Analyse quantitative et statistique de la sémantique dans un corpus technique

Ann Bertels[1], Dirk Speelman[2], Dirk Geeraerts[3]

[1] Katholieke Universiteit Leuven, ILT
ann.bertels@ilt.kuleuven.be

[1 2 3] Katholieke Universiteit Leuven, QLVL
{dirk.speelman ; dirk.geeraerts}@arts.kuleuven.be

Résumé

Cet article présente la méthodologie et les résultats d'une analyse sémantique quantitative d'environ 5000 spécificités dans le domaine technique des machines-outils pour l'usinage des métaux. Les spécificités seront identifiées avec la méthode des mots-clés (*KeyWords Method*). Ensuite, elles seront soumises à une analyse sémantique quantitative, à partir du recouvrement des cooccurrences des cooccurrences, permettant de déterminer le degré de monosémie des spécificités. Finalement, les données quantitatives de spécificité et de monosémie feront l'objet d'analyses de régression. Nous avançons l'hypothèse que les mots (les plus) spécifiques du corpus technique ne sont pas (les plus) monosémiques. Nous présenterons ici les résultats statistiques, ainsi qu'une interprétation linguistique. Le but de cette étude est donc de vérifier si et dans quelle mesure les spécificités du corpus technique sont monosémiques ou polysémiques et quels sont les facteurs déterminants.

Mots-clés : sémantique lexicale, sémantique quantitative, spécificités, polysémie, cooccurrences, analyse de régression.

Abstract

This article discusses the methodology and results of a quantitative semantic analysis of about 5000 keywords (pivotal terms) in the domain of French machining terminology. The KeyWords Method is used in order to identify the most typical words. Next, a quantitative semantic analysis of the keywords determines their degree of monosemy, which is implemented in terms of degree of overlap between co-occurrents of co-occurrents of keywords. Finally, the quantitative data is submitted to various regression analyses, in order to check the hypothesis that the most typical terms are not always the most monosemous terms. This article presents the statistical results of this semantic analysis and provides linguistic interpretation. Building on corpus data, the investigation attempts to establish in how far keywords are polysemous and which factors are most predictive.

Keywords: lexical semantics, quantitative semantics, keywords, polysemy, co-occurrences, regression analysis.

1. Introduction

Cet article s'inscrit dans le cadre d'une thèse de doctorat, qui est une étude sémantique quantitative du vocabulaire spécifique d'un corpus en français technique. Le but de cette étude est d'étudier la sémantique (monosémie *vs* polysémie) des mots et termes spécifiques du corpus technique. Les textes du corpus relèvent du domaine technique des machines-outils pour l'usinage des métaux. Un corpus électronique de textes spécialisés offre une information indispensable pour l'analyse sémantique, à savoir le contexte linguistique. Toutefois, l'exploitation de grandes quantités de textes requiert une approche quantitative et automatisée et mène à une analyse statistique de ces données quantitatives.

Afin de quantifier l'analyse sémantique des spécificités du corpus technique, nous proposons une double analyse quantitative (Bertels, 2005), en termes de degré de spécificité et de degré de monosémie. Par conséquent, la question est mesurable : elle étudie la corrélation entre le degré de spécificité d'une unité linguistique et son degré de monosémie et dès lors, elle requiert une approche scalaire par continuum. Les unités linguistiques seront situées sur un continuum de spécificité (allant des plus spécifiques aux moins spécifiques) ainsi que sur un continuum de monosémie (allant des plus monosémiques aux plus polysémiques). Nous avançons l'hypothèse de recherche que les mots (les plus) spécifiques du corpus technique ne sont pas (les plus) monosémiques, contrairement à la thèse de la terminologie traditionnelle. L'analyse se propose donc de vérifier la polysémie des mots du corpus technique, par exemple le mot *broche* (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 ».

