Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

Automatiser et quantifier l'analyse sémantique du français technique

Ann Bertels

ILT et QLVL, K.U.Leuven, Belgique
ann.bertels@ilt.kuleuven.be

1 Introduction

Cet article sur la quantification de l'analyse sémantique s'inscrit dans le cadre d'une étude sémantique du vocabulaire spécifique d'un corpus en français technique (Bertels, 2006). L'objectif principal de l'étude est d'étudier la sémantique (la monosémie ou la polysémie) des mots et termes spécifiques d'un corpus technique. Les textes du corpus (1,7 million d'occurrences) relèvent du domaine technique restreint des machines-outils pour l'usinage des métaux. Le recours aux corpus électroniques de textes spécialisés permet d'accéder à une information indispensable pour l'analyse sémantique, à savoir le contexte linguistique. Toutefois, l'exploitation de grandes quantités de textes requiert une approche automatisée et quantitative, parce qu'il est impossible d'analyser manuellement tous les contextes d'apparition de plusieurs milliers de mots spécifiques du corpus technique.

Pour étudier la sémantique dans la langue spécialisée, l'approche adoptée traditionnellement était une approche catégorielle, qui se caractérisait, d'une part, par la dichotomie entre mot et terme (ou entre langue générale et langue spécialisée) et, d'autre part, par la dichotomie entre polysémie et monosémie. Selon les partisans de la terminologie traditionnelle, les termes de la langue spécialisée sont idéalement monosémiques, tandis que la polysémie est réservée aux mots de la langue générale (Wüster, 1931).

Toutefois, récemment, la thèse traditionnelle de monosémie et son approche onomasiologique prescriptive ont été remises en question par les partisans de la terminologie descriptive (Cabré, 2000 ; Temmerman, 2000 ; Gaudin, 2003). Les deux dichotomies sont également remises en question, puisqu'elles ne s'avèrent pas toujours opérationnelles. Premièrement, sur le plan des unités linguistiques, la dichotomie entre langue générale et langue spécialisée est rejetée (Condamines et Rebeyrolle, 1997). Les termes font partie de la langue naturelle et se caractérisent par le fait qu'ils véhiculent des connaissances spécialisées (Lerat, 1995). Par ailleurs, le vocabulaire d'un corpus technique ne contient pas uniquement des mots techniques ou « termes » au sens strict, propres au domaine de spécialité, tels que *usinage*, mais également des mots du VGOS (vocabulaire général d'orientation scientifique) (Phal, 1971). Ces mots s'emploient dans plusieurs domaines scientifiques et techniques et leur sens est déterminé par les contextes spécialisés (*machine, outil*). Le vocabulaire d'un corpus technique comprend aussi des mots de la langue générale, tels que *type, modèle, permettre*, etc.

Deuxièmement, sur le plan de l'analyse sémantique, la monosémie de la langue spécialisée est remise en question notamment par la Théorie Communicative de la Terminologie (Cabré, 2000) et par la Terminologie socio-cognitive (Temmerman, 2000). En plus, on a assisté à l'émergence de vastes corpus spécialisés, qui ont permis des études sémantiques à partir du contexte linguistique et qui ont abouti à l'observation de cas de polysémie dans la langue spécialisée, même à l'intérieur d'un (sous-)domaine spécialisé (Arnzt et Picht, 1989 ; Condamines et Rebeyrolle, 1997 ; Temmerman, 2000 ; Eriksen, 2002 ; Ferrari, 2002). Ces travaux étudient, comme nous, la polysémie dans un corpus représentatif d'un domaine spécialisé, mais ils se limitent à l'analyse sémantique de quelques mots seulement. Les remises en question théoriques et les études sémantiques limitées de corpus spécialisés, nous incitent à une remise en question à plus grande échelle. Nous nous proposons dès lors d'adopter une approche alternative, c'est-à-dire une approche quantitative et scalaire, et de procéder à l'étude sémantique automatisée de plusieurs milliers de mots.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

L'objectif global de notre étude empirique est de vérifier si les unités lexicales du corpus technique sont monosémiques, comme le prétendent les monosémistes de la terminologie traditionnelle ou, par contre, s'il existe des unités lexicales polysémiques, comme le suggèrent les partisans de la terminologie descriptive. Pour évaluer la thèse monosémiste de l'approche traditionnelle, en ayant recours à la linguistique de corpus, il faut opérationnaliser la thèse monosémiste et la reformuler en une question opérationnelle et mesurable. S'il est vrai que les unités lexicales de la langue spécialisée (d'un corpus technique) sont monosémiques, ce sera d'autant plus vrai pour les unités lexicales les plus spécifiques et les plus représentatives de ce corpus technique. Par conséquent, nous nous demandons si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques.

Pour répondre à cette question, nous procédons à une double analyse quantitative (Bertels et al., 2006). Dans un premier temps, nous repérons tous les mots spécifiques du corpus technique (c'est-à-dire les « spécificités ») et nous leur accordons un « degré de spécificité ». Ensuite, nous procédons à une analyse sémantique quantitative, afin de déterminer le « degré de monosémie » de ces spécificités, en implémentant la monosémie en termes d'homogénéité sémantique. Finalement, les données de la double analyse quantitative font l'objet d'une analyse statistique, qui étudie la corrélation entre le degré de spécificité et le degré de monosémie. Celle-ci permet donc de vérifier si les unités lexicales (les plus) spécifiques du corpus technique de langue spécialisée sont (les plus) monosémiques. Le corpus technique analysé contient effectivement des mots polysémiques ou hétérogènes sémantiquement, par exemple le mot *broche* signifie (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques » et le mot *tour* signifie (1) « machine-outil pour l'usinage des pièces » et (2) « rotation » et il a un sens général (3) dans *attendre son tour / à son tour*.

L'originalité de notre étude réside principalement dans l'approche scalaire de la sémantique et dans le développement d'une mesure pour évaluer le degré de monosémie. Cette mesure permettra non seulement de quantifier la monosémie et d'automatiser l'analyse sémantique, mais également de procéder à des analyses statistiques en vue de fournir une réponse objective à la question de la corrélation entre la spécificité et la monosémie. En plus, notre analyse sémantique quantitative porte sur quelque 5000 mots du corpus technique, contrairement aux travaux antérieurs (Cf. ci-dessus). Tant l'approche scalaire que le nombre important de mots à analyser requièrent une analyse automatisée et quantitative. De par son approche, notre étude vise à réconcilier la linguistique et la technique (notamment l'informatique et la statistique). Elle recourt à la technique pour mieux comprendre certains aspects de la linguistique.

