L'importance du recoupement des cooccurrents de deuxième ordre pour étudier la corrélation entre la spécificité et la monosémie

Ann Bertels[1,2], Dirk Geeraerts[2]

[1]ILT – KU Leuven – B-3000 Leuven – Belgique
[2]QLVL – KU Leuven – B-3000 Leuven – Belgique

Abstract

This paper involves a quantitative semantic study of typical vocabulary in a technical corpus. The study aims to find out whether the most typical or specialised lexical items of the technical corpus are the most monosemous items. Hence, the main research question concerns the correlation between the typicality rank and the monosemy rank of all typical lexical items at the level of single word items.

Typical lexical items or keywords were identified using keyword analysis. The log likelihood test statistic provides a typicality coefficient (keyness), which indicates the degree of typicality. The semantic analysis was quantified by implementing monosemy in terms of "semantic homogeneity". The degree of monosemy of a keyword was calculated by means of a monosemy measure, based on the formal overlap of the second order co-occurrents. A simple regression analysis showed a negative correlation between typicality rank and monosemy rank. As a matter of fact, the most typical lexical items of the technical corpus turned out not to be the most monosemous items.

In order to fine-tune the results of the regression analysis and to enrich the overlap measure, which is mainly based on statistical co-occurrence information, the monosemy measure was refined yielding a technical monosemy measure, weighted for the typicality of the second order co-occurrents. This technical monosemy measure calculates the degree of typical or technical monosemy. Finally, the results of the two overlap measures were compared with respect to the correlation between typicality rank and (technical) monosemy rank of the typical lexical items of the technical corpus.

Résumé

Cette contribution s'inscrit dans le cadre d'une étude sémantique quantitative du vocabulaire spécifique d'un corpus de textes techniques. L'étude vise à vérifier si les unités lexicales les plus spécifiques du corpus technique sont effectivement les unités les plus monosémiques. L'objectif de l'étude est donc d'étudier la corrélation entre le rang de spécificité des unités lexicales spécifiques, au niveau des unités simples, et leur rang de monosémie.

Pour identifier les unités lexicales spécifiques et leur degré de spécificité, nous recourons à la méthode des mots-clés et à la mesure statistique du log du rapport de vraisemblance. Pour quantifier l'analyse sémantique, nous implémentons la monosémie en termes d'homogénéité sémantique. Nous calculons le degré de monosémie à partir du degré de recoupement des cooccurrents de deuxième ordre. L'analyse statistique de régression simple montre une corrélation négative entre le rang de spécificité et le rang de monosémie. En effet, les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus homogènes sémantiquement ou les plus monosémiques, au contraire.

Dans le but d'affiner les résultats de l'analyse de régression et d'enrichir la mesure de recoupement, qui s'appuie essentiellement sur des informations statistiques de cooccurrence, nous recourons à un facteur de pondération, en

fonction de la spécificité ou technicité des cooccurrents de deuxième ordre. La mesure de recoupement technique permet de déterminer le degré de monosémie technique. Finalement, les résultats des deux mesures de recoupement sont comparés du point de vue de la corrélation entre le rang de spécificité et le rang de monosémie (technique) des unités lexicales spécifiques du corpus technique.

Mots-clés : sémantique quantitative, cooccurrents de deuxième ordre, mesure de recoupement, spécificité, analyse de régression

1. Introduction et objectifs de recherche

Cet article décrit l'importance du recoupement des cooccurrents de deuxième ordre, d'abord pour automatiser et quantifier l'analyse sémantique et ensuite pour étudier la corrélation entre la spécificité et la monosémie dans le cadre d'une étude sémantique quantitative d'un corpus spécialisé. Les trois adjectifs (*sémantique*, *quantitative* et *spécialisé*) méritent un mot d'explication, puisqu'ils permettent d'expliquer et de justifier les objectifs de recherche.

L'étude sémantique quantitative est conduite sur un corpus *spécialisé*, en l'occurrence un corpus de textes relevant du domaine technique restreint des machines-outils pour l'usinage des métaux. Une étude sémantique ou linguistique, qui se focalise sur un domaine technique, soulève tout de suite des questions sur les particularités de la langue spécialisée utilisée dans le domaine en question. Dans la langue spécialisée, les besoins communicatifs des spécialistes requièrent plus de précision, ce que la terminologie traditionnelle définit comme l'univocité et la monosémie des unités terminologiques de la langue spécialisée (Wüster, 1931 et 1991). Selon les partisans de la terminologie traditionnelle, onomasiologique et prescriptive, les termes de la langue spécialisée sont idéalement monosémiques, tandis que la polysémie est réservée aux mots de la langue générale. Récemment, les adeptes de la terminologie descriptive, sémasiologique et linguistique ont remis en question cet idéal de monosémie, ainsi que la double dichotomie qui oppose les termes aux mots comme elle oppose la monosémie à la polysémie (Cabré, 2000 ; Temmerman, 2000 ; Gaudin, 2003). Des études sémantiques qualitatives et ponctuelles menées sur des corpus spécialisés ont effectivement abouti à l'observation de cas de polysémie dans la langue spécialisée, même à l'intérieur d'un domaine spécialisé (Condamines et Rebeyrolle, 1997 ; Eriksen, 2002 ; Ferrari, 2002).