Pour étudier la sémantique dans la langue spécialisée, l'approche adoptée traditionnellement est une approche catégorielle, qui se caractérise d'une part par la dichotomie *termes – mots* et d'autre part par l'opposition *monosémie – polysémie*. La terminologie traditionnelle, qui préconise la monosémie des termes d'une langue spécialisée ainsi qu'une approche onomasiologique prescriptive, a été récemment remise en question par les partisans de la terminologie descriptive. Sur le plan des unités linguistiques, la dichotomie entre langue générale et langue spécialisée a été rejetée (Condamines et Rebeyrolle, 1997). Les termes font partie de la langue naturelle et véhiculent des connaissances spécialisées (Lerat, 1995). En plus, 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* ou *broche*, 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*). Finalement, le vocabulaire d'un corpus technique comprend des mots de la langue générale, tels que *type, modèle, permettre*, etc. Sur le plan de l'analyse sémantique, la monosémie de la langue spécialisée a été remise en question notamment par la Théorie Communicative de la Terminologie (Cabré, 2000) et par la Terminologie socio-cognitive (Temmerman, 2000). Ces remises en questions nous incitent à adopter une approche scalaire par continuum, tant au niveau des unités linguistiques (continuum de spécificité) qu'au niveau sémantique (continuum de monosémie ou de sens). Pour étudier la question principale (corrélation entre le degré de spécificité et le degré de monosémie), nous proposons de recourir à une analyse statistique de régression simple, qui fait intervenir le degré de monosémie et le degré de spécificité. La recherche relative à la question centrale s'accompagne nécessairement de l'étude de plusieurs aspects subsidiaires, dans la mesure où le degré de monosémie n'est pas uniquement influencé par le degré de spécificité, mais également par la fréquence, la classe lexicale, le nombre de classes lexicales et la longueur des mots-formes. Pour étudier l'impact de toutes ces variables indépendantes sur le degré de monosémie, nous procédons à une analyse de régression multiple.

L'originalité de cette étude réside principalement dans l'approche scalaire de la sémantique, à savoir le continuum de monosémie, qui implémente la monosémie comme l'homogénéité sémantique et qui permet de la quantifier par le biais du recouvrement des cooccurrences des cooccurrences. En plus, cette analyse sémantique quantitative porte sur presque 5000 mots du corpus technique, contrairement aux travaux antérieurs (Eriksen, 2002 ; Ferrari, 2002 ; Condamines et Rebeyrolle, 1997), qui étudient également la polysémie dans un corpus représentatif d'un domaine de spécialité, mais en se limitant à quelques mots seulement.

Dans la section suivante (2), nous expliquerons les analyses quantitatives, d'abord au niveau des spécificités et ensuite au niveau de la sémantique. Dans la section 3, nous procéderons

aux analyses statistiques faisant intervenir non seulement les données quantitatives des deux axes méthodologiques, mais également les autres variables indépendantes. Nous commenterons les résultats des analyses statistiques, les interprétations linguistiques ainsi que les problèmes principaux. Dans la dernière section (4), nous présenterons la conclusion et les perspectives.

2. Analyse quantitative

Nous étudions la corrélation entre le continuum de spécificité et le continuum de monosémie dans un corpus technique. Les textes de ce corpus technique (1,7 million de mots) relèvent du domaine des machines-outils pour l'usinage des métaux. Le corpus a été étiqueté et lemmatisé par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1998 à 2001 : revues électroniques (800.000) et fiches techniques (300.000) trouvées sur Internet, normes ISO et directives (300.000) et quatre manuels numérisés (360.000). Les textes se situent à différents niveaux de normalisation et de vulgarisation. Afin de pouvoir identifier les spécificités, le corpus technique a été complété par un corpus de référence, constitué d'articles du journal *Le Monde* (1998), comprenant environ 15.300.000 mots également lemmatisés. La question centrale (corrélation entre le continuum de spécificité et le continuum de monosémie) requiert une double analyse quantitative : le calcul des spécificités et une mesure pour déterminer le degré de monosémie. À présent, notre étude identifie les spécificités au niveau des unités simples, par exemple *fraisage, commande*. Des recherches futures devront certainement porter sur les unités polylexicales, par exemple *commande numérique*, puisque la plupart des unités terminologiques se situent à ce niveau.

