les formes rares auraient tendance à être des termes « parce qu'ils sont employés dans des situations de communication où n'interviennent que des spécialistes » (id. : 7).

Rey (1979), quant à lui, présente une définition du terme basée sur le concept de « (dé)nomination », soit la référence du terme à un concept. Les termes seraient « des unités dont la valeur est relative dans un ensemble, et dont la finalité est d'assumer une relation, en principe biunivoque, avec une notion, avec une classe logique (*i.e.*, notamment, fonctionnelle) d'objets de connaissance, par le processus de la nomination » (Rey 1979 : 78). Or, Rey affirme que le travail terminologique ne doit pas s'occuper seulement des concepts et de leur réalisation linguistique, mais doit tenir compte de différents facteurs d'ordre pragmatique : « De nombreuses décisions, arbitraires, dictées par des considérations pratique impérieuses (finalité de la description, conditions de sa réalisation, dimension souhaitée), doivent être prises avant tout travail terminologique » (id. : 88).

Étant donné que les syntagmes nominaux occupent une place prépondérante dans les ressources terminologiques, Rey porte son attention sur la nature des syntagmes à retenir. Faut-il décrire seulement les syntagmes dont l'usage est si courant qu'ils font incontestablement partie du lexique? Faut-il s'intéresser uniquement aux syntagmes **non compositionnels**, c'est-à-dire ceux dont le sens ne résulte pas de la combinaison des sens de leurs constituants? « Ce qui compte, c'est que l'unité corresponde dans l'usage des spécialistes à une forme généralement acceptée et comprise – ou proposée de l'être –, et surtout à une définition qui la fasse correspondre à une notion dans un domaine » (id. : 94). De plus, même les syntagmes très longs peuvent être des termes si on peut les mettre en opposition à d'autres concepts du domaine. On peut aussi choisir de les décomposer en unités plus petites, mais celles-ci doivent obéir au même critère de mise en opposition.

Comme Rey, Goffin définit le terme en fonction de sa référence à un concept unique et son utilisation dans une sphère d'activité spécialisée : « Le terme (technique) désigne l'emploi monosémique qui est fait d'une unité lexicale dans un vocabulaire spécial utilisé dans les limites d'une activité spécialisée. Il établit une relation univoque entre la réalité extralinguistique et le signe de par sa fonction de dénomination » (Goffin, dans OLF 1979 : 159).

Pour sa part, Dubuc (2002) définit deux types d'unités terminologiques, celles ayant une valeur conceptuelle et celles ayant une valeur fonctionnelle (des collocations ou expressions propres au domaine à l'étude). Le terme en tant qu'unité conceptuelle serait défini par le lien qui existe entre le concept dénoté par le terme et son domaine d'emploi, une notion qui revient souvent dans les différentes définitions du terme.

Ahmad et al. (1994) affirment que les définitions classiques du terme, qui focalisent sur le lien entre une forme linguistique et des concepts appartenant à un domaine de spécialité, négligent des facteurs qui doivent absolument être pris en considération pour savoir si tel terme doit être inclus dans la ressource terminologique qu'on cherche à construire :

> Various définitions of 'term' are available. The common denominator seems to be that a term is a label – usually lexical – in the special language of a specific domain, designating a particular concept in the knowledge of that domain, and arguably less context-dependent with regard to its sense than a general-language word. The problems of term identification are not, however, solved by such conventional definitions of 'term', since they also beg certain questions, including the issues of domain delimitation and of communicative purpose or level of the text[…] The question 'what is a term?' needs in our view to be relativised by two complementary questions: 'who is the terminology for and what is its purpose?' (Ahmad et al. 1994 : 269)

En effet, le choix des termes à inclure dans une ressource terminologique et le contenu des descriptions des termes doit être fonction du contexte d'utilisation et des utilisateurs. Un

traducteur s'intéressera, entre autres, aux propriétés linguistiques d'un terme ainsi qu'à ses équivalents, tandis qu'un expert du domaine n'aura pas forcément ce besoin; de même, une ressource créée pour des traducteurs contiendra un nombre plus grand d'entrées, et n'inclura pas uniquement des syntagmes nominaux terminologiques, mais des collocations, des éléments de phraséologie, des unités qui posent des problèmes de traduction, etc. Dans une étude portant sur le rôle de l'application en terminologie, Estopà (2001) a montré que différents utilisateurs produisaient différentes listes de termes, en fonction de leur activité professionnelle. Quatre groupes d'utilisateurs ont dépouillé un même texte, et le nombre ainsi que le type d'unités retenues par chaque groupe présentaient des différences importantes. Le terme ne doit donc pas être défini seulement en fonction des concepts d'un domaine, mais aussi en fonction de l'application qui motive la création d'une ressource terminologique.

Pearson (1998) fait un survol des différentes définitions du terme et affirme qu'aucune ne peut être exploitée afin d'effectuer le dépouillement terminologique (voir section 1.1.3). Elle explore différents concepts de la terminologie (catégories de termes, notion de « sous-langage ») et conclut qu'ils ne fournissent pas de moyen sûr ou objectif de décider si une unité se comporte comme un terme. Il serait plus fructueux selon Pearson de relever les situations dans lesquelles les termes vont probablement être utilisés, et de présumer, à défaut du contraire, que toutes les unités utilisées sont potentiellement terminologiques.

Les définitions du terme présentées ci-dessus ne sont donc pas facilement exploitables pour qui cherche à recenser les termes afin de construire une ressource terminologique. Elles ne fournissent pas de critères directement applicables à ce travail et permettant de vérifier le statut terminologique d'une unité lexicale, et la plupart négligent l'aspect fonctionnel du travail terminologique. Dans la section suivante, nous tenterons d'identifier des critères précis permettant de déterminer quelles unités sont à retenir dans un travail terminographique. Pour l'instant, nous retiendrons la définition du terme que

l'optique lexico-sémantique nous fournit : le **terme** est une unité lexicale, ayant un sens spécifique dans un domaine de spécialité donné et se manifestant dans les textes produits dans ce domaine. Le terme est composé d'une forme linguistique et d'un sens. Le sens est appréhendé en regard d'un domaine de spécialité clairement délimité, et se définit en fonction de ses relations avec d'autres termes du domaine. Nous définirons le **terme simple** comme un terme composé d'un seul mot graphique, et le **terme complexe** comme un terme qui est composé de plus d'un mot. Les termes complexes sont composés d'une **tête** et d'un **modificateur**. En français, la tête d'un terme complexe se trouve généralement à gauche, comme le mot *moteur* dans le terme *moteur à injection.* La tête et le modificateur peuvent eux-mêmes être modifiés : dans le terme *moteur à injection directe*, le mot *directe* vient modifier *injection*. Cette modification s'applique récursivement et peut engendrer des termes très longs, qui poseront des problèmes de découpage plus importants; nous traiterons des difficultés du découpage, ainsi que des critères de repérage des termes, dans la section suivante.

1.1.3 Critères de repérage des termes

Il faut d'abord reconnaître que le statut terminologique résulte d'une analyse et d'une prise de décision : « on peut voir le terme et le concept comme le résultat d'un processus d'analyse termino-conceptuelle. Un mot ou une unité complexe n'acquiert le statut de terme que par décision » (Bourigault et al. 2004 : 93). Il importe donc d'établir des critères afin de baliser ce processus. Différents critères ont été proposés pour baliser la sélection des termes dans un corpus, un travail qu'on appelle le **dépouillement terminologique**. Le dépouillement terminologique consiste à repérer les occurrences des termes dans un corpus et de délimiter ces termes en identifiant les constituants qui en font partie. On appellera ces étapes **repérage** et **découpage** respectivement.

Rondeau (1984 : 78) distingue cing types d'unités rencontrées lors du dépouillement terminologique :

1. des mots simples appartenant à la langue commune (p. ex. *beau*);
2. des expressions syntagmatiques appartenant à la langue commune (p. ex. *prendre part*);
3. des groupements syntagmatiques de discours, appartenant aux langages spécialisés (p. ex. en chimie, *sublimation d'un corps* contiendrait en fait deux termes, qu'on retrouve fréquemment ensemble);
4. des groupements syntaxiques lexicalisés (et monoréférentiels), appartenant aux langages spécialisés (p. ex. *corps céleste* en astronomie);
5. des termes simples (p. ex. *piston* en mécanique automobile).

En ce qui concerne les unités syntagmatiques (catégories 2 à 4), elles posent des difficultés de repérage, comme les deux autres catégories, mais aussi des difficultés de découpage. Il importe donc d'énoncer des critères pouvant guider le repérage des termes et faciliter leur découpage. Nous décrirons dans les sections suivantes quelques travaux qui ont proposé de tels critères. Ces travaux seront présentés dans un ordre qui facilite l'enchaînement des concepts plutôt que dans l'ordre chronologique.

1.1.3.1 Kageura & Umino (1996)

Deux concepts importants pour déterminer le statut terminologique sont ceux de potentiel terminologique (*termhood*) et de figement syntaxique (*unithood*), qui permettent d'aborder les difficultés posées par le repérage et le découpage des termes :

> "Unithood" refers to the degree of strength or stability of syntagmatic combinations or collocations. Thus the concept "unithood" is not only relevant to simple and complex terms, but potentially to other complex units as grammatical collocations or idiomatic expressions. On the other hand, "termhood" refers to the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts. Thus "termhood" is not only relevant to complex linguistic units, but also simple units (Kageura & Umino 1996 : 260-261).

Le figement signifie essentiellement la cohésion : c'est le degré de stabilité d'un syntagme; cette notion s'apparente à celle de *lexicalisation,* qui reviendra à plusieurs

reprises. Le potentiel terminologique désigne la force du lien entre une forme donnée et des concepts propres au domaine. Ces deux critères peuvent servir à déterminer le statut terminologique, c'est-à-dire à répondre à la question : « est-ce que cette unité est un terme? »

Les critères de figement et de potentiel terminologique ont été exploités par des concepteurs d'extracteurs de terminologie, notamment par Vivaldi & Rodríguez (2007), qui évaluent trois critères afin de déterminer si une unité correspond à un terme : le figement, le potentiel terminologique (qu'ils définissent comme l'utilisation d'une unité dans un ou plusieurs domaines de spécialité) et « l'usage spécialisé » (l'écart entre l'utilisation dans la langue générale et les domaines de spécialité).

1.1.3.2 Auger (1979)

Pour Auger, le critère le plus important pour déterminer le statut terminologique est la **monoréférentialité**, c'est-à-dire la référence à un concept unique et doté d'une certaine permanence; des critères basés sur la forme des unités seraient insuffisants à cette tâche : « Il n'existe pas de critères formels décisifs permettant de distinguer le syntagme lexical (ou lexicalisé) du syntagme nominal (de discours) formé selon les circonstances du discours et donc de façon purement accidentelle » (Auger, dans OLF 1979 : 17). La monoréférentialité demeurerait donc le meilleur critère, mais celle-ci serait plus difficile à évaluer dans le cas des syntagmes qui possèdent plusieurs modificateurs. « Le terminologue risque toutefois de rester perplexe devant des exemples de syntagmes hyperdéveloppés comme celui de la *charrue pour labour à plat à traction animale sans avant-train* dans lequel ni la cohésion syntagmatique, ni la cohérence sémantique ne sont évidentes » (Auger, dans OLF 1979 : 24). Auger propose qu'un tel syntagme serait plutôt une unité de catalogage ou un élément de nomenclature plutôt qu'un terme à retenir dans une terminologie, assimilant ce genre d'unités à une paraphrase ou à une définition. Il propose un critère additionnel, basé sur la fréquence, afin de déterminer si de telles unités sont à

retenir. On ne retiendrait un tel syntagme que si on en retrouve au moins deux attestations, dans des documents distincts :

> [...] le terminologue dispose au moins dans la pratique de deux critères qui s'avèrent assez discriminants, nous l'avons vu. Le premier est celui de la relation univoque entre le terme et la notion qu'il désigne par référence à un objet unique et bien identifié, le second critère se fonde sur la fréquence du terme dans un corpus donné et permet de prime abord d'éliminer les groupes accidentels de mots qui n'ont pas de caractère de lexicalisation (id. : 25).

À propos de cette notion voulant que les syntagmes « hyperdéveloppés » soient plutôt descriptifs que dénominatifs, au point de correspondre plutôt à des définitions qu'à des termes, Sager répond que les concepts ont des réalisations linguistiques différentes selon le type de texte ou de communication : « In spoken language shorter forms will be used, in written texts longer forms may occur [...] I therefore see no justification for attributing extended terms to an artificial language. They may be communicatively effective in a catalogue, and ineffective in other types of texts » (Sager dans OLF 1979 : 50). Il n'y aurait donc aucune raison d'exclure a priori de tels syntagmes lors du dépouillement terminologique, même si leur fréquence est basse.

1.1.3.3 Dubuc (2002)

Pour Dubuc, deux critères sont particulièrement importants pour le repérage et le découpage des termes : la situation de travail et la relation entre la tête et le modificateur. Premièrement, l'activité terminologique est ancrée dans une situation concrète, et le repérage des termes doit être fait en fonction des besoins des usagers éventuels. Si l'approche de Dubuc semble tenir compte de l'aspect fonctionnel du travail terminologique, elle est d'abord conceptuelle et onomasiologique, faisant appel à la construction d'un arbre du domaine afin de bien cerner les notions propres à ce domaine, et ainsi éliminer le bruit.

Deuxièmement, la relation entre la tête (ici appelée le *déterminé*) et le modificateur (*déterminant*) jouerait un rôle important dans le découpage des termes, ce qui amène Dubuc à établir une distinction entre déterminant essentiel et accidentel :

> Les déterminants accidentels sont ceux qui ne modifient pas le sens du déterminé; tout au plus y apportent-ils des modifications d'aspect ou de circonstance. Ainsi, les *encodeurs* ou *encodeuses*, en informatique, n'effectuent pas un travail essentiellement différent selon le support utilisé : cassettes, disques ou bandes. Seul le facteur support se trouve modifié. On n'englobera pas dans l'unité terminologique les compléments *cassettes*, *disques* ou *bandes*. [...] C'est le contraire qui se produit avec les déterminants essentiels. Le déterminant modifie la nature du déterminé ou implique avec lui une relation si étroite qu'en la supprimant, on change le sens du déterminé. Prenons comme exemple l'expression *calculateur universel,* en anglais *general purpose computer*, le déterminant n'a pas qu'une portée qualificative, mais il désigne un type précis de calculateur, qui se distingue des autres : analogique, spécialisé, etc. (Dubuc 2002 : 58)

Dubuc affirme que c'est l'analyse de la cohésion entre la tête et le modificateur qui permet de bien découper les termes. Il propose 4 indices permettant de décider si un déterminant est essentiel ou accidentel :

- Le degré de lexicalisation, qui peut être indiqué par la présence ou l'absence d'un article devant le déterminant (*chef de projet* et *chef du projet*).
- La fonction de classement du déterminant, qui permet éventuellement de mettre en opposition des notions du domaine (*hourly paid job* et *salary-paid job*).
- La cooccurrence des unités : une cooccurrence fréquente dans le domaine concerné peut indiquer une forte lexicalisation.
- Les artifices typographiques tels que les caractères gras, les italiques, les guillemets et les soulignements.

1.1.3.4 Rondeau (1984)

Rondeau, pour sa part, énonce certaines caractéristiques des langues de spécialité, sur le plan textuel et sur le plan lexical. Les caractéristiques lexicales se rapportent directement aux termes, et nous aident éventuellement à identifier les termes dans les

discours spécialisés. Ces caractéristiques portent tantôt sur la forme linguistique, tantôt sur le sens. Pour ce qui est des sens, on évaluerait la place qu'occupe le concept dénoté par le terme dans le système conceptuel d'un domaine, la précision du concept, la monoréférentialité et « des rapports d'affinité avec certains mots de la langue commune » (Rondeau 1984 : 30). Quant à cette dernière caractéristique, il s'agit de regarder la combinatoire des termes, qui se combineront de façon privilégiée avec certains mots de la langue générale (p. ex. les verbes *représenter*, *former*, *appartenir*, etc.); ces rapports d'affinité sont parfois englobés sous le terme *collocation*.

En ce qui concerne les formes linguistiques, Rondeau note premièrement la tendance à la concision, phénomène par lequel les concepts sont dénotés par des formes linguistiques de plus en plus concises, par des mécanismes tels que la réduction, les sigles et les symboles. De plus, certaines langues de spécialité font appel à la dérivation morphologique de façon particulière : on fait plus fréquemment appel aux racines de langues anciennes et à l'affixation. En outre, l'aspect graphique (orthographe, utilisation du trait d'union, etc.) des termes serait fixé plus rigidement que dans la langue générale.

Rondeau présente enfin une distinction entre les groupements syntagmatiques terminologiques et non terminologiques. La distinction porte sur le degré de lexicalisation des syntagmes, qui se caractérise par une « cohésion sémantique très forte et permanente ». Dans ce cas, la suppression de la tête ou du modificateur détruit le lien entre la forme linguistique et son sens. Ce sont donc des syntagmes dits *non compositionnels*, dont le sens n'équivaut pas à la somme des sens de leurs composantes. Les syntagmes terminologiques devraient par ailleurs être attestés dans d'autres écrits traitant du même sujet. Rondeau énumère des critères qui viennent compléter ces deux critères principaux, qui seraient

cependant suffisants pour déterminer si un syntagme est non compositionnel, donc suffisamment lexicalisé pour être considéré terminologique[1] :

1. Absence d'article devant le prédicat. Exemples : *roulement à billes*; *taquet de talon*;
2. Extension adjectivale par la gauche. Exemple : *haute tension*;
3. Absence de charnière entre le sujet et le prédicat. Exemple : *tricot côte anglaise*;
4. Impossibilité d'insérer un élément adjectival ou prédicatif entre les différentes composantes du syntagme. Exemples : *machine de texturation sur arête*, *point couvrant à trois (quatre) aiguilles*;
5. Prédicat multiple. Exemples : *point de chaînette deux fils*, *poly-éthylène-glycol-téréphtalate*;
6. Représentation, dans une autre langue, de la même notion au moyen d'un terme simple. Exemple : En *finger stop* / Fr *butée* (télécommunications) (Rondeau 1984 : 80).

Ces critères sont certainement intéressants, bien qu'ils ne s'appliquent pas à toutes les langues[2], et en effet, l'analyse de Rondeau se démarque par la quantité de critères proposés pour le dépouillement terminologique, mais l'utilité de ces critères peut être remise en question. Les critères portant sur le sens des termes, tels que la monoréférentialité et la précision, sont difficiles à mettre en oeuvre concrètement; ceux qui portent sur la forme des termes s'appliquent souvent à certains domaines seulement, et sont plutôt des tendances observées dans les langues de spécialité que des critères de dépouillement. Enfin, le critère de non-compositionnalité exclurait bon nombre d'unités lexicales observées dans les textes spécialisés qui sont compositionnelles mais présentent éventuellement un intérêt pour l'utilisateur de la ressource terminologique.

---

[1] Rondeau fait une utilisation particulière des mots *sujet* et *prédicat*. *Sujet* désigne ici la tête d'un terme complexe et *prédicat*, le modificateur.

[2] Par exemple, le critère 2 ne s'applique pas à l'anglais, qui n'admet pas d'adjectif après le nom.

1.1.3.5 Sager (1990)

Sager affirme que la terminologie d'un domaine respecte davantage que la langue générale certaines règles de formation des unités lexicales : « Unlike in general language, where the arbitrariness of the sign is accepted, special languages strive to systematise principles of designation and to name concepts according to pre-specified rules or general principles » (Sager 1990 : 57). En revanche, l'ambiguïté est un phénomène très présent en langue générale, et la création de nouveaux mots est chose plutôt rare. Picht & Draskau (1985) abondent dans le même sens, affirmant que les termes se forment d'une façon régulière qui dépend des autres termes utilisés dans une langue de spécialité. De plus, Sager énonce des critères que les termes devraient idéalement satisfaire, qu'il s'agisse de nouveaux termes ou de termes existants. Ces critères peuvent éventuellement servir au dépouillement terminologique. Love (2000) s'appuie sur ces critères de formation des termes, qu'elle considère utiles, dans une certaine mesure[3], au repérage des termes :

> It constitutes a quasi-checklist of possible terminological status indicators for questionable terms. Sager's list was only used as a guide and not followed stringently, however, because, as he conceded himself, it could only be fully applied in an optimized setting. Unfortunately, most terms are not coined under such favourable circumstances (Love 2000 : 19).

Les critères de Sager (1990 : 89) sont les suivants :

1. The term must relate directly to the concept. It must express the concept clearly. A logical construction is advisable.
2. The term must be lexically systematic. It must follow an existing lexical pattern and if the words are of foreign origin, a uniform transcription must be preserved.
3. The term must conform to the general rules of word-formation and of the language which will also dictate the word order in compounds and phrases.
4. Term should be capable of providing derivatives.
5. Terms should not be pleonastic (i.e. no redundant repetition, e.g. combining a foreign word with a native word having the same meaning.)

---

[3] Voir également section 2.2.2.11.

6. Without sacrificing precision, terms should be concise and not contain unnecessary information.
7. There should be no synonyms whether absolute, relative or apparent.
8. Terms should not have morphological variants.
9. Terms should not have homonyms.
10. Terms should be monosemic.
11. The content of terms should be precise and not overlap in meaning with other terms.
12. The meaning of the term should be independent of context.

Il n'est pas aisé de voir comment ces critères peuvent aider au repérage des termes. Une fois un terme identifié, ils permettent de vérifier sa conformité à une certaine *politique* de dénomination. Mais lorsqu'il s'agit de dépouiller un texte et de repérer les termes, ces critères nous semblent très peu utiles. En effet, Sager analyse en profondeur les mécanismes de formation de nouveaux termes, et propose des critères à respecter dans la création de nouveaux termes, mais rien n'indique que ces concepts devraient aussi être utilisés pour repérer des termes existants. Sager lui-même n'affirme pas explicitement, à notre connaissance, que les mécanismes de formation des termes devraient être utilisés pour le repérage des termes, mais il reconnaît la difficulté que pose le repérage : « In practice terminologists face difficulties with the recognition of terminological units in running text, which can generally only be resolved by general or special subject knowledge » (id. : 61).

1.1.3.6 Pearson (1998)

Pearson (1998) fait un survol de différentes définitions du terme et conclut qu'elles n'offrent aucun moyen de faire la distinction entre un terme et un mot ou un syntagme non terminologique :

> For traditional terminologists, the notion of term can apply to lexical items with special reference in a restricted subject field (Sager); it can be the label or linguistic symbol for a concept (ISO, Felber); it is the equivalent of de Saussure's linguistic sign, i.e. the combination of signifiant and signifié (Rondeau). Distinctions are made between technical terms which are used in

> a single subject field and general terms which are used in more than one subject field. Distinctions are also made between terms whereby the meaning (underlying concept) of terms is agreed, and therefore protected, and words where the meaning is not protected. […] However, it is difficult to imagine how the definition of terms as offered by those who subscribe to a theory of terminology can be applied in practice. For example, il would not be possible to use the criteria proposed by any of the authors discussed to decide on whether a lexical item in a text is being used as part of general vocabulary or as a term (Pearson 1998 : 15).

En effet, Sager (1990) établit une distinction entre les termes qui ont une référence spécialisée (« special reference ») et les termes à référence générale, qui n'ont pas un sens qui est propre à un domaine de spécialité. Or, comme nous l'avons vu, ni la définition proposée par Sager, ni son analyse de la formation des termes et de la régularité des mécanismes sous-jacents ne permet de distinguer un terme d'un non-terme. Pearson explique également que les théories classiques de la terminologie supposent des délimitations claires entre les domaines, ayant chacune leur terminologie propre, et excluent les termes qui s'utilisent dans plus d'un domaine. Or, cette délimitation entre les domaines n'existe pas toujours, bien qu'elle puisse exister dans le cas des sciences pures, et le fait qu'une unité soit utilisée dans plus d'un domaine ne suffit pas à l'exclure du dépouillement terminologique.

Pearson affirme qu'un moyen de baliser le repérage des termes est fourni par le contexte communicationnel, qui permet notamment de décider si une unité qui n'est pas utilisée uniquement dans un domaine de spécialité constitue tout de même un terme. Si une unité lexicale est utilisée dans la langue générale et a éventuellement un sens spécifique dans un domaine de spécialité, ce serait le contexte communicationnel qui nous permettrait de décider si une de ses occurrences a un statut terminologique. « It will be suggested here that terms can only be considered as terms when they are used in certain contexts and that of the discussion about whether or not a term is really a term is irrelevant if the discussion is not rooted in reality » (id. : 36). En somme, étant donné la nature spécialisée d'un corpus, on peut présumer qu'une unité lexicale, même si elle est utilisée dans un autre domaine ou

dans la langue générale, peut être considérée un terme si elle a un référent précis dans le domaine qui nous intéresse.

L'approche consiste donc à présumer, étant donné le niveau de spécialisation d'un corpus, que toute unité lexicale est potentiellement terminologique, et à filtrer à l'aide de critères de sélection. Pearson en propose deux, qui sont mis à profit dans un système d'extraction semi-automatique des termes[4]. Premièrement, des candidats sont extraits automatiquement à l'aide de patrons morphosyntaxiques (cette stratégie, souvent utilisée en extraction automatique des termes, sera décrite à la section 1.2.1.1), mais ceux-ci sont identifiés manuellement pour chaque corpus traité, car Pearson ne croit pas qu'il existe un ensemble générique de patrons capable d'identifier tous les termes dans tous les corpus ou domaines. Puis, elle filtre la sortie de l'extraction en s'appuyant sur deux critères : premièrement, un terme qui contient toujours un article devant le modificateur n'aurait pas une « référence générique » et serait exclu; deuxièmement, seuls les termes qui cooccurrent au moins une fois avec un marqueur linguistique identifié au préalable peut constituer un terme. Les marqueurs utilisés sont au nombre de 5 (*called*, *known as*, *e.g.*, *the term* et les guillemets), mais cette liste n'est pas exhaustive, et il existerait d'autres types de marqueurs linguistiques permettant de repérer les termes :

> There were other signals which could also be used as part of the refinement process. For example, if a term candidate with generic reference appears at the end of a sentence and is immediately followed at the beginning of the following sentence by phrases such as This process, This method, This device, it is very likely that the term candidate is indeed a term (id. : 134).

Or, les patrons morphosyntaxiques, les marqueurs linguistiques et la présence d'un article devant le modificateur ne sont pas des critères suffisants pour repérer tous les termes

---

[4] Il est à noter que la thèse de Pearson porte particulièrement sur la définition terminologique, donc le genre de termes recherchés, ainsi que la stratégie utilisée, est fonction de cet objectif.

dans un corpus. Nous présentons dans la section suivante un ensemble de critères plus adéquat à cette tâche.

1.1.3.7 L'Homme (2004)

Selon l'optique lexico-sémantique, le terme n'est pas une simple étiquette apposée à un concept; le terme est une unité lexicale à deux faces, la forme et le sens, qui ne peuvent être séparées. Le terme est ancré dans le discours, et ne peut être étudié en réfléchissant seulement à la structuration des concepts d'un domaine, sans égard à sa place dans le discours. « Nous retenons d'abord l'idée de la terminologie textuelle selon laquelle le terme est un construit et qu'il est défini, entre autres, en fonction de la place qu'il occupe dans un corpus [...] » (L'Homme 2005 : 1123).

Par ailleurs, l'idée d'extraire automatiquement les termes à partir des textes s'inscrit dans une conception de la terminologie qui est plutôt textuelle que conceptuelle :

> Peut-on encore considérer qu'une description figée des connaissances par domaine puisse rendre compte de leur utilisation dans les textes ? Face à cette difficulté, des outils et méthodes de construction de terminologies à partir des textes sont apparus depuis quelques années, préfigurant une terminologie textuelle, plus proche des textes et donc des usages des termes. La profusion de textes spécialisés numérisés permet l'acquisition automatique de terminologie où le corpus prend alors une place centrale et où le terme est considéré dans toute sa dimension linguistique en situation (Chiao & Sta 2002 : 114).

L'Homme (2004) propose quatre critères lexico-sémantiques afin de déterminer si une unité lexicale constitue un terme. Le premier critère stipule qu'une unité lexicale correspond à un terme si sons sens est « lié à un domaine de spécialité [...] délimité au préalable pour un projet terminographique donné » (L'Homme 2004 : 64). Le deuxième critère concerne les unités prédicatives et stipule qu'elles correspondent à des termes si leurs actants sémantiques sont eux-mêmes des termes : en informatique, le verbe *adresser* admet deux actants, qui pourraient être réalisés p. ex. par *mémoire* et *système*

d'exploitation, qui sont eux-mêmes des termes en vertu du premier critère; *adresser* pourrait donc être considéré comme une unité terminologique. Selon le troisième critère, une unité qui présente un lien de dérivation morphologique avec un terme préalablement retenu correspond à un terme s'il existe également un lien sur le plan sémantique; ainsi, le terme *compilateur* donne lieu à toute une série de dérivés morphologiques qui pourraient être considérés comme des termes : *compiler*, *compilation*, *recompiler*, *compilable*, etc. Enfin, toute unité partageant un lien paradigmatique avec un terme risque fort d'en être un elle-même. Ainsi, la dénomination d'une partie dont le tout est dénoté par un terme qui a déjà été retenu correspond à un terme : si on retient le terme *interface* en vertu du premier critère, il faudra aussi retenir *menu* et *fenêtre*. Cette relation fait partie d'une classe de relations sémantiques qu'on appelle *méronymiques*; les relations paradigmatiques comprennent également la synonymie et l'antonymie.

Le premier critère est crucial pour déterminer si une unité correspond à un terme. Le sens de l'unité doit être lié au domaine de spécialité choisi, mais cela n'implique pas qu'il ne peut pas avoir un sens spécialisé dans d'autres domaines. Pearson abonde dans ce sens, et appelle *subtechnical terms* les unités qui caractérisent les langues de spécialité sans être rattachées à un domaine particulier[5] :

> subtechnical terms [are] words which have special reference but which are used in more than one subject domain. These include words such as factor, result, accuracy. To claim that such words are precluded from being classified as terms is to distort the composition of the lexicon of a special subject field (Pearson 1998 : 13).