Cette caractéristique traditionnelle de la monosémie des unités terminologiques d'un corpus spécialisé et les récentes remises en question permettent de justifier le deuxième adjectif de notre étude, à savoir *sémantique*. L'objectif principal de notre étude sémantique est de vérifier si les unités lexicales du corpus technique sont monosémiques, comme le prétendent les monosémistes traditionnels 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 l'analyse de corpus, il est nécessaire de la reformuler en une question opérationnelle et mesurable, ce qui permet de justifier le troisième et dernier aspect de notre étude, à savoir la dimension *quantitative*. 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. Ceci implique l'idée de gradation et de continuum, contrairement à l'approche dichotomique traditionnelle, parce que les degrés

de spécificité et de monosémie permettent de situer les unités lexicales sur un continuum de spécificité et sur un continuum de monosémie. Nous procédons dès lors à une double analyse quantitative, qui consiste d'une part à déterminer les unités lexicales spécifiques du corpus technique et leur degré de spécificité et d'autre part à calculer leur degré de monosémie.

En réponse à la question de recherche quantitative, nous avançons l'hypothèse que les unités lexicales (les plus) spécifiques du corpus technique ne sont pas (les plus) monosémiques. Notre étude vise à mettre à l'épreuve la thèse monosémiste traditionnelle et les premières observations nous incitent plutôt à la réfuter qu'à la consolider. En effet, certaines unités lexicales spécifiques du corpus technique sont des mots à sens multiples. Citons en guise d'exemple le mot *tour*, qui signifie notamment (1) « machine-outil pour l'usinage des pièces » et (2) « rotation » ou le mot *découpe*, qui signifie (1) « action de découper » et (2) « résultat de la découpe ». Afin d'étayer notre hypothèse, nous procédons à une analyse statistique de régression simple, qui permet de vérifier la corrélation entre les données quantitatives de spécificité et de monosémie. De par son approche, notre étude vise donc à réconcilier la linguistique, l'informatique et la statistique. Elle recourt à des techniques informatiques, quantitatives et statistiques pour mieux comprendre et expliquer certains aspects linguistiques.

Dans cet article, nous expliquons d'abord le premier volet de la double analyse quantitative, qui consiste à identifier les unités lexicales spécifiques du corpus technique et à déterminer leur degré de spécificité (section 2). La section 3 décrit les principes de l'analyse sémantique quantitative et le développement de la mesure de recoupement ou de monosémie. Dans la section 4, nous discutons les résultats de l'analyse statistique de régression simple, plus particulièrement la corrélation entre le rang de spécificité et le rang de monosémie des unités lexicales spécifiques. Dans le but d'affiner ces résultats, nous les comparons aux résultats d'une analyse de régression simple faisant intervenir le rang de monosémie technique, déterminé à partir de la mesure de recoupement technique (section 5).

2. Le corpus technique et les unités lexicales spécifiques

Le corpus technique permet non seulement d'identifier les unités lexicales spécifiques, mais il fournit également les contextes d'apparition indispensables à l'analyse sémantique quantitative. Le corpus technique que nous avons constitué dans le cadre de cette étude compte environ 1,7 million d'occurrences et il relève du domaine spécialisé restreint des machines-outils pour l'usinage des métaux. Le corpus technique a été étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1996 à 2002, à savoir des revues électroniques (800.000 occurrences), des fiches techniques (300.000 occurrences), des normes ISO et directives (300.000 occurrences) et quatre manuels numérisés (360.000 occurrences). Le corpus de référence de langue générale compte environ 15,3 millions d'occurrences lemmatisées et il est constitué d'articles du journal *Le Monde* de la même période (1998). Les fichiers étiquetés par Cordial se composent de trois colonnes, avec une occurrence par ligne : (1) la forme fléchie ou la forme graphique, (2) le lemme ou la forme canonique et (3) le code Cordial, qui est comparable à un POS-tag (*Part-Of-Speech*) et qui indique la classe lexicale.