2.1. Quantifier la spécificité

La première analyse quantitative est celle des spécificités et permet d'établir le continuum de spécificité. Les spécificités ou mots-clés ne sont pas les mots les plus fréquents du corpus technique, mais les mots les plus caractéristiques et les plus représentatifs[1]. Afin d'identifier les spécificités, le corpus technique est comparé à un corpus de référence de langue générale. En termes relatifs, les spécificités sont significativement plus fréquentes dans le corpus technique que dans le corpus de référence de langue générale. Pour dresser la liste de mots-clés, nous recourons à la méthode des mots-clés (*KeyWords Method*), implémentée dans le logiciel Abundantia Verborum Frequency List Tool[2] et basée sur la statistique du LLR (*log likelihood ratio)* (log de vraisemblance). D'autres logiciels et méthodologies sont également disponibles, notamment WordSmith Tools[3] pour la méthode des mots-clés et Lexico3[4] pour le calcul des spécificités. Les trois logiciels aboutissent à des résultats similaires. Après suppression des hapax et après avoir filtré les mots grammaticaux et les noms propres, nous recensons 4717 spécificités dans le corpus technique, tant des termes au sens strict tels que *fraisage*, *usinage* que des mots tels que *type*, *permettre*. La mesure statistique du LLR indique le degré de spécificité et permet de classer les mots-clés par ordre de spécificité décroissante et par conséquent, de les situer sur un continuum en fonction de leur rang de spécificité. Les

[1] À titre d'expérimentation, les 4717 mots les plus spécifiques du corpus technique ont été comparés aux 4757 mots les plus fréquents du corpus technique (fréquence technique absolue ≥ 18). Le recouvrement est important : parmi les 4717 mots spécifiques, 2548 mots appartiennent à la liste des 4757 mots les plus fréquents (54 %), les autres 2169 spécificités étant moins fréquentes (<17).

[2] Abundantia Verborum : http://wwwling.arts.kuleuven.ac.be/genling/abundant/obtain/ (Date de consultation : le 16/02/06).

[3] WordSmith Tools version 3 : http://www.lexically.net/wordsmith/ (Date de consultation : le 16/02/06).

[4] Lexico3 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/ (Date de consultation : le 16/02/06).

mots avec un degré de spécificité identique, c'est-à-dire une valeur de LLR identique, auront le même rang de spécificité. Les mots les plus spécifiques, à savoir *machine, outil, usinage, pièce, mm, vitesse, coupe,* … reflètent clairement la thématique du domaine.

2.2. Quantifier la sémantique

La deuxième analyse quantitative sert à établir le continuum de monosémie des spécificités. À cet effet, nous avons développé une mesure de monosémie permettant de déterminer le degré de monosémie d'une spécificité (ou mot-clé), à partir du recouvrement des cooccurrents de ses cooccurrents. Pourquoi recourir aux cooccurrents de deuxième ordre ? Le caractère monosémique ou polysémique d'une unité linguistique se caractérise par des contextes sémantiquement homogènes, si elle est monosémique, ou sémantiquement hétérogènes si elle est polysémique (Schütze, 1998 ; Véronis, 2003). L'accès à la sémantique des cooccurrents se fait par leurs cooccurrents et plus particulièrement par le recouvrement de ces cooccurrents de deuxième ordre. Si les cooccurrents de deuxième ordre ont un degré de recouvrement élevé, ils sont fortement partagés par les cooccurrents, ce qui est une indication de l'homogénéité sémantique des cooccurrents. Le degré de ressemblance ou de similarité lexicale des coocccurrents d'un mot de base étant proportionnel au degré de monosémie de ce mot, un fort recouvrement des cooccurrents de deuxième ordre signale ainsi un degré plus important de monosémie. La formule de recouvrement (cf. ci-dessous) est basée sur le recouvrement formel des cooccurrents des cooccurrents (cc), tenant compte (1) de la fréquence d'un cc dans la liste des cc, (2) du nombre total de c et (3) du nombre total de cc. Un cc sera d'autant plus important pour le recouvrement total s'il figure plus souvent dans la liste des cc, donc si sa fréquence dans la liste des cc est plus élevée ou s'il est plus partagé par les cooccurrents (c).

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

Considérons en guise d'exemple un cc fortement partagé, par exemple par 5 c des 7 c au total. 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 recouvrement est donc exprimé par la fraction 5/7. En exprimant pour chaque cc le recouvrement 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 recouvrement) et 1 (recouvrement important ou parfait) et sera facilement interprétable. Nous considérons les c et cc au niveau des formes graphiques et non pas au niveau des lemmes (formes canoniques). De cette façon, la mesure de recouvrement permet de faire la distinction entre, par exemple, *pièce usinée* et *pièce à usiner*. La mesure d'association utilisée pour déterminer les cooccurrences pertinentes est la statistique du LLR (log de vraisemblance). Comme le seuil de significativité est très sévère (une p-valeur < 0.0001), on relève les cooccurrents sémantiquement pertinents. L'algorithme et les scripts en Python[5] permettent de définir les paramètres (fenêtre d'observation, seuil de significativité, etc.) au niveau des cooccurrents (premier ordre) et au niveau des cooccurrents des cooccurrents (deuxième ordre). Les scripts génèrent une grande base de données indexée avec toutes les informations statistiques pertinentes (LLR, seuil de significativité, etc.). Pour les 4717 spécificités du corpus technique, nous calculons ainsi le degré de recouvrement 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.