L'Homme affirme que le fait de considérer que le terme est d'abord une unité lexicale a comme conséquence de retenir surtout des termes simples, et de ne retenir les termes complexes que si leur sens n'est pas compositionnel. Mais les critères nous

---

[5] Ces unités ont intéressé de nombreux chercheurs. Mentionnons le *vocabulaire général d'orientation scientifique* de Phal (1971), la *academic word list* de Coxhead (2000) et le *lexique scientifique transdisciplinaire* de Drouin (2007).

semblent tout aussi utiles pour le repérage des termes si on part du principe que toutes les unités lexicales spécialisées, même celles dont le sens est compositionnel, doivent être décrites. Un travail comme celui-ci exige que nous procédions de cette façon.

1.1.3.8 Synthèse

Ce survol des critères proposés pour le repérage et le découpage des termes mettent en lumière la diversité des approches à ce problème, et l'absence d'un cadre unificateur pour ces tâches. Dans le cadre de ce travail, nous retiendrons les critères lexico-sémantiques proposés par L'Homme (à l'exception de celui qui concerne les unités prédicatives, puisque nous nous intéressons seulement aux termes nominaux, et en particulier ceux qui dénotent des entités). Nous retiendrons également l'idée que la fonction de la ressource terminologique à construire, et les besoins de ses usagers, sont déterminants quant au repérage des termes et à leur description.

1.2 Les extracteurs de termes

Les connaissances évoluent de plus en plus rapidement, et les textes utilisés pour véhiculer ces connaissances font de même; cette évolution s'accompagne d'un renouvellement constant de la terminologie des domaines de spécialité. Rondeau (1984) donne plusieurs raisons pour la profusion croissante des termes spécialisés, notamment l'avancement des sciences et des techniques; le développement des médias, des rapports politiques internationaux, du commerce international; l'essor des multinationales, l'avènement des standards et de la normalisation et l'intervention de l'état dans les questions linguistiques. Étant donné la profusion sans cesse des textes et des nouvelles idées, une assistance automatique dans le travail terminologique s'avère indispensable : « Le cycle d'acquisition terminologique entièrement manuel est beaucoup trop long pour répondre aux besoins terminologiques de la documentation d'aujourd'hui » (Chiao & Sta 2002 : 113-114).

L'**extraction automatique de termes** vise à recenser tous les termes contenus dans un corpus, la nature des termes recherchés étant dépendante d'une application (terminographie, traduction, indexation, etc.). Elle peut être monolingue ou bilingue; l'extraction bilingue cherche à identifier non seulement des termes, mais aussi leurs équivalents dans une ou plusieurs autres langues. Dans ce travail, nous nous intéressons uniquement à la dimension monolingue de l'extraction. De plus, certains extracteurs de termes intègrent d'autres fonctionnalités : certains identifient des relations sémantiques entre les termes, par exemple. Ces autres fonctionnalités ne seront pas abordées dans ce travail, seulement l'extraction monolingue des termes contenus dans un corpus.

En plus de fournir aux terminologues et autres utilisateurs une assistance indispensable dans le dépouillement terminologique, les extracteurs de termes présentent de nombreux avantages. D'abord, l'outil peut traiter une quantité importante de texte en très peu de temps, permettant à l'utilisateur de parcourir une liste triée de candidats termes plutôt qu'une masse de documents. Il permet également d'observer des phénomènes qu'il serait difficile d'observer sans l'aide d'un outil automatique : « L'outil automatique est nécessaire là où l'extraction manuelle ne permet pas de mettre en valeur la diversité des usages que l'on peut trouver dans différents corpus de grande taille et la variation terminologique qui y est perceptible » (Calberg Challot et al. 2008 : 196). Ces avantages, la possibilité de traiter des quantités importantes de texte rapidement et d'observer de nombreux phénomènes terminologiques, se manifestent peu importe la nature du travail qui motive un utilisateur à faire appel à ce genre d'outils, qu'il s'agisse de la traduction, la terminologie, la lexicographie ou l'indexation.

L'extraction de termes est fortement reliée à l'extraction automatique de mots-clés pour l'indexation de documents, une tâche qui intéresse les chercheurs en recherche documentaire depuis les années 1950. Les deux tâches sont apparentées notamment en ce qui concerne certains moyens statistiques utilisés pour identifier les unités linguistiques

pertinentes (candidats termes et mots-clés respectivement) et les moyens utilisés pour évaluer les résultats (des métriques telles que la précision et le rappel).

Les travaux sur l'extraction de termes ont débuté au début des années 1990, leur genèse étant liée aux travaux en linguistique de corpus (Nazarenko et al. 2009 : 258). De nombreuses recherches ont été entreprises afin de concevoir des techniques d'extraction. Cette effervescence semble avoir diminué depuis quelques années, et on affirme souvent que les extracteurs ont atteint un certain niveau de maturité, bien qu'il soit difficile de le démontrer, comme nous le verrons au chapitre 2.

> Sur le plan de la recherche, le domaine est moins actif qu'il ne l'a été au cours des années 1990 comme si le problème de l'acquisition terminologique était résolu ou, du moins, comme si un palier de performance avait été atteint. Il est d'autant plus important d'établir une cartographie des méthodes et techniques proposées pour déterminer s'il reste des marges de progression et mettre en correspondance méthodes et types de besoins. Un effort collectif d'évaluation devrait permettre à terme de mieux valoriser les résultats de la terminologie computationnelle (Nazarenko et al. 2009 : 260-261).

En plus d'être utilisés seuls, les extracteurs de termes s'intègrent fréquemment dans d'autres outils du traitement automatique de la langue :

> Automatic term recognition in particular is much needed because a simple but coherently built terminology is the starting point of many applications such as human or machine translation, indexing, thesaurus construction, knowledge organisation, etc., and because manual efforts cannot keep up with the rapid growth of technical terms (Kageura & Umino 1996 : 259-260).

Les applications, manuelles ou automatisées, qui profitent de l'extraction de termes sont nombreuses. Parmi celles-ci, on peut nommer la traduction, la rédaction technique, la localisation, la construction de thesaurus pour l'indexation et la recherche d'information, la construction d'index pour les documents techniques, la veille technologique, la construction

d'ontologies pour le web sémantique, le résumé automatique et la recherche d'information multilingue (Mustafa El Hadi & Chaudiron 2007 : 170).

Étant donné les difficultés que posent le repérage et le découpage des termes, même pour un terminologue humain (voir section 1.1), les listes produites par un extracteur doivent toujours être validées par l'utilisateur; ainsi appelle-t-on les unités recensées par un extracteur des **candidats termes.** L'une des raisons qui explique la difficulté de l'extraction est que, sur le plan formel, les termes ne se distinguent pas des mots de la langue générale; dans les deux cas, on traite des chaînes de caractères. En outre, comme nous l'avons mentionné, la définition de ce qui constitue un terme ne fait pas consensus, et dépend de l'application envisagée (voir section 1.1.2). De plus, les termes ne se réalisent pas toujours de la même façon, et différents phénomènes de variation terminologique (voir section 1.3) modifient la façon dont les termes se manifestent dans le discours. « Toute la difficulté de ces techniques d'extraction est liée à la grande variabilité du terme en corpus. Tout au plus peuvent-elles déterminer un surensemble de termes, composé d'expressions qui, faute de statut précis, sont appelées candidats termes » (Chiao & Sta 2002 : 115).

1.2.1 Stratégies utilisées

L'Homme affirme qu'un certain nombre de présupposés sous-tendent l'extraction de termes, à savoir :

> a) Les textes spécialisés comportent beaucoup de termes qui servent de véhicules privilégiés des connaissances spécialisées;
>
> b) Un terme significatif sera utilisé à plusieurs reprises dans un texte spécialisé;
>
> c) La très grande majorité des termes sont de nature nominale;
>
> d) La plupart de ces termes sont complexes, c'est-à-dire qu'il [sic] sont composés de plusieurs mots par ailleurs utilisés isolément (ex. pression artérielle; intelligence artificielle, aigle à tête blanche)

> e) Les termes complexes se construisent au moyen d'un nombre fini de séquences de catégories grammaticales (L'Homme 2002 : 13).

Ces présupposés sont reflétés dans les stratégies mises en place afin de repérer automatiquement les termes. Il existe deux catégories principales de stratégies, et les extracteurs privilégient généralement l'une ou l'autre, bien que la plupart des extracteurs font appel aux deux types, ces extracteurs étant parfois qualifiés d'*hybrides*.

1.2.1.1 Les techniques linguistiques

La première catégorie comprend les techniques linguistiques, qui reposent sur le degré de figement syntaxique des termes, ou sur certains aspects de leur structure interne, tels que les mécanismes de formation des termes ou leur composition morphosyntaxique. On exploite notamment le fait que la majorité des termes sont des syntagmes nominaux complexes correspondant à un nombre restreint de patrons morphosyntaxiques. Par exemple, pour le français, il s'agit généralement d'un nom modifié par un autre nom, un adjectif, un syntagme prépositionnel, ou une combinaison de ceux-ci. Nous décrirons ci-dessous quelques techniques linguistiques utilisées pour l'extraction de termes.

Une technique linguistique répandue est la recherche de patrons morphosyntaxiques. Cette technique est basée sur l'observation que la grande majorité des termes peuvent être décrits par un petit nombre de séquences de parties du discours. Cette technique a été appliquée de différentes façons. Par exemple, Daille et al. (1994) supposent que la majorité des termes sont des syntagmes nominaux composés de deux mots pleins (p. ex. nom+préposition+nom pour le français). Un automate identifie les séquences de parties du discours pertinentes, puis des filtres statistiques sont appliqués pour faire ressortir les candidats termes les plus intéressants. Justeson & Katz (1995) proposent un extracteur qui exploite deux des caractéristiques (supposées) des termes, à savoir que leur forme est fixe et ne varie pas en contexte, et qu'ils sont construits sur un nombre fixe de patrons morphosyntaxiques correspondant à des syntagmes nominaux. Ils utilisent donc

l'étiquetage morphosyntaxique afin d'extraire des séquences correspondant à ces patrons, qu'ils filtrent à l'aide d'un seuil de fréquence.

D'autres indices linguistiques ont été exploités en extraction de termes. Bourigault (1992) identifie des unités qui ne font jamais ou presque jamais partie d'un terme (les conjonctions ou les signes de ponctuation, par exemple). Une liste de ces *frontières de termes* est dressée, puis toute séquence contenue dans le texte et se trouvant entre des frontières de termes est considérée un candidat terme. D'autres chercheurs se sont intéressés à la formation des termes et aux mécanismes linguistiques qui sont en jeu lorsque des nouveaux termes sont créés, et ont tenté d'exploiter cette information pour repérer automatiquement les termes.

Les techniques linguistiques font généralement appel à l'analyse morphosyntaxique. Or, l'utilisation de l'analyse morphosyntaxique en extraction de termes présente des limites :

> [...] freely formed phrases often have the same syntactic structure as complex terms. Hence syntactic analysis can at best only recognize promising candidate terms. To proceed further with the separation of complex terms from freely formed phrases, either analysis at a higher (i.e., semantic or pragmatic) level or statistical evidence (i.e., the frequency of a candidate term in the text) is necessary (Lauriston 1994 : 154).

Ainsi, d'autres moyens seront souvent mis en place afin de ne retenir que les syntagmes qui ont un potentiel terminologique élevé. Ces moyens sont généralement de nature statistique, et seront décrits ci-dessous.

1.2.1.2 Les techniques statistiques

L'extraction de termes exploite parfois des calculs statistiques d'abord conçus pour l'extraction de mots-clés dans le domaine de la recherche d'information (Kageura & Umino 1996). Ces calculs sont basés sur :

- la présence d'un mot dans un document;
- sa fréquence dans le document;
- sa présence dans un nombre restreint de documents dans un corpus;
- sa fréquence dans un document en comparaison avec sa fréquence dans le reste du corpus;
- sa distribution dans le corpus.

Certains de ces calculs ont été adaptés à la tâche d'extraction de termes; d'autres calculs ont été conçus spécifiquement à cette fin. Nous présenterons ci-dessous quelques-uns des calculs qui sont utilisés pour l'extraction. D'abord, mentionnons que les techniques statistiques sont souvent utilisées pour faire de la comparaison de corpus (Chung 2003), c'est-à-dire pour comparer un **corpus d'analyse** (le corpus spécialisé dont on cherche à extraire les termes) à un **corpus de référence** (un corpus de langue générale ou une collection de textes appartenant à différents domaines) afin d'identifier des termes propres au corpus d'analyse. L'hypothèse sur laquelle reposent ces techniques est la suivante : les unités qui sont fréquentes dans un domaine, qui sont présentes seulement dans un domaine ou qui sont plus fréquentes dans un domaine que dans l'usage général, ont une forte probabilité d'être des termes (un grand potentiel terminologique). Les techniques statistiques peuvent aussi servir à mesurer le degré de figement syntaxique des syntagmes.

Un des calculs statistiques qu'on utilise pour l'extraction de termes est **l'information mutuelle**. Ce terme désigne la probabilité que deux mots apparaissent à proximité l'un de l'autre dans un document par rapport à la probabilité qu'ils apparaissent séparément. Frantzi & Ananiadou (1995) utilisent, entre autres, une version de l'information mutuelle, basée sur les travaux de Damerau (1993), qui calcule :

- l'information mutuelle dans un corpus d'analyse;
- l'information mutuelle dans un corpus de référence;
- la différence entre les logarithmes de ces deux taux.

Il effectue donc de l'analyse contrastive de corpus à l'aide du calcul de l'information mutuelle.

TF·IDF est une mesure fréquemment utilisée en recherche documentaire, qui peut aussi servir à comparer des corpus dans le cadre de l'extraction de termes. Elle est basée sur la fréquence d'une unité dans un document et sur l'inverse de sa fréquence dans le reste de la collection (un ensemble de documents). Cette mesure a été formulée de différentes façons, mais le dénominateur commun de ces formules est que plus un terme est fréquent dans le corpus d'analyse, et moins il est fréquent dans le corpus de référence, plus son potentiel terminologique est élevé.

Le taux de log-vraisemblance est un autre indice statistique basé sur la fréquence d'une unité dans un corpus d'analyse et sa fréquence dans le corpus de référence. Il peut être utilisé pour faire de l'analyse contrastive, car « comparisons made between a large corpus of general text and a domain-specific text can be used to produce lists consisting only of words and bigrams characteristic of the domain-specific texts » (Dunning 1993). Daille et al. (1994) l'utilisent plutôt afin de mesurer le degré de figement syntaxique entre deux mots. Cette technique, basée sur la présupposition que les termes sont tous des syntagmes nominaux à deux mots pleins, leur fournit un moyen de classer leurs candidats termes en fonction de leur potentiel terminologique.

Il est à noter que ces mesures statistiques sont toutes basées sur la fréquence. Un des désavantages de cette approche est que, dans tout texte spécialisé, on retrouve des termes de fréquence 1 (qu'on appelle **hapax**) ou de très faible fréquence, que ces techniques vont occulter dans leur recherche de termes fréquents. Or, les hapax peuvent représenter une proportion importante des termes contenus dans un corpus. En effet, Calberg-Challot et al. (2008 : 197) ont montré que 23 % des termes communs à une liste de candidats termes et à un dépouillement manuel avaient une fréquence égale à 1.

Contrairement aux techniques statistiques décrites ci-dessus, la **C-value** (Frantzi et al. 1998) n'exploite pas uniquement la fréquence, et elle a été conçue spécifiquement pour l'extraction de termes. La méthode proposée par Frantzi et al. consiste à identifier des candidats à l'aide de patrons morphosyntaxiques, à les filtrer à l'aide d'une liste de

frontières de termes, puis à les trier en fonction de leur C-value. Cet indice est basé non seulement sur la fréquence des candidats, mais aussi sur la fréquence à laquelle ils apparaissent imbriqués à l'intérieur d'autres candidats :

> The [C-value] is built using statistical characteristics of the candidate string. These are:
>
> 1. The total frequency of occurrence of the candidate string in the corpus.
>
> 2. The frequency of the candidate string as part of other longer candidate terms.
>
> 3. The number of these longer candidate terms.
>
> 4. The length of the candidate string (in number of words) (Frantzi et al. 1998 : 589).

Les hypothèses qui sous-tendent ces trois derniers calculs peuvent être formulées ainsi : si une chaîne candidate apparaît presque toujours imbriquée dans une chaîne candidate plus longue, elle ne forme probablement pas un terme; si une chaîne candidate apparaît imbriquée dans beaucoup de chaînes candidates plus longues, il s'agit probablement d'un terme; les longs termes apparaissent moins fréquemment, donc le calcul devrait normaliser la fréquence en fonction de la longueur de la chaîne. Un deuxième indice, la NC-value, vient intégrer plus d'information contextuelle sur les candidats, en observant certains mots qui se combinent de façon privilégiée avec les candidats termes.

La technique proposée par Enguehard & Pantera (1994) repose sur des hypothèses semblables. Cette technique consiste d'abord à extraire des termes simples identifiés par fréquence relative, puis à utiliser des heuristiques (ainsi que la fréquence) pour trouver des termes complexes : si deux termes connus apparaissent souvent ensemble, ils constituent un terme complexe; si un mot apparaît souvent accompagné d'un terme connu, ils constituent un terme complexe; si un mot apparaît souvent avec des termes connus dans certaines configurations précises, on l'ajoute aux termes simples. L'hypothèse ici est donc que les termes complexes sont formés à partir de termes simples.

Les dernières années ont vu le développement de nouvelles approches à l'extraction, faisant appel aux techniques de l'apprentissage-machine, entre autres. Par exemple, Patry & Langlais (2005) décrivent une méthode d'extraction où l'utilisateur fournit au système un exemple de corpus dans lequel tous les termes ont été identifiés par un humain. Le système exploite par la suite un modèle de langue afin de calculer la probabilité qu'un patron morphosyntaxique corresponde à un terme. Puis, différents calculs statistiques sont réalisés et utilisés, avec la probabilité du patron, comme *features* dans un algorithme de classification qui vise à identifier automatiquement des candidats termes.

La majorité des extracteurs modernes utilisent une combinaison de techniques statistiques et linguistiques. Par exemple, la méthode de Daille (1996) repère certains patrons morphosyntaxiques, puis filtre les résultats à l'aide de calculs statistiques. Daille compare différents filtres, et conclut que le taux de log-vraisemblance, et les fréquences brutes, sont les meilleurs indices du statut terminologique. Drouin (2003) extrait d'abord des *pivots lexicaux spécialisés* en comparant leur fréquence dans le corpus d'analyse à leur fréquence dans un corpus général; l'indice statistique utilisé est appelé *spécificité.* Il accole aux pivots tous les mots qui apparaissent entre le pivot et une frontière de terme et qui sont aussi des pivots. Les listes de candidats ainsi produites sont filtrées afin d'enlever les candidats termes qui apparaissent seulement à l'intérieur d'un autre candidat, puis triées en fonction du potentiel terminologique des candidats (estimé à l'aide d'un indice statistique tel que la spécificité).

1.2.2 Le rôle de l'application

Kageura et al. (1999) affirment que les stratégies mises en place par les concepteurs sont fonction de leur définition du terme. Un système qui produit une liste courte de candidats, qui a une meilleure chance d'avoir une précision élevée, serait conçu pour extraire la terminologie fondamentale, un ensemble qui correspond à ce que contiendrait l'index d'une encyclopédie du domaine. Certaines stratégies visent à extraire le plus de

termes possible, quitte à introduire du bruit dans la sortie; d'autres, à éliminer le bruit dans la mesure du possible, quitte à passer sous silence certains termes véritables contenus dans le corpus. Autrement dit, certaines stratégies favorisent la précision; d'autres, le rappel (ces notions seront décrites au chapitre 2). Le choix des stratégies utilisées dépendra de certains présupposés théoriques sur la notion de terme et de considérations pratiques telles que l'application qui motive la recherche de termes.

Un des critères importants pour évaluer la performance d'un extracteur de termes est la pertinence des listes produites à une application précise : « the performance of a given system must be evaluated according to the application it was designed for [...] A translator may have specific needs that will not necessarily be shared by terminologists and vice-versa » (L'Homme et al. 1996 : 294). En effet, l'élaboration d'une ressource terminologique est guidée par une « double contrainte de pertinence » :

> - pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques au domaine et stables dans le corpus ;
>
> - pertinence vis-à-vis de l'application visée. Les unités finalement retenues doivent l'être en fonction de leur utilité dans l'application visée, qui s'exprime en termes d'économie, de cohérence interne et d'efficacité, et de leur pertinence pour l'utilisateur (Bourigault et al. 2004 : 93).

Ainsi, si un extracteur vise à produire une liste de candidats termes pertinents, les unités qu'il repère doivent être non seulement propres à un domaine ou à un corpus, elles doivent également être pertinentes à l'application qui motive l'utilisation de l'extracteur. Même lorsque le dépouillement terminologique est effectué par un humain, les termes retenus, ainsi que leur description, sont fonction de l'application : « Le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas *une* terminologie, qui représenterait le savoir sur le domaine, mais autant de ressources terminologiques ou ontologiques que d'applications dans lesquelles ces ressources sont utilisées » (id. : 89). Le rôle de l'application a également été démontré par Estopà (voir section 1.1.2). En somme, le dépouillement terminologique doit tenir compte non seulement de la conception qu'on

peut avoir du terme, mais de l'application qui motive l'utilisateur à faire appel à l'extracteur. De même, l'évaluation d'un extracteur doit être faite en adéquation avec une application précise. La façon dont nous prenons ce facteur en compte sera décrite au chapitre 3.

1.3 La variation terminologique

Les termes sont susceptibles de voir leur forme transformée en contexte pour différentes raisons. Lorsqu'on insère un terme dans un énoncé, on doit ajuster la casse s'il figure en début de phrase, la flexion s'il est utilisé au pluriel. Des modifications plus importantes (ajout ou suppression de constituants) peuvent être opérées pour différentes raisons, qui feront l'objet de cette section.

Il n'existe pas de cadre théorique unificateur pour la description de la variation terminologique (Cabré et al. 2005), mais différentes caractérisations et typologies ont été proposées. Comme le souligne Carreño Cruz (2004 : 7) : « chaque groupe de chercheurs détermine sa propre classification en fonction des objectifs poursuivis, des langues traitées, de l'application visée ainsi que les types de termes envisagés ». Carreño Cruz propose sa propre typologie de la variation, basée sur certains des travaux que nous décrirons ci-dessous.

Daille (2005) propose un survol des descriptions de la variation terminologique et montre également qu'elles sont fonction de l'application finale de ces travaux (par exemple, l'indexation ou la veille technologique) et de facteurs pragmatiques : « Several typologies of variations have been established which depend on the application, but also on computer techniques involved or the kind of data (mono- , bi- or multilingual) » (Daille 2005 : 183). Elle présente 4 typologies, dont la sienne, conçue pour l'extraction de termes. Cette typologie présuppose que les *termes de base* contiennent deux mots pleins et correspondent tous à des syntagmes comprenant une tête nominale et un modificateur nominal ou adjectival. Une forme peut être considérée une variante d'un terme de base

seulement si elle dénote le même concept. Daille définit ainsi les types des variantes suivants :

1. Une variante graphique est identique au terme de base, exception faite des différences de casse ou de la présence d'un trait d'union : *power steering* ⇔ *power-steering*.
2. Une variante flexionnelle est identique au terme de base mais présente une flexion différente : *fuel injector* ⇔ *fuel injectors*.
3. Les variantes syntaxiques de surface contiennent des différences quant aux mots-outils (modification de la préposition, ajout ou omission d'une préposition ou d'un déterminant, utilisation de la forme attributive de l'adjectif) : *fixation azote* ⇔ *fixation d'azote* ⇔ *fixation de l'azote*.
4. Les variantes syntaxiques profondes modifient la structure interne du terme de base (ajout d'un modificateur, coordination ou permutation des éléments) : *lait de brebis* ⇔ *lait cru de brebis; protéine végétale* ⇔ *protéine d'origine végétale; alimentation humaine* ⇔ *alimentation animale et humaine; hand function* ⇔ *function of the hand*.
5. Les variantes morphosyntaxiques, en plus de modifier la structure interne du terme, provoquent des changements morphologiques (remplacement d'une préposition par un préfixe, dérivation morphologique) : *pourrissement après récolte* ⇔ *pourrissement post-récolte; acidité du sang* ⇔ *acidité sanguine*.
6. La variation paradigmatique voit un des mots pleins du terme (ou les deux) se faire remplacer par un synonyme sans que la structure syntaxique soit modifiée : *battery box* ⇔ *battery case*.
7. Les variantes anaphoriques comprennent les formes réduites des termes complexes ainsi que les sigles et acronymes : *anti-lock brake system* ⇔ *brake system* ⇔ *ABS*.

Les trois autres typologies présentées par Daille, crées pour l'indexation automatique (Jacquemin & Tzoukermann 1999), la veille technologique (Ibekwe-SanJuan & SanJuan 2002) et la traduction assistée par ordinateur (Carl et al. 2004), décrivent des concepts semblables. Celle de Daille offre la meilleure couverture des phénomènes de variation. Par ailleurs, elle ne permet pas de variations qui dénoteraient des concepts différents du terme de base, contrairement à celle de Ibekwe-SanJuan & SanJuan (2002),

par exemple[6]. Or, la typologie de Daille suppose que les termes de base sont tous composés de deux mots pleins. Nous ne procèderons pas à partir de ce postulat, puisqu'il néglige l'existence de termes simples et de termes de base comportant plus de deux mots pleins. La plupart des types de variantes définis par Daille s'applique toujours si on enlève cette contrainte, mais quelques-uns seraient à redéfinir. La typologie de Jacquemin & Tzoukermann (1999) part du même présupposé, mais pas celle de Carl et al. (2004). Basée sur les travaux de Daille (1996) et de Jacquemin (2001), cette dernière typologie a été élaborée afin d'extraire les termes et leurs variantes dans un corpus bilingue parallèle. Les mécanismes de variation les plus fréquents seraient l'omission, l'insertion, la permutation, la coordination, la synonymie, les variantes typographiques et la dérivation. Carl et al. (2004 : 106 et suiv.) définissent et illustrent les sept types de la façon suivante :

1. l'omission : omission d'un ou plusieurs constituants d'un terme; *c3a1 sniper rifle* ⇔ *c3a1 rifle*.
2. l'insertion : ajout d'un ou plusieurs constituants (préposition, déterminant, modificateur); *prone position* ⇔ *prone supported position*.
3. la permutation : changement de l'ordre linéaire des constituants, souvent accompagné d'insertions; *rifle butt* ⇔ *butt of a rifle*.
4. la coordination : insertion d'une conjonction de coordination, suppression des constituants communs aux termes coordonnés, ajout de virgules dans certains cas; *elevation adjustment* ⇔ *elevation and windage adjustment*; *visual acuity* ⇔ *visual ability and acuity*.
5. la synonymie : substitution d'un constituant par un de ses synonymes; *spotting telescope* ⇔ *spotting scope*.
6. la variation typographique: différences dans l'utilisation de l'espace ou du trait d'union dans les termes, ou ajout d'autres ponctuations telles que les guillemets; *hand stop* ⇔ *handstop*.
7. la dérivation : utilisation d'un mot dérivé de la même racine qu'un des constituants du terme; *dégagement de l'œil* ⇔ *dégagement oculaire*.

---

[6] Il faut se rappeler que l'application qu'ils visent est la veille technologique, donc le fait que leur définition de la variation permette des variantes qui pointent vers des concepts différents du terme de base est tout à fait justifié.

Cette typologie nous fournit toutes les balises dont nous avons besoin pour ce travail d'annotation de corpus, et ne suppose pas de minimum ou de maximum de constituants des termes.

En ce qui concerne l'omission, que nous appellerons *réduction*, nous devons en distinguer deux types. En effet, Haralambous & Lavagnino (2011) opposent deux types de réduction des termes complexes, la réduction anaphorique et la réduction lexicale :

> La réduction anaphorique est un processus discursif et textuel, tandis que la réduction lexicale est générée par des conditions internes au syntagme plein (caractéristiques morphosyntaxiques, notionnelles, statut terminologique des constituants) ou par des conditions externes (niveau de spécialité du texte, typologie textuelle) (Haralambous & Lavagnino 2011 : 44).

Ils associent à la réduction anaphorique une valeur contextuelle et à la réduction lexicale une valeur synonymique. Par exemple, le terme *mode de production biologique* peut donner lieu aux réductions anaphoriques *mode de production* et *mode*, qui sont difficilement lexicalisables; elles n'apparaîtront généralement que si le terme de base est utilisé ailleurs dans le texte. En revanche, la réduction lexicale *mode biologique*, où la tête est toujours liée au constituant *de production biologique*, peut être considérée un synonyme. Collet (1997) fait également la distinction entre réduction anaphorique et lexicale[7] :

> Nous établissons une distinction entre la *réduction à caractère lexical* (RL) pouvant transcender le milieu contextuel immédiat (ex. station terrienne de navire ⇔ station de navire, station terrienne côtière ⇔ station côtière), et la *reprise anaphorique* (RA), qui est purement contextuelle (ex. service mobile maritime par satellite ⇔ service) (Collet 1997 : 198).

[7] Les exemples donnés par Collet montrent que ce qui est considéré ici la réduction d'un terme complexe (station terrienne de navire ⇔ station de navire) serait vu d'un tout autre point de vue suivant d'autres typologies de la variation; en effet, selon la typologie de Daille, il s'agirait plutôt de l'insertion d'un constituant à un terme de base.