Pour identifier les unités spécifiques d'un corpus d'analyse (spécialisé), en le confrontant à un corpus de référence, plusieurs approches méthodologiques sont envisageables. Ces approches permettent de générer une liste d'unités spécifiques, pourvues d'une indication de leur degré de

spécificité. Les différences les plus importantes résident dans la méthodologie et les mesures statistiques sous-jacentes. La première approche méthodologique, à savoir le calcul des spécificités, est basée sur la distribution hypergéométrique (Lafon, 1984 ; Labbé et Labbé, 2001). Elle est implémentée notamment dans les logiciels Lexico3[1], Hyperbase[2] et TermoStat[3]. Du point de vue méthodologique, le calcul des spécificités procède par comparaison partie-tout. Une partie (ou une section) d'un corpus est comparée au corpus entier dans le but d'identifier les mots spécifiques de la section. La deuxième approche méthodologique, à savoir l'analyse des mots-clés (*Keyword Analysis*) (Scott et Tribble, 2006), est implémentée dans les logiciels WordSmith[4], AntConc[5], TermoStat et Abundantia Verborum Frequency List Tool[6]. Elle vise à comparer les fréquences relatives des mots dans un corpus de langue spécialisée à celles dans un corpus de référence de langue générale, compte tenu de la taille des deux corpus, dans le but d'identifier les mots significativement plus fréquents dans le corpus spécialisé. L'analyse des mots-clés s'appuie sur la mesure statistique du log du rapport de vraisemblance (*Log-Likelihood Ratio* ou LLR ou encore G^2) (Dunning, 1993) ou sur d'autres mesures statistiques, telles que le chi-carré. La dernière approche est celle de l'analyse des marqueurs lexicaux stables ou *Stable Lexical Marker Analysis* (SLMA) (Speelman *et al.*, 2006 ; Speelman *et al.*, 2008), qui s'inspire de l'analyse des mots-clés de Scott (Scott et Tribble, 2006). Toutefois, la comparaison des deux corpus s'effectue à partir de plusieurs listes de fréquence par corpus, dans le but d'identifier les différences lexicales, qui sont consistantes ou stables entre les deux corpus.

Afin de repérer les unités lexicales spécifiques de notre corpus technique, nous recourons à l'analyse des mots-clés et à la mesure statistique du log du rapport de vraisemblance (LLR). Dans le logiciel Abundantia Verborum Frequency List Tool, nous confrontons la liste de fréquence des lemmes du corpus technique à la liste de fréquence des lemmes du corpus de référence de langue générale. Les listes de fréquence des lemmes des deux corpus sont réalisées à l'aide de scripts en Python. Ensuite, le logiciel génère une liste de mots-clés ou de lemmes spécifiques du corpus technique avec l'indication de leur degré de spécificité, à savoir la valeur du LLR (*keyness*), et avec une valeur p associée. Pour le corpus technique entier, nous relevons ainsi 4717 lemmes spécifiques ($p<0,05$), après suppression des mots grammaticaux, des noms propres et des hapax. Le degré de spécificité permet de situer ces 4717 unités lexicales spécifiques sur un continuum de spécificité et de leur accorder un rang de spécificité. Les unités avec un degré de spécificité identique, c'est-à-dire une valeur de LLR identique, se voient accorder le même rang de spécificité. Les unités lexicales les plus spécifiques du corpus technique, à savoir *machine, outil, usinage, pièce, mm, vitesse, coupe, plaquette, fraisage, tour, broche, découpe,* etc. reflètent clairement la thématique du domaine spécialisé des machines-outils pour l'usinage des métaux.

Il est à noter que l'analyse des mots-clés se situe à présent au niveau des unités simples. Nos recherches futures devront certainement porter sur les unités polylexicales, puisque la plupart

1 SYLED – CLA2T, Paris3 : http://www.tal.univ-paris3.fr/lexico/.

2 Hyperbase : http://ancilla.unice.fr/~brunet/pub/hyperbase.html.

3 TermoStat : http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html.

4 WordSmith Tools : http://www.lexically.net/wordsmith/.

5 AntConc : http://www.antlab.sci.waseda.ac.jp/software.html.

6 Abundantia Verborum Frequency List Tool : http://wwwling.arts.kuleuven.be/genling/abundant/obtain. htm.

des unités terminologiques d'un domaine spécialisé se situent au niveau des unités complexes (par exemple *machine à fraiser* et *commande numérique*). Toutefois, pour l'instant, il n'est pas possible de déterminer le degré de spécificité des unités complexes de façon fiable et statistiquement significative, principalement en raison du fait que la plupart d'entre elles sont absentes dans un corpus de référence de langue générale.

Dans la section suivante, les 4717 unités lexicales spécifiques font l'objet d'une analyse sémantique quantitative, dans le but de leur attribuer un degré de monosémie. Celui-ci permet de procéder, finalement, à une analyse statistique de régression simple et d'étudier la corrélation entre le rang de spécificité et le rang de monosémie (section 4).

3. L'analyse sémantique quantitative

Pour répondre à la question de recherche quantitative, il est très important de quantifier et d'automatiser l'analyse sémantique. Le défi à relever consiste donc à développer une mesure de monosémie, qui permet de quantifier la monosémie et de situer les unités lexicales analysées sur un continuum de monosémie, à l'instar du continuum de spécificité. La mesure permet en même temps d'automatiser l'étude sémantique, parce que l'on fait l'économie d'une analyse manuelle de plusieurs milliers de concordances et de contextes d'apparition des unités lexicales spécifiques.