[5] http://www.python.org/ (Date de consultation : le 16/02/06).

3. Analyse statistique

À l'aide d'une analyse de régression simple dans le logiciel statistique R[6], nous évaluons l'impact du rang de spécificité sur le rang de monosémie, ce qui constitue la réponse à la question de recherche principale. Nous procédons également à une analyse de régression multiple, faisant intervenir toutes les variables indépendantes (spécificité, fréquence, classe lexicale, nombre de classes lexicales, etc.). Dans cette section, nous commenterons les résultats des analyses de régression simple et multiple, les interprétations linguistiques, ainsi que les problèmes techniques, tels que l'hétéroscédasticité et la multicollinéarité.

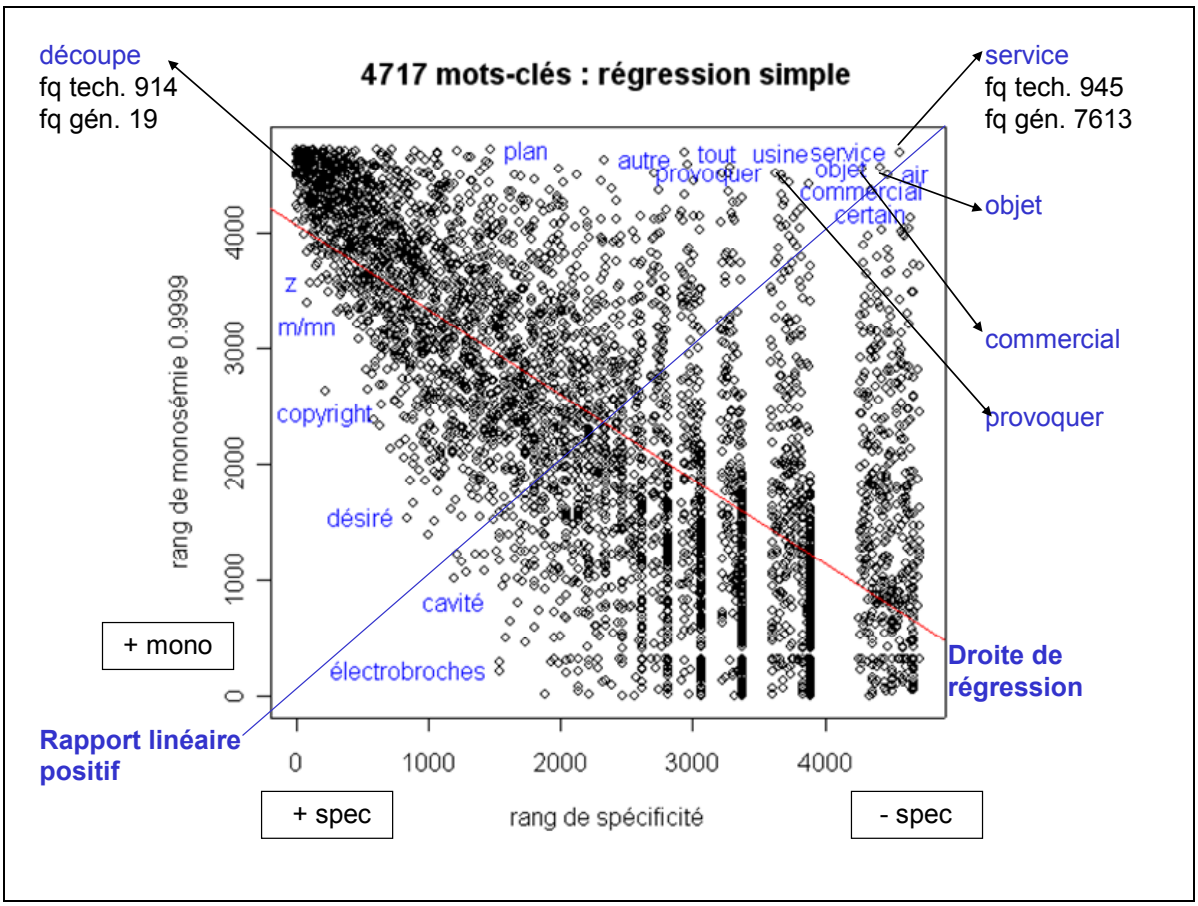

Figure 1. Visualisation de l'analyse de régression simple

3.1. Analyse de régression simple

L'analyse de régression simple vise à évaluer l'impact du rang de spécificité (VI : variable indépendante ou explicative) sur le rang de monosémie (VD : variable dépendante ou expliquée). Le coefficient de corrélation Pearson (-0,71) indique 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. La visualisation ci-dessous (cf. figure 1) indique clairement la tendance linéaire négative, visualisée par la droite de régression descendante, contrairement au rapport linéaire positif

[6] http://www.r-project.org/ (Date de consultation : le 16/02/06).

des monosémistes traditionnels. Notons d'emblée que la figure 1 soulève la question de la pertinence de la régression linéaire, abordée ci-dessous. Les résultats de l'analyse de régression simple ainsi que la visualisation permettent d'infirmer la thèse traditionnelle évoquée plus haut. En effet, les mots les plus spécifiques de notre corpus technique d'analyse ne sont pas les plus monosémiques, mais au contraire, ce sont les plus polysémiques[7]. En plus, les mots les moins spécifiques sont plutôt monosémiques, à quelques exceptions près, notamment *service, objet, commercial,* etc. qui se situent en haut à droite. Ils ont des résidus positifs importants : ils sont plus polysémiques qu'attendu en fonction de leur rang de spécificité. Leur rang de monosémie estimé en fonction de leur rang de spécificité se situe sur la droite de régression, plus bas.