À la différence de la réduction anaphorique, la réduction lexicale produit des variantes « susceptibles de devenir des membres permanents de la terminologie du domaine du [syntagme terminologique] plein » (Collet 1997 : 198). Elle distingue celles-ci des synonymes, qui doivent comporter des unités lexicales différentes au niveau de la tête, du modificateur ou des deux. Sur le plan formel, la réduction anaphorique se démarquerait de la réduction lexicale par le nombre limité de modifications qu'elle admet. Dans le cas de la réduction anaphorique, le modificateur est supprimé au complet, qu'il soit simple ou complexe, tandis que la tête est conservée en entier ou en partie. En revanche, la réduction lexicale admet une plus grande variété d'effacements.

Collet classe la réduction des termes complexes parmi cinq « mécanismes intraphrastiques et interphrastiques qui perturbent la linéarité du [syntagme terminologique] en le rendant discontinu, en le dissolvant ou en effaçant un ou plusieurs de ses constituants » (id. : 193). Ces cinq mécanismes discursifs sont :

- La coordination.
- La prédication : « elle dissout le [syntagme terminologique en un sujet et un prédicat, […] le plus souvent la copule être » (id. : 194); p. ex. ce répéteur est à double changement de fréquence.
- L'insertion : elle insère entre la tête et le modificateur d'un terme complexe un élément qui « ne correspond pas à un trait pertinent distinctif de la notion dénommée par le [syntagme terminologique] » (ibid.); p. ex. *antenne du type Cassegrain*.
- La dénomination : insertion de *dit*, *appelé*, etc.
- La réduction : suppression de un ou plusieurs constituants du terme.

La réduction anaphorique est un phénomène très fréquent, bien qu'il soit difficile d'obtenir une vraie mesure de son importance, car la question des réductions anaphoriques a reçu très peu d'attention, comme l'ont affirmé L'Homme (2004) et Daille (2005). S'il est difficile de savoir à quel point la réduction anaphorique est fréquente, c'est tout aussi vrai de la variation terminologique en général. Daille (2003) montre que différentes études

indiquent que 15 à 35 % des termes participent à des relations de variation. Ibekwe-SanJuan & SanJuan (2004 : 491) obtiennent des résultats très différents : « The variation identification program linked 41 058 [sic] terms which were involved in the six variation relations described above, that is 87% of the term candidates ». Or, cette différence pourrait être attribuable à leur définition de la variation, qui est moins contraignante que celle utilisée par d'autres chercheurs. Puisque différentes définitions de la variation sont utilisées, il est difficile d'effectuer des comparaisons à ce sujet.

Une réduction anaphorique peut parfois revêtir la même forme qu'un terme à part entière, dans lequel cas il faut distinguer l'anaphore du terme, et décrire celui-ci en tant que terme :

> D'abord, page (signifiant « document hypertexte … ») peut s'utiliser seul et il ne s'agit pas simplement d'un raccourci anaphorique (page s'emploie aussi avec d'autres modificateurs – ~ d'accueil, ~ personnelle – et revêt dans ces syntagmes le même sens). Donc, il devrait être décrit comme un terme à part entière (L'Homme 2005 : 1117).

Or, lorsqu'il s'agit effectivement d'une réduction anaphorique, un lien doit être fait entre la réduction et la forme pleine pour bien cerner ce phénomène de variation terminologique. C'est, entre autres, ce genre de liens de variation que nous chercherons à décrire lors de l'annotation de notre corpus (voir chapitre 3).

1.4 Conclusion

Ce travail porte sur l'annotation d'un corpus spécialisé aux fins de l'évaluation des extracteurs de termes. Dans le cadre de cette tâche, les termes de base seront choisis en fonction de critères précis. Tout d'abord, ce choix sera fait en fonction d'une application spécifique, puisque l'évaluation d'un extracteur doit tenir compte de l'application qui motive une personne à utiliser un extracteur (voir section 1.2.2). Cette application sera la compilation d'un dictionnaire spécialisé de la mécanique automobile, portant surtout sur la

structure de l'automobile. Ainsi, nous repérerons des termes qui correspondent à des parties de l'automobile, ainsi que quelques concepts connexes (voir chapitre 3); nous nous intéresserons donc seulement aux unités nominales.

Le choix des termes sera aussi balisé par les critères énoncés par L'Homme (voir section 1.1.3.7), à l'exception du critère portant sur les unités prédicatives, puisque nous nous intéresserons uniquement à des termes nominaux dénotant des concepts concrets. Nous ne retiendrons pas l'idée que les syntagmes longs, à plusieurs déterminants, doivent avoir au moins deux attestations dans des documents distincts, comme le recommande Auger (voir section 1.1.3.2), entre autres. Suivant les idées de Pearson (voir section 1.1.2), nous annoterons plutôt toute unité nominale potentiellement terminologique selon les critères de L'Homme et des critères thématiques dictés par notre application. Nous montrerons au chapitre 4 qu'il est possible de paramétrer la liste de termes que nous compilerons afin d'éliminer les hapax, et éventuellement retenir, au sein des hapax, des unités plus petites pour autant qu'elles apparaissent ailleurs dans le corpus. En outre, comme L'Homme et Pearson, nous ne limiterons pas notre choix des termes à ceux qui sont utilisés uniquement dans le domaine à l'étude, celui de la mécanique automobile.

Dans les travaux sur l'extraction de termes, la variation doit être prise en compte, car une représentation juste de la fréquence des termes passe nécessairement par un regroupement des variantes terminologiques. « In terminology extraction, […] the crucial part of the handling of term variations is to obtain an optimised representativity of the candidate term occurrences. The conflating of term variants identifies inside the hapax candidate term set 10% which are hapax which would otherwise be omitted » (Daille 2005 : 192). Ainsi, des formes peu fréquentes mais potentiellement pertinentes, que des techniques statistiques simples basées sur la fréquence et la cooccurrence passeraient sous silence, pourraient être extraites si on identifie le lien entre un terme de base et ses variantes. L'identification des liens de variation terminologique permet également de lever une partie des ambiguïtés présentes dans les textes spécialisés. De plus, la variation terminologique est

un des aspects moins bien traités par les outils de traitement automatique de la langue : « Plusieurs problèmes sont mal résolus aujourd'hui. Il s'agit de phénomènes linguistiques tels que l'enchâssement des termes, l'élision, la coordination, etc. » (Chiao et Sta 2002 : 116). La description de la variation dans un corpus étalon comme celui qui fait l'objet de ce travail est une étape importante pour rendre compte de l'importance des phénomènes de variation terminologique sur l'extraction de termes.

Ainsi, nous annoterons également les variantes terminologiques (voir section 1.3) des termes de base; nous poserons une condition à leur inclusion, à savoir qu'elles doivent dénoter le même concept que le terme de base. Nous emprunterons la typologie de Carl et al. (2004), mais ne retiendrons pas tous les types décrits; notamment, les variantes par permutation ne seront pas recensées, car les unités qui ont la forme de *butt of a rifle* seront exclues. Nous nous intéresserons, entre autres, aux variantes par insertion (que nous appellerons *surcomposition* pour les distinguer des insertions décrites par Collet) et par dérivation ainsi qu'aux variantes typographiques. En ce qui concerne les variantes par omission, celles qui ont une valeur anaphorique seront appelées *réductions anaphoriques*, et celles qui ont une valeur lexicale seront traitées comme des synonymes. Nous emprunterons aussi à la typologie de Collet les variantes par insertion, où un mot comme *type* est inséré entre les constituants d'un terme complexe. Toutes les variantes et tous les synonymes seront liés à un terme de base. Ceux-ci seront choisis en fonction de la fréquence des occurrences dans le corpus, et en cherchant le terme qui figure le plus souvent comme vedette dans des ressources terminologiques. Aucune limite ne sera posée au nombre de constituants des termes de base ou des variantes. Notre méthodologie d'annotation et de traitement des variantes sera décrite au chapitre 3.

Dans le chapitre qui suit, nous recenserons les travaux sur l'évaluation des extracteurs afin de mieux caractériser notre approche à ce problème.

2. État de la question

Dans ce chapitre, nous recensons les travaux qui ont touché à l'évaluation des extracteurs de termes. Nous commencerons par expliquer brièvement quelques concepts utilisés pour caractériser cette évaluation. Puis, nous traiterons un certain nombre d'études sur l'extraction qui comportent une part d'évaluation et décrirons quelques protocoles d'évaluation qui ont été mis de l'avant. Ensuite, nous décrirons les campagnes qui ont été organisées sur le thème de l'évaluation des extracteurs.

2.1 L'évaluation des extracteurs de termes

2.1.1 Typologie de l'évaluation des logiciels

Certains concepts utilisés pour caractériser l'évaluation des logiciels sont nécessaires pour bien saisir les travaux qui ont touché à l'évaluation des extracteurs. En effet, l'évaluation des logiciels s'articule autour de différents axes (manuelle ou automatique, verticale ou horizontale, orientée système ou centrée sur l'utilisateur, etc.). Un bon point de départ est la norme ISO 9126, qui fixe des balises pour l'évaluation des logiciels. Cette norme, qui vise les logiciels en général, sans égard à leur application spécifique, a été adaptée à différents domaines et différentes tâches. Notamment, le protocole EAGLES (King et al. 1996) a modifié cette norme afin de l'appliquer aux outils d'aide à la traduction et à la rédaction; les auteurs envisageaient, à plus long terme, d'étendre le protocole à tout un ensemble d'outils de traitement automatique de la langue.

La norme ISO porte surtout sur la définition des qualités à évaluer. On y distingue globalement trois types d'évaluation, en fonction des caractéristiques qui en font l'objet :

> L'approche ISO distingue […] trois types de caractéristiques de qualité : internes, externes et à l'usage. Les qualités internes peuvent être mesurées sans exécution du logiciel – évaluations dites « en boîte de verre » – alors que les qualités externes doivent être mesurées en faisant fonctionner le

> logiciel – évaluations dites « en boîte noire » où l'on s'intéresse aux résultats produits. Enfin, la qualité à l'usage doit être mesurée en plaçant le système dans un contexte d'utilisation, expérimental ou final, et en observant dans quelle mesure le système aide ses utilisateurs à accomplir leurs tâches (Popescu-Belis 2007 : 70).

Plus précisément, l'évaluation en boîte noire « porte sur le jugement des performances globales du système à partir seulement des ressources fournies en entrée (Input) et des résultats produits en sortie (Output), sans examiner le traitement intermédiaire des données effectué par les divers modules du système » (Timimi 2007 : 145). En revanche, l'évaluation en boîte de verre (ou boîte transparente) porte sur les différents modules qui composent le système global, les algorithmes qui sont implémentés, l'optimalité du code, etc.

On distingue également l'évaluation objective et subjective (Paroubek et al. 2007 : 10). L'évaluation subjective fait intervenir la perception des humains de la performance d'un logiciel, tandis que l'évaluation objective mesure directement les données fournies par le logiciel. De même, l'évaluation qualitative attribue une description de la qualité du comportement d'un logiciel, tandis que l'évaluation quantitative porte strictement sur les valeurs de différentes variables que l'on cherche à mesurer.

De plus, l'évaluation peut porter sur différentes versions d'un même logiciel, dans lequel cas on parlera d'une évaluation verticale ou de progression, ou elle peut chercher à déterminer dans quelle mesure un logiciel répond à des besoins particuliers ou se compare à d'autres outils (évaluation horizontale).

On distingue également les approches automatique et manuelle, et ces deux types ont été utilisés pour l'évaluation des extracteurs de termes. Dans le cas de l'évaluation manuelle, on peut, par exemple, faire inspecter la sortie d'un extracteur par un expert et lui demander son jugement sur la validité de chaque candidat terme. Les qualifications de l'expert sont éventuellement déterminées en fonction de l'application que l'on donne à l'extracteur; il peut s'agir d'un terminologue, mais aussi d'un traducteur, un

documentaliste, un expert du domaine qui caractérise le corpus de test, etc. Cette approche, bien qu'elle soit facile à implémenter, présente des désavantages, à savoir que le travail peut être laborieux et fastidieux, surtout lorsque les listes de candidats sont longues, et que les résultats sont influencés par la subjectivité de l'évaluateur humain (Mustafa El Hadi et al. 2006 : 948); en outre, une fois l'évaluation effectuée, il n'est pas possible de la reproduire. L'approche automatique, quant à elle, consiste à comparer la sortie d'un extracteur à un étalon quelconque afin de mesurer le degré de correspondance.

Un dernier type d'évaluation mérite notre attention, à savoir l'évaluation centrée sur l'utilisateur, qui met l'accent sur la façon dont de vrais utilisateurs utilisent les outils de TAL, plutôt que sur différentes caractéristiques des systèmes eux-mêmes (Paroubek et al. 2007 : 11). À l'autre bout du spectre, on retrouve l'approche orientée système ou *technocentrée*, qui caractérise la majorité des travaux sur l'évaluation des extracteurs.

2.1.2 Les métriques

L'évaluation des outils informatiques fait souvent appel à des métriques, surtout lorsqu'il s'agit d'une évaluation automatique. Les métriques ou mesures utilisées pour évaluer quantitativement les extracteurs sont généralement basées sur les notions de précision et de rappel[8], d'abord utilisées dans le domaine de la recherche d'information. La précision correspond à la proportion de termes valides présents dans une liste de candidats termes, tandis que le rappel correspond à la proportion de termes que l'extracteur a su recenser, parmi tous les termes valides dans le corpus. À l'inverse, le bruit correspond à la proportion de candidats non terminologiques dans une liste de candidats termes, et le silence dénote la proportion de termes qu'un extracteur n'a pas identifiés, parmi tous les termes valides dans le corpus. Un tableau de contingence (Tableau I) formalise ces quatre concepts.

---

[8] D'autres métriques ont été proposées, entre autres par Nazarenko et al. (2009). Voir section 2.2.2.10.

	Terme	Non-terme
Extrait	VP (vrais positifs)	FP (faux positifs)
Pas extrait	FN (faux négatifs)	VN (vrais négatifs)

Tableau I : Tableau de contingence utilisé pour calculer la précision et le rappel

Formellement, $P = VP/(VP+FP)$ et $R = VP/(VP+FN)$, où P désigne la précision et R, le rappel. Chacune de ces métriques ne peut, à elle seule, rendre compte de la correspondance entre une liste de candidats termes et une liste de référence; elles doivent être utilisées ensemble :

> L'exemple classique de distances imparfaites mais complémentaires sont les métriques de rappel et de précision [...], définies à l'origine pour la recherche documentaire, mais applicables à toute tâche visant à identifier des éléments pertinents parmi un ensemble d'éléments candidats. Aucune des deux métriques ne mesure à elle seule la distance entre l'ensemble des éléments identifiés par le système et l'ensemble correct ; c'est leur moyenne harmonique, ou f-mesure, qui est en général utilisée (Popescu-Belis 2007 : 84).

En effet, la *F-mesure* (Van Rijsbergen 1979 : 174), combine ces deux mesures en une seule : $F = 1 / (\alpha(1/P) + (1-\alpha)(1/R))$, où α est un coefficient qui spécifie le poids que l'on accorde à la précision (et $1-\alpha$, le poids attribué au rappel). Si l'on attribue des pondérations égales à la précision et au rappel, l'on obtient la F-mesure équilibrée, parfois notée F_1 : $F_1 = 2PR/(P+R)$. On utilise généralement le terme *F-mesure* pour désigner plus spécifiquement la mesure F_1.

Mustafa El Hadi (2004) juge que les métriques héritées de la recherche d'information sont utiles et adéquates à l'évaluation des extracteurs tant qu'elles sont mesurées en fonction de tâches de contrôle spécifiques telles que la terminographie, la traduction ou l'indexation. Elle propose d'ajouter à celles-ci le gain de temps et l'utilité pour

l'utilisateur final. Or, ces dernières techniques correspondent à un modèle d'évaluation centré sur l'utilisateur plutôt que la performance du système.

2.1.3 La liste de référence

Une dernière notion est essentielle pour comprendre les travaux sur l'évaluation des extracteurs, celle de « liste de référence ». Un moyen simple d'évaluer la performance d'un extracteur est de comparer la liste de candidats termes qu'il produit à une liste produite par un humain. En outre, lorsqu'on souhaite mesurer non seulement la précision mais le rappel, il est nécessaire d'avoir une telle liste, dont on présume qu'elle contient tous les termes que l'extracteur devrait idéalement identifier. Cette simplification est un peu réductrice, puisqu'aucun extracteur ne peut extraire tous les termes d'une référence quelconque et uniquement ces termes, et que tout dépouillement terminologique manuel comporte une part de subjectivité : différents annotateurs produiront différentes terminologies, tel que démontré par Estopà (2001), et si on demande à un annotateur de dépouiller deux fois le même corpus, il est fort à parier que les listes produites seront différentes (Drouin 2003 : 107). Il ne peut donc pas exister une référence unique pour l'extraction, mais l'utilisation de critères spécifiques de sélection de termes permet d'éliminer une part de cette subjectivité.

Sur le plan épistémologique, Popescu-Belis affirme que, pour pouvoir formuler une question de traitement automatique de la langue, il faut pouvoir définir une métrique d'évaluation adéquate et des données de référence, qu'il s'agisse d'une réponse correcte ou d'un espace, potentiellement vaste et soumis à une variabilité, de réponses acceptables :

> Étant donné le rôle central de l'évaluation, il ne serait pas absurde de conclure que le problème épistémologique central de toute question étudiée en TAL est la définition de métriques d'évaluation ainsi que l'élaboration de données de référence. La définition de chaque question apparaît en effet inséparable de la définition d'une métrique d'évaluation (Popescu-Belis 2007 : 87).

La liste de référence peut être construite par divers moyens : on peut la construire manuellement à partir du corpus d'analyse qui servira à tester l'extracteur, ou l'extraire de ressources terminologiques existantes, dans lequel cas il peut être nécessaire de l'enrichir pour compenser l'inadéquation entre les termes de la référence et ceux contenus dans le corpus de test (Mustafa El Hadi 2004 : 156). Une dernière approche, qui peut être utilisée pour les évaluations comparatives, consiste à créer une référence consensuelle à partir de l'intersection ou de l'union des sorties de différents outils, l'hypothèse voulant que les candidats qui se retrouvent le plus souvent dans la sortie de ces outils aient une plus forte probabilité de correspondre effectivement à des termes; on pourrait également pondérer les candidats termes en fonction du nombre de listes dans lesquelles ils apparaissent.

2.2 Travaux sur l'évaluation des extracteurs

Dans cette section, nous résumons des travaux qui ont touché à l'évaluation des extracteurs. Cet échantillon a pour but de montrer le peu d'attention que reçoit la liste de référence dans la littérature. Nous ne retiendrons que les travaux qui comportent un volet quantitatif. Nous ne traiterons pas, par exemple, l'étude de Estopà (1999) qui évalue l'adéquation des extracteurs aux attentes de leurs utilisateurs; cette évaluation est notable parce qu'elle est centrée sur l'utilisateur, mais ne présente pas de volet quantitatif, et ne présente donc pas de lien direct avec la présente recherche. Les travaux comportant une évaluation quantitative contiennent souvent un volet qualitatif, où l'on analyse les sources d'erreur, les forces et les faiblesses des systèmes, etc. Nous ne nous attarderons pas sur cet aspect dans le survol présenté ci-dessous.

2.2.1 Évaluations ad hoc

Lorsque les chercheurs ont commencé à s'intéresser à l'extraction de termes, une bonne part des écrits traitant l'évaluation des extracteurs était composée d'évaluations ad hoc qui accompagnaient la description de nouvelles techniques d'extraction. Ces évaluations sont peu informatives, car peu de renseignements sont fournis sur la

méthodologie d'évaluation : « There are many other TEs where the information regarding evaluation is limited to mentioning a 'global' figure of precision » (Vivaldi & Rodríguez 2007 : 233).

Les exemples sont nombreux. Ladouceur & Cochrane (1996) évaluent une technique d'extraction et affirment que les « performances du moteur d'analyse sont excellentes. Nous l'avons testé sur une série de textes de diverses longueurs et avec des densités terminologiques différentes, et dans tous les cas, les résultats sont très bons » (Ladouceur & Cochrane 1996 : 53). On fournit des mesures de précision, mais aucun renseignement n'est donné sur la méthode d'évaluation, mis à part ceci : « Une fois les textes traités par la machine, l'indexeur humain a épuré les listes de résultats afin de ne conserver que les termes du domaine. Ce travail, effectué à l'aide d'un chiffrier, se fait en une vingtaine de minutes » (id. : 54). Cette méthode leur permet d'affirmer que leur système ne produit aucun silence et que le bruit « est assez bien contrôlé » (ibid.). Pour le reste, ils comparent la sortie de leur outil à un dépouillement fait par un humain sur le même corpus, et affirment que l'outil a un meilleur rappel. On représente donc un extracteur qui offre une très bonne précision sans générer de silence, mais nous ne savons pas comment il fonctionne, et nous en savons très peu sur leur méthodologie d'évaluation; l'expérience est donc impossible à reproduire. De même, Auger et al. (1996) décrivent une technique d'extraction de termes complexes, implémentée dans l'outil Filtact, qui obtiendrait une précision de 40 %, mais on ne décrit pas comment la validation des candidats termes est effectuée.

Dans Ahmad et al. (1994), une technique d'extraction contrastive est évaluée en analysant sommairement des échantillons de candidats complexes générés par le système implémenté. Les auteurs affirment simplement que « the majority of pairs listed are promising » (Ahmad et al. 1994 : 275). Ils soulignent qu'il serait difficile d'évaluer le rappel, mais supposent que le silence est moins important que le bruit :

> There is no easy way of knowing at the moment how many ‘false negatives’ the technique produces, ie how many real compound terms are missed, but our experience so far indicates that overgeneration (false positives) is a greater problem than undergeneration (false negatives) in statistically-based techniques (ibid.).

Dagan & Church (1994), pour leur part, évaluent la composante monolingue de leur extracteur bilingue en fonction du gain de temps seulement. Bien qu’ils affirment que l’extracteur double la vitesse à laquelle ils peuvent construire des terminologies, aucune évaluation de la précision de la composante monolingue de leur outil n’est effectuée.

Lauriston (1994), dans un article portant sur la version 1.2.3 de l’extracteur TERMINO, affirme que la seule évaluation quantitative de TERMINO qui a été publiée est celle de Otman (1991), qui cite un rappel de 70 % pour la version 1.0 et une amélioration indéterminée pour la version 1.1. En effet, (Otman 1991 : 93) présente une évaluation très sommaire de TERMINO, effectuée en prenant comme étalon le dépouillement d’un terminologue sur 4 échantillons du corpus d’analyse. Aucun renseignement supplémentaire n’est donné sur la méthodologie d’évaluation. Lauriston souligne d’autres lacunes de cette évaluation, notamment l’absence d’un volet qualitatif et le fait que la précision n’a pas été évaluée :

> The study gave no qualitative report about the remaining 30% of terms that went undetected. Were they single-word terms, acronyms, complex terms that were not synapsies or synapsies that the software failed to detect? Another factor which limits the usefulness of this report is that neither the size of the text sample nor the amount of noise (synapsies that were not terms) are mentionned in Otman’s assessment (Lauriston 1994 : 161-162).

Lauriston présente donc sa propre évaluation, qui fait appel à une liste de référence de 592 termes complexes extraits manuellement d’un corpus de 8500 mots. Par contre, il ne précise pas comment les termes ont été sélectionnés, ni par qui, ni pour quelle application. En revanche, il précise que deux types de correspondances sont prises en compte : dans un premier temps, on considère qu’un candidat est valide s’il partage au moins un constituant

avec un terme de la référence; puis, on considère un candidat valide seulement s'il correspond exactement à un terme de la référence. Lauriston obtient un rappel de 0,74 et une précision de 0,72 avec l'évaluation plus indulgente; et de 0,51 et 0,48 respectivement avec l'évaluation rigoureuse.

Gaussier (2001) effectue une évaluation quantitative graduée d'une méthode d'extraction bilingue. Il évalue les *n* premières (100 à 1000) paires de candidats termes en les comparant à une liste de référence extraite de la banque de terminologie EURODICAUTOM et en calculant la précision et le rappel. Il n'est pas clair comment le rappel a été évalué avec une liste de référence qui n'est pas construite à partir du corpus d'analyse, mais étant donné les résultats affichés (rappel de 0,57 sur les 1000 premiers candidats), on peut présumer que la liste de référence a été filtrée afin de ne retenir que les termes présents dans le corpus d'analyse.

Savary (2001) effectue une évaluation sur deux systèmes, ACABIT, un extracteur de termes, et LexProTerm, un système de reconnaissance de termes. Savary évalue la précision sur un corpus de 280 000 mots, mais ne calcule pas le rappel parce qu'il serait trop long d'annoter toutes les occurrences des termes dans le corpus[9]. Elle fait par contre une approximation du silence en observant les termes extraits seulement par un des deux systèmes. Quant à la précision, elle l'obtient en analysant manuellement les listes de candidats termes, jouant le rôle du terminologue et d'expert à la fois. Les critères de sélection de termes ne sont pas donnés, et aucune mention n'est faite de la variation terminologique.

[9] Nous pouvons confirmer qu'un tel travail est très long à effectuer.

Vivaldi & Rodríguez (2001) proposent de combiner différentes techniques d'extraction, chacune utilisant un type différent d'information, puis évaluent ces techniques à l'aide de deux corpus d'analyse différents. Un expert du domaine valide manuellement les candidats extraits du premier, et trois experts analysent ceux du deuxième, ce qui n'est pas sans causer des problèmes : « It should be noted that this task is very tedious and not error-free. Disagreement was found regarding what is to be considered a term between specialists on the one hand and between specialists and terminologists on the other hand » (Vivaldi & Rodríguez 2001 : 34). Les deux listes de références obtenues par validation manuelle des candidats permettent de mesurer (ou plutôt d'estimer) la précision et le rappel. Puis, ils combinent les techniques de différentes façons, et la précision et le rappel de chaque combinaison sont présentés, sur une échelle graduée en fonction de la proportion de candidats pris en compte. Ces résultats sont certainement intéressants, mais les renseignements fournis sur la méthodologie d'évaluation laissent toujours à désirer, surtout en ce qui concerne la sélection des termes et le traitement de la variation terminologique.

Pour ne citer qu'un dernier exemple, Vintar (2004) évalue la C-value (voir section 1.2.1.2) et son apport à l'extraction de termes. L'évaluation est effectuée manuellement par un terminologue. Vintar calcule la précision et le rappel avant et après que les candidats termes imbriqués, identifiés au moyen de la C-value, aient été éliminés. L'auteur affirme que les résultats sont décevants, mais que la méthodologie d'évaluation est rigoureuse « in the sense that it allowed only candidates of a very high quality and terminological significance » (Vintar 2004 : 56). Or, la validité de cette dernière affirmation n'est pas démontrée, puisque la sélection des termes n'est pas décrite.

2.2.2 Travaux saillants

Nous présentons ci-dessous quelques travaux où l'attention portée à l'évaluation des extracteurs est plus conséquente. Dix travaux seront présentés, en ordre chronologique, puis nous porterons en dernier lieu une attention particulière à la méthodologie de Love (2000).

2.2.2.1 L'Homme et al. (1996)

Dans une réflexion sur l'évaluation des extracteurs, L'Homme et al. (1996) proposent une grille d'évaluation des extracteurs faisant appel à deux types de critères. La première catégorie comprend des critères de pré-évaluation, qui servent à aligner le choix d'un extracteur avec les besoins de l'utilisateur. Elle comprend des critères tels que le type de techniques déployées par l'extracteur, le type et le nombre de listes fournies en sortie, les langues prises en compte, etc. Le deuxième type de critères concerne la performance de l'extracteur. Les auteurs proposent d'utiliser le bruit et le silence pour mesurer la performance, et de faire appel à une liste de référence pour calculer ces mesures : « A simple method for evaluating the performance of term-extraction systems is to compare a list of terms produced by a human to the one produced by the system. The human list is taken as a reference and is assumed to be perfect, although humans may omit some terms » (L'Homme et al. 1996 : 302).

Les auteurs mettent leur protocole en application en évaluant deux extracteurs, ATAO et LOGOS, en fonction de leurs critères de performance. Ils soumettent différents types de textes aux deux outils, afin de mesurer comment la performance varie en fonction du type de texte, et comparent les sorties à des listes de référence. Par contre, peu d'information est fournie quant aux listes de référence utilisées, mis à part qu'elles ont été construites manuellement.

2.2.2.2 Fulford (2001)

Dans un survol des travaux touchant à l'évaluation des extracteurs, Fulford souligne l'insuffisance des descriptions portant sur la tâche d'évaluation. Elle appuie cette observation avec une énumération d'évaluations ad hoc semblables à celles que nous avons mentionnées (Fulford 2001 : 268). Fulford propose donc une méthode d'évaluation basée sur une liste de référence, qu'elle privilégie à cause de sa réutilisabilité et parce qu'elle permet de comparer différents outils. Elle l'applique à une technique d'extraction implémentée par l'outil Textprobe, et montre que « Textprobe matched on average 80% of

the terms selected by the domain experts; and again it demonstrated a tendency to over- rather than under-generate, and to expand rather than truncate terms » (id. : 272). La méthode qu'elle propose fait appel à une liste de référence produite par un expert du domaine. Or, les critères utilisés pour sélectionner les termes ne sont pas énoncés; on se fie simplement à l'expertise de la personne et son intuition terminologique :

> Working on the assumption that an expert's domain knowledge will include a familiarity with, and knowledge of, the terms of his/her domain, an expert's manual scanning output is deemed to be reliable, and thus constitutes the benchmark against which Textprobe output can be measured (id. : 268).

Puis, elle compare la performance de son système *ainsi que celles de terminologues* à cette liste de référence; c'est-à-dire qu'elle se sert de l'étalon créé par l'expert pour mesurer non seulement la performance de son outil, mais celles de terminologues humains. Elle affirme que son outil « performs slightly better than the human terminologists with regard to term matching » (id. : 271) parce que son outil repère exactement les termes choisis par un expert plus souvent qu'un terminologue. Il est intéressant de noter que Fulford croit qu'une liste de référence doit être construite par un expert du domaine (ou préférablement plusieurs) plutôt qu'un terminologue, une opinion que partagent d'autres chercheurs qui ont évalué des extracteurs (voir section 2.2.2.3).