Le développement de la mesure de monosémie s'inscrit dans le cadre de la sémantique distributionnelle, plus particulièrement dans le contexte de l'analyse des cooccurrences (Grossmann et Tutin, 2003 ; Condamines ; 2005 ; Blumenthal et Hausmann, 2006). Afin de quantifier et objectiver l'analyse sémantique, nous proposons d'implémenter la monosémie en termes d'homogénéité sémantique (Habert *et al.*, 2005). 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 plutôt 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). Toutefois, en sémantique distributionnelle et contextuelle, deux problèmes 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 qui permettent de classer les mots est souvent très éparpillée (Habert *et al.*, 2004). Pour remédier à ces problèmes, on peut recourir aux cooccurrents de deuxième ordre (Grefenstette, 1994). Les cooccurrents de premier ordre sont généralement des cooccurrents syntagmatiques du mot de base. Les cooccurrents de deuxième ordre se caractérisent principalement par des relations paradigmatiques avec le mot de base (hyponymes, hyperonymes, synonymes, antonymes) (Pezik, 2005) et dès lors ils sont plus intéressants pour caractériser sémantiquement le mot de base. Les cooccurrents de deuxième ordre ont déjà permis de mettre en évidence des relations de synonymie (Martinez, 2000). L'analyse de l'axe syntagmatique, effectuée à deux reprises, contribue ainsi à la découverte de l'axe paradigmatique. Il est clair que les différents synonymes d'un mot de base sont des indices sémantiques précieux dans la perspective de la quantification de l'analyse sémantique.

Comme nous l'avons expliqué ci-dessus, une unité lexicale monosémique se caractérise par des cooccurrents qui appartiennent à des champs sémantiques plutôt similaires. L'accès à la sémantique de ces cooccurrents pourrait se faire à partir de leurs cooccurrents et donc à partir

des cooccurrents de deuxième ordre. En effet, 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. 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 sont (très) peu partagés par les cooccurrents de premier ordre. Dès lors, ces derniers sont sémantiquement plus diversifiés et 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). Le cooccurrent *usine* dans la première phrase indique clairement le sens « machine-outil », tandis que le cooccurrent *minute* dans la deuxième phrase indique le sens « rotation ». Pour accéder au(x) sens des cooccurrents, on pourra analyser leurs cooccurrents, c'est-à-dire les cooccurrents de deuxième ordre, non seulement dans la même phrase (par exemple *alésage* dans la première phrase comme cooccurrent d'*usine*), mais aussi dans tous les autres contextes (par exemple *pièces* et *outils* dans la troisième phrase comme cooccurrents d'*usine*). Plus les cooccurrents de deuxième ordre 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. ...*
2. *... broches pouvant monter jusqu'à quinze mille **tours** par **minute**, ...*
3. *Un tour CNC équipé d'outils modulaires Capto **usine** les pièces en question avec…*

La mesure de recoupement que nous avons développée permet de déterminer le degré de monosémie d'un mot de base. Elle s'appuie sur le recoupement formel des cooccurrents de deuxième ordre (ou cc) d'un mot de base et fait intervenir les paramètres suivants :

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

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

Figure 1 : Formule de la mesure de recoupement

Verbalisons la formule de la mesure de recoupement en prenant l'exemple d'un cc partagé par 3 c des 5 c au total. Cela veut dire que 3 c des 5 c apparaissent avec ce cc en question, ce qui indique un recoupement plutôt important. Dans le numérateur de la formule, nous incluons le nombre de c qui ont ce cc en commun (fq cc), en l'occurrence 3, dans le dénominateur nous incluons le nombre total de c différents (au niveau des *types*), en l'occurrence 5. Le recoupement est donc exprimé par la fraction 3/5. 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

est facilement interprétable. Comme on somme pour tous les cc, le dénominateur comprend aussi le nombre total de cc, car on considère tous les cc (*tokens*) évidemment avec les doublons responsables du recoupement formel.

Nous considérons les cooccurrents de premier et de deuxième ordre au niveau des formes graphiques (ou formes fléchies), afin de tenir compte de la différence sémantique entre, par exemple, *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage »). Le mot de base sur lequel portent les analyses est le lemme, pour assurer l'appariement ultérieur des informations sémantiques aux informations de spécificité. La mesure d'association utilisée est le LLR (Cf. section 2). Le seuil de significativité très sévère (valeur $p<0,0001$) permet de relever uniquement les cooccurrents sémantiquement très pertinents. Les cooccurrents sont identifiés dans une fenêtre d'observation (ou *span*) de 5 mots à gauche et à droite. Elle apporte suffisamment d'informations sémantiques pertinentes, sans introduire trop de bruit, et elle permet un traitement informatique efficace. Les mots vides sont conservés dans les listes de cooccurrents relevés, dans la mesure où ils sont susceptibles d'apporter des informations sémantiques (par exemple *pendant* pour indiquer qu'il s'agit d'un processus).