Ces mots peu spécifiques, polysémiques et à résidus importants se caractérisent par une fréquence technique importante et par une fréquence générale beaucoup plus importante en termes relatifs. À titre de comparaison, la fréquence technique de *service* est de 945 et sa fréquence générale de 7613, tandis qu'un mot en haut à gauche polysémique, tel que *découpe*, a une fréquence technique comparable de 914, pour une fréquence générale de 19. Les mots à résidus positifs importants, comme *service*, sont des mots généraux : ils sont très peu spécifiques en raison de leur fréquence générale importante. Toutefois, ils sont polysémiques en raison de leur nombre important de c et cc (et en raison de leur fréquence technique importante). Les mots à résidus négatifs importants, tels que *électrobroches* (cf. figure 1), se retrouvent en dessous de la droite de régression, car ils sont plus monosémiques qu'attendu en fonction de leur rang de spécificité. Ces mots sont plutôt spécifiques en raison de leur fréquence générale limitée ou zéro et ils sont plus monosémiques qu'attendu en raison du très faible nombre de cc différents, et cela en dépit de leur fréquence technique considérable.

Le problème auquel on est confronté dans cette analyse de régression simple est le problème de l'hétéroscédasticité des résidus, parce que les résidus (ou les distances entre les valeurs observées et les valeurs estimées) ne sont pas répartis selon une distribution normale. Certains mots à résidus positifs importants se situent (très) loin de la droite de régression. Toutefois l'on n'observe pas autant de mots à résidus négatifs aussi importants en dessous de la droite de régression. L'hétéroscédasticité peut être détectée à l'aide du test statistique de Goldfeld-Quandt, implémenté dans R. Les solutions techniques adoptées généralement consistent soit en transformations polynomiales ou logarithmiques soit en une régression pondérée. Les transformations logarithmiques ne permettent pas de résoudre le problème. La transformation polynomiale ($y^2 \sim \sqrt{x}$) aboutit à un pourcentage de variation expliquée R^2 de 57 % et à l'homoscédasticité des résidus. Du point de vue technique, le problème est résolu, mais du point de vue linguistique, il est difficile d'interpréter le carré du rang de monosémie et la racine carrée du rang de spécificité. La deuxième solution technique, l'analyse de régression pondérée, est basée sur la méthode des moindres carrés pondérés. Elle consiste à accorder moins d'importance aux mots à résidus importants et plus d'importance aux mots à résidus limités. Le résultat de la régression pondérée est un pourcentage de variation expliquée R^2 de 62,51 %. La figure 2 ci-dessous montre la tendance linéaire négative plus claire. Les mots sont moins dispersés (d'où le R^2 plus élevé), mais ils sont comprimés. Le mot *service* se retrouve au même niveau de spécificité, mais à un niveau plus monosémique.