Un des aspects intéressants de l'étude de Fulford est une réflexion sur la façon de mesurer le degré de similitude entre la liste de candidats et la liste de référence, et les différents types de correspondance. Elle en décrit cinq : correspondances complètes entre un candidat terme et un terme de la référence, correspondances partielles avec constituants manquants ou ajoutés, silence et bruit. Mentionnons que Nazarenko et al. (2009) proposent une façon différente de mesurer la correspondance entre la liste de candidats et la liste de référence, basée sur la notion de distance d'édition (voir section 2.2.2.10).

2.2.2.3 Collier et al. (2001)

Collier et al. (2001) présentent un système probabiliste d'identification et de classification de termes basé sur l'apprentissage-machine. Ils visent particulièrement à identifier des termes appartenant à certaines classes conceptuelles déterminées de façon automatique et de les associer à leur classe conceptuelle (p. ex. la classe des protéines).

L'annotation sur laquelle repose la liste de référence utilisée pour l'évaluation consiste à découper les termes et leur assigner une classe conceptuelle. Les auteurs utilisent un corpus de 100 sommaires annotés par un expert du domaine, dont 80 servent à l'entraînement et 20 à l'évaluation. Tout comme Fulford (2001), ils choisissent un expert du domaine pour construire le corpus annoté, mais cette décision comporte son lot de difficultés :

> Although the expert had no formal training in computational linguistics or terminology, she had considerable knowledge about term boundaries and classes gained from reading and research experience so that we considered her annotations to be our 'gold standard', even though they sometimes called for unsatisfactory choices to be made (Collier et al. 2001 : 243).

Ils affirment, par exemple, que l'annotatrice hésite à annoter les variantes réduites en raison d'une coordination, et associe de tels termes réduits à une classe conceptuelle différente, comme s'ils étaient utilisés de façon autonome (notre traitement des réductions par coordination est décrit à la section 3.3.2.1). À ce type de difficulté s'en ajoutent d'autres liées à la méthode d'annotation, qui repose sur un simple balisage pour découper les termes et les associer à une classe. Par exemple, les auteurs notent la difficulté que posent les sigles imbriqués à l'intérieur d'un terme dans la tâche d'annotation, à cause d'une règle qui empêche d'annoter des termes imbriqués (notre traitement de ces disjonctions est décrit à la section 3.3.2.2). Ils soulignent l'opportunité d'une annotation plus riche qui permettrait de mieux traiter de tels cas.

Les problèmes observés lors de l'annotation ont un effet défavorable sur la performance de leur système, puisque celui-ci fonctionne par apprentissage-machine et

exploite le corpus annoté. Les auteurs attribuent cela, entre autres, à un manque de cohérence dans les annotations, ce qui montre l'intérêt d'avoir une méthodologie d'annotation permettant de rendre compte des termes dans toutes leurs diverses manifestations.

2.2.2.4 Tiedemann (2001)

Tiedemann (2001) est une étude qui met en valeur l'importance et la difficulté de l'évaluation. L'auteur décrit une technique qui permet d'améliorer l'extraction monolingue de termes complexes (que Tiedemann appelle *phrase extraction*) au moyen de l'alignement des mots dans un bitexte. Puis, il évalue sa technique en utilisant 3 procédés : comparaison avec une liste de référence de termes, comparaison avec une liste de référence de patrons morphosyntaxiques, et évaluation manuelle d'échantillons. Il décrit brièvement les pour et les contre de chaque méthode, à savoir que l'approche avec étalon est reproductible, mais coûteuse : « However, creating reference data for evaluating phrase extraction is very time-consuming and requires detailed guidelines » (Tiedemann 2001 : 200). Quant à l'utilisation d'une ressource existante, puisque la référence ne provient pas du même corpus que les données de test, on ne peut pas évaluer le rappel précisément, mais on peut comparer les différentes techniques afin d'estimer le rappel; on ne peut pas évaluer la précision exactement non plus, aussi l'auteur emploie-t-il d'autres techniques d'évaluation. La dernière technique d'évaluation utilisée est originale : on compare tous les candidats extraits (par des moyens statistiques) à une liste de patrons morphosyntaxiques valides. L'auteur obtient ainsi une meilleure idée de la précision de sa technique d'extraction et de ses différentes composantes.

Cette évaluation est aussi intéressante parce qu'elle tient compte de l'application : « Evaluating precision and recall of such lists is far from trivial. Evaluation depends on the purpose of the application, i.e. on the type of result that is aimed at. This investigation focuses on the extraction of phrasal terms for building terminology databases » (id. : 206).

L'application fixée pour cette évaluation est donc la construction d'une banque de terminologie.

2.2.2.5 Sauron (2002)

Sauron (2002) a adapté la méthodologie EAGLES, elle-même basée sur la norme ISO 9126 et destinée aux outils d'aide à la traduction et à la rédaction, cette fois aux fins de l'évaluation des extracteurs de terminologie. Contrairement à la norme ISO 9126, qui définit des qualités intrinsèques et extrinsèques à évaluer, la méthode que propose Sauron est de type boîte noire, ne portant que sur les caractéristiques du logiciel qui sont visibles à l'utilisateur. Comme la méthode de EAGLES, celle de Sauron est centrée sur les besoins d'un utilisateur : « a particular evaluation is essentially a function of the needs of a user, who wants to know what the software being evaluated might give him in terms of support for a particular task to be accomplished » (Sauron 2002 : 2).

Sauron conserve un certain nombre des caractéristiques de qualité proposées par l'ISO (la fonctionnalité, la facilité d'utilisation, la fiabilité et l'efficacité), décompose chacune d'elle en attributs, et en dérive des critères susceptibles d'être mesurés :

- Présence de toutes les fonctions décrites dans la documentation
- Cohérence terminologique de l'interface
- Formats de fichiers pris en charge
- Langues prises en charge
- Performance (F-mesure)
- Compatibilité avec d'autres outils de traduction assistée par ordinateur
- Fonction d'enregistrement automatique
- Disponibilité de : documentation écrite, aide interactive, tutoriels, forums de discussion, groupes d'utilisateurs
- Exhaustivité et clarté de la documentation
- Convivialité de l'interface

Une pondération spécifie le poids de chacun des critères. Celui qui reçoit le plus de poids selon la pondération de Sauron est la précision des listes produites par l'extracteur,

qu'elle mesure en comparant la sortie avec une liste de référence et en calculant la F-mesure.

Un des points intéressants de l'étude de Sauron est qu'elle fournit quelques renseignements sur la construction de la liste de référence, chose rare dans la littérature :

> In order to reduce subjectivity, a certain number of rules were followed during the manual extraction process. We considered uniterms, including acronyms, as well as complex words to be valid terminological resources. We also allowed expressions which were not strictly terms but which did present translation difficulties. We put no maximum length on the terms to be extracted, and specified no particular syntactic patterns (id. : 13).

Elle fournit également un renseignement important sur la façon dont le degré de correspondance entre la liste de candidats termes et la liste de référence est évalué, à savoir qu'un terme complexe qui n'a été extrait que partiellement est tout de même considéré valide; Sauron justifie cette décision en affirmant qu'un candidat partiellement extrait peut toujours être utile à l'utilisateur, qui peut retourner aux contextes afin de vérifier s'il participe à un terme. En effet, les différentes façons de mesurer la correspondance ont une incidence sur les résultats de l'évaluation, et cela devrait être précisé dans tout travail d'évaluation des extracteurs.

2.2.2.6 Enguehard (2003)

Enguehard (2003) propose une démarche collaborative pour l'évaluation des systèmes de reconnaissance de termes, qui pourrait éventuellement être appliquée aux extracteurs de termes. L'approche consiste à comparer la sortie de différents outils et de construire une référence consensuelle. La sortie de chaque outil est ensuite comparée à cette référence afin d'évaluer sa performance. Cette approche est intéressante parce qu'elle est automatisable et très peu coûteuse, mais elle évacue du processus d'évaluation tout jugement humain de la pertinence des termes.

2.2.2.7 Drouin (2003)

Drouin (2003) évalue manuellement une technique d'extraction, implémentée dans TermoStat. Il utilise une approche en boîte transparente, observant les *pivots lexicaux spécialisés* qui sont à la base de sa technique (voir section 1.2.1.2), plutôt que les candidats qui sont produits en sortie du système. Ces pivots lexicaux spécialisés (PLS, en anglais SLP) sont analysés par trois terminologues qui travaillent dans le domaine qui caractérise le corpus : « The terminologists were told to consider an SLP as valid if the item was representative of the domain or the main topic of the corpus. Thus, this is not an evaluation of the precision from a terminological point of view but rather of the relevance of the SLPs » (Drouin 2003 : 102). Il évalue également dans quelle mesure la précision globale du système est influencée par une contrainte voulant que seuls les candidats contenant une PLS soient retenus. De plus, il évalue la *stabilité* des listes de PLS produites par TermoStat en faisant varier la taille du corpus de référence utilisé pour calculer la spécificité des PLS.

Enfin, Drouin effectue une double évaluation (automatisée et manuelle) de la précision de TermoStat. La version automatique consiste à comparer les candidats termes à une liste de termes extraits d'une base de données terminologiques; cette base, ainsi que le corpus de test, est fournie par une entreprise de télécommunications, ce qui assure une certaine cohérence entre le corpus d'analyse et la liste de référence. L'absence d'un recoupement total motive l'auteur à ajouter une étape de validation manuelle : les candidats termes qui n'apparaissent pas dans la liste de référence sont soumis à une validation manuelle par les terminologues.

> As with any validation process that relies on humans, we cannot assert that the results obtained are free of subjectivity. We strongly believe that different terminologists will identify different terms in the same document and that the same phenomenon could be observed with one terminologist looking at the same corpus over a period of time. It would be very interesting to be able to take into account the human influence over the validation process [...] (id. : 107).

Ainsi, Drouin souligne la part de subjectivité que comporte toute évaluation qui s'appuie sur les jugements d'un humain, en l'occurrence des terminologues.

2.2.2.8 Lemay et al. (2005)

Lemay et al. (2005) présentent une comparaison de deux techniques d'extraction de termes simples. Ces techniques ayant pour but de faciliter le travail terminographique (la confection de dictionnaires spécialisés), le choix des listes de référence dans l'évaluation de ces techniques est fait en conséquence : il s'agit en effet de deux dictionnaires spécialisés. Le choix des dictionnaires est justifié : ceux-ci contenaient seulement des termes simples, leur couverture était la meilleure et la date de parution était proche de celle des textes dans leur corpus de test. Trois échantillons sont retenus, composés des entrées commençant par les lettres A, C et P. En plus de comparer les candidats termes à la liste de référence, les auteurs comparent les candidats extraits par les deux méthodes entre eux. Ce deuxième type d'évaluation leur permet de conclure que les candidats qui sont fournis par les deux méthodes d'extraction ont une meilleure chance d'apparaître dans leurs listes de référence (Lemay et al. 2005 : 233).

Les auteurs évaluent le rappel de deux façons. Ils comparent d'abord les candidats à la liste de référence complète, puis aux entrées de la liste qui se retrouvent également dans le corpus d'analyse. Cette distinction est importante si on se sert d'une ressource extérieure comme liste de référence, plutôt qu'une liste élaborée à partir du corpus d'analyse lui-même. Un extracteur ne peut évidemment pas identifier des termes qui ne se retrouvent pas dans le corpus qu'on lui soumet. Les auteurs pèsent les avantages et inconvénients de leur approche à l'évaluation :

> We are aware that this evaluation procedure has a number of weaknesses, since most dictionaries have not been completely compiled according to observations derived from corpora and their contents rely on decisions made by specialized lexicographers. However, we are interested in assessing the value of our comparisons in a terminological setting. We believe that the contents of specialized dictionaries are a reflection – albeit imperfect – of the

> needs of terminologists. In addition, this evaluation allows us to take into account recall and precision (id. : 232).

Entre autres, ils soulignent l'importance de prendre en compte l'application lors de la confection de la liste de référence.

2.2.2.9 Vivaldi & Rodríguez (2007)

Vivaldi & Rodriguez (2007) présentent une synthèse des travaux sur l'évaluation des extracteurs, en soulignant les lacunes des techniques utilisées dans le passé. Ils proposent quelques idées pour combler ces lacunes, et les appliquent à un extracteur qu'ils ont conçu, YATE. La lacune principale qu'ils identifient est que les concepteurs d'extracteurs choisissent souvent d'évaluer seulement la précision, étant donné qu'une évaluation du rappel repose sur l'existence d'un corpus d'analyse dont toutes les unités terminologiques ont été recensées manuellement. Au coût d'une telle entreprise on doit ajouter la difficulté d'obtenir un consensus sur le contenu de la liste de référence :

> The main difficulty in evaluation comes from the impossibility of building a fair gold standard against which the results of the system we wish to evaluate can be compared. This difficulty is due to the very low agreement among human evaluators when faced with non-trivial decisions. This lack of agreement comes, in turn, from the difficulty of defining the set of measurable properties that contribute to the system's quality (Vivaldi & Rodríguez 2007 : 227).

Or, on peut augmenter l'accord entre annotateurs en établissant des critères spécifiques de sélection des termes. Les auteurs en proposent trois auxquels tout candidat doit répondre pour être considéré un terme, à savoir la cohérence interne de l'unité, son potentiel terminologique, et son degré de spécialisation (un terme utilisé dans un seul domaine aura une plus forte probabilité d'être un terme), mais ces critères peuvent être difficiles à mesurer.

L'approche qu'ils adoptent consiste à dépouiller manuellement un texte qui sera ensuite soumis à un extracteur afin d'évaluer la précision et le rappel. Les termes sont

recensés par des experts du domaine en question. Puisque les listes établies par les experts ne se correspondent pas, les auteurs choisissent de considérer comme valide tout candidat qui a été retenu par au moins un expert. Puis, ils mesurent la précision et le rappel de leur extracteur, en utilisant une approche en boîte transparente : ils n'évaluent pas seulement les données présentées à l'utilisateur, mais les données fournies par différentes modules de leur programme, afin d'identifier les sources d'erreurs.

Les auteurs concluent que l'évaluation des extracteurs souffre de l'absence des ressources d'évaluation qu'on retrouve dans différentes branches du traitement automatique de la langue, tels que les corpus annotés :

> the only way to improve term extraction is to define, as in other areas of NLP, some kind of gold standard to which different TEs can compare. Such a standard should include at least decisions about the corpus and the criteria used to design it, metrics to be used and evaluation protocols for the terms included in the corpus. Due to the great variability of TE techniques and the low agreement between terminologists and domain experts on what term candidates should be treated as terms, such a gold standard should be highly parameterizable […] (id. : 244)

Les auteurs suggèrent donc qu'un moyen de compenser le faible degré d'accord entre annotateurs est de créer une liste de référence qui peut être paramétrée. Une méthodologie comme celle que nous présentons, basée sur une annotation fine des termes contenus dans un corpus, permet justement de paramétrer la référence en fonction de différents critères, tels que le type de termes (simples ou complexes) et de variantes terminologiques à retenir, ainsi que des indices statistiques tels que la fréquence des unités. Nous nous pencherons sur cette question dans l'analyse de nos résultats, au chapitre 4.

2.2.2.10 Nazarenko et al. (2009)

Un protocole novateur d'évaluation a été proposé par Nazarenko et al. (2009). Les auteurs proposent d'abord de décomposer l'extraction en trois tâches fondamentales, l'extraction proprement dite, le regroupement des variantes terminologiques et la

structuration de candidats termes; cette approche a d'ailleurs été adoptée dans le cadre de la campagne CESART (voir section 2.2.3.2). Les auteurs cherchent à évaluer les trois fonctions de façon autonome, ce qui faciliterait la comparaison de différents extracteurs, qui n'offrent pas toujours les mêmes fonctionnalités.

En ce qui concerne l'extraction de candidats termes, les auteurs proposent une redéfinition des métriques qui servent à évaluer cette tâche, des propositions qui « ne préjugent pas du choix ou du mode de construction de la [liste de] référence » (Nazarenko et al. 2009 : 267). En effet, il ne peut exister une liste de référence unique et absolue pour les extracteurs de termes, entre autres parce que son contenu dépend de l'application donnée à l'extracteur; il importerait donc d'ajuster la référence en fonction de la sortie de chaque outil évalué, pour trouver une correspondance maximale. Les métriques classiques de précision et de rappel supposent un statut binaire (oui/non) aux termes à valider, alors que la pertinence d'un candidat n'est pas aussi simple à évaluer. La métrique proposée pour l'extraction évalue la pertinence des candidats sur une échelle graduée plutôt que par un choix binaire, en exploitant la notion de distance d'édition (le nombre d'opérations à effectuer pour que deux chaînes de caractères se correspondent), et elle transforme la sortie de l'extracteur « pour trouver sa correspondance maximale avec la référence, ce qui revient à ajuster la sortie au type de la référence » (Nazarenko et al. 2009 : 274).

> Comme plusieurs termes de la sortie peuvent correspondre au même terme de la référence, on peut les considérer en bloc et nous proposons de calculer les mesures de précision et de rappel non pas directement sur S [la sortie] mais sur une partition de S qui est définie relativement à R [la référence]. Cette partition P(S) est telle que toute partie p de P(S) est soit un ensemble de termes de S qui se rapprochent du même terme de R avec une distance inférieure au seuil σ, soit composée d'un terme [unique] (ibid.).

Ce partitionnement de la liste de candidats en ensembles de candidats qui partagent un certain degré de similarité avec un terme de la référence, calculée au moyen de la distance d'édition, permet d'obtenir une correspondance maximale entre la sortie et la référence. Cette technique pallie la relativité de la notion de liste de référence, et permet

d'utiliser différents types de listes de référence sans favoriser arbitrairement un extracteur plutôt qu'un autre (Nazarenko & Zargayouna 2009 : 301).

2.2.2.11 Love (2000)

Love (2000) présente une évaluation technocentrée comparative de deux extracteurs faisant appel à une liste de référence construite manuellement. L'évaluation est quantitative et qualitative, car une fois la référence construite et les métriques calculées, Love analyse les sources d'erreurs, les points forts et faibles de chaque système. L'accent ici est placé sur l'évaluation proprement dite, plutôt que sur la construction de la référence. Globalement, sa méthodologie consiste à entrer les termes sélectionnés manuellement dans une banque à l'aide d'un logiciel qu'elle-même a conçu, puis une fois l'extraction effectuée, à créer une nouvelle fiche pour chaque candidat terme qui ne fait pas partie de la référence (bruit). La fiche de chaque terme contient sa graphie (incluant la casse et les éléments superflus), un numéro d'identification, le ou les documents où le terme se trouve, son domaine d'usage (tiré de la banque de terminologie Termium et éventuellement validé par un expert lorsqu'il ne s'agissait pas de l'informatique), sa structure syntaxique, le nombre de mots qu'il contient, ainsi que divers commentaires. Pour les candidats valides, on indique quel outil les a repérés. Enfin, pour les termes de la référence qui n'ont pas été extraits par l'un ou l'autre des logiciels, on indique qu'il s'agit du silence. L'analyse des résultats peut alors être effectuée. Love calcule les métriques afin de quantifier la performance de chaque outil, puis analyse les données afin d'effectuer une comparaison qualitative des outils.

L'aspect du travail de Love qui nous intéresse particulièrement est la description de la sélection des termes et du traitement des variantes, deux aspects qui sont très rarement abordés dans la littérature. En ce qui concerne la sélection des termes, plusieurs points retiennent notre attention. D'abord, des critères, empruntés de Sager (voir section 1.1.3.5), sont fixés pour la sélection des termes, mais Love affirme ne pas les avoir suivis à la lettre, puisqu'ils concernent la formation de nouveaux termes, et qu'il valait mieux ne pas occulter les termes qui n'y adhéraient pas. Ses critères principaux sont énumérés, et concernent des

notions telles que la transparence d'un terme, son degré de motivation, son mode de formation et la cohérence de l'ensemble des termes recensés. De plus, Love ne restreint pas son choix des termes à un domaine particulier (son travail portant sur des textes du domaine de l'informatique), donc des termes d'autres domaines sont inclus dans la référence. Elle justifie cette approche en affirmant qu'elle s'apparente à celle des traducteurs; d'ailleurs, les deux extracteurs évalués font partie de suite d'outils de traduction (un de traduction automatique et l'autre de traduction assistée par ordinateur). On peut donc supposer que le choix des termes est le reflet d'une application, soit la traduction. Par ailleurs, Love ne retient pas uniquement des syntagmes nominaux, mais ne précise pas exactement toutes les structures qu'elle considère valides.

Pour ce qui est de la variation terminologique, Love présente 9 types de *redondance* (qu'on peut assimiler à la variation) qu'elle essaie d'éviter (Love 2000 : 91). Les variantes sont donc consignées à un champ réservé aux commentaires dans les fiches qu'elle élabore. « The potential for term redundancy was an issue to which we paid special attention because our corpus of manually extracted terms cannot contain doubles or even variations of the same term, since such variations would skew our results » (ibid.). Les neuf sources de redondance (variation) sont examinées pour voir comment chacune est gérée par les deux extracteurs. Love touche brièvement au concept de la réduction des termes complexes (voir section 1.3), mais ne s'y attarde pas, et ce type de variation ne semble pas avoir été décrit dans sa base de données. Par ailleurs, les termes complexes coordonnées sont reconstruits; c'est donc la forme complète qui est consignée dans la banque de termes. Lors du calcul des métriques, un traitement particulier des termes complexes coordonnés est effectué, parce qu'un des deux outils (ATAO) propose des candidats termes coordonnés en bloc.

L'analyse des résultats de son évaluation contient quelques points intéressants : notamment, elle évalue le rappel autour de 35 à 40 %, ce qui concorde avec ce qu'on trouve dans d'autres ouvrages où on évalue les extracteurs en question. En ce qui concerne le

calcul des métriques, il est à noter que Love considère certaines correspondances partielles valides (absence ou ajout d'un trait d'union ou la présence du *-s* indiquant le pluriel).

Love (2000) propose quelques améliorations qui pourraient être faites à sa méthodologie. Elle affirme qu'exploiter des renseignements sur la fréquence des termes et des candidats, ce qui n'a pas été fait, pourrait être avantageux (notamment pour identifier les causes du bruit et du silence). Elle affirme que cela nécessiterait un plus gros corpus; en outre, sa méthode de gestion des termes (programme conçu sur mesure, base de données de type Access) ne permet pas facilement de le faire. Puisque la référence se présente sous la forme d'une banque de termes plutôt que d'un corpus annoté, il n'est pas possible de paramétrer la référence en exploitant des renseignements sur la fréquence, entre autres.

2.2.3 Campagnes

Dans de nombreuses branches du traitement automatique de la langue, des campagnes d'évaluation ont lieu régulièrement pour mettre à l'épreuve les dernières techniques. Lors de ces campagnes, des concepteurs soumettent leur système à un protocole d'évaluation, et différentes ressources sont mises à la disposition des participants. Bien qu'il n'existe pas de campagne périodique pour l'extraction de terminologie, quelques-unes ont tout de même eu lieu. Dans cette section, nous décrivons les trois campagnes d'évaluation qui ont été mises sur pied.

2.2.3.1 ARC A3

La première campagne d'évaluation des extracteurs a eu lieu dans le cadre des Actions de recherche concertées (ARC) de l'Agence universitaire de la francophonie, de 1996 à 2000 :

> L'objectif de l'action était de mettre au point un cadre méthodologique et un protocole pour l'évaluation des différents systèmes sur des données communes. La campagne a commencé en octobre 1996 et sa première édition (campagne à blanc en 1997) a fait l'objet de nombreuses publications et communications. La dernière campagne a été lancée en 2000 (Timimi & Mustafa el Hadi 2008 : 72).

La méthodologie élaborée concernait non seulement des extracteurs de termes, mais différents systèmes d'acquisition de ressources terminologiques, notamment des extracteurs de relations sémantiques. Deux corpus d'analyse ont été compilés et un protocole d'évaluation défini :

> Nous pouvons décrire l'évaluation entreprise lors de la denière campagne, en 2000, comme une évaluation multidimensionnelle : qualitative, comparative, opaque (boîte noire) et par adéquation relative à trois champs d'applications. Notre protocole d'évaluation était fondé sur l'appariement entre les résultats produits par les systèmes et les listes de référence existantes. Outre l'appariement effectué par des juges (humains) sur un échantillon des résultats, nous avons conçu un programme (*EvalTerm*) qui établit un appariement de tous les résultats fournis par les logiciels et des listes de référence proposées par nos experts ou déjà existantes (cas du thésaurus *Motbis* ou de la liste *Francis* fournie par l'INIST). Les métriques retenues pour évaluer la performance des systèmes ont été les traditionnelles mesures de rappel et de précision (id. : 73).

Les listes de référence utilisées sont donc de deux types. Des listes ont été extraites des ressources terminologiques existantes, et d'autres ont été compilées par des experts dans le champ d'application donné. Par exemple, deux listes ont été établies pour la traduction, l'une par un novice et l'autre par un traducteur professionnel chevronné. Selon les organisateurs, l'utilité des listes de référence variait d'une application à l'autre : « As far as indexing is concerned the interest of these lists is quite limited and we think that a lot of time has been lost in drawing them up and even grooming them » (Mustafa El Hadi et al. 2001 : 5).

Les organisateurs ont noté de nombreux facteurs qui ont rendu leur entreprise difficile, notamment le fait que les systèmes pouvait appartenir à différentes catégories et

fournissaient donc des données de types différents (termes, relations ou réseaux sémantiques). En outre, les extracteurs de termes n'ordonnançaient pas leur sortie de la même façon, et n'offraient généralement pas une interface de gestion des résultats, ce qui compliquait l'appariement nécessaire pour mesurer la performance. Les organisateurs ont donc « adopté une approche essentiellement qualitative (grâce à une expertise humaine) » (Timimi & Mustafa el Hadi 2008 : 73). De plus, ils soulignent que des modifications devaient parfois être apportées aux systèmes afin que les données d'entrée soient les mêmes pour tous les systèmes, ce qui pouvait avoir un effet favorable ou défavorable sur la performance d'un système (Mustafa El Hadi et al. 2006 : 947). Ils soulignent également que, puisque le protocole d'évaluation résulte d'une entente entre les organisateurs et les participants, les résultats ne peuvent pas être considérés comme un indicateur fiable de la performance des systèmes. Enfin, la quantité importante de données fournies par les systèmes et le recours à des juges humains ont fait en sorte que les données de sortie devaient être échantillonnées.

2.2.3.2 CESART

À la lumière des observations faites lors de la campagne ARC A3, une deuxième campagne appelée CESART a été mise sur pied en 2003 dans le cadre du projet de recherche Technolangue. Étant donné que les outils évalués pouvaient offrir différentes fonctionnalités et viser une variété d'applications, les organisateurs ont décidé d'élaborer différents protocoles d'évaluation en fonction du type d'outil concerné. « Les enseignements que nous avons tirés à partir de ces deux constats ont eu un impact sur le choix du protocole CESART, à savoir la mise en place d'un protocole par catégorie d'outils et par rapport à des usages possibles » (Timimi & Mustafa el Hadi 2008 : 73). En effet, le rôle de l'application dans l'évaluation des extracteurs est un des facteurs importants ayant guidé l'élaboration du protocole de CESART : « L'élément le plus saillant a été de définir des contextes d'usage, élément qui était totalement absent lors de l'ARC A3 » (id. : 74). Trois applications ont donc été fixées, à savoir l'extraction de termes pour l'enrichissement de ressources terminologiques, l'indexation contrôlée et l'enrichissement d'index, et

l'extraction de relations sémantiques (Timimi 2006 : 899). Des méthodes d'évaluation ont été fixées pour chacune de ces trois tâches. Nous ne décrirons que la méthode portant sur l'enrichissement de ressources terminologiques.

Le protocole de CESART a été conçu pour quatre catégories d'outils (bien que toutes les catégories n'aient pas été représentées, faute de participants) : les extracteurs de termes, les extracteurs de relations sémantiques, les extracteurs de relations morphosyntaxiques et les éditeurs d'ontologies. Seule la langue française a été traitée. Trois corpus ont été construits, dont un a servi pour l'entraînement des systèmes participants qui fonctionnaient par apprentissage-machine, et les deux autres pour les tests. Les corpus sont construits en fonction de critères spécifiques :

> Le corpus doit vérifier d'après (Pincement, 1999), trois types de conditions : *signifiance, acceptabilité et exploitabilité* en plus de la *pertinence* par rapport à un objectif d'analyse. L'ensemble de ces conditions est nécessaire pour sa réutilisabilité. De même, la production de ces corpus doit respecter la règle d'*homogénéité :* les documents retenus doivent être homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors de ces critères de choix (Timimi 2007 : 149).

Comme dans le cas de l'ARC A3, les données fournies par les systèmes ont été échantillonnées afin de faciliter l'appréciation des résultats par des juges humains. Les 1000 premiers candidats[10] fournis par chaque système ont donc été évalués manuellement par des experts du domaine concerné ayant également des connaissances en indexation ou en documentation (ils étaient en fait affiliés aux organisations ayant produit les ressources à

[10] Il est à noter que les différents travaux portant sur la campagne CESART ne s'accordent pas sur tous les points. Par exemple, la signification des cotes de pertinence données par les évaluateurs humains est différente dans Mustafa el Hadi et Chaudiron (2006) et dans Timimi & Mustafa el Hadi (2008). De plus, dans le premier ouvrage, on affirme que les 10 000 premiers candidats sont évalués manuellement lorsqu'ils ne sont pas recensés dans la liste de référence, alors que le deuxième affirme que chacun des 1000 premiers candidats est évalué manuellement. Il semble pourtant qu'ils traitent la même évaluation, puisque le nombre de termes extraits par chaque système est le même dans les deux ouvrages. Bref, nous nous fions à Timimi & Mustafa el Hadi (2008) en cas de désaccord.

la base des listes de référence). Ils attribuent à chaque candidat « une note allant de 0 à 4 selon son exactitude » (Timimi & Mustafa el Hadi 2008 : 79). Cette cote permet de calculer la précision à différentes granularités : termes présents dans les listes de référence, termes absents mais pertinents, termes partiellement extraits, etc. De plus, une évaluation automatique de tous les candidats termes a été effectuée par appariement avec des listes de référence préétablies.