La mesure de recoupement a été implémentée à l'aide de scripts en Python, qui permettent de définir les paramètres (fenêtre d'observation, seuil de significativité, etc.) au niveau du repérage des cooccurrents de premier ordre et de deuxième ordre. Pour les 4717 unités lexicales spécifiques du corpus technique, nous calculons ainsi le degré de recoupement et donc le degré de monosémie, qui nous permet de les situer 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).

Comme nous ne disposons pas de listes de sens préétablis, ni d'autres mesures sémantiques comparables, nous avons procédé à une validation manuelle de la mesure de recoupement à partir de l'analyse manuelle des cooccurrents, ainsi qu'à une validation externe au moyen de dictionnaires. Les résultats de ces validations pour un échantillon de 50 unités lexicales spécifiques représentatives confirment les résultats de notre mesure de recoupement. Il convient tout de même de signaler que les mots les plus fréquents, tels que *machine* et *outil*, entrent très souvent dans la composition d'unités polylexicales (*machine à fraiser, machine à usiner*), ce qui pourrait en partie expliquer leur hétérogénéité sémantique. Comme nous l'avons déjà indiqué, l'analyse des unités polylexicales fera certainement l'objet de nos recherches futures. Notons également que des recherches supplémentaires s'imposent pour examiner la relation précise entre, d'une part, notre mesure de recoupement, qui implémente la monosémie en termes d'homogénéité sémantique, et, d'autre part, ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure, dans le but de développer un critère mesurable et de quantifier l'analyse sémantique.

4. La corrélation entre la spécificité et la monosémie

Afin de répondre à la question de recherche quantitative, nous étudions la corrélation entre le rang de spécificité et le rang de monosémie et nous procédons à une analyse statistique de régression simple. Celle-ci permet d'étudier l'impact d'une variable indépendante ou explicative, en l'occurrence le rang de spécificité, sur une variable dépendante ou expliquée, le rang de monosémie. Le résultat d'une analyse de régression simple est le coefficient de

détermination ou le pourcentage de variation expliquée R^2. Il représente le pourcentage de la variation du rang de monosémie que l'on pourra expliquer ou prédire à partir de la variation du rang de spécificité des 4717 unités lexicales spécifiques. Le résultat R^2 de l'analyse de régression comprend toujours une valeur p, indiquant la significativité statistique du modèle de régression et donc la fiabilité de sa capacité prédictive.

Les résultats de l'analyse de régression simple pour les 4717 unités lexicales spécifiques du corpus technique montrent une corrélation négative entre le rang de spécificité et le rang de monosémie (coefficient de corrélation Pearson de -0,72). La visualisation (Cf. figure 2) montre effectivement que la droite de régression s'incline vers le bas. Il s'avère donc que les unités lexicales les plus spécifiques du corpus technique, à gauche de la visualisation, 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, à droite de la visualisation, sont les plus homogènes sémantiquement (par exemple *rationnellement, télédiagnostic*), à quelques exceptions près. Le pourcentage de variation expliquée R^2 est de 51,57%, ce qui veut dire que la variation du rang de spécificité permet d'expliquer 51,57% de la variation du rang de monosémie. L'analyse de régression simple est hautement significative ($p<2.2e-16$).

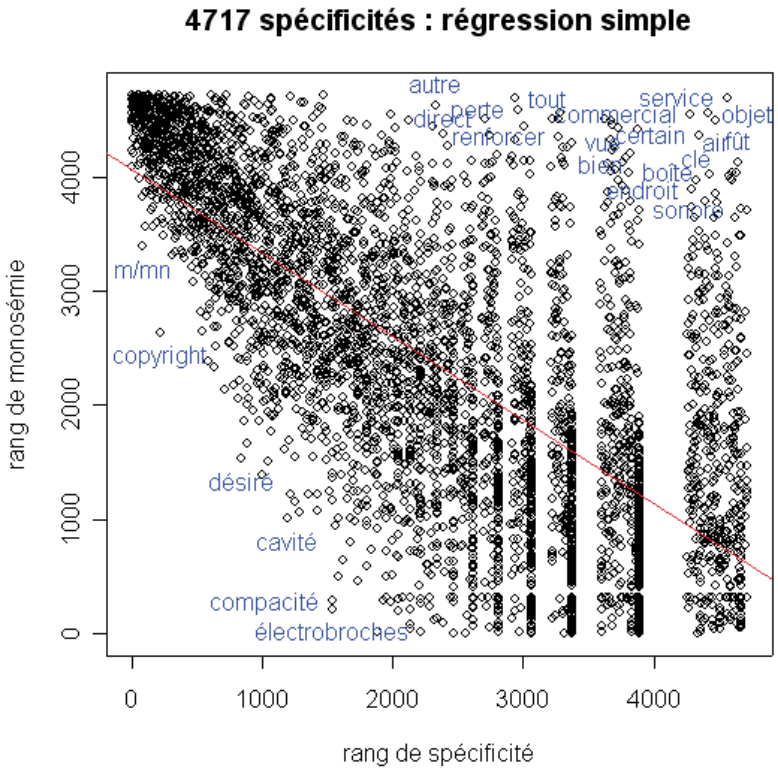

Figure 2 : Visualisation : corrélation entre le rang de spécificité et le rang de monosémie