[7] Il est à noter que les mots spécifiques se retrouvent en grande partie dans la liste des mots les plus fréquents du corpus technique. Ils entrent souvent dans la composition des syntagmes nominaux et des unités polylexicales (terminologiques et monosémiques). Comme ces unités polylexicales ont des distributions hétérogènes, il pourrait en résulter que les mots les plus fréquents (et constituants des unités polylexicales) aient des cooccurrents très différents. La polysémie des mots les plus spécifiques et les plus fréquents pourrait donc s'expliquer par le fait qu'ils entrent dans la composition de nombreuses unités polylexicales.

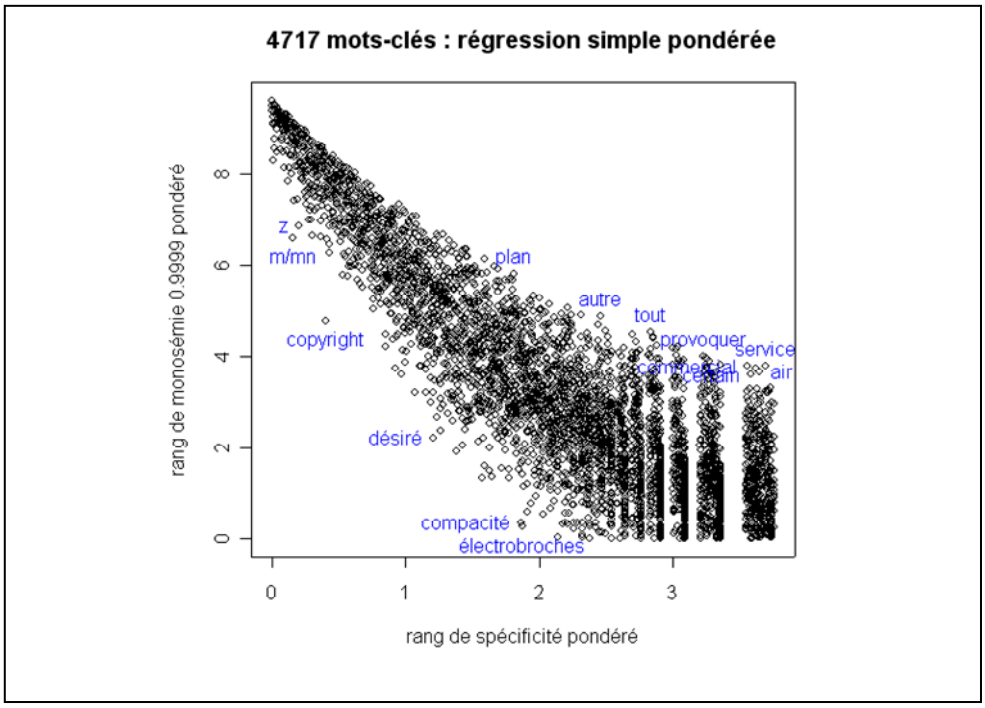

Figure 2. Visualisation de l'analyse de régression simple pondérée

Force est de constater que si la régression pondérée revient à sous-estimer ou méconnaître les caractéristiques des mots (fréquence technique ou générale), elle semble néanmoins confirmer notre hypothèse initiale. Il se pourrait qu'un autre facteur, que nous n'avons pas envisagé ici, entre en ligne de compte. Nous tenterons de répondre à cette question en élaborant une mesure de monosémie technique pondérée, qui tient compte de la spécificité (ou technicité) des cc en intégrant leur LLR dans le calcul du recouvrement, par le biais d'un facteur de pondération.