Dans cet article, nous nous concentrons sur l'analyse sémantique quantitative. Dans les sections suivantes, nous situerons la méthodologie par rapport aux travaux antérieurs (section 2) et nous expliquerons notre mesure de monosémie (section 3). Ensuite, nous discuterons les résultats de l'analyse sémantique quantitative (section 4) et la mise au point de la mesure de monosémie (section 5). Nous terminerons par les conclusions et les perspectives de recherche (section 6).

2 Quantifier l'analyse sémantique des spécificités à partir des cooccurrences

2.1 Quelles sont les spécificités ?

Le premier volet de la double analyse quantitative consiste à identifier les spécificités ou les unités linguistiques spécifiques du corpus technique et à déterminer leur degré de spécificité. Les spécificités ne sont pas les unités les plus fréquentes, mais ce sont les unités les plus représentatives du corpus technique. En termes relatifs, les spécificités sont significativement plus fréquentes dans le corpus technique que dans un corpus de référence de langue générale. Etant donné que plusieurs méthodes et outils[1] sont disponibles à cet effet, nous n'entrerons pas dans les détails ici. Ces outils permettent non seulement de repérer les spécificités[2], mais aussi de déterminer leur degré de spécificité, grâce à une mesure statistique, par exemple le LLR (*log likelihood ratio* ou log de vraisemblance) (Dunning, 1993). Plus une unité linguistique est spécifique dans le corpus technique par rapport au corpus de référence de langue générale,

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

plus son degré de spécificité sera élevé. Le degré de spécificité permet ensuite d'ordonner les spécificités, de leur accorder un rang de spécificité[3] et de les situer sur un continuum de spécificité. Les unités les plus spécifiques sont généralement très fréquentes dans le corpus technique (par exemple *machine, usinage*, *broche*) et elles reflètent clairement la thématique du domaine. Après avoir supprimé les hapax, les mots grammaticaux et les noms propres, nous recensons 4717 unités lexicales spécifiques dans le corpus technique. A présent, nous étudions uniquement les unités lexicales simples, mais nous projetons d'étudier ultérieurement aussi les unités polylexicales.

2.2 Comment quantifier l'analyse sémantique ?

Le deuxième volet de la double analyse quantitative vise à quantifier l'analyse sémantique, afin de pouvoir déterminer le degré de monosémie des 4717 spécificités ou unités lexicales spécifiques du corpus technique. Pour y arriver, nous recourons à l'analyse des cooccurrences (Grossmann et Tutin, 2003 ; Condamines, 2005 ; Blumenthal et Hausmann, 2006). Celle-ci permet de quantifier la monosémie en l'implémentant en termes d'homogénéité sémantique (Habert et al., 2005). En effet, une unité lexicale monosémique apparaît dans des contextes plutôt homogènes sémantiquement, c'est-à-dire qu'elle se caractérise par des cooccurrents qui appartiennent à des champs sémantiques similaires. Par contre, une unité lexicale polysémique se caractérise par des cooccurrents plus hétérogènes sémantiquement, qui appartiennent à des champs sémantiques différents (Véronis, 2003 ; Habert et al., 2004).

Habert et al. (2004) cherchent à détecter les mots qui ont plusieurs sens ou qui sont employés simultanément avec des sens divergents dans différentes parties du corpus. Afin de détecter ces mots aux sens mouvants, les auteurs proposent de recourir aux cooccurrents de ces mots, étant donné que ces mots changent souvent de voisins. Ils avancent l'hypothèse qu'un mot à sens multiple « aurait des voisins moins proches entre eux qu'un mot univoque » (Habert et al., 2004 : 570). Il s'ensuit que les mots homonymiques, polysémiques et vagues auraient des cooccurrents sémantiquement plus hétérogènes. En effet, les homonymes ont des contextes d'emploi souvent très différenciés. Toutefois, les sens des mots polysémiques, sémantiquement apparentés, ont plus de chances de se retrouver « dans des contextes proches » (Habert et al., 2004 : 566). Dès lors, les cooccurrents des mots polysémiques seront moins hétérogènes que ceux des mots homonymiques. Ces observations confirment notre intention d'adopter l'idée d'un continuum d'homogénéité sémantique avec plusieurs degrés, en fonction des cooccurrents plus ou moins homogènes.

Or, en sémantique distributionnelle et contextuelle, deux problèmes majeurs se posent. D'une part, la distribution des différents sens d'un mot dans le corpus est souvent irrégulière et, d'autre part, « la répartition des traits permettant de classer les mots est souvent très éparpillée » (Habert et al., 2004 : 573). Pour remédier à ces problèmes, Habert et al. (2004) suggèrent de recourir aux cooccurrents et aux similarités de deuxième ordre. De plus, les cooccurrents de premier ordre sont généralement des cooccurrents syntagmatiques du mot de base et parfois des cooccurrents paradigmatiques. Par contre, les cooccurrents de deuxième ordre se caractérisent principalement par des relations paradigmatiques avec le mot de base (hyponymes, hyperonymes, synonymes, antonymes) et dès lors ces derniers sont plus intéressants pour caractériser sémantiquement le mot de base.

Les cooccurrents de deuxième ordre ou les cooccurrents des cooccurrents permettent entre autres de mettre en évidence des relations de synonymie (Martinez, 2000). Pour un mot de base (ou « pôle ») tel que *mesures*, Martinez (2000) calcule d'abord tous les cooccurrents de *mesures*, comme *nouvelles, unilatérales, concrètes, adopter, prises*. L'étape suivante consiste à calculer les cooccurrents des cooccurrents les plus spécifiques[4] (*nouvelles, prises*), ce qui revient à déterminer les cooccurrents de deuxième ordre, par exemple *décisions, dispositions, initiatives, monétaires*. Comme Martinez cherche les synonymes du pôle initial (*mesures*), il retient uniquement des cooccurrents de deuxième ordre qui apparaissent à la fois avec *nouvelles* et avec *prises* (Cf. figure 1).