> Certes, cette approche a le défaut de ne traiter que des chaînes de caractères, mais elle a le mérite de traiter l'intégralité des résultats donnés par les systèmes (et de ne pas se contenter des échantillons) et reste toutefois un indicateur sur le comportement des systèmes face à l'ensemble du corpus. L'évaluation automatique permet également d'approcher la valeur du *rappel,* une mesure impossible avec un travail humain (id. : 76).

L'évaluation repose sur des métriques, mais les organisateurs ajoutent « quelques éléments de cadrage » (id. : 80) afin de pallier le côté réducteur de la comparaison avec des listes de référence : ils évaluent ainsi l'accord entre les juges humains, utilisent une variété de listes de référence, et créent un « référentiel de consensus à partir des résultats communs à la plupart des systèmes » (ibid.).

Les listes de référence ont été construites à partir de ressources terminologiques existantes, un thésaurus du domaine des sciences de l'éducation et une terminologie du domaine médical. Les organisateurs ont noté quelques problèmes que posent les listes de référence :

> la constitution de ces listes de référence et la méthode d'évaluation elle-même, posent différents types de problèmes […] : d'une part, ces listes contiennent des éléments qui ne sont pas forcément extraits par les systèmes (verbes et collocations par exemple) ; d'autre part, il y a des différences entre les listes produites par les utilisateurs humains (Timimi & Mustafa el Hadi 2008 : 77).

À ces difficultés s'ajoute une variabilité, en fonction de l'application visée, du nombre d'unités qu'on vise à extraire et de leur statut (représentativité d'un domaine dans le cas de

la terminologie et d'un document dans le cas de l'indexation). Malgré ces difficultés, les organisateurs expliquent que les listes de référence demeurent très utiles, parce qu'elles sont bien adaptées aux évaluations en boîte noire et aux évaluations où on a seulement accès aux listes de candidats termes fournis par un système.

Une des observations qui ressort de cette campagne est qu'elle s'inscrit dans un paradigme d'évaluation des logiciels qui est fortement technocentré. Les organisateurs soulignent l'intérêt de se tourner vers un modèle centré sur l'utilisateur, sur son interaction avec le logiciel, sur son degré de satisfaction ou de résistance face à un outil, etc. Ils affirment que la prise en compte des besoins de l'utilisateur est l'un des apports les plus importants de cette campagne. On présume qu'il est question ici d'une expérience qui a été entreprise afin d'évaluer l'adéquation des ressources enrichies à l'aide des extracteurs aux besoins d'utilisateurs potentiels. Or, cette expérience n'est pas bien documentée. À notre connaissance, l'explication la plus complète qu'on en donne se résume à ceci :

> Concerning the medical corpus the Rouen Hospital University team is assessing the extracted terminology for two tasks (see above). The idea was to measure the adequacy of the tools to their daily work, which is free indexing, enriching the *CISMeF* database and the related thesaurus. The objective of *CISMeF* is to describe and index the main French-language health resources to assist health professionals and consumers in their search for electronic information available on the Internet
>
> As for the second corpus, specialists in educational science are testing the adequacy of the tools in their daily work, i.e. updating and enriching the terminology for CNDP (*Centre National de Documentation Pédagogique*) and the related indexing tools.
>
> In these two use cases the idea to assess to what extent the tested tools are adapted to accomplishing theses tasks though they are not really designed for them. They are considered as generic tools. It would be however interesting to measure users satisfaction when using these tools […] (Mustafa el Hadi et al. 2006 : 947-948)

2.2.3.3 TMREC

De 1998 à 1999, dans le cadre de la conférence NTCIR Workshop 1, s'est organisé un workshop articulé autour de trois tâches connexes : l'extraction de termes, l'extraction de mots-clés (indexation automatique) et l'analyse de rôles (extraction du sujet d'un article scientifique et de la méthode mise en application). Le point focal de ce workshop, appelé *TMREC* (pour *term recognition*), était l'extraction de termes, et c'est cette tâche qui a attiré le plus grand nombre de participants : 8 équipes ont participé, surpassant le nombre de participants à ARC A3 et à CESART. L'objectif n'était pas de déterminer la meilleure technique, mais de réaliser une analyse comparative des différentes techniques utilisées, en mettant à la disposition des participants un jeu de test commun, et de favoriser les échanges entre les chercheurs : « our intention is to stimulate the constructive discussion of both technical and conceptual aspects of the automatic processing of terms and terminology » (Kageura et al. 1999a : 415). L'évaluation est donc qualitative, car les organisateurs cherchent à décrire les listes de candidats produites par chaque système plutôt que leur assigner une cote. Elle comporte tout de même une évaluation de la précision. Celle-ci repose sur deux listes de référence : l'une est construite manuellement à même le corpus d'analyse, et enrichie à l'aide de l'index d'une encyclopédie du domaine concerné (l'intelligence artificielle); ces deux catégories représentent 8834 et 671 termes respectivement (Hisamitsu et al. 2000 : 223); la deuxième liste est établie en sélectionnant les candidats qui sont communs à un nombre *n* des listes produites par les extracteurs, la présupposition étant que les candidats qui sont extraits par plusieurs extracteurs ont une plus forte probabilité d'être des termes (une technique semblable à celle proposée par Enguehard (2003)). En ce qui concerne les termes sélectionnés manuellement, on ne décrit pas le protocole de sélection des termes, mais on précise que, la définition de « terme » étant toujours floue, ils n'ont pas un statut définitif : « Our Manual-Candidates and Index-Candidates are reflections of a *possible* conceptualisation of 'terms' and 'terminology', but we do not claim that we can give any special status to it. Rather, it is intended to be used for stimulating the further discussion » (Kageura et al. 1999b : 431-432). Les organisateurs insistent sur le fait que les listes de référence n'ont pas un caractère absolu, vu que la

terminologie d'un domaine dépend de différentes présuppositions théoriques et considérations pratiques.

> It should be emphasised that the very aim of the TMREC task, though it was 'contest'-style, is to promote discussion, and not to see which methods are better. We nevertheless prepared the candidate sets because they give a concrete basis on which the participants can constructively discuss which methods are suitable for what kind of purpose, etc. This in turn means that these two Candidates do not in any sense have a prescriptive status (Kageura et al. 2000 : 159-160).

Ils appellent donc les termes sélectionnés manuellement des *candidats termes*, ajoutant à cette appellation une majuscule pour les distinguer des candidats termes produits par les extracteurs. « This, however, in no way implies that the candidate lists prepared by the TMREC group have any prescriptive status » (Kageura et al. 1999b : 419).

Une collection de plus de 330 000 textes (des sommaires d'articles présentés dans des conférences, dont plus de la moitié sont des textes parallèles en japonais et en anglais) a été préparée dans le cadre de NTCIR 1 et distribuée pour les chercheurs participant aux différents workshops. Seule une petite partie de ce corpus a été adaptée à la tâche d'extraction de termes, mais les participants étaient libres d'utiliser tout le corpus (Kageura et al. 2000 : 158-159). La construction du corpus est décrite dans Koyama et al. (1998), l'accent étant placé sur le problème de la segmentation en mots des textes japonais.

Il est intéressant de noter le domaine que les organisateurs ont choisi pour le corpus de TMREC, à savoir l'intelligence artificielle. Ce choix était effectivement motivé (Kageura et al. 2000 : 158) : les discussions entre les participants seraient plus constructives du fait qu'ils avaient tous une bonne connaissance du domaine; il serait aussi plus facile de construire manuellement une liste de référence, les organisateurs jouant à la fois le rôle de terminologue et d'expert du domaine.

Bien que les listes de référence produites pour TMREC n'aient pas été conçues dans le but d'offrir une référence absolue, elles permettent d'évaluer la performance des extracteurs en mesurant différentes métriques. Par exemple, Mima & Ananiadou (2000)

évaluent leur technique d'extraction qui exploite la C-value et la NC-value en mesurant la précision de différentes façons : précision absolue, précision graduée en fonction des *n* meilleurs candidats, précision en fonction du rappel. Ils comparent également leurs résultats à ceux obtenus en considérant seulement la fréquence des unités. Nakagawa (2000), entre autres, illustre qu'un corpus d'analyse et une liste de référence peuvent servir à une évaluation de progression en boîte transparente : cette méthode lui permet de calculer le rappel et la précision qu'on obtient avec différents patrons morphosyntaxiques et différentes façons de calculer le score de chaque candidat. Nakagawa montre les résultats obtenus par 4 systèmes, dont le sien, en termes de précision, de rappel et de f-mesure, en considérant tantôt les correspondances exactes seulement, tantôt les correspondances partielles.

2.2.3.4 Synthèse

Les campagnes d'évaluation des extracteurs ne sont pas à la hauteur des efforts déployés dans d'autres branches du TAL. En outre, les campagnes qui ont eu lieu ont attiré un nombre relativement faible de participants; aussi n'existe-t-il pas actuellement une campagne périodique qui vise à mesurer les progrès qui se font dans le domaine de l'extraction de termes. D'une façon plus générale, il n'existe pas de cadre d'évaluation standard pour l'extraction de termes, ce qui nous empêche d'avoir une idée claire des progrès qui ont été réalisés :

> Malgré les années de recul et d'expériences accumulées, il est difficile de se faire une idée claire de l'état de maturité des recherches en terminologie computationnelle. À la différence de nombreux autres pans du TAL, il y a eu peu d'effort collectif pour définir un cadre d'évaluation adapté à ce type de travaux. Les raisons sont diverses. Issue au départ de pays francophones, la terminologie computationnelle n'a pas bénéficié de l'impulsion américaine pour tout ce qui touche aux tâches d'évaluation (Nazarenko et al. 2009 : 258).

Ils ajoutent des difficultés liées à la tâche d'extraction, et les différences qu'on observe entre les outils (ordonnancement des listes, fonctionnalités offertes) et la diversité des

applications de l'extraction, tous des facteurs qui rendent l'évaluation et les comparaisons difficiles. « Seule l'accumulation des évaluations faites sur des bases comparables (sur les mêmes tâches et avec les mêmes métriques) peut permettre de faire un état des lieux global de la terminologie computationnelle » (id : 279). L'établissement d'une méthodologie de construction de la liste de référence constitue une étape importante. Cela fera l'objet du chapitre 3.

2.3 Conclusion

Les techniques d'extraction ont évolué considérablement depuis 20 ans; de même, les techniques utilisées pour évaluer les extracteurs se sont multipliées, mais la construction de la liste de référence n'a pas encore été traitée en profondeur, surtout en ce qui concerne la sélection des termes et le traitement des variantes terminologiques, comme nous l'avons montré à la section 2.2. Ce travail vise donc la création d'une liste de référence compilée en fonction de critères précis et qui rend compte de l'importante variation qu'on observe quant aux réalisations des termes en contexte. Bien que la construction d'une telle référence soit coûteuse, comme l'ont souligné Vivaldi & Rodriguez (2007) et Tiedemann (2001), elle demeure le meilleur moyen d'évaluer automatiquement et objectivement la performance d'un extracteur de termes.

Pour situer notre méthodologie par rapport aux types d'évaluation présentés à la section 2.1, soulignons d'abord qu'elle visera une évaluation de type automatique. Cette approche a de nombreux avantages, notamment son objectivité et sa reproductibilité :

> une évaluation automatique (basée sur un algorithme d'appariement à des référentiels externes), semble être a priori un procédé d'évaluation plus efficace que l'expertise humaine dans la mesure où, plus encore que l'expertise humaine et l'alignement linguistique, elle garantit la reproductibilité de l'expérience et par là même, l'obtention de résultats objectifs […] et offre surtout une possibilité de procéder à des évaluations horizontales en raison de l'extensibilité du protocole (Timimi 2006 : 901).

Si l'on désire obtenir des résultats objectifs et pouvoir reproduire l'expérience à souhait, il semble donc avantageux de d'opter pour une évaluation automatique. L'évaluation automatique fournira un premier portrait objectif de la performance d'un extracteur, dont les parties saillantes pourront être analysées manuellement par la suite.

À la différence des méthodologies d'évaluation qui sont centrées sur l'utilisateur, telles que celles de L'Homme et al. (1996; voir section 2.2.2.1) et de Sauron (2002; voir section 2.2.2.5), notre méthodologie est technocentrée; elle est également de type objectif, portant sur la performance d'un logiciel plutôt qu'une appréciation subjective de son utilité. De, plus, elle vise une évaluation en boîte noire : elle permet de mesurer l'impact de modifications apportées à un extracteur, mais seulement à partir des listes de candidats termes produites en sortie; elle ne peut pas mesurer la performance à des étapes intermédiaires du traitement (lemmatisation, étiquetage morphosyntaxique, etc.). Enfin, elle est caractérisée par la prise en compte de l'application qui guide le dépouillement terminologique, comme celle de Tiedemann (2001), notamment. Cette application est la compilation d'un dictionnaire spécialisé de la mécanique automobile.

L'utilisation d'un corpus annoté en XML (voir chapitre 3) et la richesse des renseignements que nous y consignons au sujet de la variation permet de paramétrer la référence de différentes façons, contrairement à d'autres méthodologies, comme celle de Love (2000; voir section 2.2.2.11), et permettrait éventuellement de voir quel type de termes ou de variantes est moins bien traité par les extracteurs. Un programme d'évaluation basé sur cette référence pourrait facilement, au lieu de simplement calculer les métriques, énumérer tous les termes qui n'ont pas été retenus par l'extracteur et de regrouper automatiquement ces termes passés sous silence en classes selon leur fréquence et selon le type de variation, par exemple. Par ailleurs, contrairement à l'approche de Love, notre corpus annoté rendra compte de la réduction des termes complexes d'une façon systématique.

Comme Collier et al. (2001; voir section 2.2.2.3), notre balisage des occurrences de termes sera linéaire et ne permettra pas l'imbrication de termes à l'intérieur d'autres termes. Par contre, un traitement basé sur des *attributs* permettra de décrire différents types de variation terminologique, notamment les réductions anaphoriques ou par coordination ainsi que les disjonctions provoquées par l'insertion d'un sigle, d'un synonyme ou d'une paraphrase à l'intérieur d'un terme complexe. Ce type d'annotations permettra également de paramétrer la liste de référence, notamment pour spécifier si différentes sortes de variantes devraient être incluses dans celle-ci.

Le prochain chapitre portera sur notre méthodologie. Nous y décrirons comment nous avons procédé pour annoter un corpus afin de produire une liste de référence pour l'évaluation des extracteurs de termes.

3. Méthodologie

Ce chapitre porte sur le travail d'annotation de corpus que nous avons réalisé et qui nous a permis de construire une liste de référence pour l'évaluation des extracteurs de termes. Nous décrirons d'abord le corpus lui-même. La section 3.2 portera sur le repérage des unités terminologiques. Puis, nous décrirons le découpage et l'annotation des termes dans la section 3.3, qui montrera, entre autres, comment les variantes terminologiques ont été traitées. La section 3.4 portera sur la banque de termes que nous avons construite, qui contient notamment les liens de variation et de synonymie entre les unités annotées dans le corpus; c'est cette banque qui permettra de produire une liste de référence afin d'évaluer un extracteur de termes.

3.1 Le corpus

Le corpus que nous avons annoté est constitué des trois manuels de mécanique automobile suivants :

- HCW : Newton, T. (1999). How Cars Work. Vallejo : Black Apple Press, 96 p.
- CCB : Florence, M. & Blumer, R. (2002). The Everything Car Care Book: How to Maintain Your Car and Keep It Running Smoothly. Avon : Adams Media Corporation, 289 p.
- AF : Stockel, M. W., Stockel, M. T. & C. Johanson. (2000). Auto Fundamentals. Illinois : The Goodheart-Willcox Company, Inc., 607 p.

La langue des trois ouvrages est l'anglais américain, et leur niveau de spécialisation est moyen : ils sont rédigés par des experts à l'intention d'initiés qui veulent éventuellement devenir des experts. La situation de communication en est donc une d'enseignement plutôt que de communication entre spécialistes. Ainsi, les textes n'ont pas une densité élevée de termes très spécialisés, et fournissent des explications et des paraphrases qui ne se retrouveraient pas dans des textes rédigés par des experts pour des experts. Or, comme l'a

souligné Pearson (1998), il s'agit tout de même d'un des types de textes qui contiennent de nombreuses unités terminologiques.

Ces trois manuels totalisent 224 159 mots, comme le montre le Tableau II. Un extrait du corpus est présenté à l'annexe 1. Les ouvrages imprimés ont été numérisés, puis un traitement de reconnaissance optique des caractères a été effectué afin de convertir les images numérisées en texte. Toutes les figures ont été supprimées, afin de ne conserver que le texte. Puis, une entête XML a été ajouté à chacun des documents, puisque c'est ce format qui a été choisi pour l'annotation (voir section 3.3).

Ouvrage	Nombre de mots
HCW	10 851
CCB	75 087
AF	138 221
Total	224 159

Tableau II : Répartition des mots dans le corpus

3.2 Repérage des unités terminologiques

Comme nous l'avons mentionné au chapitre 1, des critères précis ont été fixés pour le repérage des unités terminologiques contenues dans le corpus. Ces critères sont d'ordre linguistique, formel et thématique. Dans les sections suivantes, nous décrirons chacun des critères utilisés.

Tout d'abord, mentionnons que notre approche consiste globalement à présumer que, étant donné la nature spécialisée des textes observés, toute unité est potentiellement terminologique, puis à filtrer les candidats en fonction de nos critères de sélection. En ce qui concerne le domaine d'usage des termes, nous n'exigeons pas que les termes soient utilisés exclusivement dans le domaine de l'automobile; le simple fait que ces termes soient utilisés dans des textes spécialisés portant sur la mécanique automobile suffit pour que l'on les considère potentiellement terminologique. Par exemple, le terme *piston* est utilisé dans

d'autres domaines tels que le génie mécanique, mais son utilisation dans le domaine de la mécanique automobile fait en sorte que nous le considérons comme un terme.

La liste des 50 termes de base les plus fréquents est présentée à l'annexe 3. Dans les sections suivantes, nous décrirons les critères que nous avons utilisés pour le repérage et le découpage des unités terminologiques dans le corpus.

3.2.1 Critères thématiques

Tout dépouillement terminologique est fait en fonction d'une application précise; de même, tout étalon utilisé pour évaluer automatiquement un extracteur doit être compilé en fonction d'une application (voir section 1.2.2). L'application ayant guidé le dépouillement terminologique dans le cadre de ce travail est la confection d'un dictionnaire spécialisé de la mécanique automobile. La thématique de ce dictionnaire serait la structure de l'automobile; ainsi, la grande majorité des termes retenus dénoteront des parties de l'automobile. Ainsi, les unités retenues doivent dénoter soit :

- un objet tangible faisant partie de la structure de l'automobile (p. ex. *piston* et *fuel injector*);
- un concept lié étroitement au fonctionnement du moteur (p. ex. *top dead center* et *power stroke*);
- une sorte d'automobile (p. ex. *sport utility vehicle* et *pickup truck*);
- un produit nécessaire au fonctionnement de l'automobile (p. ex. *oil* et *fuel*); nous excluons cependant les types spécifiques de tels produits (p. ex. *10W40* ou *synthetic automatic transmission fluid*).

En revanche, nous excluons les termes qui dénotent :

- les émissions de la voiture lorsqu'elle est en mouvement (p. ex. *carbon monoxide*);
- les unités de mesure (p. ex. *miles per hour* et *horsepower*);
- les dommages, l'entretien et les outils (p. ex. *dent*, *oil change* et *torque wrench*);
- les sources d'énergie qui ne sont pas spécifiques à l'automobile (p. ex. *kerosene*).

Ces critères thématiques sont liés au premier critère de sélection de termes énoncé par L'Homme (2004 : 64), qui stipule que l'unité lexicale doit avoir un sens qui est lié au domaine préalablement délimité pour le projet terminographique. Il n'est pas nécessaire que

son sens soit exclusif au domaine choisi (voir l'exemple de *piston* ci-dessus). Ce qui compte, c'est que l'unité ait un sens précis dans le domaine qui nous intéresse, et qu'elle soit attestée dans les écrits de ce domaine.

À ces critères thématiques s'ajoutent des critères de nature linguistique, qui seront décrits dans la section suivante.

3.2.2 Critères linguistiques

Les critères thématiques énoncés ci-dessous nous portent à ne retenir que des unités de nature nominale : les concepts concrets sont généralement dénotés par des unités nominales; par ailleurs, de manière générale, les dictionnaires spécialisés contiennent surtout des noms et des syntagmes nominaux.

Donc, nous n'annotons que des noms et des syntagmes nominaux. De plus, nous retenons deux critères lexico-sémantiques énoncés par L'Homme (2004) :

- Une unité qui présente un lien de dérivation morphologique avec un terme retenu en vertu du premier critère peut également être retenue, à condition qu'elle présente un lien sur le plan sémantique. De tels cas ne semblent pas être nombreux lorsqu'on s'intéresse uniquement aux unités nominales, du moins dans le domaine de l'automobile. On peut tout de même citer l'exemple de *cooling pump* et *coolant pump*, qui dénotent le même concept et dont les modificateurs entretiennent un lien de dérivation morphologique.
- Une unité qui partage un lien paradigmatique avec une unité retenue en fonction des deux autres critères est considérée comme un terme. Les relations paradigmatiques comprennent la synonymie, l'antonymie et la méronymie (relation partie-tout), entre autres. Ce critère a été particulièrement utile, étant donné que nous nous intéressons à la structure de l'automobile, qui fait notamment intervenir de nombreux liens de méronymie. Si le terme *air conditioning system* est retenu en vertu du premier critère, nous nous intéresserons à ses composantes : *condenser*, *compressor*, *evaporator*, etc.

Nous ne posons aucune exigence quant à la compositionnalité des termes. Une unité dont le sens peut être déduit du sens de ses composantes est tout de même admissible. Encore une fois, ce choix est motivé par notre application. En effet, la plupart des

dictionnaires spécialisés contiennent des unités compositionnelles et non compositionnelles. Toute partie d'un syntagme qui est porteuse de sens et qui ne brise pas la référence de l'unité à un concept précis de l'automobile peut être incluse dans le découpage.

Les paraphrases et les expressions descriptives appartenant plutôt à la langue générale, utilisées pour expliquer plutôt que dénoter, sont exclues. Cela vaut même si le sens de l'expression équivaut en contexte à celui d'un terme. Prenons par exemple la phrase suivante : *Small metal pipes called brake lines carry the hydraulic fluid*. Ici, nous ne considérons pas que *pipes* est un terme : il fait partie d'une paraphrase qui décrit le concept « brake lines », et nous ne l'annotons pas.

De plus, si une forme est une variante terminologique d'un terme préalablement sélectionné, nous l'annotons, à condition qu'elle dénote le même concept que le terme de base. Les types de variantes (voir section 1.3) que nous retenons sont les suivants :

- Variante orthographique : la présence ou l'absence d'un trait d'union ou d'un espace.
- Réduction anaphorique : la suppression d'un ou plusieurs constituants d'un terme complexe utilisé ailleurs dans le corpus.
- Réduction lexicale : lorsque la forme réduite a un degré élevé de lexicalisation, elle est traitée simplement comme un synonyme; le degré de lexicalisation est évalué en fonction de la fréquence, de la présence du terme de base ailleurs dans le corpus et de la présence de la forme réduite dans des ouvrages de référence.
- Coordination : lorsque deux ou plusieurs termes complexes qui partagent un constituant sont coordonnés, on assiste généralement à la suppression du constituant commun.
- Disjonction : insertion de signes de ponctuation (voir section 3.3.2.2).
- Surcomposition : ajout d'un constituant à un terme de base.
- Insertion : ajout d'un mot comme *type* à l'intérieur d'un terme complexe.
- Sigle ou acronyme : l'utilisation de la siglaison; celle-ci peut être seulement partielle dans le cas des termes complexes, c'est-à-dire que dans certains cas, ce n'est pas tous les constituants qui sont remplacés par le sigle.

Le Tableau III présente un exemple pour chacun des types de variantes retenues.

Type	**Terme de base**	**Variante**
Variante orthographique	*fuel injection system*	*fuel-injection system*
Réduction anaphorique	*exhaust valve*	*valve*
Réduction lexicale	*ignition system*	*ignition*
Coordination	*compression stroke*	*compression and exhaust strokes*
Disjonction	*big end bearing*	*lower, or big end, bearing*
Surcomposition	*cooling fan*	*engine cooling fan*
Insertion	*rotary engine*	*rotary type engine*
Sigle ou acronyme	*throttle valve cable*	*TV cable*

Tableau III : Exemples de variantes terminologiques

Dans tous les cas où plusieurs formes pointent vers un même concept, qu'il s'agisse de synonymie ou de variation terminologique, un terme de base est choisi, et toutes les variantes et les synonymes pointent vers le terme de base au moyen d'un système de renvois intégré dans notre banque de termes (voir section 3.4). Le terme de base est choisi en fonction de la fréquence de chacune des formes dans le corpus, et en fonction de la forme qui apparait le plus souvent comme vedette dans nos ouvrages de référence (voir section 3.2.4). Dans la banque de termes, les variantes par insertion et par surcomposition ainsi que les réductions lexicales seront simplement appelées des synonymes; les réductions anaphoriques, les sigles et les variantes orthographiques sont appelées comme telles.

En ce qui concerne les réductions anaphoriques, nous annotons non seulement les cas où la réduction suit, en contexte, le terme de base, mais aussi les cas où elle le précède; dans tous les cas, la proximité du terme est prise en compte, mais il faut savoir qu'une reprise anaphorique peut parfois apparaître bien loin du terme de base. Par contre, nous n'annotons pas les mots très polysémiques parfois utilisés pour reprendre un ou plusieurs concepts : si on se sert de *system* pour reprendre un ou plusieurs termes, sans que *system*

soit en fait la réduction d'un terme complexe, nous ne l'annotons pas. Prenons cette phrase : *Almost all master cylinders are divided into two separate systems.* Ici, *system* n'est pas la réduction d'un terme complexe, et il n'est pas monoréférentiel, car il ne dénote pas un concept précis, donc nous ne l'annotons pas.

Quelques types de variation terminologique n'ont pas été retenus, notamment les variantes par permutation, où l'ordre des constituants est modifié, souvent accompagnées par l'insertion d'autres mots à l'intérieur du syntagme. En anglais, une forme courante se manifeste ainsi : *piston head* ⇔ *the head of the piston*. De telles formes, parce qu'elles ne figureraient pas dans un dictionnaire, sont exclues, mais nous annotons tout de même les termes *head* (comme variante réduite de *piston head*) et *piston*. Si nous observons la variante, *the piston's head*, le même traitement est appliqué (voir section 3.3.2.3).

3.2.3 Critères formels

Un critère formel concernant la longueur des unités mérite notre attention. Comme l'a souligné Auger (voir section 1.1.3.2), entre autres, certaines unités potentiellement terminologiques contiennent de nombreux modificateurs, et peuvent poser des problèmes de découpage. Par exemple, notre corpus contient des formes telles que *electronic sequential multi-port fuel injection, five-cylinder variable displacement compressor* et *magnetic rack-and-pinion spool valve assembly*. Nous ne voyons aucune raison de les exclure sous prétexte qu'elles agissent plutôt comme des descriptions, des paraphrases ou des définitions métalinguistiques. Comme l'a souligné Sager (voir section 1.1.3.2), ces unités remplissent leur fonction communicative dans ce type de texte, et nous croyons qu'elles méritent d'être retenues en vertu de nos critères thématiques. Si l'on souhaite filtrer ces unités en exigeant qu'elles soient attestées plus d'une fois ou dans plus d'un ouvrage, nous montrerons au chapitre 4 qu'il est facile de paramétrer la liste de référence que nous avons créée en utilisant des seuils de fréquence ou de répartition, entre autres.

Ces longues unités peuvent contenir des formes qui correspondent elles-mêmes à des termes. Par exemple, la forme *magnetic rack-and-pinion spool valve assembly* contient le terme *spool valve*. Il faut alors se demander si l'on retient seulement l'une ou l'autre forme, ou les deux. Pour des raisons liées au format XML, il est plus aisé de ne pas permettre l'imbrication de termes à l'intérieur d'autres termes. Ainsi, nous annoterons seulement les unités de longueur maximale, pour autant qu'elles satisfassent à nos autres critères. Tout élément du syntagme qui est porteur de sens et qui contribue à la référence à un concept précis du domaine fait partie du terme. Par exemple, dans *computerized antilock brake system*, le mot *computerized* est redondant[11], mais il est porteur de sens; nous retenons donc tout le syntagme. Ce traitement a l'avantage de permettre de retrouver par la suite des termes imbriqués dans des termes plus longs, tandis que si nous annotions seulement les formes ayant le plus grand potentiel terminologique, il serait beaucoup plus difficile de revenir en arrière et de reconstruire ces longs termes.

Dans le cas où nous rejetons une forme complexe en vertu d'un de nos critères thématiques ou linguistiques, il faut parfois se demander si la forme contient un terme plus court qu'il serait bon de retenir. Dans ces cas, il faut évaluer la compositionalité de la forme complète. L'unité plus petite doit avoir conservé tout son sens, sans quoi elle n'est pas retenue. Par exemple :

- La forme *full throttle,* exclue parce qu'elle ne satisfait pas à nos critères thématiques, contient le mot *throttle,* qui peut être une réduction lexicale de *throttle valve* ou de *throttle pedal.* Or, le mot *throttle* n'a pas le même sens que *throttle valve* ou *throttle pedal* dans ce contexte, puisque ni un ni l'autre de ces termes ne peut être modifié par *full,* donc nous excluons la forme *full throttle* au complet.
- Dans *engine modification*, forme exclue en vertu des critères thématiques, *engine* a le même sens que le terme simple *engine,* et nous l'annotons ainsi.