3.2. Analyse de régression multiple

L'analyse de régression multiple évalue l'impact combiné et simultané de plusieurs variables indépendantes sur la variable dépendante, en l'occurrence le rang de monosémie. Ces variables indépendantes ou explicatives servent à prévoir ou à expliquer la variation de la variable dépendante ou expliquée. Malheureusement, ces variables indépendantes ne sont pas toujours indépendantes les unes des autres. Parfois, deux ou plusieurs variables indépendantes sont corrélées les unes avec les autres, c'est-à-dire qu'elles expliquent en grande partie la même variation de la variable dépendante. C'est le problème de la multicollinéarité, qui entraîne deux conséquences importantes. Premièrement, la multicollinéarité mène à une augmentation des écarts-types des estimations de coefficient dans le modèle de régression multiple. Par conséquent, on trouvera moins vite des rapports significatifs entre les variables indépendantes et la variable dépendante. Lorsqu'on procède à des tests *t* pour déterminer la significativité des coefficients particuliers, on risque de trouver qu'aucune des variables n'est significative, tandis que le test *F* du modèle de régression multiple dénote une relation (hautement) significative. Deuxièmement, la multicollinéarité rend le modèle de régression multiple peu fiable, car elle accroît l'erreur sur les valeurs estimées de la variable dépendante. Compte tenu de ces corollaires, il importe de détecter la multicollinéarité et de la résoudre.

Pour détecter des problèmes de multicollinéarité faisant intervenir deux ou plusieurs variables indépendantes, on fait appel aux facteurs d'inflation de la variance (VIF ou *Variance Inflation*

Factor). Le VIF d'une variable indépendante est calculé en considérant cette variable comme variable dépendante d'une analyse de régression multiple particulière avec toutes les autres variables indépendantes[8] comme variables indépendantes. Si cette variable est caractérisée par des rapports linéaires avec les autres variables, son coefficient de corrélation (R^2) ou pourcentage de variation expliquée sera élevé. Le calcul des VIF[9] est implémenté dans R et se fait simultanément pour toutes les variables indépendantes. Un VIF supérieur à 10 signale un problème de multicollinéarité et, le cas échéant, toutes les variables impliquées dans le rapport colinéaire auront un VIF très (ou trop) élevé. La solution du problème de multicollinéarité consiste à exclure du modèle de régression multiple une des variables indépendantes responsables de la multicollinéarité et ayant un VIF trop élevé.

La variable dépendante du modèle de régression multiple de notre analyse est le rang de monosémie au seuil de significativité de 0,9999. Les variables indépendantes sont le rang de spécificité, le log du LLR, le rang de fréquence dans le corpus technique, le rang de fréquence dans le corpus général, la fréquence technique absolue, la fréquence générale absolue, la classe lexicale (comme variable catégorielle), le nombre de classes lexicales et la longueur du mot-clé. La matrice des corrélations montre un coefficient de corrélation Pearson trop élevée (supérieur à 0,90) entre le rang de spécificité et le log du LLR, qui sont clairement intercorrélés. Le calcul des VIF signale un problème de multicollinéarité pour trois variables : le log du LLR (VIF 36,26), le rang de spécificité (VIF 26,32) et le rang de fréquence technique (VIF 14,72). La suppression du rang de spécificité ne permet pas de résoudre tout le problème de multicollinéarité, car le VIF du rang de fréquence technique reste trop élevé. En raison de la corrélation très importante du rang de fréquence technique et la variable dépendante (rang de monosémie), nous préférons le garder. Mais sa corrélation importante avec le rang de fréquence générale, permet de supprimer le rang de fréquence générale. Par conséquent, la multicollinéarité est résolue et le rang de fréquence technique est maintenu dans le modèle de régression multiple. Afin de tout de même maintenir la différence (ou l'écart) du rang de fréquence générale par rapport au rang de fréquence technique, nous envisageons d'intégrer dans nos recherches futures une variable indépendante supplémentaire, à savoir l'écart du rang de fréquence. Si le rang de fréquence générale est supprimé, cette nouvelle variable permettra de reprendre partiellement l'information perdue, sans le problème de multicollinéarité.

La régression multiple fait donc intervenir le rang de monosémie comme variable dépendante et les variables indépendantes citées ci-dessus, sans le rang de spécificité ni le rang de fréquence générale. Les variables indépendantes significatives (cf. figure 3) expliquent 80,68 % de la variation du rang de monosémie : le rang de fréquence technique, le log du LLR (log_LLR), la longueur et le nombre de classes lexicales. La colonne des valeurs estimées (*estimate*) montre que le rang de fréquence technique a un rapport de corrélation négative avec le rang de monosémie. Plus les mots sont fréquents dans le corpus technique, moins ils sont monosémiques. Le coefficient du log_LLR est positif : plus le degré de LLR est élevé (c'est-à-dire plus les mots-clés sont spécifiques), plus ils sont polysémiques (avec des rangs de monosémie près de 4717). Le coefficient de la longueur indique un rapport de corrélation négative : plus les mots sont longs, plus ils sont monosémiques. Finalement, nous observons un léger impact du nombre de classes lexicales : si un mot-clé appartient à plusieurs classes lexicales, il sera plus polysémique. La dernière colonne de la valeur p (significativité) montre que le rang de fréquence technique et le log_LLR sont les facteurs les