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

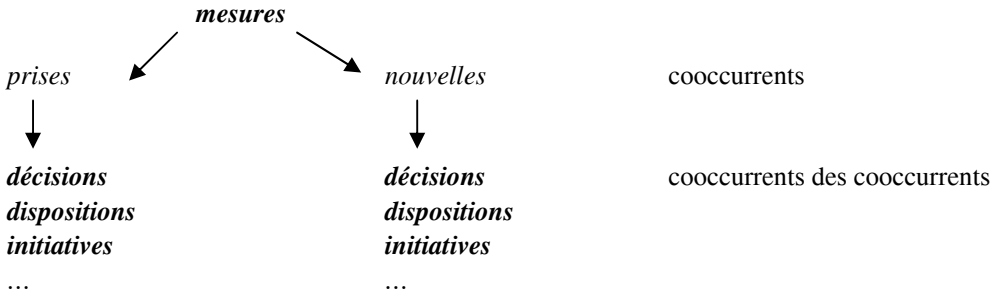

Figure 1 Cooccurrents des cooccurrents pour la détection de synonymes (Cf. Martinez, 2000)

Martinez (2000) identifie tant les cooccurrents que les cooccurrents des cooccurrents au niveau des formes fléchies, ce qui lui permet de préserver la distinction entre le singulier et le pluriel. L'analyse de l'axe syntagmatique, effectuée à deux reprises, contribue ainsi à la découverte de l'axe paradigmatique (les synonymes). Il est clair que les différents synonymes d'un mot de base sont des indices sémantiques précieux dans la perspective de la mesure que nous envisageons de développer.

Comme nous l'avons évoqué ci-dessus, une unité lexicale monosémique apparaît dans des contextes plutôt homogènes sémantiquement, tandis qu'une unité lexicale polysémique se caractérise par des cooccurrents plus hétérogènes sémantiquement. L'accès à la sémantique des cooccurrents d'un mot de base pourrait se faire à partir de leurs cooccurrents, c'est-à-dire à partir des cooccurrents de deuxième ordre. Si les cooccurrents d'un mot de base (ou les cooccurrents de premier ordre) partagent beaucoup de cooccurrents de deuxième ordre, ces derniers se recoupent formellement, ce qui est une indication de l'homogénéité sémantique des cooccurrents de premier ordre et, dès lors, du mot de base (Cf. Martinez, 2000). Le degré de ressemblance lexicale ou de similarité lexicale des cooccurrents d'un mot de base est donc proportionnel au degré de monosémie de ce mot de base. La similarité distributionnelle reflète clairement la similarité sémantique. Par conséquent, un recoupement important des cooccurrents de deuxième ordre révèle un degré plus important de monosémie du mot de base. Par contre, si les cooccurrents de deuxième ordre sont formellement (très) différents, ils se recoupent (très) peu et ils sont (très) peu partagés. Par conséquent, les différents cooccurrents sont sémantiquement plus diversifiés, car une structure formelle de cooccurrence différente indique un sens différent. Si les cooccurrents sémantiquement diversifiés appartiennent à plusieurs champs sémantiques, le mot de base aura moins de chances d'être monosémique.

Regardons quelques phrases-exemples avec le mot *tour* (Cf. phrases 1 et 2 ci-dessous). Pour avoir une idée de la sémantique de *tour*, on regarde son contexte : le cooccurrent *usine* (Cf. phrase 1) indique clairement le sens « machine-outil », tandis que le cooccurrent *minute* (Cf. phrase 2) indique le sens « rotation ». Un être humain sait très bien interpréter le sens de ces cooccurrents, mais pour une machine ou un ordinateur, ce n'est pas clair. Il faut donc analyser les cooccurrents de ces cooccurrents, dans la même phrase (par exemple *alésage* dans la première phrase comme cooccurrent d'*usine*) et dans tous les autres contextes (par exemple *pièces* et *outils* dans la troisième phrase comme cooccurrents d'*usine*). Plus les cooccurrents des cooccurrents d'un mot de base (*alésage, pièces, outils, ...*) se recoupent, plus le mot de base (*tour*) sera homogène sémantiquement.

(1) *La première est un* ***tour*** *sur lequel on* ***usine*** *l'alésage central. ...*

= « machine-outil pour l'usinage des pièces »

(2) *... broches pouvant monter jusqu'à quinze mille* ***tours*** *par* ***minute****, ...*

= « rotation, révolution »

(3) *Un tour CNC équipé d'outils modulaires Capto usine les pièces en question avec une vitesse de ...*

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

3 La mesure de monosémie

Afin de calculer le degré de recoupement des cooccurrents des cooccurrents et dès lors le degré de monosémie d'un mot de base, nous avons développé une mesure de recoupement ou de monosémie (Cf. figure 2). Elle est basée sur le recoupement formel des cooccurrents des cooccurrents (ou cc) d'un mot de base et tient compte des paramètres suivants :

- la fréquence d'un cc dans la liste des cc (= le nombre de cooccurrents (ou c) apparaissant avec ce cc)
- le nombre total de c (= cooccurrents de premier ordre)
- le nombre total de cc (= cooccurrents de deuxième ordre)

Un cc sera d'autant plus important pour le recoupement total s'il figure plus souvent dans la liste des cc, c'est-à-dire si sa fréquence dans la liste des cc est plus élevée ou s'il est plus souvent partagé par les cooccurrents ou c.

$$\sum_{\text{cc}} \frac{\text{fq } cc}{\#\text{ total } c \cdot \#\text{ total } cc}$$

Figure 2 : Formule de recoupement des cooccurrents des cooccurrents

Considérons en guise d'exemple un cc fortement partagé : il est partagé par exemple par 5 c des 7 c au total. Cela veut dire que 5 c des 7 c apparaissent avec ce cc en question, ce qui indique un recoupement important. Dès lors, nous proposons d'inclure dans le numérateur le nombre de c qui ont ce cc en commun (fq cc), par exemple 5, et d'inclure dans le dénominateur le nombre total de c, par exemple 7. Le recoupement est donc exprimé par la fraction 5/7. En exprimant pour chaque cc le recoupement par la fraction « nombre de c avec le cc » (ou fq cc) divisé par « nombre total de c », le résultat se situe toujours entre 0 (pas ou peu de recoupement) et 1 (recoupement important ou parfait) et par conséquent, le résultat sera facilement interprétable. Si on somme pour tous les cc, il faut ajouter dans le dénominateur le « nombre total de cc » (au niveau des tokens), car on considère tous les cc, avec les doublons responsables du recoupement. Il ne s'agit pas du nombre de cc différents, mais du nombre total de cc, à savoir tous les mots qui apparaissent avec tous les c.