[11] Le constituant *computerized* est redondant parce que les systèmes de freinage antiblocage comprennent forcément un microprocesseur.

3.2.4 Ouvrages de référence

Tout au long de ce travail, nous avons utilisé de multiples ressources terminologiques et lexicographiques. Celles-ci comprennent des banques de terminologie telles que Termium et le Grand Dictionnaire terminologique, ainsi que des dictionnaires spécialisés de mécanique automobile. Ces ouvrages nous ont aidé à évaluer le degré de lexicalisation des unités, ainsi qu'à choisir le terme de base lorsqu'un ensemble d'unités terminologiques dénotaient le même concept. Les références bibliographiques des principaux ouvrages de référence que nous avons consultés sont présentées à l'annexe 4.

3.3 Annotation du corpus

L'annotation du corpus est entièrement manuelle; nous n'avons pas eu recours à un extracteur de termes afin de pré-annoter le corpus ou de dégager des pistes quant au repérage des termes. L'annotation du corpus se fait directement au sein du texte, au moyen des balises du langage XML, afin de faciliter certains traitements automatiques telles que des transformations XSLT (voir section 3.5). Nous utilisons l'éditeur Oxygen afin d'effectuer l'annotation.

3.3.1 Déclaration de type de document (DTD)

Dans le langage XML, les déclarations de type de document (DTD) sont utilisées pour déclarer tous les *éléments* et les *attributs* qui seront utilisés dans une tâche d'annotation donnée. Dans la terminologie du XML, des données qui sont encadrées par des balises forment un *élément.* Le XML permet de définir soi-même la nature des éléments qui seront annotés dans un document. Dans le cadre de ce travail, ces éléments dont des termes, et sont encadrés par la paire de balises <term>...</term>. Les balises qui encadrent les *éléments* peuvent être assorties *d'attributs,* qui servent à décrire le contenu des éléments. Les *valeurs* attribuées à ces attributs doivent toujours être encadrées de guillemets ("").

La DTD doit définir tous les éléments qui seront utilisés pour l'annotation, ainsi que les attributs qui seront associés à chaque élément, dont la valeur doit être précisée chaque fois qu'un élément est annoté. Ces attributs peuvent être facultatifs ou obligatoires, ce qu'on doit préciser dans la DTD. On précise également la nature des données qui sont balisées (dans ce cas-ci, les éléments *term* contiennent simplement des chaînes de caractères), et les valeurs que peuvent prendre les attributs. Dans le cadre de ce travail, les attributs précisent certaines caractéristiques des termes annotés : on distingue ainsi les termes simples des termes complexes, les termes disjoints ou coordonnés des termes apparaissant dans leur forme canonique, etc.

La DTD permet de vérifier automatiquement la cohérence du document annoté, notamment pour s'assurer que tous les attributs qui doivent obligatoirement être déclarés pour un élément donné le sont, qu'aucun élément ou attribut inconnu n'est utilisé, etc.

La DTD que nous avons définie pour l'annotation du corpus est présentée dans la Figure 2. Dans la première ligne, nous déclarons que chaque document dans notre corpus est composé de termes et de chaînes de caractères qui ne sont pas des termes. Puis, nous déclarons que chaque terme est lui-même composé de caractères, et que les termes ont un certain nombre d'attributs parmi les cinq suivants : numéro d'identification (*id*), type (simple, complexe ou sigle), structure (un attribut facultatif qu'on assigne si le terme est coordonné ou disjoint, ou s'il constitue une réduction anaphorique) et langue, ainsi que des notes utilisées pendant l'annotation pour les cas problématiques. Ces attributs, et les valeurs qui peuvent leur être affectées, feront l'objet d'explications ci-dessous.

```
<!ELEMENT document (#PCDATA | term)*>
<!ELEMENT term (#PCDATA)>
<!ATTLIST term
    id CDATA ""
    type (simp | comp | sigacr) "simp"
    struct (coord | disj | anaphore) #IMPLIED
    lang (en | fr) "en"
    note (terme | domaine | decoupage | sens) #IMPLIED>
```

Figure 2 : DTD du corpus annoté

3.3.2 Balisage XML

L'annotation du corpus repose sur la paire de balises <term>...</term>, qui sert à identifier et à découper les unités terminologiques retenues en vertu des critères présentés dans la section précédente. La balise ouvrante <term> identifie le début d'une unité terminologique, et la balise </term>, sa fin. La Figure 3 illustre comment cette paire de balises est utilisée pour encadrer les termes repérés dans le corpus; un extrait annoté du corpus est présenté à l'annexe 2.

```
This drives the <term>piston</term> back down through the
<term>cylinder</term> with great force, transmitting the energy of the
expanding gas to the <term>crankshaft</term>.
```

Figure 3 : Utilisation de balises pour encadrer les termes

Dans de nombreux cas, comme nous l'avons vu, le découpage des termes est moins aisé que dans cet exemple simple. Nous décrirons ci-dessous quelques cas où le découpage peut poser problème, et comment nous avons choisi de traiter ces cas en ce qui concerne l'utilisation des balises.

3.3.2.1 Réduction par coordination

Les termes complexes réduits en raison d'une coordination sont balisés séparément[12], comme le montre la Figure 4.

```
On <term>compression</term> and <term>exhaust strokes</term>, the <term>rings</term> will tend to slip.
```

Figure 4 : Balisage des termes complexes coordonnés

Un attribut viendra déclarer que le terme *compression stroke* a été réduit en raison d'une coordination (voir section 3.3.3.3), et un lien sera fait dans la banque de termes entre la forme tronquée et le terme de base.

3.3.2.2 Disjonction

Lorsque des signes de ponctuation, des paraphrases, des synonymes ou des variantes terminologiques sont imbriqués dans un terme complexe, la structure linéaire du terme est brisée afin d'insérer l'élément en question. Voici des exemples tirés de notre corpus qui illustrent la variété des éléments qui peuvent causer une telle disjonction :

- Synonyme entre virgules : *the lower, or big end, bearing rotates*
- Synonymes entre parenthèses : *the I-head (overhead valve or valve-in-head) engine*
- Sigle entre parenthèses : *a throttle valve (TV) linkage*
- Paraphrase entre parenthèses : *the pistons in a multicylinder (more than one) engine*
- Virgules entre plusieurs modificateurs : *a typical downdraft, single-barrel carburetor*
- Guillemets et synonymes : *a "glass pack" or "steel pack" muffler*

On peut envisager au moins trois façons de traiter de tels termes disjoints. La première, qui consiste tout simplement à reformuler le texte pour éliminer la disjonction, nous semble à éviter. La deuxième consiste à permettre l'imbrication des éléments de type *term*, comme le montre la Figure 5.

---

[12] Du point de vue syntaxique, nous traitons les termes complexes coordonnés comme un cas d'omission.

```
Many manufacturers have introduced <term>full time FWD drive</term> or
<term>all wheel drive (<term>AWD</term>) systems</term>
```

Figure 5 : Traitement possible des termes complexes disjoints

Par contre, les parenthèses seraient toujours présentes dans un des éléments *term,* que nous pouvons représenter ainsi, en faisant abstraction de l'autre terme qu'il renferme : <term>all wheel drive **()** systems</term>.

Il y aurait moyen d'exclure les parenthèses en définissant d'autres éléments XML (en plus de *term*), mais cela compliquerait la représentation de la structure des termes. Nous avons donc décidé de traiter les termes disjoints de manière linéaire, en excluant dans la mesure du possible les éléments perturbateurs (p. ex. dans la Figure 6, la parenthèse ouvrante ne fait pas partie du découpage de l'unité *AWD systems*).

```
Many manufacturers have introduced <term>full time FWD drive</term> or
<term>all wheel drive</term> (<term>AWD) systems</term>
```

Figure 6 : Découpage des termes complexes disjoints

Ainsi, le découpage (et les attributs qui seront ajoutés par la suite) montre qu'un premier terme est tronqué en raison de l'imbrication d'un sigle, et qu'un deuxième terme est disjoint par un signe de ponctuation, la parenthèse fermante. Les deux formes balisées renverront tous deux à un lemme dans sa forme canonique; ce lien se fait dans la banque de fiches, qui sera décrite à la section 3.4.

3.3.2.3 Autres problèmes de découpage

L'anglais se sert du *-s* génitif pour indiquer la possession ou la détermination; p. ex. *the piston's head*. Comme nous l'avons mentionné, nous n'annotons pas ce genre de

variante terminologique, qui n'apparaitrait pas dans un dictionnaire spécialisé[13]. Or, ce marqueur peut aussi indiquer le pluriel dans le cas des sigles et des acronymes, entre autres; p. ex. *ECU's are generally contained in one casing*. L'utilisation de l'apostrophe dans ces cas est flottante, parfois dans le même ouvrage, voire pour un même sigle. Lorsqu'on retrouve un sigle ou un acronyme suivi de l'apostrophe et du *-s*, il faut donc inclure ces deux caractères dans le découpage, mais seulement s'ils servent à marquer le pluriel.

3.3.3 Attributs XML

Dans ce corpus, les éléments de type *term* peuvent avoir jusqu'à cinq attributs. Deux d'entre eux sont obligatoires : *id* et *type*.

3.3.3.1 id

L'attribut *id* est un numéro d'identification unique à chaque terme (voir Figure 7). Il est indépendant du document ou des documents dans lesquels nous l'avons repéré. Les numéros sont attribués l'un après l'autre à mesure que les textes sont dépouillés : le premier terme du premier texte reçoit la valeur 1. Le numéro est attribué à un terme dans sa forme lemmatisée (p. ex. *engine*); ses variantes flexionnelles (*engines*) reçoivent le même identificateur.

```
<term id="178">Fuel</term> burns in <term id="1096">combustion
chambers</term> inside an <term id="6">engine</term>.
```

Figure 7 : Utilisation de l'attribut *id*

[13] Tout porte à croire que cette forme est plutôt rare, du moins en ce qui concerne les termes : « most studies reject this pattern insofar as such a use of the Saxon genitive is rare. The only example of Saxon genitive we found in our corpus, occurring only once, is earth's curvature (courbature de la terre). Furthermore, all the dictionaries we looked at propose curvature of the earth, and not earth's curvature » (Gaussier 2001 : 170).

3.3.3.2 type

L'attribut *type* sert à distinguer les termes simples, les termes complexes et les sigles ou acronymes, comme le montre la Figure 8. Lorsqu'un terme complexe contient un sigle ou un acronyme, on donne à l'attribut *type* la valeur *comp*, non pas *sigacr*. Cependant, un élément dans la fiche de cette unité (voir section 3.4) précisera qu'il s'agit s'une variante par siglaison.

```
A unique <term id="3804" type="simp">transaxle</term> design is the
<term id="1196" type="comp">continuously variable transmission</term>
(<term id="1197" type="sigacr">CVT</term>).
```

Figure 8 : Utilisation de l'attribut *type*

3.3.3.3 struct

L'attribut *struct* est utilisé pour certains types de variantes terminologiques. La valeur *coord* est affectée à l'attribut *struct* des termes complexes qui sont réduits en raison d'une coordination ou d'une juxtaposition, comme le montre la Figure 9.

```
A <term id="2017" type="comp">throttle body</term> or <term id="4416"
type="simp" struct="coord">plate</term> is used on all <term id="1227"
type="comp">fuel injection systems</term>.
```

Figure 9 : Utilisation de la valeur *coord* de l'attribut *struct*

Dans certains cas, nous avons jugé que le trait d'union, la barre oblique ou la parenthèse avait la même fonction que la conjonction de coordination. Prenons le cas de *standard (manual) transmision,* où les mots *standard* et *manual* sont des synonymes. Ici, nous donnons la valeur *coord* à *standard,* une variante réduite de *standard transmission*, coordonnée en quelque sorte à *manual transmission*, la parenthèse ayant ici une fonction semblable à la conjonction *or*. L'annotation résultante est présentée dans la Figure 10.

```
<term struct="coord">standard</term> (<term struct="disj">manual)
transmision</term>
```

Figure 10 : Cas spécial de coordination

On retrouve parfois (quoique rarement) des coordinations complexes telles que *front and rear differentials and drive shafts*. Pour ne pas compliquer outre mesure notre représentation de la structure des termes, et pour observer la règle qui veut que seules les unités de longueur maximale soient incluses, nous avons décidé de baliser, à l'intérieur de ces syntagmes, les unités qui correspondaient à un terme de base (*rear differential*). Ensuite, il ne reste qu'à renvoyer chaque élément tronqué à un terme. Or, chaque forme tronquée renvoie à plus d'un terme : dans l'exemple, le mot *front* participe aux termes *front differential* et *front drive shaft*. Pour éviter les renvois multiples, nous avons décidé de renvoyer à la première potentialité qui se réalise (*front differential*). Un commentaire est aussi ajouté dans la banque de termes pour rappeler que la variante réduite participe également à un autre terme.

La valeur *anaphore* est affectée à l'attribut *struct* des réductions anaphoriques, qui obtiennent par ailleurs leur propre numéro d'identification (voir Figure 11).

```
This will drive a <term id="1149" type="comp">timing chain</term> or
<term id="20" type="comp">timing belt</term>. Some <term id="6"
type="simp">engines</term> use a combination of <term id="1825"
type="simp" struct="anaphore">chain</term> and <term id="1152"
type="simp" struct="anaphore">belt</term>.
```

Figure 11 : Utilisation de la valeur *anaphore* de l'attribut *struct*

Dans certains cas, différents types de réductions se succèdent. Par exemple, lorsqu'on a plusieurs termes réduits en raison d'une coordination, si le dernier est lui-même une réduction anaphorique, nous donnons la valeur *coord* à toutes les formes sauf la dernière, qui reçoit la valeur *anaphore*.

La valeur *disj* est appliquée à toute réalisation d'un terme qui contient des signes de ponctuation (guillemets, parenthèses, virgules, etc.), comme le montre la Figure 12. La forme disjointe reçoit le même numéro d'identification que le terme de base. Par ailleurs, si un signe de ponctuation la précède immédiatement (un guillemet, une parenthèse ouvrante), nous plaçons la balise après le signe, pour exclure autant que possible la ponctuation[14].

```
A <term id="4322" type="comp" struct="coord">four wheel drive</term>
(<term id="4323" type="comp" struct="disj">FWD) vehicle</term> has two
<term id="1102" type="simp">differentials</term> to supply power to all
four <term id="132" type="simp">wheels</term>.
```

Figure 12 : Utilisation de la valeur *disj* de l'attribut *struct*

3.3.3.4 lang

Cet attribut sert uniquement à indiquer la langue à laquelle le terme appartient. Il servirait éventuellement à décrire les termes empruntés à une autre langue, mais notre corpus ne comporte pas de tels emprunts.

3.3.3.5 note

Cet attribut est utilisé par l'annotateur pour indiquer les cas qui posent problème de façon à effectuer un retour systématique sur ces unités lors des périodes de révision.

3.4 La banque de termes

Au fur et à mesure que de nouvelles unités terminologiques sont repérées dans le corpus, qu'il s'agisse de termes de base ou de leurs synonymes ou variantes[15], elles sont

---

[14] Par contre, si une forme tronquée se termine par un trait d'union faisant partie du terme de base, nous incluons le trait d'union dans le découpage, puisqu'il fait partie du terme.

[15] Nous n'insisterons pas sur la différence entre un synonyme et une variante. Mentionnons tout de même que les variantes par insertion et par surcomposition seront traitées comme des synonymes. Un traitement particulier, décrit dans cette section, est réservé aux

consignées dans une banque de termes. Elle contient une fiche pour le lemme de chaque forme annotée. Cette banque recense tous les termes retrouvés dans les textes annotés, indépendamment du document ou des documents dans lesquels ils ont été repérés. Elle fournit ainsi un moyen rapide de récupérer toutes les unités terminologiques du corpus dans leur forme lemmatisée. Elle contient par ailleurs des renseignements qui encombreraient les annotations du corpus, tels que les liens des synonymie ou de variation entre les unités, ainsi que des définitions. Chaque fiche contient dix champs (ou éléments), représentés dans la Figure 13.

```
<fiche>
    <ID></ID>
    <entree></entree>
    <var></var>
    <SourceLem></SourceLem>
    <defmod></defmod>
    <def></def>
    <SourceDef></SourceDef>
    <IDvar></IDvar>
    <commentaire></commentaire>
    <statut></statut>
</fiche>
```

Figure 13 : La structure des fiches terminologiques

Le champ *ID* contient un numéro d'identification unique à chaque terme. C'est l'inscription d'un nouveau terme dans cette banque qui permet de déterminer son numéro d'identification. Ce dernier ne doit pas être séquentiel par rapport à la lecture des textes, il doit tout simplement être unique, afin qu'on puisse repérer facilement toutes les occurrences d'une unité terminologique dans le corpus.

Le champ *entree* contient la forme lemmatisée associée aux diverses variations flexionnelles identifiées dans le corpus. Ainsi, les termes *engines* et *engine* pointent sur le lemme *engine*. Les unités regroupées sous un même lemme sont les variantes flexionnelles

---

réductions (anaphoriques ou par coordination), aux variantes orthographiques et aux sigles et acronymes.

et les disjonctions (voir section 3.3.2.2) uniquement; les autres types de variantes terminologiques sont consignées dans des fiches différentes et pointeront sur le terme de base au moyen du champ *IDvar*. Par exemple, le sigle *ECU* possède sa propre fiche, où le champ *IDvar* contient le numéro d'identification du terme de base *electronic control unit*. Le choix du terme de base (effectué en fonction de la fréquence des unités, du nombre de variantes qu'elles possèdent et de la vedette que privilégient les ouvrages de référence consultés) peut changer à mesure que se construit la banque de fiches et que de nouvelles unités terminologiques sont recensées. Ainsi, il peut devenir ardu de s'assurer qu'un ensemble donné de formes pointe toujours vers le bon terme.

Le champ *var* contient le type de variation terminologique observée. Si ce champ est rempli, le champ *IDvar* doit obligatoirement contenir un lien vers un terme de base. Les valeurs possibles sont *Synonyme*, *Sigle ou acronyme, Troncature*, et *Variation orthographique*. Les réductions lexicales, les variantes par surcomposition et par insertion ainsi que les véritables synonymes reçoivent la valeur *Synonyme*. Toute forme qui contient un sigle ou un acronyme d'un terme reçoit la valeur *Sigle ou acronyme*. Les variantes orthographiques reçoivent la valeur *Variation orthographique*. Enfin, les réductions anaphoriques et les réductions par coordination reçoivent la valeur *Troncature*. Puisque ces deux types de réduction peuvent être observés dans le même corpus, la meilleure façon de procéder semble être de décrire la raison pour laquelle un terme a été tronqué directement dans le corpus, et de signaler simplement dans la banque de termes qu'une ou plusieurs réductions ont été observées[16].

[16] Bien que les réductions de termes complexes n'apparaîtraient pas dans un dictionnaire spécialisé, nous les incluons dans la banque pour avoir une représentation plus précise de la fréquence des termes et de leurs variantes. Cela permettrait par ailleurs d'évaluer la capacité d'un outil de relier les réductions à leur forme pleine. En outre, le paramétrage de la liste de référence (voir section 4.3) permet facilement d'éliminer ces unités au besoin.

Le champ *statut* indique si le terme est *valide* ou s'il a été *reconstruit*. Ce dernier statut est utilisé pour les termes qui n'ont pas été recensés dans le corpus, mais dont les variantes ont été relevées. Par exemple, si un terme n'apparaît qu'une fois et sous une forme réduite en raison d'une coordination, le champ *IDvar* de cette variante sera le numéro d'identification du terme dit *reconstruit*.

Le champ *def* des fiches des termes de base (seulement) contient, dans bien des cas, une définition. La définition est transcrite directement, et nous notons sa source dans le champ *SourceDef***.** Si la forme est une variante ou un synonyme d'un terme de base, aucune définition n'est requise. Si la définition prise dans un ouvrage de référence doit être adaptée parce qu'elle ne respecte pas les règles de rédaction d'une définition ou qu'elle comporte des éléments superflus, nous notons la nouvelle définition dans le champ *defmod*. Par ailleurs, si nous ne trouvons dans aucun ouvrage de référence la définition d'un terme, nous nous inspirons du contexte ou de renseignements puisés dans Internet pour forger une définition; en l'occurrence nous saisissons la définition directement dans le champ *defmod*. Dans tous les cas, nous notons la source de l'ouvrage de référence ou des sites Web.

Les champs *SourceLem* et *SourceDef* contiennent un code qui renvoie à un fichier qui lie chaque code à la référence complète d'un ouvrage de référence. Le champ *SourceLem* indique un ouvrage de référence dans lequel le lemme est attesté, mais le fait qu'une unité ne soit attestée dans aucun ouvrage de référence ne nous empêche pas de la retenir si elle satisfait à nos critères de sélection. Dans certains cas, ce champ peut indiquer quelle source nous a permis d'établir un lien de synonymie ou de variation.

Enfin, le champ *commentaire* nous permet d'inscrire au besoin des commentaires sur le contenu de la fiche ou l'annotation des occurrences du terme dans le corpus.

L'annexe 5 contient la fiche du terme *air conditioner* ainsi que les fiches de tous les synonymes et variantes que nous avons recensés dans le corpus.

3.5 Transformations XSLT

Au moyen du langage XSLT, nous avons élaboré des *transformations* qui génèrent, à partir de nos fichiers XML, des fichiers HTML plus faciles à consulter pour un utilisateur ponctuel. En effet, les balises XML rendent la lecture du corpus et la consultation de la banque plutôt ardues (voir les annexes 2 et 5). Les transformations XSLT permettent de cibler des éléments spécifiques, de les extraire et de les afficher dans un format beaucoup plus convivial.

Deux transformations très simples ont été élaborées pour la banque de fiches. Elles affichent chacune un tableau, trié alphabétiquement ou par numéro d'identification, qui contient toutes les formes consignées dans la banque, assorties de différents renseignements (statut, type de variante, etc). Les transformations peuvent aisément être modifiées pour filtrer la banque ou afficher des éléments différents.

Compression rings are designed to prevent leakage between the **piston** and the **cylinder**, Figure 2-31.

The idea is to create an internal stress within the **ring**. This stress will tend to cause the **ring** to twist in such a fashion that the lower edge of the **ring** is pressed against the **cylinder wall** on the **intake stroke**. This will cause the **ring** to act as a mild scraper. The scraping effect will assist in the removal of surplus **oil** that may have escaped the **oil control rings**, Figure 2-32.

On **compression** and **exhaust strokes**, the **rings** will tend to slip lightly over the **oil** film. This will prolong the life of the **ring**, Figure 2-33. On the **power stroke**, pressure of the burning gases will force the top edge of the **ring** downward. This causes the **ring** to rub the **wall** with full face contact and provides a good seal for the enormous pressure generated by the **power stroke**.

Figure 14 : Affichage convivial du corpus annoté

Quant au corpus, nous avons créé une transformation qui efface toutes les balises qu'il contient et transforme chaque forme annotée en un lien hypertexte. Un exemple de sortie est présenté dans la Figure 14. Toutes les formes annotées sont mises en relief, et lorsque nous pointons sur une forme avec la souris, une boite apparaît et affiche soit la définition (dans le cas des termes de base), soit le terme de base et la définition correspondante (dans le cas des variantes et synonymes). Lorsqu'aucune définition n'est disponible, on indique simplement s'il s'agit d'un terme de base ou d'une variante. Cette transformation est un peu plus complexe, parce qu'elle repose sur le lien créé entre les deux fichiers XML, à savoir le corpus et la banque.

Ces transformations ont été utiles pour réviser le travail d'annotation et s'assurer de la cohérence des annotations et des renvois créés dans la banque de termes, comme nous le montrerons au chapitre 4.

3.6 Conclusion

Dans ce chapitre, nous avons décrit la méthodologie d'annotation de corpus qui est au coeur de ce travail. Celle-ci repose sur l'insertion de balises XML dans le corpus pour encadrer les termes, qui sont identifiés en fonction de critères thématiques, linguistiques et formels. Les unités terminologiques retenues sont annotées au sein du corpus, et une fiche est créée pour chaque forme dans une banque de termes. Dans le chapitre 4, nous décrirons comment exploiter le corpus annoté et la banque de termes afin de produire une liste de référence, et comment utiliser cette liste pour évaluer un extracteur. Nous chercherons également à valider notre travail d'annotation.

4. Analyse des résultats

Dans ce chapitre, nous analysons le résultat du travail d'annotation de corpus présenté au chapitre 3. Dans la section 4.1, nous cherchons à vérifier la validité de notre travail d'annotation. La section 4.2 porte sur les caractéristiques que nous pouvons dégager du corpus annoté et de la banque de termes quant aux termes qu'ils contiennent. Nous montrons comment il est possible de paramétrer la liste de référence dans la section 4.3. Enfin, quelques exemples d'utilisation de la liste de référence sont présentés dans la section 4.4.

4.1 Validation de l'annotation

Le langage XML fournit un moyen simple de confirmer la validité d'un document annoté en vérifiant sa conformité à la déclaration de type de document (DTD) qui encadre le projet d'annotation. Or, cette validation ne vient que confirmer que toutes les balises et attributs utilisés ont été préalablement définis, que le contenu des éléments correspond au type de données attendu, et que les attributs obligatoires ont reçu une valeur, en somme. Il ne s'agit donc pas d'un moyen suffisant pour confirmer la validité de nos résultats.

4.1.1 Relecture du corpus et programme de validation

Un premier travail de validation a été réalisé en relisant le corpus annoté au complet, dans un format convivial produit par une transformation XSLT (voir section 3.5, Figure 14), et en observant chaque forme annotée. Ce travail de relecture visait à vérifier qu'aucune forme correspondant à nos critères de dépouillement n'avait été occultée, et que le contenu des annotations et les renvois à la banque de fiches étaient cohérents. De nombreuses erreurs ont ainsi été corrigées.

Après de nombreuses relectures partielles faites au cours de l'annotation et cette relecture complète effectuée une fois l'annotation terminée, nous avons cherché un autre

moyen de valider la cohérence de nos annotations, étant donné qu'avec un si gros corpus, une validation manuelle ne peut se faire sans erreurs. Nous avons donc développé un programme à cette fin.

Le programme de validation que nous avons conçu est un script écrit dans le langage Python qui exploite la librairie xml.dom, qui nous permet de parcourir la banque de termes et le corpus annoté et d'extraire facilement les éléments et les attributs qui nous intéressent. Le programme nous permet de vérifier non seulement les différents liens à l'intérieur de la banque de termes, mais aussi les liens créés entre le corpus annoté et la banque de termes au moyen de l'attribut *ID* des balises *term*. Il nous permet également d'effectuer des tests portant sur la fréquence des formes annotées.

4.1.2 Validation des annotations

Tout d'abord, nous nous intéressons aux annotations dans le corpus. Notre programme parcourt le corpus et, pour chaque forme annotée, consulte la fiche correspondante dans la banque de termes pour vérifier certains aspects de l'annotation. Nous vérifions qu'aucune forme annotée n'a un numéro d'identification qui est absent de la banque, puis, nous comparons la forme annotée au lemme de la fiche correspondante, afin de nous assurer que chaque forme annotée a reçu le bon numéro d'identification. À l'aide d'une distance d'édition[17], nous vérifions que les deux formes se correspondent à l'intérieur d'une certaine marge, allouée pour les formes au pluriel ou ayant une majuscule initiale, par exemple. Le calcul effectué normalise la distance d'édition (nombre d'opérations) par la longueur de la forme annotée. Nous avons d'abord fixé un seuil arbitraire de 0,25 : lorsque la distance était supérieure à ce seuil, nous avons vérifié manuellement que la forme annotée équivalait au lemme de la fiche correspondante (la validation des numéros

[17] La distance d'édition est fonction du nombre d'opérations nécessaires pour transformer une chaîne en une autre. Les opérations permises ici sont la suppression d'un caractère, l'insertion d'un caractère et la substitution d'un caractère par un autre. Chaque opération entraîne la même pénalité.

d'identification est illustrée dans le Tableau IV). Cela nous a donné un nombre raisonnable de paires de formes à comparer, soit 505. Sur ce nombre, 19 étaient des erreurs (dont 14 concernaient toutes la forme *regulator*), que nous avons corrigées.

Forme annotée	**Lemme de la fiche correspondante**	**ID**	**Distance d'édition**	**Erreur**
ECU's	ECU	3307	0.4	non
IC's	IC	3318	0.5	non
Tie Rods	tie rod	1401	0.375	non
Rings	ring	1098	0.4	non
batteries	battery	35	0.333	non
head	cylinder head	802	0.692	oui
filters	return valve	2341	0.917	oui
venturi system	four wheel drive vehicle	2536	0.833	oui
regulator	voltage regulator	479	0.471	oui
Lights	idiot light	304	0.3333	oui

Tableau IV : Validation des numéros d'identification

Puis, afin de vérifier si un seuil plus bas nous permettrait de relever plus d'erreurs, nous l'avons baissé à 0,18, ce qui a plus que quadruplé le nombre de paires à vérifier. Or, aucune erreur supplémentaire n'a été relevée, ce qui confirme notre intuition que les formes annotées auxquelles on attribue un mauvais numéro d'identification auront une distance d'édition élevée par rapport aux entrées correspondantes dans la banque de termes. À défaut de vérifier manuellement plus de 28 000 paires de formes, la distance d'édition semble un bon moyen de vérifier que le bon numéro d'identification est associé à chaque forme annotée.

4.1.3 Validation des renvois dans la banque de termes

Deuxièmement, nous nous intéressons au système de renvois qui relie, à l'intérieur de la banque de termes, les termes de base à leurs (éventuels) synonymes et variantes.