[8] Le calcul des facteurs d'inflation de la variance ou des VIF prend en considération uniquement des variables numériques, donc pas des variables catégorielles.

[9] $VIF = 1/(1-R^2)$

plus significatifs et les plus importants pour prévoir le rang de monosémie des spécificités du corpus technique.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4157.43062	67.14828	61.914	< 2e-16	***
log_LLR	**156.50174**	**19.90169**	**7.864**	**4.59e-15**	***
rang_v_freq1	**-0.85304**	**0.01121**	**-76.101**	**< 2e-16**	***
nbr_claslex	69.30520	23.46212	2.954	0.00315	**
long	**-19.84211**	**2.73853**	**-7.246**	**5.01e-13**	***

Figure 3. Résultats de l'analyse de régression multiple

4. Conclusion et perspectives

Les résultats de l'analyse statistique de régression simple ont démontré et visualisé clairement la corrélation négative, même si elle n'est pas tout à fait linéaire. Plus les mots-clés sont spécifiques dans le corpus technique, plus ils sont polysémiques, c'est-à-dire sémantiquement hétérogènes. L'analyse de régression multiple a permis de détecter les facteurs influant significativement sur la variation du rang de monosémie : il s'agit surtout du rang de fréquence dans le corpus technique, du degré de spécificité et de la longueur et, dans une moindre mesure, du nombre de classes lexicales. Toutefois, des recherches supplémentaires s'imposent pour examiner la relation précise entre notre mesure de monosémie, implémentant la monosémie comme l'homogénéité sémantique, et ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure de monosémie ou de recouvrement, dans le but de développer un critère opérationnalisable et mesurable. Sans recherches supplémentaires, nous ne pourrons pas prétendre que notre mesure de monosémie correspond parfaitement à ce que les traditionalistes considèrent comme polysémie. Nous envisageons également de procéder à une validation manuelle de la mesure de recouvrement à partir d'une analyse manuelle des collocations. En plus, la mesure de recouvrement technique pondérée, intégrant la spécificité des cc dans le calcul du recouvrement, permettra de nuancer les résultats des analyses de régression simple et multiple présentées ci-dessus.

Nous nous proposons finalement d'effectuer des analyses de régression détaillées, c'est-à-dire par classe lexicale (substantifs / adjectifs / verbes / adverbes) et par sous-catégorie (par exemple les substantifs déverbaux, les abréviations et sigles). D'ailleurs, il serait intéressant également de vérifier la corrélation entre le rang de spécificité et le rang de monosémie dans les différents sous-corpus (revues / fiches techniques / normes / manuels). On peut en effet se demander si il n'y aurait pas une meilleure corrélation ou une corrélation positive entre le rang de spécificité et le rang de monosémie dans les normes, censées être prescriptives ?

Références

BERTELS A. (2005). « À la découverte de la polysémie des spécificités du français technique ». In *Actes de RÉCITAL 2005* : 575-584.

CABRÉ M.T. (2000). « Terminologie et linguistique : la théorie des portes ». In *Terminologies nouvelles* 21 : 10-15.

CONDAMINES A., REBEYROLLE J. (1997). « Point de vue en langue spécialisée ». In *Meta* 42 (1) : 174-184.

ERIKSEN L. (2002). « Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder : Zur Terminologie der ‚Sache' im Deutschen ». In *Journal of Linguistics* 28 : 211-222.

FERRARI L. (2002). « Un caso de polisemia en el discurso jurídico ? ». In *Terminology* 8 (2) : 221-244.

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

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

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

SCHÜTZE H. (1998). « Automatic Word Sense Discrimination ». In *Computational Linguistics* 24 (1) : 97-123.

VERONIS J. (2003). « Cartographie lexicale pour la recherche d'informations ». In *Actes de TALN 2003* : 265-274.