Remarquons que nous considérons les c et les cc au niveau des formes graphiques et non pas au niveau des lemmes (formes canoniques). De cette façon, la mesure de recoupement permet de faire la distinction entre, par exemple, *pièce usinée* et *pièce à usiner*. La mesure d'association utilisée pour déterminer les cooccurrents statistiquement significatifs ou pertinents est la statistique du LLR (log de vraisemblance). Comme le seuil de significativité est très sévère (valeur $p < 0{,}0001$), on relève uniquement les cooccurrents sémantiquement pertinents. L'algorithme et les scripts en Python permettent de définir les paramètres (fenêtre d'observation, seuil de significativité, etc.) au niveau des cooccurrents et au niveau des cooccurrents des cooccurrents. Les scripts génèrent une grande base de données indexée avec toutes les informations statistiques pertinentes (LLR, seuil de significativité, etc.). La base de données sera alors interrogée pour analyser le recoupement des cooccurrents de deuxième ordre pour chaque unité lexicale spécifique comme base (par exemple *tour, machine, outil, usinage, broche* …). A cet effet, la fonction Python de l'algorithme prévoit les paramètres suivants : la base (spécificité à analyser), le seuil de significativité pour les cooccurrents de premier ordre ($p < 0.0001$) et de deuxième ordre ($p < 0.0001$) et la base de données à interroger. Il y a plus de recoupement, si plus de cooccurrents (c) partagent le même cc, ce qui signifie un poids plus lourd pour ce cc (score près de 1). Un cc moins/pas partagé indique donc peu/pas de recoupement (score près de 0).

Pour les 4717 spécificités ou unités lexicales spécifiques du corpus technique, nous pouvons ainsi calculer le degré de recoupement et donc le degré de monosémie, qui nous permet de situer les spécificités sur un continuum de monosémie, en fonction de leur rang de monosémie. Les mots avec un degré de monosémie identique auront le même rang de monosémie, par analogie avec le rang de spécificité (Cf. section 2.1).

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

4 Discussion des résultats

4.1 Le corpus technique

Le corpus technique qui fait l'objet de l'analyse des spécificités et de l'analyse sémantique quantitative comprend environ 1,7 million d'occurrences et il est constitué de textes techniques spécialisés du domaine des machines-outils pour l'usinage des métaux. Il a été étiqueté par Cordial 7 Analyseur[5] et consiste en 4 sous-corpus, datant de 1996 à 2002 : des revues électroniques (800.000 occurrences), des fiches techniques (300.000), des normes ISO et directives (300.000) et quatre manuels numérisés (360.000). Les textes des quatre sous-corpus se situent à différents niveaux de normalisation et de vulgarisation, ce qui assure la représentativité et la qualité du corpus technique. Afin d'identifier les unités spécifiques du corpus technique, nous avons aussi recours à un corpus de référence de langue générale, constitué d'articles du journal *Le Monde* (1998), comprenant environ 15,3 millions d'occurrences lemmatisées. Les fichiers étiquetés par Cordial se composent de trois colonnes, avec un mot par ligne : (1) la forme fléchie ou forme graphique, (2) le lemme ou forme canonique et (3) le code Cordial, qui est comparable à un POS-tag (*Part-Of-Speech*) et qui indique la classe lexicale.

4.2 Les premiers résultats de la mesure de monosémie

La comparaison des données quantitatives des rangs de spécificité et de monosémie permet de tirer les premières conclusions. Il s'avère que les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais, au contraire, les plus hétérogènes sémantiquement (par exemple *machine, pièce, tour*). En plus, les unités lexicales les moins spécifiques du corpus technique sont les plus homogènes sémantiquement (par exemple *rationnellement, télédiagnostic*), à quelques exceptions près, comme *service* et *objet*. Les résultats de l'analyse statistique de régression simple, qui permet d'évaluer la corrélation entre le rang de spécificité des 4717 spécificités et leur rang de monosémie, confirme ces premières observations. Le coefficient de corrélation Pearson (-0,72) démontre une corrélation négative entre le rang de spécificité et le rang de monosémie : les mots les plus spécifiques sont les moins monosémiques. L'analyse de régression simple est hautement significative ($p < 2.2e\text{-}16$) et le pourcentage de variation expliquée R^2 est de 51,57%. La variation du rang de spécificité permet donc d'expliquer 51,57% de la variation du rang de monosémie. Ces résultats quantitatifs permettent d'infirmer la thèse monosémiste traditionnelle et de confirmer les observations des études de corpus antérieures. Nous n'entrerons pas dans les détails de l'analyse statistique, car nous préférons approfondir l'analyse sémantique quantitative.

4.3 La validation de la mesure de monosémie

Une des conséquences du caractère novateur de notre mesure de monosémie ou de recoupement est qu'il n'existe pas de mesure de référence ou de Gold Standard pour évaluer les résultats quantitatifs de notre mesure. Nous avons dès lors procédé à une comparaison manuelle des cooccurrents les plus saillants et les plus pertinents d'un certain nombre d'unités lexicales spécifiques, intuitivement polysémiques ou homonymiques (*machine, outil, tour, avance, arête*) et intuitivement monosémiques (*m/min, Iso*). Pour les mots hétérogènes sémantiquement, *machine, outil, tour, avance* et *arête*, l'hétérogénéité des cooccurrents statistiquement très significatifs reflète effectivement l'hétérogénéité sémantique du mot de base. Ainsi, on retrouve pour le mot *tour*, d'une part *minute, mille* (sens : « rotation ») et d'autre part *centre, horizontal, bi-broche* (sens : « machine-outil pour l'usinage de pièces »). Il est à noter que pour *machine*, les unités polylexicales se manifestent clairement à travers les cooccurrents très significatifs (*machine + à + meuler / scier / rectifier*).

Nous avons également procédé à une validation externe de notre mesure de monosémie au moyen de dictionnaires, puisque nous ne disposons pas de listes de sens préétablis, ni d'autres mesures sémantiques comparables. Les résultats confirment les résultats de notre mesure de recoupement pour un échantillon de 50 spécificités. Il convient de signaler que les mots les plus fréquents, tels que *machine* et *outil*, entrent

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

très souvent dans la composition d'unités polylexicales (*machine à fraiser, machine à usiner*, …), ce qui pourrait en partie expliquer leur hétérogénéité sémantique.

4.4 Les mesures alternatives

Comme nous l'avons expliqué ci-dessus, la mesure de monosémie permet de quantifier l'analyse sémantique en implémentant la monosémie en termes d'homogénéité sémantique. La mesure de monosémie ou de recoupement détermine à quel point les cooccurrents de deuxième ordre se recoupent ou à quel point ils sont partagés par les cooccurrents de premier ordre. On peut se demander s'il est important de tenir compte, dans le dénominateur de la formule, du nombre de c (cooccurrents de premier ordre) et quel serait l'impact sur le rang de monosémie, si ce facteur était exclu de la formule.