Comme le montrent les exemples de fiches présentés dans l'annexe 5 (qui renvoient toutes au terme de base *air conditioner*), il n'est pas aisé de parcourir la banque de fiches pour vérifier la cohérence des renvois. Ainsi, le programme de validation parcourt la banque de termes, extrait les fiches, et rassemble les entrées liées par des renvois. Une fois ces ensembles isolés, le programme vérifie certaines contraintes, à savoir que chaque fiche a un numéro d'identification unique, et que dans chaque ensemble, il y a seulement un terme de base, et les synonymes pointent tous sur celui-ci; les variantes, quant à elles, peuvent pointer sur n'importe quel type d'unité.

```
Terme de base : power steering (22)
Synonymes : power steering system (15), power steering unit (3)
Var. orthographique : power steering system -> power-steering system (6)
Réduction : power steering system -> power system (1)
Réduction : power steering system -> system (7)
Réduction : power steering -> steering (1)
Var. orthographique : power steering -> power-steering (2)
Réduction : power steering unit -> power unit (1)
Réduction : power steering unit -> unit (5)
```

Figure 15 : Exemple de sortie du programme de validation

Le programme affiche tous les ensembles ainsi identifiés, comme le montre la Figure 15. Chaque forme est suivie de sa fréquence dans le corpus; la mention *reconstruit* indique qu'il s'agit d'un terme complexe qui n'apparaît nulle part dans le corpus, mais dont au moins une variante réduite a été recensée, comme le montre la Figure 16.

```
Terme de base : crankcase ventilation system (reconstruit)
Réduction : crankcase ventilation system -> ventilation system (1)
Réduction : crankcase ventilation system -> crankcase ventilation (1)
```

Figure 16 : Réductions d'un terme de base reconstruit

Cet affichage nous a permis de vérifier manuellement la cohérence de tous les ensembles. Plus précisément, nous avons vérifié que le bon type de variation avait été attribué à toutes les variantes (variantes orthographiques, sigles et réductions), et que celles-ci pointaient toutes sur la bonne fiche. Nous avons également confirmé que toutes les

formes annotées répondaient bien aux critères de dépouillement énoncés dans la section 3.2. Les connaissances que nous avons acquises au cours de l'annotation sur les concepts et les termes de la mécanique automobile ont grandement facilité cette étape. En effet, bien que nous ne soyons pas expert dans ce domaine, les lectures répétées du corpus et la consultation des ouvrages de référence nous ont fourni une bonne compréhension de la mécanique automobile, condition nécessaire pour effectuer le dépouillement terminologique.

La validation a mis en lumière de nombreux cas où deux formes identiques pointaient sur le même terme de base, mais en passant par des renvois différents. Par exemple, la forme *coupling* apparait comme réduction anaphorique de *fluid coupling*, mais aussi comme réduction de *hydraulic coupling*, un synonyme de *fluid coupling*. En fait, ces doublons ne sont pas des erreurs, car une réduction anaphorique devrait pointer sur une forme attestée à l'intérieur du même document, à tout le moins. En effet, c'est le contexte qui nous a indiqué le terme de base des réductions anaphoriques. Un de ces doublons constituait toutefois une erreur. Le sigle *ECU* désigne *electronic control unit*, mais dans un des documents, il accompagnait, entre parenthèses, la forme *engine control unit*, qui dénote le même concept, mais n'est attestée qu'une fois dans un seul document, et n'apparait dans aucun des ouvrages de référence que nous avons consultés. Nous avons conclu que la forme *ECU* est un sigle de *electronic control unit,* et seulement de cette forme. Il est possible que le sigle lui-même ait porté l'auteur du document en question à croire qu'il était plutôt dérivé de *engine control unit*; quoi qu'il en soit, le lien entre le sigle et la forme vraisemblablement erronée n'a pas été consigné dans la banque de fiches.

4.1.4 Validation de l'adéquation de la banque de termes au corpus

Troisièmement, nous nous intéressons à la fréquence des formes consignées dans la banque de termes, afin de nous assurer qu'elles sont toutes attestées dans le corpus (à l'exception des termes reconstruits). Nous avons d'abord extrait de la banque de termes toutes les formes qui avaient une fréquence de 0 dans notre corpus, et avons vérifié si elles

avaient bien le statut *reconstruit*; quelques erreurs ont ainsi été corrigées, à savoir des unités dont nous avions décidé de ne pas annoter les occurrences, mais dont nous avions oublié de supprimer la fiche. Puis, nous avons vérifié la fréquence cumulative des ensembles constitués d'un terme de base et de ses synonymes et variantes. La fréquence minimale d'un ensemble serait de 1 : il pourrait s'agir d'un ensemble qui contient un seul terme ayant une seule occurrence, ou bien d'un ensemble constitué d'un terme de base reconstruit et d'une variante dont la fréquence est de 1. La présence d'un ensemble dont la fréquence est nulle indiquerait une erreur.

Deux tels ensembles existaient. L'un résultait de notre traitement des coordinations complexes (voir section 3.3.3.3). Dans la forme *three- and four-speed automatic transmissions and transaxles*, la forme *three-* constitue une réduction de *three-speed automatic transmission*, mais aussi de *three-speed automatic transaxle*. Ce dernier a été reconstruit et consigné dans la banque de termes, mais il formait un ensemble à un seul membre, parce qu'aucune autre variante ne pointait sur lui, et que le renvoi de *three-* à ce dernier avait été consigné à un commentaire; la fréquence de l'ensemble était naturellement de 0. Nous avons donc ajusté manuellement sa fréquence à 1 dans le programme de validation.

L'autre cas était effectivement une erreur, à savoir une unité dont nous avions décidé de ne pas annoter les occurrences, mais que nous avions oublié de supprimer de la banque. Des oublis de la sorte sont faciles à faire dans ce genre de travail d'annotation, d'où l'importance de valider l'adéquation de la banque au corpus et vice-versa, ce que notre programme de validation a permis de faire.

4.1.5 Synthèse

Ce processus de validation nous a permis d'améliorer la banque de termes de différentes façons. Dans plusieurs cas, nous avons corrigé le choix de terme du base en fonction de la fréquence et du nombre de variantes des unités; p. ex. *engine control*

computer a une fréquence plus élevée et plus de variantes (dont un sigle) que *on-board engine control computer*, qui était jusqu'alors le terme de base, sans doute parce qu'il avait été rencontré en premier – la validation nous a permis de le traiter plutôt comme une variante par surcomposition. Surtout, nous avons pu ajouter de nombreux liens de variation qui n'avaient pas été repérés pendant le processus d'annotation; en effet, tous ces liens ne peuvent être identifiés simplement en essayant de se rappeler, chaque fois qu'on rencontre une nouvelle forme, si le concept associé est déjà dénoté par une forme présente dans la banque. Dans plusieurs cas, deux ensembles désignaient le même concept, mais le lien entre le terme de base des deux ensembles n'avait pas été identifié jusqu'alors. Ces ensembles isolés ont pu être reliés, et nous obtenons ainsi une image plus précise de la variation au sein du corpus.

Le programme que nous avons conçu a grandement facilité le processus de validation; certaines de améliorations réalisées n'auraient pas été possibles sans un outil de cette sorte. Il nous a aussi permis de dégager les caractéristiques des termes, que nous présentons dans la section suivante.

4.2 Caractéristiques du corpus annoté et de la banque de termes

La banque de fiches, le corpus annoté et notre programme de validation nous permettent de dégager certaines caractéristiques des termes présents dans notre corpus. La banque de termes contient 5478 entrées. Parmi celles-ci, 2613 (moins de la moitié) correspondent à des termes de base, 1288 à des synonymes[18], et le reste à des variantes, à savoir : 1449 réductions anaphoriques ou par coordination, 69 variantes orthographiques et 59 sigles ou acronymes. 176 termes complexes ont été reconstruits. La proportion élevée de réductions mérite d'être soulignée. De toute évidence, le phénomène de la réduction des

[18] Rappelons que les entrées dont le champ *var* reçoit la valeur *Synonyme* comprennent les variantes par surcomposition et par insertion. Voir section 3.4.

termes complexes est très important, et cela a forcément des conséquences quant à l'extraction de termes.

Aux 2613 termes de base sont associés un nombre équivalent d'ensembles de synonymes et de variantes, tous reliés à un terme de base. Le nombre moyen d'éléments par ensemble est de 2,1, c'est-à-dire que les termes ont en moyenne environ un synonyme ou une variante. Ce taux élevé de variation pourrait s'expliquer par le degré de spécialisation relativement faible du corpus. Freixa (2002 : 222), dans une étude portant sur la variation dénominative, montre que les textes moins spécialisés présentent un taux de variation plus élevé. En effet, elle observe des taux de variation de 2,02 dans un corpus de textes de vulgarisation et de 1,58 dans un corpus de textes très spécialisés.

Un exemple qui illustre bien la quantité et la variété de synonymes et de variantes que peut avoir un terme de base est celui de *air conditioner* (voir l'annexe 5, qui contient les fiches de toutes les unités reliées à ce terme) qui comporte neuf synonymes (*air conditioning*, *air conditioning system*, *auto air conditioning system*, *automotive air conditioner*, *automotive air conditioning system*, *vehicle air conditioner*, *automobile air conditioning system*, *air*, *automotive air conditioning*) et les variantes suivantes :

- Réduction : *air conditioning system* ⇔ *system*
- Réduction : *air-conditioning system* ⇔ *system*
- Sigle : *air conditioning system* ⇔ *A/C system*
- Sigle : *air conditioning* ⇔ *A/C*
- Variante orthographique : *air conditioning system* ⇔ *air-conditioning system*
- Variante orthographique : *air conditioner* ⇔ *air-conditioner*
- Variante orthographique : *air conditioning* ⇔ *air-conditioning*

Le nombre total d'occurrences des termes de la banque dans le corpus est de 28 656, ce qui donne une fréquence moyenne de 5,23 occurrences par terme. Sur les 5478 formes dans la banque, 2643 ont une fréquence de 1 dans le corpus[19]. Or, si on calcule le nombre de hapax sur les ensembles (tels que celui présenté ci-dessus) plutôt que les formes, ce nombre tombe à 776. Les occurrences de termes sont réparties dans les trois documents de la façon suivante : 20 334 occurrences dans le document AF, 6449 dans le document CCB et 1873 dans le document HCW[20]. La densité terminologique de chaque document est donc de 0,15, 0,09 et 0,17 unités terminologiques par mot respectivement. La densité relativement faible du document CCB semble concorder avec son niveau de spécialisation, que nous jugeons plus faible que celui des deux autres ouvrages.

Si l'on représente la fréquence des termes de base et des ensembles (constitués d'un terme de base et de ses synonymes et variantes) en fonction de leur rang de fréquence, comme le montre la Figure 17 pour les 1000 unités les plus fréquentes, on voit que les données sont conformes à la loi de Zipf. Cette dernière stipule que la fréquence d'une unité est en relation à peu près constante avec son rang. Or, si l'on compare la fréquence d'un terme à la fréquence de toutes les unités qui renvoient à celui-ci, on constate une variabilité très importante, comme le montre la Figure 18 pour les 200 unités les plus fréquentes.

---

[19] Rappelons que les sens des termes polysémiques sont traités séparément, ce qui augmente le nombre de hapax considérablement.
[20] Voir la section 3.1 pour la référence de chacun des documents.

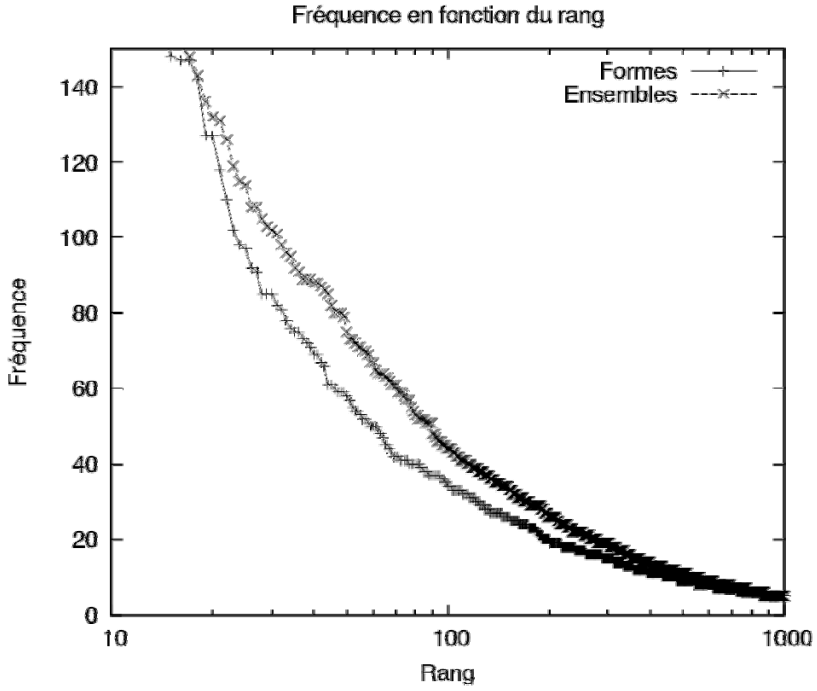

Figure 17 : Fréquence des formes et des ensembles en fonction de leur rang

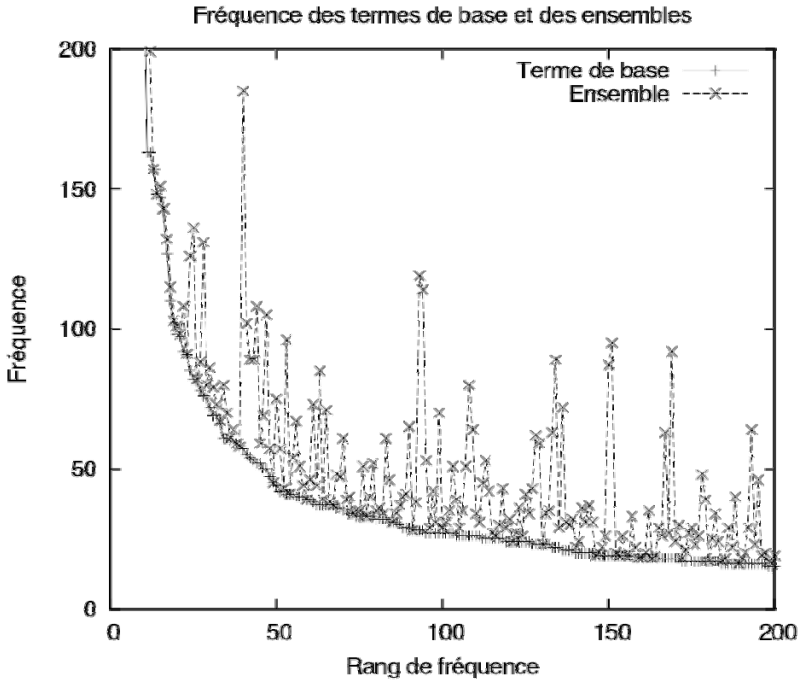

Figure 18 : Fréquence des termes de base et des ensembles correspondants

Nous nous sommes demandé ce qui pouvait expliquer cette grande variabilité. Est-ce qu'elle résulte du processus d'annotation, ou d'une propriété des termes eux-mêmes? Afin d'explorer cette question, nous avons isolé les termes de base simples et les termes de base complexes; la banque en contient 272 et 2341 respectivement. Si l'on observe la *fréquence de variation* de chacun de ces deux types de termes, on voit que la variabilité observée est beaucoup moins prononcée parmi les termes simples, comme le montrent la Figure 19 et la Figure 20.

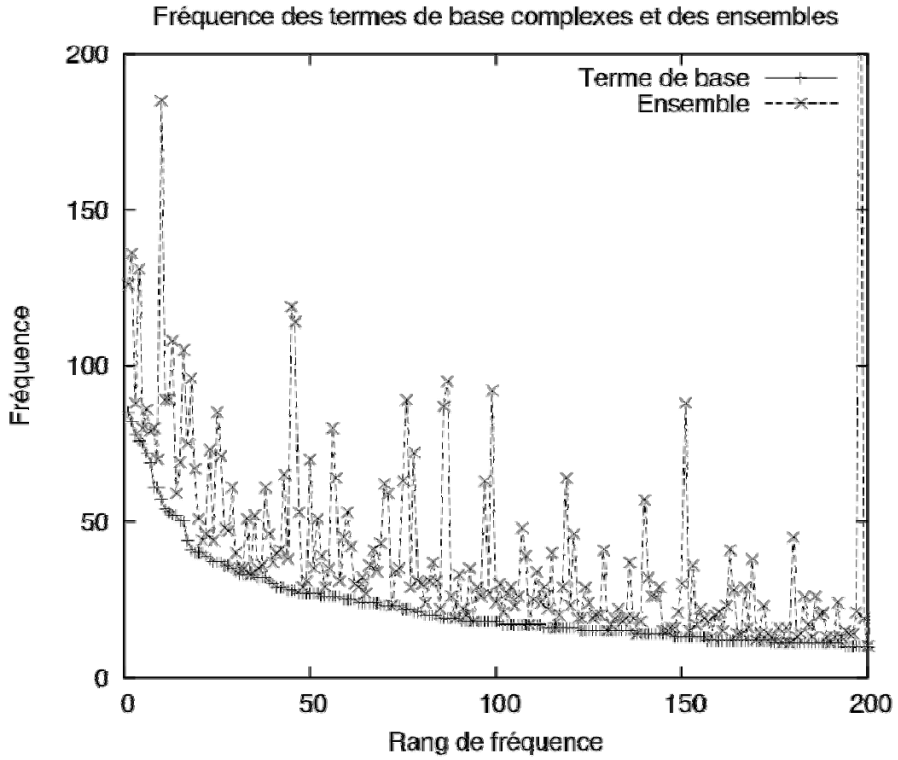

Figure 19 : Fréquence des termes de base complexes et de leur ensemble

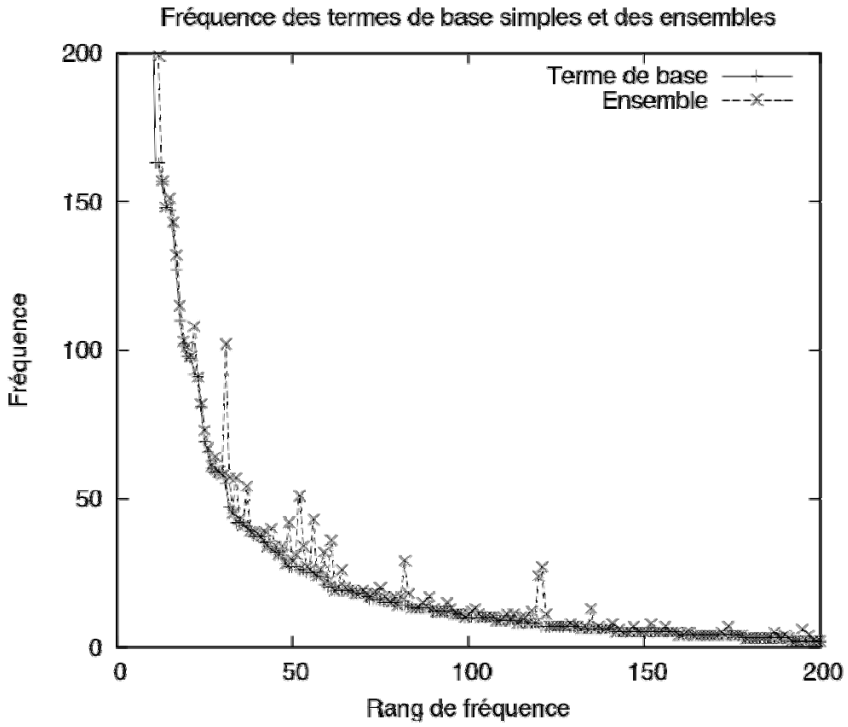

Figure 20 : Fréquence des termes de base simples et de leur ensemble

La fréquence moyenne de variation, que nous calculons en divisant la fréquence de l'ensemble moins celle du terme de base par celle du terme de base, est de 19,94 % pour les termes de base simples en moyenne, et de 102,56 % pour les termes de base complexes. Il semble donc que la variabilité de fréquence de variation observée soit attribuable, du moins en partie, aux propriétés des termes eux-mêmes : les termes simples seraient moins propices à la synonymie et la variation, comme l'on pourrait s'y attendre. En effet, les termes complexes offrent des possibilités de réduction, contrairement aux termes simples, et comme nous l'avons vu ci-dessus, la réduction est un phénomène très important. Une analyse plus poussée des propriétés des termes qui favorisent la variation pourrait être effectuée : l'on pourrait isoler tous les termes qui génèrent un nombre anormalement élevé de variantes ou dont les variantes sont très fréquentes, et chercher à identifier des caractéristiques qui les distinguent des autres termes, mais cela est hors de la portée de ce

mémoire. D'autres analyses pourraient également être envisagées, notamment en ce qui concerne la polysémie.

Les quelques caractéristiques présentées dans cette section montrent qu'un corpus étalon comme celui-ci, conçu d'abord et avant tout aux fins de l'évaluation des extracteurs, peut être exploité afin de dégager certaines propriétés des termes eux-mêmes.

4.3 Paramétrage de la liste de référence

La banque de termes peut être utilisée telle quelle afin d'évaluer un extracteur. Il suffit d'extraire les entrées de toutes les fiches dans la banque, puis de comparer la liste à celle que produit l'outil évalué lorsqu'on lui soumet le corpus. Nous décrivons ce processus à la section 4.4.

Or, la richesse des renseignements qui sont consignés dans les annotations et dans la banque de termes permet de paramétrer la liste de référence de différentes façons. Une liste sur mesure peut être extraite de la banque de termes en fonction des champs des fiches terminologiques, des attributs contenus dans les balises des formes annotées dans le corpus, ou de calculs basés sur le corpus annoté. On peut ainsi exclure les termes simples ou les termes complexes si on désire évaluer la capacité d'un extracteur à identifier chacun de ces types de termes. On peut également exclure différents types de variantes, comme les disjonctions, les réductions anaphoriques ou par coordination, les sigles, les variantes orthographiques, etc. L'extraction des entrées peut se faire par différents moyens : on peut utiliser une transformation XSLT ou un script exploitant un parseur XML, par exemple.

À l'aide d'un outil comme le programme que nous avons conçu pour la validation de nos résultats, on peut également filtrer les termes en fonction de leur fréquence ou de leur répartition (nombre de documents où un terme est attesté). On peut ainsi exclure les hapax, ou ne retenir que les termes ayant une certaine fréquence ou une certaine répartition.

4.4 Exploitation de la liste de référence

Pour évaluer un extracteur à l'aide de la liste de référence, il suffit de soumettre une version non annotée du corpus à l'extracteur, et de comparer la sortie qu'il produit à la liste de référence. La liste de référence peut être extraite de la banque de fiches telle quelle, ou paramétrée sur mesure, comme nous l'avons vu à la section 4.3.

Pour automatiser la comparaison de la sortie de l'extracteur à la référence, nous utilisons des métriques (voir section 2.1.1). La précision, le rappel et la F-mesure sont les métriques les plus fréquemment utilisées pour évaluer les extracteurs. L'écriture d'un programme qui calcule les métriques est chose aisée. Or, il faut d'abord définir ce qui constitue un *vrai positif*, c'est-à-dire une correspondance entre un candidat terme et un des termes présents dans la liste de référence. On peut ne retenir que les correspondances exactes, ou permettre un certain degré de différence entre le candidat terme et le terme de référence. Par exemple, on peut stipuler qu'un candidat est considéré valide si une partie du candidat correspond à un terme de la référence, ou si le candidat correspond à une partie d'un terme de la référence.

D'autres métriques peuvent être utilisées. Par exemple, les métriques proposées par Nazarenko et al. (2009) sont implémentées dans l'outil Termometer. Ces métriques visent à trouver une correspondance maximale entre la sortie de l'extracteur et la liste de référence en partitionnant la liste de candidats termes et en exploitant la notion de distance d'édition, afin ne pas favoriser arbitrairement une stratégie d'extraction plutôt qu'une autre (voir section 2.2.2.10).

Afin d'illustrer la mise en application du corpus étalon, ainsi que le possibilités qu'il offre quant au paramétrage de la liste de référence, nous procédons à une évaluation de l'extracteur TermoStat (Drouin 2003). Lorsqu'on lui fournit un corpus, TermoStat produit une liste de candidats termes, triés en fonction de leur potentiel terminologique, comme le montre la Figure 21. TermoStat permet d'extraire non seulement des syntagmes nominaux,

mais des unités simples de différentes parties du discours. Étant donné que notre liste de référence ne contient que des unités nominales, nous demandons à TermoStat d'extraire seulement ce type d'unités.

Candidat de regroupement	Fréquence	Score (Spécificité)	Variantes orthographiques	Matrice
master cylinder	96	54.41	*master cylinder* *master cylinders*	Nom Nom
intake manifold	85	50.23	*intake manifold*	Nom Nom
combustion chamber	78	48.99	*combustion chamber* *combustion chambers*	Nom Nom
fuel injection	78	48.01	*fuel injection*	Nom Nom
throttle valve	65	44.66	*throttle valve* *throttle valves*	Nom Nom
exhaust system	65	43.6	*exhaust system* *exhaust systems*	Nom Nom
drive shaft	64	43.59	*drive shaft* *drive shafts*	Nom Nom

Figure 21 : Exemple de sortie de l'extracteur TermoStat

Dans un premier temps, nous produisons une liste de référence en conservant les 5478 entrées de notre banque de termes. Nous nettoyons le corpus annoté de ses balises et le soumettons à l'extracteur, qui nous offre une liste de 3971 candidats termes. Nous comparons ensuite la sortie de l'extracteur à la liste de référence, en ne retenant que les correspondances exactes entre un candidat terme et un terme de la référence; ainsi, seuls les candidats qui se retrouvent tels quels dans la référence sont considérés comme valides. La Figure 22 présente la précision de l'extraction sur une échelle graduée en fonction du nombre de candidats termes qui sont pris en compte. Par exemple, la précision se situe à 0,632 pour les 1000 premiers candidats, ce qui signifie que 632 de ces candidats se retrouvent dans la liste de référence. Cette précision diminue à mesure qu'on augmente le nombre de candidats pris en compte.

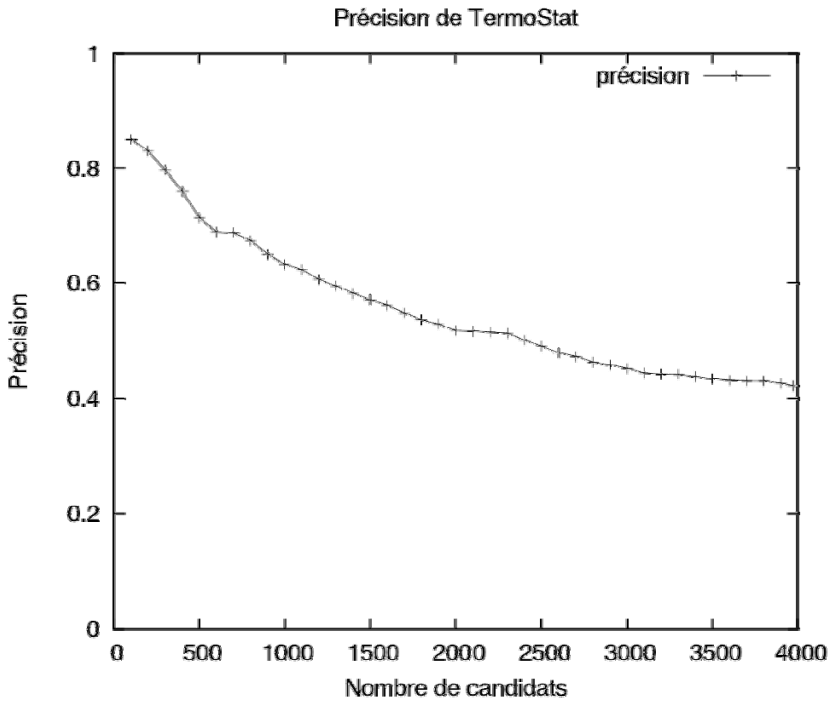

Figure 22 : Précision de l'extraction effectuée par TermoStat sur le corpus étalon

Afin d'illustrer la notion de paramétrage de la référence, nous vérifions quel effet l'inclusion des variantes réduites dans la liste de référence peut avoir sur la précision de l'extraction. Cette fois-ci, en plus de la liste de référence qui contient toutes les entrées de la banque, nous produisons une liste contenant seulement les unités qui ne sont pas des réductions (anaphoriques ou par coordination) d'un terme complexe; cette liste contient 4029 unités. Puis, nous comparons la sortie de TermoStat à chacune de ces deux listes en termes de précision. La Figure 23 présente le résultat de cette évaluation. Elle suggère qu'une proportion non négligeable des formes recensées par l'extracteur ne sont pas des termes de base, mais des réductions anaphoriques ou par coordination. Ainsi, l'évaluation d'un extracteur serait moins stricte si on inclut des réductions de termes complexes dans la référence. L'évaluation est d'ailleurs moins stricte si on considère comme valides des correspondances partielles entre un candidat terme et un terme de la référence, ou si on fait appel à des métriques telles que celles proposées par Nazarenko et al. (2009).