5. La corrélation entre la spécificité et la monosémie technique

Dans le but de préciser les résultats de la mesure de recoupement et d'aboutir éventuellement à une granularité plus fine, nous avons enrichi la mesure en y intégrant la technicité des cooccurrents de deuxième ordre. Ce sont précisément ces derniers qui sont responsables du recoupement et qui influencent le plus le calcul du degré de recoupement. Nous avons développé une mesure de recoupement technique, qui permet de déterminer le degré de monosémie technique et dès lors le rang de monosémie technique. Nous nous demandons entre autres si ces résultats permettent de conduire à une distinction opérationnelle entre la monosémie technique et la monosémie générale (section 5.1.) et quel sera leur impact sur la corrélation entre le rang de spécificité et le rang de monosémie technique (section 5.2.).

5.1. La mesure de recoupement ou de monosémie technique

La nouvelle mesure repose sur un principe très simple : les cc techniques, c'est-à-dire les cc spécifiques du corpus technique, pèseront plus lourd sur le recoupement total de tous les cc que les cc généraux ou non techniques. Un mot de base avec plus de cc techniques qui se recoupent aura un degré de recoupement technique plus élevé. La nouvelle mesure de recoupement technique permet ainsi de déterminer dans quelle mesure le mot de base se caractérise par l'homogénéité sémantique technique.

Pour l'implémentation, nous recourons à un facteur de pondération, en fonction de la spécificité des c, qui est déterminée à partir de la liste de toutes les formes graphiques spécifiques du corpus technique. La nouvelle mesure de recoupement technique s'appuie sur le LLR pondéré (*weighted LLR* ou *wllr*). Elle prend en considération tous les c et tous les cc (dans le dénominateur de la fraction), mais elle effectue une pondération pour le recoupement des cc (dans le numérateur). Les cc très techniques ou très spécifiques du corpus technique se caractérisent par une valeur de LLR très élevée et par une valeur p très faible ($p<0,01$). Pour le calcul du recoupement technique, seuls les cc les plus spécifiques, dont le complément de la valeur p (ou 1-p) est supérieur ou égal à 0,9999, auront le poids intégral de 1, c'est-à-dire le facteur de pondération de 1. Les cc un peu moins techniques seront comptabilisés au poids de 0,9 (pour $(1\text{-}p) \geq 0,99$) et ainsi de suite, en fonction du complément de la valeur p. Les cc qui ne sont pas spécifiques ($p>0,05$) sont inclus également, mais au poids très faible de 0,01. Si ces cc non techniques se recoupent, leur apport au recoupement technique total sera limité. Etant donné que la nouvelle mesure de recoupement technique n'exclut aucun cc lors du calcul de recoupement, le dénominateur de la formule reste inchangé. Le numérateur intègre le facteur de pondération *wllr*, comme le montre la figure 3.

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

Figure 3 : Formule de la mesure de recoupement technique pondéré

Le degré de recoupement technique sera toujours inférieur au degré de recoupement de base (Cf. figure 1). En effet, pour le calcul du recoupement de base, tous les cc sont pris en considération au poids théorique de 1. De manière générale, plus le degré de recoupement technique est élevé, plus il s'approche du degré de recoupement de base et plus le recoupement

se fait par des cc techniques, qui sont plus fréquents. La comparaison croisée de la fréquence d'un cc et de son facteur de pondération (*wllr*), qui reflète sa spécificité dans le corpus technique, permet de distinguer quatre cas de figure (Cf. tableau 1).