Afin de juger la pertinence des facteurs repris dans la mesure de monosémie (Cf. figure 2), nous proposons de comparer plusieurs mesures alternatives. Les expérimentations sont conduites sur le corpus technique entier, pour un petit échantillon de 50 spécificités représentatives des 4717 spécificités. L'échantillon comprend des mots intuitivement homonymiques ou polysémiques, tels que *tour* et *avance*, des mots intuitivement monosémiques, tels que *Fig*, *m* et *m/min*, des mots très fréquents, moins fréquents et très peu fréquents dans le corpus technique, ainsi que des mots très spécifiques, moins spécifiques et très peu spécifiques du corpus technique. Cette diversité d'homogénéité, de fréquence et de spécificité permettra de vérifier à fond l'impact des différentes mesures sur le degré et le rang de monosémie des spécificités de l'échantillon.

Plusieurs mesures sont comparées, généralement des variations sur le thème de la mesure de monosémie de base (ou la mesure de recoupement) (Cf. figure 2).

- (1) M_monosémie : $\sum$ *(fq cc / (nbr total c * nbr total cc))* (Cf. figure 2)
- (2) M_cc_diff : *-log (nbr cc différents / nbr total cc)*
- (3) M_fq_cc : $\sum$ *(fq cc / nbr total cc)*
- (4) M_cc-types : $\sum$ *(fq cc-t / (nbr total c * nbr total cc-t))*
  *= nbr total cc / (nbr total c * nbr cc différents)*

Du point de vue méthodologique, la mesure de monosémie de base, M_monosémie (1), s'oppose aux deux mesures suivantes, à savoir M_cc_diff (2) et M_fq_cc (3), parce que celles-ci ne tiennent pas compte du nombre total de c. La dernière mesure, M_cc-types (4), se situe au niveau des cc-types (c'est-à-dire des cc différents ou uniques), car elle tient compte de la fréquence des cc-types (*cc-t*) et du nombre total de cc-types (*cc-t*), au lieu des occurrences de cc (*cc-tokens*) de la mesure de base (1). La somme de la fréquence de tous les cc-types équivalant au nombre total de cc (occurrences ou *cc-tokens*), la mesure (4) équivaut à la formule simplifiée mentionnée plus bas. Pour les 50 spécificités de l'échantillon, les quatre mesures alternatives permettent de générer quatre rangs de monosémie ; le rang de monosémie de la mesure de base (1) est indiqué en gris clair (Cf. tableau 1).

Premièrement, on observe que la mesure de monosémie de base, M_monosémie (1) accorde des rangs de monosémie entre 1 et 10 (les plus homogènes sémantiquement) à des mots peu fréquents et peu spécifiques, visualisés en bas de liste. Pour des raisons évidentes, les mots les moins fréquents auront moins de chances d'apparaître dans des contextes sémantiquement très hétérogènes. Comme les deux mesures M_cc_diff (2) et M_fq_cc (3) ne prennent pas en compte le nombre de c, elles accordent les rangs de monosémie les plus bas (entre 1 et 15) ou les plus monosémiques aux mots les plus fréquents, qui correspondent toutefois à des mots intuitivement plutôt hétérogènes sémantiquement, tels que *machine*, *outil* et *avance*. Par contre, pour les mots intuitivement plutôt monosémiques, les rangs de monosémie accordés par ces deux mesures alternatives correspondent bien à l'intuition (3 et 1 pour *Fig*, 7 et 3 pour *m* et 6 et 7 pour *m/min*).

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

N°	spécificité	(1)	(2)	(3)	(4)
1	*machine*	49	1	2	49
2	*outil*	48	2	6	48
3	*permettre*	44	5	12	43
4	*système*	46	8	8	47
5	*travail*	43	13	16	40
6	*Fig*	30	3	1	46
7	*précision*	29	14	11	32
8	*arête*	37	17	14	37
9	*avance*	32	4	9	36
10	*usiner*	36	19	20	30
11	*effectuer*	33	10	19	28
12	*tour*	45	20	13	45
13	*puissance*	39	9	10	38
14	*technique*	47	18	18	41
15	*m*	38	7	3	44
16	*table*	41	12	15	39
17	*valeur*	40	15	17	33
18	*etc*	35	33	21	31
19	*abrasif*	42	21	22	35
20	*Iso*	31	16	4	42
21	*électroérosion*	27	22	24	25
22	*tonne*	22	11	5	34
23	*concept*	21	25	27	20
24	*m/min*	19	6	7	29
25	*emboutissage*	26	23	23	27
26	*variable*	24	30	25	23
27	*commander*	23	27	33	21
28	*externe*	28	32	29	24
29	*maîtriser*	14	39	40	13
30	*verre*	25	37	35	22
31	*meulage*	34	31	26	26
32	*fraiser*	17	44	44	16
33	*collaboration*	18	36	31	19
34	*assembler*	20	28	34	18
35	*raboutage*	15	38	36	15
36	*mm/s*	16	26	28	17
37	*insérer*	11	34	37	9
38	*cloison*	12	35	32	14
39	*numériquement*	10	41	39	10
40	*réfrigération*	5	47	46	6
41	*présérie*	13	42	41	11
42	*endommagement*	--	--	--	--
43	*extérieurement*	1	24	38	1
44	*numérotation*	4	45	45	4
45	*réutilisable*	6	46	47	5
46	*microbiologique*	9	29	30	12
47	*puisard*	8	40	42	8
48	*commuter*	7	49	49	7
49	*vidangeur*	3	43	43	3
50	*batch*	2	48	48	2

Tableau 1 : Echantillon de 50 spécificités : rangs alternatifs de monosémie

Deuxièmement, la mesure M_cc-types (4) repose sur les *cc-types* et non sur les *cc-tokens*. Méthodologiquement, le recoupement d'un cc-type (fréquence de ce cc-type) pèse moins lourd sur le résultat final que le recoupement d'un cc-token, car il est compté une fois, alors que le recoupement du cc-token sera compté autant de fois que la fréquence du cc-token . La comparaison des résultats, c'est-à-dire des rangs de monosémie des mesures (1) et (4), montre peu de différences, à première vue. En effet, les mots hétérogènes sémantiquement dans les résultats de la mesure (1) le sont également dans ceux de la mesure (4), ce qui est visualisé dans la dernière colonne. Toutefois, les mots intuitivement monosémiques, tels que *Fig*, *Iso* et *m/min*, se retrouvent à des rangs considérablement plus polysémiques (car plus élevés et plus près de 50) pour la mesure M_cc-types (4). Intuitivement, les résultats de la mesure (4) sont donc moins plausibles que les résultats de la mesure de base (1), ce qui s'explique par le calcul du recoupement des cc (*cc-tokens* (1) versus *cc-types* (4)).