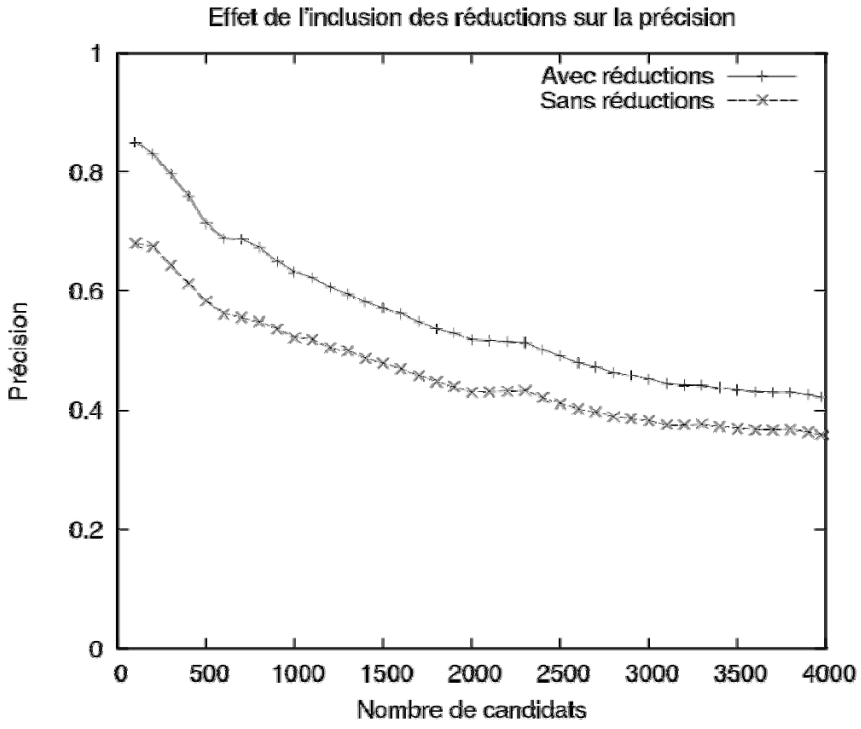

Figure 23 : Effet de l'inclusion des réductions sur la précision de TermoStat

Ces deux exemples ne servent qu'à illustrer comment la liste de référence est mise en application pour évaluer un extracteur. Ce genre de tests peut servir à mesurer divers aspects de la qualité d'un extracteur, ou à effectuer une évaluation comparative de différentes stratégies d'extraction.

Nous avons entrepris des démarches afin de rendre le corpus disponible à la communauté, mais ne savons pas pour le moment si cela sera possible.

4.5 Conclusion

Dans cette section, nous avons décrit comment nous avons procédé pour valider, de façons manuelle et automatisée, le travail d'annotation décrit au chapitre 3. De plus, nous avons analysé quelques aspects du corpus et de la banque de termes et montré qu'ils offrent des renseignements intéressants sur les termes, notamment quant à la variation

terminologique. Enfin, nous avons montré comment exploiter le corpus, et comment il est possible de paramétrer la liste de référence.

Conclusion

L'objectif de ce travail était de construire un corpus étalon pour l'évaluation des extracteurs de termes. Cet étalon est constitué d'un corpus dans lequel toutes les occurrences des termes ont été annotées. Les termes ont été choisis en fonction d'une application spécifique, à savoir la compilation d'un dictionnaire spécialisé de la mécanique automobile, et en fonction d'un ensemble de critères de dépouillement terminologique. Les unités annotées comprennent non seulement les termes apparaissant dans leur forme habituelle, mais aussi des variantes terminologiques telles que les réductions de termes complexes, les sigles et acronymes et les variantes orthographiques.

Le corpus annoté, ainsi que la banque dans laquelle nous avons consigné toutes les unités terminologiques, sont utilisés pour produire une liste de référence. On peut inclure dans cette liste toutes les entrées dans la banque de termes, ou produire une liste sur mesure en filtrant les unités en fonction du type de terme (simple ou complexe), du type de variation terminologique ou de la fréquence. Les possibilités de paramétrage qu'offre le corpus ajoutent une dimension importante à la tâche d'évaluation des extracteurs.

Le corpus étalon est facile à exploiter. Une fois la liste de référence produite, il suffit d'enlever du corpus les balises que nous avons utilisées pour annoter les unités terminologiques, de soumettre le corpus à un extracteur, et de comparer la sortie de l'extracteur à la liste de référence au moyen de métriques. Cela permet une évaluation automatique des extracteurs qui est objective, paramétrable et reproductible. Ainsi, nous pouvons évaluer un extracteur en particulier, comparer différentes techniques d'extraction ou vérifier comment des modifications apportées à un système influent sur sa performance. S'il n'existe toujours pas de cadre unificateur pour l'évaluation des extracteurs, nous croyons que l'existence d'un corpus étalon comme celui-ci constitue une étape importante vers un modèle d'évaluation qui permettra de rendre compte des progrès qui ont été réalisés

dans le domaine de l'extraction de termes et de favoriser l'essor des techniques d'extraction.

Le but premier de ce travail d'annotation de corpus était d'élaborer un corpus étalon permettant d'évaluer la performance des extracteurs de termes. Nous avons montré comment il est possible d'alimenter un extracteur avec un corpus et de comparer les résultats obtenus avec une version du corpus qui a été dépouillée et annotée manuellement dans le langage XML. Étant donné que la liste de référence provient du corpus qui sert pour l'évaluation, elle permet de mesurer non seulement la précision, mais aussi le rappel. Il serait possible d'évaluer l'extracteur sur un autre corpus du domaine de la mécanique automobile en utilisant la même liste de référence, mais on aurait seulement une idée approximative de la précision et du rappel dans ce cas.

Comme nous l'avons montré, il est également possible d'utiliser une transformation XSLT afin de générer, à partir du corpus annoté, un texte dynamique qui permet à l'utilisateur d'obtenir à tout moment la définition d'un terme ou de savoir à quel terme renvoie une variante donnée. Ces textes dynamiques pourraient éventuellement être utiles à d'autres applications, notamment l'apprentissage des langues de spécialité et des savoirs spécialisés ou l'enseignement de la terminologie.

Le travail d'annotation que nous avons effectué est le reflet d'une application possible d'un extracteur, à savoir la compilation d'un dictionnaire spécialisé. Cette application a guidé le repérage et le découpage des termes dans le corpus. Or, les extracteurs sont utilisés à des fins diverses, telles que l'indexation, la recherche d'information, et la traduction. Il serait donc intéressant d'annoter le corpus en couches multiples, en fonction de différentes applications. Ce travail impliquerait nécessairement une redéfinition des critères de dépouillement; l'on pourrait, selon l'application, être amené à annoter d'autres parties du discours, des collocations, des unités qui posent des problèmes de traduction, etc. Une telle entreprise augmenterait les possibilités de paramétrage de la

liste de référence, et permettrait une évaluation des extracteurs en fonction de différentes applications.

Dans le cadre de ce travail, l'annotation et la validation de celle-ci ont été réalisées par la même personne. Dans un travail futur, il serait intéressant de mettre en place un système de validation externe, c'est-à-dire une méthodologie où une personne s'occupe de l'annotation et une autre, de la validation; il pourrait s'agir d'un terminologue ou d'un expert du domaine, par exemple. Il serait aussi intéressant de faire annoter le même corpus par plus d'une personne, ce qui permettrait de mesurer l'accord entre annotateurs.

Le corpus annoté nous fournit de nombreux renseignements sur différents phénomènes de variation terminologique, notamment sur la réduction des termes complexes. Le corpus pourrait donc servir à une étude de la réduction anaphorique, qui viserait à identifier des régularités dans le processus de réduction, ce qui permettrait éventuellement d'identifier en contexte et de reconstruire les réductions anaphoriques automatiquement.

Bien que la taille du corpus ne s'y prête pas facilement, il serait envisageable d'utiliser des techniques d'apprentissage-machine afin de pré-annoter de nouveaux corpus, appartenant éventuellement à des domaines différents.

Pour ne mentionner qu'une dernière piste de recherche, le corpus pourrait alimenter une réflexion sur les notions de « terme », de « synonyme » et de « variante terminologique ». Nous n'avons pas insisté sur la différence entre ces trois notions, mais le corpus annoté permettrait d'observer certaines caractéristiques des différentes unités terminologiques quant à leur forme, leur fréquence, leur comportement dans le corpus, etc. Existe t-il des variantes qui sont plus fréquentes que le terme de base qu'on consigne dans les dictionnaires? Est-ce que certaines propriétés des termes, formelles ou autres, favorisent l'utilisation de variantes ou de synonymes? Quels facteurs favorisent l'implantation des variantes terminologiques? Des éléments de réponse à ces questions pourraient être dégagés d'une analyse plus approfondie du corpus annoté.

Bibliographie

Ahmad, K. & Rogers, M. (2001). Corpus Linguistics and Terminology Management. In *Handbook of Terminology Management, S.E. Wright & Budin, G. (dir.)*. Amsterdam/Philadelphie : John Benjamins, p. 725–760.

Ahmad, K., Davies, A., Fulford, H., & Rogers, M. (1994). What Is a Term? The Semi-Automatic Extraction of Terms from Text. In *Translation Studies: An Interdiscipline*, p. 267–278.

Auger, P., Drouin, P. & Auger, A. (1996). Filtact© : un automate d'extraction des termes complexes. *Terminologies nouvelles*, (15), p. 48–49.

Bourigault, D., Aussenac-Gilles, N. & Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1), p. 87–110.

Cabré, T., Condamines, A., & Ibekwe-SanJuan, F. (2005). Introduction: Application-driven terminology engineering. *Terminology*, 11(1), p. 1–19.

Cabré, T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2), p. 163–199.

Calberg-Challot, M., Candel, D., Bourigault, D., Dumont, X., Humbley, J. & Joseph, J. (2008). Une analyse methodique pour l'extraction terminologique dans le domaine du nucleaire. *Terminology*, 14(2), p. 183–203.

Carl, M., Rascu, E., Haller, J. & Langlais, P. (2004). Abducing term variant translations in aligned texts. *Terminology*, 10(1), p. 101–130.

Carreño Cruz, S. I. (2004). Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue. Mémoire de maîtrise. Université de Montréal.

Chiao, Y. C. & Sta, J. D. (2002). Accès à l'information multilingue et terminologie. Paris : Hermès, p. 111–127.

Collet, T. (1997). La réduction des unités terminologiques complexes de type syntagmatique. *Méta : journal des traducteurs*, 42(1), p. 193–206.

Collier, N., Nobata, C. & Tsujii, J. (2001). Automatic Acquisition and Classification of Terminology Using a Tagged Corpus in the Molecular Biology Domain. *Terminology*, 7(2), p. 239–257.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), p. 213–238.

Dagan, I. & Church, K. (1994). Termight: Identifying and Translating Technical Terminology. *Proceedings of the Fourth Conference on Applied Natural Language Processing*, p. 34–40.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Resnik, P. & Klavans, J. (dir.).* Cambridge : MIT Press, p. 49–66.

Daille, B. (2003). Conceptual Structuring through Term Variations. *Proceedings of the ACL 2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment,* p. 9–16.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1), p. 181–197.

Daille, B., Gaussier, É. & Langé, J.-M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the Fifteenth International Conference on Computational Linguistics (Coling 1994).* p. 515–521.

Damerau, F. J. (1993). Generating and Evaluating Domain-oriented Multi-word Terms from Texts. *Information Processing and Management*, 29(4), p. 433–447.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), p. 99–115.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2), p. 45–64.

Dubuc, R. (2002). Manuel pratique de terminologie, 4e éd. Brossard : Linguatech.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), p. 61–74.

Enguehard, C. (2003). CoRRecT : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. *Actes de la 10e conférence annuelle sur le Traitement Automatique des Langues (TALN 2003),* p. 339–345.

Enguehard, C. & Pantera, L. (1994). Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics*, 2(1), p. 27–32.

Estopà, R. (1999). Eficiencia en la extracción automática de terminología. *Perspectives: Studies in Translatology*, 7(2), p. 277–286.

Estopà, R. (2001). Les unités de signification spécialisées: élargissant l'objet du travail en terminologie. *Terminology*, 7(2), p. 217–237.

Frantzi, K. T. & Ananiadou, S. (1995). Statistical Measures for Terminological Extraction. *Proceedings of the Third International Conference on Statistical Analysis of Textual Data*, p. 297–308.

Frantzi, K. T., Ananiadou, S. & Tsujii, J. (1998). The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. *Proceedings of ECDL '98*, p. 585–604.

Freixa, J. (2002). La variació terminològica. Anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient. Thèse de doctorat. Universitat de Barcelona.

Fulford, H. (2001). Exploring Terms and Their Linguistic Environment in Text: A Domain-Independent Approach to Automated Term Extraction. *Terminology*, 7(2), p. 259–279.

Gaudin, F. (2003). Socioterminologie: une approche sociolinguistique de la terminologie, 1re éd. Bruxelles : De Boeck & Larcier.

Gaussier, É. (2001). General Considerations on Bilingual Terminology Extraction. In *Recent Advances in Computational Terminology, Bourigault, D. et al. (dir.).* Amsterdam/Philadelphie : John Benjamins, p. 167–183.

Guilbert, L. (1973). La spécificité du terme scientifique et technique. *Langue française*, 17(1), p. 5–17.

Haralambous, Y. & Lavagnino, E. (2011). La réduction de termes complexes dans les langues de spécialité. *TAL*, 52(1), p. 37–68.

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. & Takano, A. (2000). Extracting terms by a combination of term frequency and a measure of term representativeness. *Terminology*, 6(2), p. 211–232.

Ibekwe-SanJuan, F. & SanJuan, E. (2002). From term variants to research topics. *Knowledge Organisation*, 29(3/4), p. 181–197.

Jacquemin, C. (2001). Spotting and Discovering Terms through Natural Language Processing. Cambridge : MIT Press.

Jacquemin, C. & Tzoukermann, E. (1999). NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In *Natural Language Information Retrieval, Strzalkowski, T. (dir.)*. Boston : Kluwer, p. 25–74.

Justeson, J. S. & Katz, S. M. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, 1(1), p. 9–27.

Kageura, K. et al. (1999a). Overview of the TMREC Tasks. *NTCIR Workshop 1: Proceedings of the first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, p. 415.

Kageura, K. et al. (1999b). Evaluation of the Term Recognition Task. *NTCIR Workshop 1: Proceedings of the first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, p. 417–434.

Kageura, K. & Umino, B. (1996). Methods of Automatic Term Recognition: A Review. *Terminology*, 3(2), p. 259–289.

Kageura, K., Yoshioka, M., Takeuchi, K., Koyama, T., Tsuji, K. & Yoshikane, F. (2000). Recent advances in automatic term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology*, 6(2), p. 151–173.

King, M., Maegaard, B., Schütz, J., des Tombes, L., Bech, A., Neville, A., Arppe, A. et al. (1996). *EAGLES Evaluation of Natural Language Processing Systems*. En ligne.

<http://www.issco.unige.ch/en/research/projects/ewg95/node1.html>. Consulté le 15 mai 2012.

Koyama, T., Yoshioka, M. & Kageura, K. (1998). The construction of a lexically motivated corpus: The problem with defining lexical units. *Proceedings of First International Conference on Language Resources and Evaluation*, p. 1015–1019.

L'Homme, M.-C. (2002). Nouvelles technologies et recherche terminologique : Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. In *L'impact des nouvelles technologies sur la gestion terminologique (18 août 2001, Université York, Toronto)*. En ligne. <http://olst.ling.umontreal.ca/?page_id=75>. Consulté le 4 avril 2012.

L'Homme, M.-C. (2004). La terminologie : principes et techniques. Montréal : Presses de l'Université de Montréal.

L'Homme, M.-C. (2005). Sur la notion de « terme ». *Meta*, 50(4), p. 1112–1132.

L'Homme, M.-C., Benali, L., Bertrand, C. & Lauduique, P. (1996). Definition of an Evaluation Grid for Term-Extraction Software. *Terminology*, 3(2), p. 291–312.

L'Homme, M.-C., Heid, U. & Sager, J. C. (2003). Terminology during the Past Decade (1994-2004): An Editorial Statement. *Terminology*, 9(2), p. 151–161.

Ladouceur, J. & Cochrane, G. (1996). Termplus, système d'extraction terminologique. *Terminologies nouvelles*, (15), p. 52–56.

Lauriston, A. (1994). Automatic Recognition of Complex Terms: Problems and the TERMINO Solution. *Terminology*, 1(1), p. 147–170.

Lemay, C., L'Homme, M.-C. & Drouin, P. (2005). Two Methods for Extracting "Specific" Single-Word Terms from Specialized Corpora. *International Journal of Corpus Linguistics*, 10(2), p. 227–255.

Love, S. (2000). Benchmarking the Performance of Two Automated Term-Extraction Systems: LOGOS and ATAO. Mémoire de maîtrise. Université de Montréal.

Mima, H. & Ananiadou, S. (2000). An Application of the C/NC-Value Approach for the Automatic Term Recognition of Multi-Word Units in Japanese. *Terminology*, 6(2), p. 175–194.

Mustafa El Hadi, W. (2004). L'évaluation d'outils d'acquisition de ressources terminologiques. In *Évaluation des systèmes de traitement de l'information*. Paris : Hermès, p. 149–169.

Mustafa El Hadi, W. & Chaudiron, S. (2007). L'évaluation d'outils d'acquisition de ressources terminologiques : problèmes et enjeux. *Actes de TOTh 2007*, p. 163–179.

Mustafa El Hadi, W., Timimi, I., Béguin, A. & Dabbadie, M. (2001). The ARC A3 Project: Terminology Acquisition Tools: Evaluation Method and Tasks. *Proceedings of Evaluation Methodologies for Language and Dialogue Systems Workshop, ACL/EACL*, p. 41–50.

Mustafa El Hadi, W., Timimi, I., Dabbadie, M., Choukri, K., Hamon, O. & Chiao, Y.-C. (2006). Terminological Resources Acquisition tools : Towards a User-Oriented Evaluation Model. *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 945–948.

Nakagawa, H. (2000). Automatic Term Recognition Based on Statistics of Compound Nouns. *Terminology*, 6(2), p. 195–210.

Nazarenko, A. & Zargayouna, H. (2009). Evaluating Term Extraction. *Actes de RANLP 2009*, p. 299–304.

Nazarenko, A., Zargayouna, H., Hamon, O. & van Puymbrouck, J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *TAL*, 50(1), p. 257–281.

OLF (Office de la langue française). (1979). Table ronde sur les problèmes de découpage du terme (26 août 1978). Montréal/Québec : Éditeur officiel du Québec.

Otman, G. (1991). Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur. *La Banque des mots*, (no spécial 4), p. 59–96.

Paroubek, P., Chaudiron, S. & Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *TAL*, 48(1), p. 7–31.

Patry, A. & Langlais, P. (2005). Corpus-Based Terminology Extraction. *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)*.

Pearson, J. (1998). *Terms in Context*. Amsterdam : J. Benjamins.

Phal, A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique*. Paris : Didier.

Picht, H. & Draskau, J. (1985). *Terminology: An Introduction*. Guildford, England : University of Surrey, Department of Linguistic and International Studies.

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *TAL*, 48(1), p. 67–91.

Rey, A. (1979). La terminologie : noms et notions. Collection Que sais-je? Paris : Presses universitaires de France.

Rondeau, G. (1984). Introduction à la terminologie, 2e éd. Chicoutimi : Gaëtan Morin Éditeur.

Sager, J. C. (1990). A Practical Course in Terminology Processing. Amsterdam/Philadelphie : John Benjamins.

Sauron, V. (2002). Tearing out the Terms: Evaluating Terms Extractors. *Proceedings of the Aslib conference Translating and the Computer 24*.

Savary, A. (2001). Étude comparative de deux outils d'acquisition de termes complexes. *Actes des Quatrièmes Rencontres « Terminologie et Intelligence Artificielle », INIST-CNRS*, p. 129–139.

Tiedemann, J. (2001). Can bilingual word alignment improve monolingual phrasal term extraction? *Terminology*, 7(2), p. 199–215.

Timimi, I. (2006). Évaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques. *Actes des 8e journées internationales d'Analyse statistique des Données Textuelles (JADT 2006),* p. 895–906.

Timimi, I. (2007). Peut-on faire confiance aux outils de terminologie? L'évaluation entre un souci de normalisation et une complexité de modélisation. *Actes de TOTh 2007*, p. 143–162.

Timimi, I. & Mustafa El Hadi, W. (2008). CESART : une campagne d'évaluation de systèmes d'acquisition de ressources terminologiques. In *L'évaluation des technologies en traitement de la langue : Les campagnes.* Paris : Hermès, p. 71–91.

Van Rijsbergen, C. J. (1979). Information Retrieval, 2e éd. London : Butterworths.

Vintar, S. (2004). Comparative evaluation of C-value in the treatment of nested terms. *Proceedings of Memura 2004 — Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC 2004)*, p. 54–57.

Vivaldi, J. & Rodríguez, H. (2001). Improving term extraction by combining different techniques. *Terminology*, 7(1), p. 31–48.

Vivaldi, J. & Rodríguez, H. (2007). Evaluation of Terms and Term Extraction Systems: A Practical Approach. *Terminology*, 13(2), p. 225–248.

Annexe 1 : Extrait brut du corpus

Piston Rings

The piston must have some clearance in the cylinder. If the skirt has .001 "" to .002" (0.025 to 0.05 mm) clearance and the head has .030" to .040" (0.76 to 1.02 mm) clearance, it is obvious that the piston cannot seal the cylinder effectively. See Figure 2-23.

To solve the leakage problem, piston rings are used. A properly constructed and fitted ring will rub against the cylinder wall with good contact all around the cylinder. The ring will ride in a groove that is cut into the piston head. There will be a slight clearance between the sides of the ring and the edges of the groove. This clearance, which is known as side clearance is generally around .002" (0.05 mm).

The rings do not contact the bottom of the ring grooves. Actually, the ring will rub the cylinder wall at all times but will be solidly fastened to the piston at any one point. See Figure 2-24.

Ring Gap

The ring is built so it must be squeezed together to place it in the cylinder. This will cause the ring to exert an outward pressure, which keeps it tightly against the cylinder wall. See Figure 2-25. The ring is not solid all the way around, but is cut through in one spot. This cut spot forms what is called the ring gap. See Figure 2-26.

When the ring is in the cylinder, the cut ends must not touch. When the ring heats up, it will lengthen. Since it cannot expand outwardly, it will close the gap. See Figure 2-27. If there is not enough gap clearance, the ends will touch as the ring expands. As the ring continues to lengthen, it will break up into several pieces. This can quickly ruin an engine.

A general rule for ring gap clearance is to allow .003" to .004" per inch (0.07 to 0.10 mm per 25.4 mm) of cylinder diameter. For example, a 4" (101.6 mm) diameter cylinder would require from .012" to .016" (0.30 to 0.41 mm) clearance in the gap.

Many different types of joints have been used in an endeavor to stop leakage through the ring gap. This leakage is commonly referred to as blow-by. It has been found that the common butt joint is effective and is simple to adjust. Figure 2-28 illustrates a few of the joints that have been used.

The ring is placed in the groove by expanding it until it will slip over the piston head and slide down and into the ring groove. Figure 2-29 illustrates how a ring fits the piston groove. Note that there is both side clearance and back clearance.

Types of Rings

There are two distinct types of rings. One is called a compression ring, and the other is called an oil control ring. Most engines have three rings: two compression rings at the top and one oil control ring at the bottom, Figure 2-30.

Compression Rings

Compression rings are designed to prevent leakage between the piston and the cylinder, Figure 2-31.

Various types of grooves, bevels, and chamfer shapes are used to achieve this goal. The idea is to create an internal stress within the ring. This stress will tend to cause the ring to twist in such a fashion that the lower edge of the ring is pressed against the cylinder wall on the intake stroke. This will cause the ring to act as a mild scraper. The scraping effect will assist in the removal of surplus oil that may have escaped the oil control rings, Figure 2-32.

Annexe 2 : Extrait annoté du corpus

<term id="1140" type="comp">Compression Rings</term>
<term id="1140" type="comp">Compression rings</term> are designed to prevent leakage between the <term id="224" type="simp">piston</term> and the <term id="78" type="simp">cylinder</term>, Figure 2-31.
Various types of grooves, bevels, and chamfer shapes are used to achieve this goal. The idea is to create an internal stress within the <term id="1098" type="simp" struct="anaphore">ring</term>. This stress will tend to cause the <term id="1098" type="simp" struct="anaphore">ring</term> to twist in such a fashion that the lower edge of the <term id="1098" type="simp" struct="anaphore">ring</term> is pressed against the <term id="1137" type="comp">cylinder wall</term> on the <term id="128" type="comp">intake stroke</term>. This will cause the <term id="1098" type="simp" struct="anaphore">ring</term> to act as a mild scraper. The scraping effect will assist in the removal of surplus <term id="41" type="simp">oil</term> that may have escaped the <term id="1115" type="comp">oil control rings</term>, Figure 2-32.
On <term id="181" type="simp" struct="coord">compression</term> and <term id="133" type="comp">exhaust strokes</term>, the <term id="1098" type="simp" struct="anaphore">rings</term> will tend to slip lightly over the <term id="41" type="simp">oil</term> film. This will prolong the life of the <term id="1098" type="simp" struct="anaphore">ring</term>, Figure 2-33. On the <term id="58" type="comp">power stroke</term>, pressure of the burning gases will force the top edge of the <term id="1098" type="simp" struct="anaphore">ring</term> downward. This causes the <term id="1098" type="simp" struct="anaphore">ring</term> to rub the <term id="1138" type="simp" struct="anaphore">wall</term> with full face contact and provides a good seal for the enormous pressure generated by the <term id="58" type="comp">power stroke</term>. See Figure 2-34.

Annexe 3 : Liste des 50 termes de base les plus fréquents

- engine (1354)
- car (1077)
- fuel (584)
- oil (531)
- tire (420)
- battery (327)
- wheel (272)
- coolant (268)
- cylinder (243)
- piston (240)
- crankshaft (163)
- gasoline (163)
- carburetor (157)
- gear (148)
- valve (147)
- brake (142)
- transmission (127)
- alternator (110)
- radiator (102)
- camshaft (98)
- fuse (97)
- circuit (92)
- refrigerant (91)
- intake manifold (85)
- cooling system (82)
- bearing (81)
- combustion chamber (78)
- brake fluid (76)
- master cylinder (75)
- sensor (72)
- clutch (69)
- crankcase (69)
- thermostat (66)
- exhaust system (61)
- compressor (61)
- hood (60)
- flywheel (59)
- frame (59)
- axle (58)
- transmission fluid (57)
- spark plug (55)
- brake pedal (54)
- manual transmission (53)
- automatic transmission (52)
- steering wheel (52)
- drive shaft (50)
- throttle valve (50)
- rotor (47)
- muffler (45)
- differential (44)

Annexe 4 : Ouvrages de référence

Bureau de la traduction. (2006). Termium Plus. En ligne. <http://www.btb.termiumplus.gc.ca>. Consulté le 2 mai 2012.

Dinkel, J. (2000). The Road & Track Illustrated Automotive Dictionary. Cambridge : Bentley Publishers, 253 p.

Dwiggins, B. H. & South, D. W. (1997). Delmar's Automotive Dictionary. New York : Delmar Publishers, 281 p.

Goodsell, D. (1995). Dictionary of Automotive Engineering, 2e éd. Oxford : Butterworth-Heinemann, 265 p.

Lane, K. (2002). Automotive A-Z: Lane's Complete Dictionary of Automotive Terms. Dorchester : Veloce Publishing, 352 p.

Motorera. (2003). Dictionary of Automotive Terms. En ligne. <http://www.motorera.com/dictionary>. Consulté le 2 mai 2012.

Office de la langue française. (2006). Grand Dictionnaire terminologique. En ligne. <http://www.granddictionnaire.com/>. Consulté le 2 mai 2012.

Popular Mechanics. (2005). Complete Car Care Manual. New York : Hearst Books, 346 p.

Ramsey, D. (2003). The Complete Idiot's Guide to Car Care and Repair Illustrated. Indianapolis : Alpha, 207 p.

Sclar, D. (1999). Auto Repair for Dummies. Hoboken : Wiley Publishing, 567 p.

Annexe 5 : Exemples de fiches

```
<fiche>
      <ID>541</ID>
      <entree>air conditioner</entree>
      <var></var>
      <SourceLem>DELMAR'S AUTOMOTIVE DICTIONARY</SourceLem>
      <defmod>A device used for the automatic control of the temperature,
      humidity, cleanness, and movement of air in a car.</defmod>
      <def>A device used for the automatic control of the temperature,
      humidity, cleanness, and movement of air in a given space.</def>
      <SourceDef>DELMAR'S AUTOMOTIVE DICTIONARY</SourceDef>
      <IDvar></IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>676</ID>
      <entree>air-conditioner</entree>
      <var>Variante orthographique</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>1340</ID>
      <entree>air conditioning system</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>1515</ID>
      <entree>automobile air conditioning system</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
```

```
<fiche>
      <ID>1518</ID>
      <entree>vehicle air conditioner</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>5373</ID>
      <entree>automotive air conditioning system</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>5376</ID>
      <entree>automotive air conditioner</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>5511</ID>
      <entree>auto air conditioning system</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
```

```
<fiche>
      <ID>5525</ID>
      <entree>automotive air conditioning</entree>
      <var>Synonyme</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>689</ID>
      <entree>air</entree>
      <var>Synonyme</var>
      <SourceLem>DICTIONARY OF AUTOMOTIVE TERMS</SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>275</ID>
      <entree>air conditioning</entree>
      <var>Synonyme</var>
      <SourceLem>AUTOMOTIVE A-Z</SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>541</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>5654</ID>
      <entree>system</entree>
      <var>Troncature</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>1340</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
```

```
<fiche>
      <ID>301</ID>
      <entree>air-conditioning system</entree>
      <var>Variante orthographique</var>
      <SourceLem>TERMIUM</SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>1340</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>680</ID>
      <entree>A/C system</entree>
      <var>Acronyme ou Sigle</var>
      <SourceLem>GDT</SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>1340</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>412</ID>
      <entree>system</entree>
      <var>Troncature</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>301</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
<fiche>
      <ID>302</ID>
      <entree>A/C</entree>
      <var>Acronyme ou Sigle</var>
      <SourceLem>DICTIONARY OF AUTOMOTIVE TERMS</SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>275</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
```

```
<fiche>
      <ID>4320</ID>
      <entree>air-conditioning</entree>
      <var>Variante orthographique</var>
      <SourceLem></SourceLem>
      <defmod></defmod>
      <def></def>
      <SourceDef></SourceDef>
      <IDvar>275</IDvar>
      <commentaire></commentaire>
      <statut>valide</statut>
</fiche>
```