1. Si la fréquence du cc est plutôt élevée (plus de recoupement) et si le cc est plutôt technique ou spécifique (*wllr* de 1 ou 0,9), sa contribution au degré de monosémie technique sera importante. Un mot de base avec beaucoup de cc techniques fréquents se caractérisera par un degré de monosémie technique très élevé et dès lors par un rang de monosémie technique plutôt bas ou plus près de 1 (c'est-à-dire plus monosémique).
2. Si la fréquence du cc est minimale (pas de recoupement) et si le cc est plutôt technique (*wllr* de 1 ou 0,9), sa contribution au degré de monosémie technique sera tout de même faible. Le degré de monosémie technique sera donc faible globalement et conduira à un rang de monosémie technique plutôt élevé (ou polysémique).
3. Si la fréquence du cc est minimale (pas de recoupement) et si le cc est général (*wllr* de 0,01), sa contribution au degré de monosémie technique sera bas à l'extrême. Le degré de monosémie technique global sera très faible et conduira à un rang de monosémie technique encore plus élevé (ou polysémique).
4. Si la fréquence du cc est plutôt élevée (plus de recoupement) et si le cc est général (*wllr* de 0,01), sa contribution au degré de monosémie technique sera très limitée, en dépit de sa fréquence importante. En plus, le facteur de pondération très faible de 0,01 génère la différence la plus grande possible entre le degré de monosémie et le degré de monosémie technique. Par conséquent, si un mot de base a beaucoup de cc généraux fréquents, son degré de monosémie technique sera beaucoup plus faible que son degré de monosémie (Cf. tableau 1). Le rang de monosémie technique sera plutôt élevé.

fréquence du cc	facteur de pondération (wllr) du cc (spécificité)	contribution au degré de monosémie technique	contribution au degré de monosémie	conclusion
élevée (p.ex. 6)	élevé (1 ou 0,9)	$6^2 \times 0{,}9 = 32{,}4$	$6^2 = 36$	mono tech.
minimale (p.ex. 1)	élevé (1 ou 0,9)	$1^2 \times 0{,}9 = 0{,}9$	$1^2 = 1$	poly tech.
minimale (p.ex. 1)	limité (0,01)	$1^2 \times 0{,}01 = 0{,}01$	$1^2 = 1$	poly gén.
élevée (p.ex. 6)	limité (0,01)	$6^2 \times 0{,}01 = 0{,}36$	$6^2 = 36$	mono gén.

Tableau 1 : Comparaison croisée de la fréquence et de la spécificité du cc

Bien évidemment, les cc d'un mot de base ne se situent pas tous dans le même cas de figure. Toutefois, les caractéristiques des cc donnent une indication fiable du type de monosémie du mot de base. Si les cc d'un mot de base sont majoritairement des cc techniques, spécifiques du corpus technique, et s'ils sont plutôt fréquents, le mot de base se caractérise par la monosémie technique. Si, en revanche, les cc d'un mot de base se situent principalement dans un des autres

cas de figure, le calcul de la mesure de monosémie technique conduira à un degré de monosémie technique plus bas. Toutefois, celui-ci ne coïncide pas toujours avec la polysémie technique (cc peu fréquents). En effet, si les cc sont majoritairement généraux, un degré de monosémie technique plutôt bas ou même très bas cache respectivement de la monosémie générale (cc fréquents) ou de la polysémie générale (cc peu fréquents). Signalons en guise d'exemple que les mots *tour* et *avance* se caractérisent par une polysémie plutôt générale. Ils s'utilisent effectivement dans plusieurs sens généraux dans le corpus technique.

5.2. Les résultats des analyses statistiques

Le rang de monosémie technique se prête aussi à des analyses statistiques de corrélation et de régression simple, qui permettent de déterminer si et dans quelle mesure le rang de spécificité d'un mot explique ou prédit son rang de monosémie technique. Le coefficient de corrélation Pearson (-0,65) montre également une corrélation négative, mais elle est moins convaincante que celle entre le rang de spécificité et le rang de monosémie de base (-0,72) (Cf. section 4). L'analyse de régression simple pour le rang de monosémie technique est hautement significative ($p<2.2e^{-16}$) et le pourcentage de variation expliquée R^2 est de 42,74%. La variation du rang de spécificité permet donc d'expliquer 42,74% de la variation du rang de monosémie technique, tandis qu'elle explique 51,57% de la variation du rang de monosémie de base (Cf. section 4). Il s'ensuit que, pour les 4717 unités lexicales spécifiques, le rang de spécificité est une variable explicative ou prédictive moins bonne pour le rang de monosémie technique que pour le rang de monosémie.

La visualisation ci-dessous (Cf. figure 4) montre clairement la corrélation négative entre le rang de spécificité et le rang de monosémie technique, parce que les mots les plus spécifiques du corpus technique, à gauche de la visualisation, sont les moins monosémiques techniquement. Etant donné que les mots les moins spécifiques se situent principalement en bas de la visualisation, à droite, ils se caractérisent plutôt par la monosémie technique. Toutefois, les résultats pour le rang de monosémie technique sont moins concluants que ceux pour le rang de monosémie de base. En effet, la comparaison des résultats visualise un faible déplacement de la droite de régression : un peu plus monosémique à gauche pour les mots plus spécifiques et un peu plus polysémique à droite pour les mots moins spécifiques (Cf. figure 4). Ce léger déplacement de la droite de régression par rapport à la droite de régression précédente (Cf. figure 2), visualisée ici en pointillé, pourrait s'interpréter comme un léger effet de la thèse monosémiste. Apparemment, les mots les plus spécifiques, ayant probablement beaucoup de cc techniques, sont un peu plus monosémiques techniquement. Néanmoins, la tendance observée pour le rang de monosémie technique s'oppose aussi à la corrélation positive préconisée par les monosémistes.