Reprenons finalement les mesures (2) et (3) qui ne prennent pas en considération le nombre de c, afin d'expliquer pourquoi elles produisent des résultats contre-intuitifs. La mesure (2) est basée sur le rapport entre le nombre de cc uniques (ou différents) (=cc-types) et le nombre total de cc (=cc-tokens). Ce rapport augmente, si les mots sont moins spécifiques et moins fréquents et s'ils ont moins de c et de cc. Même si le nombre total de cc d'un mot de base augmente (cc-tokens), le nombre de cc différents (cc-types) de ce mot n'augmentera pas dans la même mesure, ce qui correspond grosso modo au TTR (*Type-Token Ratio*) des mots d'un corpus. En effet, si le nombre d'occurrences (*tokens*) augmente, le nombre de types (*types*) n'augmente pas dans la même mesure. Ainsi, deux cas de figure se distinguent pour la mesure M_cc_diff (2). Si un cc est partagé par 2 c sur 390 c (par exemple pour *machine*), il est bel est bien partagé et il n'est pas unique, mais pour l'image globale de ce mot, c'est une très faible indication de monosémie. Si, par contre, un cc est partagé par 2 c sur 2 ou 3 c, il est aussi partagé, mais pour l'image globale de ce mot,

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

c'est une plus forte indication de monosémie. Autrement dit, la polysémie obtenue en regardant uniquement le nombre de cc uniques ou différents n'est qu'apparente. Par conséquent, il faut également tenir compte du nombre de fois que chaque cc est partagé, donc du nombre de c ou de cooccurrents avec lesquels ce cc apparaît. En effet, il faut inclure le nombre total de c, car l'exclure revient à la mesure (3) et génère également des résultats peu intuitifs.

De ce qui précède, il ressort que la mesure de recoupement de base (1) est une mesure plus intuitive, en dépit du fait que les mots peu fréquents se voient attribuer des rangs de monosémie inférieurs à 10. Il s'ensuit que les mots peu fréquents dans le corpus technique et homogènes sémantiquement relèguent les autres mots intuitivement plutôt monosémiques, tels que *Fig, m, m/min*, à des rangs un peu plus polysémiques (19 ou 30). Force est de constater que la mesure de recoupement de base accorde les rangs les plus polysémiques (entre 40 et 50 dans l' échantillon) aux mots les plus fréquents ayant beaucoup de c. C'est une première indication que la mesure de recoupement semble être sensible à la fréquence du mot spécifique et à son nombre de c.

5 Mise au point de la mesure de monosémie : intégration de la classe lexicale

La mesure de monosémie pourra être enrichie, si on y intègre plus d'informations linguistiques. Nous nous proposons dès lors de tenir compte des informations de classe lexicale, d'abord pour déterminer les unités lexicales spécifiques du corpus technique et ensuite pour calculer leur degré de monosémie. Jusqu'à présent, les unités lexicales spécifiques ou les spécificités étaient déterminées à partir des fréquences des lemmes dans le corpus technique et dans le corpus de référence, en fonction de critères formels. Toutes les occurrences des lemmes *accessoire* (nom) et *accessoire* (adjectif), par exemple, étaient considérées comme appartenant au même lemme formel *accessoire*, qui était relevé comme spécificité (ou unité lexicale spécifique). Par contre, la prise en compte de la classe lexicale, fournie dans les fichiers étiquetés par Cordial, permettra de considérer ces occurrences comme des occurrences de deux lemmes différents, à savoir le lemme du nom *accessoire* et le lemme de l'adjectif *accessoire*. Dans la liste des spécificités, on regroupera ainsi les occurrences du nom sous la spécificité *accessoire*|*nom* et celles de l'adjectif sous la spécificité *accessoire*|*adj*. Par conséquent, chaque spécificité sera « enrichie », car munie d'une indication de classe lexicale. Nous optons pour les codes[6] *adj*, *nom*, *verbe*, *adv*, *fadj* (adj. dém./poss./num.), *art*, *conj*, *prep*, *interj*, *pron*, *npr* (nom propre) et *ponc*. La prise en compte de la classe lexicale pourra également se faire pendant le calcul du degré de monosémie, c'est-à-dire pendant le calcul du degré de recoupement des cooccurrents des cooccurrents de ces spécificités enrichies. En effet, lors de la génération de la base de données des mots de base (spécificités) et des cooccurrents de premier et de deuxième ordre, on pourra intégrer la nouvelle indication de classe lexicale. Cette mise au point à deux niveaux (identification des spécificités et calcul du degré de monosémie) devra permettre de dissocier les occurrences des homonymes. En plus, elle aura probablement des effets sur les résultats de l'analyse sémantique quantitative et sur les résultats de l'analyse statistique de régression simple, qui évalue la corrélation entre le rang de spécificité et le rang de monosémie.

A l'aide de scripts en Python, nous rassemblons les fichiers étiquetés par Cordial dans un grand fichier miroir, qui contient des indications de classe lexicale (p.ex. ...|*nom*), aussi bien dans la colonne des formes fléchies (p.ex. *accessoires*|*nom*), que dans la colonne des lemmes (p.ex. *accessoire*|*nom*). A l'instar de la liste précédente des 4717 spécificités (Cf. section 2.1), nous dressons une nouvelle liste de spécificités du corpus technique à partir de la colonne des lemmes enrichis du fichier miroir. La nouvelle liste de spécificités enrichies contient 5207 spécificités munies de leur classe lexicale. Cette liste est plus longue, parce que certaines spécificités (essentiellement homonymiques) y figurent plus d'une fois si elles appartiennent à plus d'une classe lexicale, par exemple *aéronautique*|*nom* et *aéronautique*|*adj* ou *annexe*|*nom* et *annexe*|*adj*. Le grand fichier miroir permet également de générer une nouvelle base de données avec des informations de cooccurrences enrichies, car munies de l'indication de classe lexicale. En effet, au moment de déterminer les cooccurrents statistiquement pertinents d'un mot de base (p.ex. *aéronautique*|*nom*), le script prend en considération la classe lexicale des cooccurrents (p.ex. *standard*|*adj*

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

versus *standard*|*nom*) et ne mettra plus toutes les occurrences de la même forme (p.ex. *standard*) dans le même sac. Il en va de même pour les cooccurrents de deuxième ordre. Ceux-ci sont rattachés à un cooccurrent de premier ordre enrichi et ils sont eux-mêmes également enrichis, car ils comprennent une indication de classe lexicale. La prise en compte de la classe lexicale permettra ainsi d'affiner les résultats de l'analyse sémantique quantitative. Elle améliorera la précision et la fiabilité des informations sémantiques et, par conséquent, la précision du calcul du degré de recoupement.

Quel est l'effet de la prise en compte de la classe lexicale ? Dans un premier temps, nous vérifions l'effet global, c'est-à-dire l'effet sur les résultats de l'analyse statistique de régression simple. Ensuite, nous comparons le degré de monosémie précédent et le nouveau degré de monosémie pour un certain nombre de spécificités intéressantes. Comme nous l'avons expliqué ci-dessus pour la liste des 4717 spécificités (Cf. section 4.2), l'analyse de régression étudie la corrélation entre le rang de spécificité et le rang de monosémie. La nouvelle analyse de régression pour les 5207 spécificités, qui tient compte de la classe lexicale pour le repérage des spécificités et pour le calcul du degré de recoupement, confirme les résultats de l'analyse de régression précédente (pour les 4717 spécificités). La nouvelle analyse est également hautement significative ($p < 2.2e\text{-}16$) et le pourcentage de variation expliquée R^2 est de 51,63% (Cf. 51,57% pour les 4717 spécificités). Il s'ensuit que l'effet de la prise en compte de la classe lexicale est très limité, voire négligeable. Si certains lemmes (enrichis) deviennent plus homogènes sémantiquement et se caractérisent par un nouveau degré de monosémie plus élevé, il s'avère que leur nouveau rang de monosémie n'est pas fortement affecté.

Afin d'évaluer l'impact de la prise en compte de la classe lexicale sur le degré de monosémie, nous procédons à une comparaison détaillée des résultats du calcul du degré de recoupement, pour les 4717 spécificités de base (Cf. section 2.1) et pour les 5207 spécificités enrichies. Dans cette liste de spécificités enrichies, les spécificités sont dissociées en fonction de leur classe lexicale. Il est à noter que les occurrences des 4717 spécificités étaient regroupées sous le lemme de la classe lexicale la plus fréquente, qui rassemblait les fréquences de toutes les occurrences. Ainsi, dans la liste des 4717 spécificités, le lemme *standard* se caractérise par la fréquence cumulée 902 dans le corpus technique, tandis que dans la liste des 5207 spécificités enrichies, les lemmes enrichis *standard*|*adj* et *standard*|*nom* ont les fréquences 485 et 417 respectivement.

Le tableau ci-dessous (Cf. tableau 2) visualise la comparaison du degré de monosémie, avec et sans prise en compte de la classe lexicale. La liste des 4717 spécificités contient les lemmes sans indication de classe lexicale et leur degré de monosémie, sans prise en compte de la classe lexicale lors du calcul du degré de recoupement. La liste des 5207 spécificités, par contre, contient les lemmes avec indication de classe lexicale et leur degré de monosémie, compte tenu de la classe lexicale des lemmes et des cooccurrents de premier et de deuxième ordre. La dernière colonne (en gris clair) visualise la différence entre les deux degrés de monosémie (mono - mono_claslex). Si la différence est négative, les lemmes avec indication de classe lexicale sont plus homogènes sémantiquement (ou plus monosémiques), lorsqu'on prend en considération la classe lexicale. Si la différence est positive, les lemmes avec classe lexicale sont plus hétérogènes sémantiquement, compte tenu de la classe lexicale.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

4717_ lemme	4717_ degré mono	5207_ lemme	5207_ lemme|claslex	fréquence	5207_ degré mono	mono - mono_claslex
abrasif	0,03130209	*abrasif*	*abrasif|adj*	372	0,0326374	-0,001335316
		abrasif	*abrasif|nom*	151	0,04913793	-0,017835844
accessoire	0,04527177	*accessoire*	*accessoire|nom*	247	0,04970057	-0,004428799
		accessoire	*accessoire|adj*	79	0,05562685	-0,010355078
aimant	0,10513006	*aimant*	*aimant|nom*	19	0,21164021	-0,106510154
		aimant	*aimant|adj*	6	0,15306122	-0,047931166
standard	0,03605286	*standard*	*standard|adj*	485	0,05527809	-0,019225229
		standard	*standard|nom*	417	0,04203992	-0,005987055
m³/heure	0,41666667	*m³/heure*	*m³/heure|nom*	4	0,42647059	-0,009803921
t/m	0,35973597	*t/m*	*t/m|nom*	3	0,35294118	0,006794798
avance	0,04701646	*avance*	*avance|nom*	1832	0,04319039	0,003826074
tour	0,0284628	*tour*	*tour|nom*	1476	0,02698511	0,001477684
usinage	0,03491978	*usinage*	*usinage|nom*	6720	0,03279069	0,00212909

Tableau 2 : Extrait de la comparaison du degré de monosémie des 4717 et des 5207 spécificités

Il ressort de cette comparaison que les lemmes homonymiques dissociés, par exemple *standard|adj* et *standard|nom*, deviennent plus homogènes sémantiquement (Cf. différence négative entre les degrés de monosémie). Notons que le lemme enrichi le plus fréquent (par exemple *accessoire|nom*) se caractérise par la différence la plus petite, puisque la liste précédente des 4717 spécificités rattachait toutes les fréquences sous le lemme le plus fréquent, en l'occurrence le nom. Pour les lemmes où les fréquences des deux classes lexicales sont équivalentes (par exemple *standard|adj* (485) et *standard|nom* (417)), l'homogénéité sémantique des deux lemmes homonymiques dissociés est plus importante. Notre mesure de monosémie et la variante enrichie, qui prend en considération la classe lexicale des spécificités et des cooccurrents de premier et de deuxième ordre, permettent donc de détecter aussi bien l'hétérogénéité sémantique des homonymes que l'homogénéité sémantique des lemmes homonymiques dissociés.

Le tableau ci-dessus (Cf. tableau 2) montre également que les mots intuitivement monosémiques (par exemple *m/min/mm, m³/heure, t/m*) restent plutôt monosémiques, lorsqu'on prend en considération la classe lexicale. En effet, leur degré de monosémie change à peine. Les mots intuitivement polysémiques (*avance, tour*) ou les mots vagues (*usinage*), qui sont hétérogènes sémantiquement dans la liste des 4717 spécificités, se caractérisent également par de faibles différences entre les deux degrés de monosémie. Par ailleurs, la différence positive indique que la prise en compte de la classe lexicale renforce leur hétérogénéité sémantique. Lorsque le calcul du degré de monosémie tient compte de la classe lexicale, l'hétérogénéité sémantique de ces mots intuitivement polysémiques est plus importante, mais elle reste cachée. La dissociation des lemmes en fonction de la classe lexicale ne permet pas de dissocier les lexies des mots polysémiques, telles que *tour* « machine-outil » et *tour* « rotation » pour le substantif masculin *tour*. On pourrait peut-être remédier à ce problème de polysémie cachée en prenant en considération des éléments syntaxiques lors du repérage des spécificités et lors du calcul du degré de monosémie. Pour l'instant, notre mesure réussit déjà à détecter l'hétérogénéité sémantique des mots polysémiques, mais nous n'arrivons pas encore à dissocier les sens polysémiques.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Sémantique
DOI 10.1051/cmlf08202