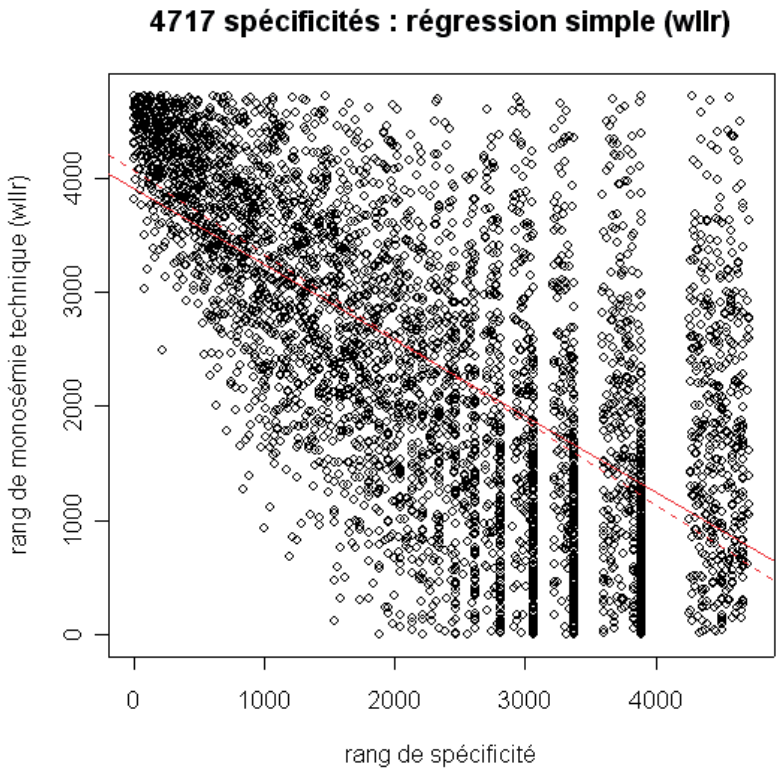

Figure 4 : Visualisation : corrélation entre le rang de spécificité et le rang de monosémie technique

6. Conclusion

L'originalité de notre étude réside principalement dans le développement d'une mesure de recoupement, afin d'automatiser et de quantifier l'analyse sémantique. A cet effet, nous avons implémenté la monosémie en termes d'homogénéité sémantique et nous avons calculé le degré de monosémie à partir du degré de recoupement des cooccurrents de deuxième ordre. Dans le but d'enrichir cette mesure, nous avons intégré un facteur de pondération, en fonction de la technicité des cooccurrents de deuxième ordre. Les analyses statistiques de régression simple ont permis d'étudier la corrélation entre, d'une part, le rang de spécificité des 4717 unités lexicales spécifiques du corpus technique et, d'autre part, leur rang de monosémie et leur rang de monosémie technique respectivement. Les résultats nous ont permis de répondre à la question de recherche quantitative et d'ébranler la thèse monosémiste traditionnelle.

Références

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

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

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

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

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, vol.(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,* vol.(28): 211-222.

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

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

Grefenstette G. (1994). Corpus-derived first, second and third-order word affinities. In Martin, W., Meijs, W. e.a. editors, *Proceedings of Euralex '94. International Congress on Lexicography, Amsterdam*, pp. 279-290.

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

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

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

Labbé C. and Labbé D. (2001). Que mesure la spécificité du vocabulaire ? *Lexicometrica,* vol.(3): http://cavi.univ-paris3.fr/lexicometrica/article/numero3/specificite 2001.PDF

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

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),* pp. 78-84.

Pezik P. (2005). You shall know a word by the company it keeps. A comparative study of co-occurrence statistics. Paper presented at *PALC 2005, Practical applications in language and computers*, Lodz, Poland.

Scott M. and Tribble C. (2006). *Textual Patterns. Key words and corpus analysis in language education.* Studies in Corpus Linguistics, vol.(22). Amsterdam: Benjamins.

Speelman D., Gondelaers S. and Geeraerts D. (2006). A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. In Wilson, A., Archer, D. and Rayson, P., editors, *Corpus Linguistics around the World*. Amsterdam: Rodopi.

Speelman D., Gondelaers S. and Geeraerts D. (2008). Variation in the choice of adjectives in the two main national varieties of Dutch. In Kristiansen, G. and Dirven, R. editors, *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin/New York: Mouton de Gruyter.

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. *Actes de TALN 2003 (10ème Conférence sur le Traitement Automatique des Langues Naturelles)*, pp. 265-274.

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

Wüster E. (1991). Einführung in die allgemeine Terminologielehre und terminologische Lexikographie (3. Aufl.). Bonn : Romanistischer Verlag.