6 Conclusions et perspectives

Dans cet article, nous avons présenté la méthodologie et les résultats d'une étude sémantique quantitative menée sur un corpus technique. L'originalité de l'étude réside principalement dans l'approche quantitative et scalaire de la spécificité et de la sémantique. Nous avons implémenté la monosémie en termes d'homogénéité sémantique et nous avons développé une mesure de monosémie basée sur le recoupement des cooccurrents des cooccurrents. L'analyse sémantique quantitative de quelque 5000 mots spécifiques du corpus technique a permis d'ébranler la thèse monosémiste traditionnelle et de démontrer une corrélation négative entre le rang de spécificité et le rang de monosémie. En effet, plus les unités lexicales sont spécifiques et représentatives dans le corpus technique, plus elles sont hétérogènes sémantiquement. Afin d'affiner les résultats, nous avons enrichi la mesure de monosémie en y intégrant des indications de classe lexicale, tant pour les spécificités que pour les cooccurrents de premier et de deuxième ordre. La prise en compte de la classe lexicale a permis de préciser les résultats de l'analyse sémantique quantitative des homonymes.

Pour nos recherches futures, nous poursuivons la mise au point de notre mesure de monosémie et nous envisageons une analyse similaire dans d'autres corpus spécialisés et l'analyse sémantique quantitative des unités polylexicales dans des corpus spécialisés. Nous projetons de compléter notre mesure de monosémie par l'intégration d'informations syntaxiques et par des analyses statistiques multivariées de regroupement (*cluster analysis*). Celles-ci permettraient de regrouper les cooccurrents (ou c) d'un mot de base (unité spécifique) à partir des cc qu'ils partagent. Les analyses de regroupement conduiraient peut-être à mieux comprendre encore le phénomène de l'hétérogénéité sémantique et à opérer des distinctions sémantiques plus fines entre la polysémie, l'homonymie et le vague. Finalement, notre analyse sémantique quantitative mérite d'être appliquée à d'autres corpus spécialisés, à un corpus de langue générale et à d'autres unités, en particulier aux unités polylexicales de notre corpus technique. L'analyse des cooccurrences serait parfaitement transposable, par le biais du recoupement des cooccurrents de troisième ordre. L'analyse des spécificités, quant à elle, constitue un défi majeur, principalement en raison de l'absence des unités polylexicales spécialisées dans le corpus de référence de langue générale.

Références bibliographiques

Arntz, R., Picht H. (1989). *Einführung in die Terminologiearbeit*. Hildesheim : Georg Olms Verlag.

Bertels, A. (2006). *La polysémie du vocabulaire technique. Une étude quantitative*. Thèse de doctorat non publiée. Université de Leuven, Belgique.

Bertels, A., Speelman D., Geeraerts D. (2006). Analyse quantitative et statistique de la sémantique dans un corpus technique. In Mertens P., Fairon C., Dister A., Watrin P. (éd.), *Actes de TALN 2006, Verbum ex machina (13ème Conférence sur le Traitement Automatique des Langues Naturelles)*, Louvain-la-Neuve : Presses Universitaires de Louvain, 73-82.

Blumenthal, P., Hausmann F.J. (éds) (2006). Collocations, corpus, dictionnaires. *Langue française, 150*.

Cabré, M.T. (2000). Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles, 21,* 10-15.

Condamines, A (éd.) (2005). *Sémantique et corpus*. Paris : Hermes-Science.

Condamines, A., Rebeyrolle J. (1997). Point de vue en langue spécialisée. *Meta, 42-1,* 174-184.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19-1,* 61-74.

Eriksen, L. (2002). Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder : Zur Terminologie der ‚Sache' im Deutschen. *Hermes – Journal of Linguistics, 28,* 211-222.

Ferrari, L. (2002). Un caso de polisemia en el discurso jurídico? *Terminology, 8-2,* 221-244.

Gaudin, F. (2003). *Socioterminologie : une approche sociolinguistique de la terminologie*. Bruxelles : Duculot.

Grossmann, F., Tutin A. (éds) (2003). Les collocations, analyse et traitement. *Travaux et Recherches en linguistique appliquée, Série E, n° 1*.

Habert, B., Illouz G., Folch H. (2004). Dégrouper les sens : pourquoi ? comment ? In Purnelle G., Fairon C., Dister A. (éds), *Actes de JADT 2004 (7es Journées internationales d'Analyse statistique des Données Textuelles)*, Louvain-la-Neuve : Presses Universitaires de Louvain, 565-576.

Habert, B., Illouz G., Folch H. (2005). Des décalages de distribution aux divergences d'acception. In Condamines, A. (éd.), *Sémantique et corpus*, Paris : Hermes-Science, 277-318.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris : Slatkine-Champion.

Lerat, P. (1995). *Les langues spécialisées*. Paris : PUF.

Martinez, W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. In *Actes de JADT 2000 (5es Journées internationales d'Analyse statistique des Données Textuelles)*, 78-84.

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

Temmerman, R. (2000). *Towards new ways of terminology description. The sociocognitive approach*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Veronis, J. (2003). Cartographie lexicale pour la recherche d'informations. In *Actes de TALN 2003 (10ème Conférence sur le Traitement Automatique des Langues Naturelles)*, 265-274.

Wüster, E. (1931). *Internationale Sprachnormung in der Technik : besonders in der Elektrotechnik*. Berlin : VDI-Verlag.

---

[1] Citons notamment le Calcul des spécificités (Lafon, 1984), implémenté dans Lexico3 (http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/) et la Méthode des mots-clés (*Keywords Method*), implémentée notamment dans WordSmith Tools (http://www.lexically.net/wordsmith/ et http://www.oup.com).

[2] Nous adoptons le terme « spécificités » pour désigner les mots les plus représentatifs et les plus caractéristiques du corpus technique, indépendamment de la méthode utilisée (*Calcul des spécificités* versus *Méthode des mots-clés*).

[3] Les unités spécifiques avec le même degré de spécificité se voient accorder un rang de spécificité identique.

[4] Le logiciel Lexico3 permet d'indiquer la spécificité des cooccurrents pendant l'analyse des cooccurrences.

[5] Cordial Analyseur : www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm.

[6] Ces codes remplacent les codes attribués par Cordial, trop nombreux et de granularité trop fine, parce que Cordial attribue, par exemple, des codes différents aux adjectifs masculins et féminins, singuliers et pluriels.