Sémantique quantitative et corpus technique : des analyses statistiques aux interprétations linguistiques

Ann Bertels

ILT et QLVL – K.U.Leuven – B-3000 Leuven – Belgique

Abstract

This article discusses the results of a quantitative semantic analysis of pivotal terms (or keywords) in the domain of machining terminology in French. The investigation attempts to find out whether, and to what extent, pivotal lexical items in a technical domain are monosemous or polysemous. Since about 5 000 keywords are included in the semantic analysis, automation and quantification are required. Hence, we conducted a double quantitative analysis, including both the identification of pivotal lexical items and their semantic analysis, resulting in a statistical analysis of the quantitative data. In this article, we explain how to go from a qualitative semantic question to a quantitative semantic analysis and next to statistical analyses, in order to finally draw quantitative and linguistic conclusions.

Résumé

Cet article présente les résultats d'une étude sémantique quantitative du vocabulaire spécifique d'un corpus en français technique. Les textes du corpus relèvent du domaine technique des machines-outils pour l'usinage des métaux. L'objectif de l'étude est de vérifier si et dans quelle mesure les unités lexicales spécifiques de ce domaine technique sont monosémiques ou polysémiques. Comme l'analyse sémantique porte sur quelque 5 000 unités lexicales du corpus technique, l'automatisation et la quantification s'imposent. Nous procédons à une double analyse quantitative (analyse des spécificités et analyse sémantique), qui mène à une analyse statistique des données quantitatives. Dans cet article, nous expliquerons comment on pourra passer d'une question sémantique qualitative à une analyse sémantique quantitative et ensuite à des analyses statistiques, afin de formuler finalement des conclusions quantitatives et linguistiques.

Mots-clés : sémantique quantitative, corpus technique, spécificités, polysémie, analyses de régression.

1. Approche méthodologique

Nous procédons à une étude *sémantique quantitative* d'un corpus *spécialisé* (Bertels, 2006), en l'occurrence un corpus relevant du domaine technique des machines-outils pour l'usinage des métaux. Les trois adjectifs (sémantique, quantitative et spécialisé) méritent un mot d'explication, puisque notre étude vise à remettre en question la thèse *sémantique* de monosémisme défendue par l'approche traditionnelle, au moyen d'une étude *quantitative* d'un corpus *spécialisé*.

Une étude linguistique qui se focalise sur un domaine technique *spécialisé* soulève tout de suite des questions sur les particularités de la langue spécialisée. 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). 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. Toutefois, récemment, les partisans de la terminologie descriptive ont remis en question cet idéal de monosémie et d'univocité. 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 domaine spécialisé (Arnzt et Picht, 1989 ; Condamines et Rebeyrolle, 1997 ; Temmerman, 2000 ; Eriksen, 2002 ; Ferrari, 2002). Ces travaux antérieurs étudient, comme nous, la polysémie dans un corpus représentatif d'un domaine spécialisé, mais ils se limitent à l'analyse de quelques mots seulement. Ainsi, Condamines et Rebeyrolle (1997) étudient un corpus de textes spécialisés du domaine de l'espace. Leur analyse consiste à classer les contextes d'apparition d'un terme (par exemple *satellite*) afin de vérifier si ces contextes peuvent être considérés comme sémantiquement homogènes ou non. Ferrari (2002) analyse les termes espagnols *distinción* et *discriminación* dans un corpus juridique spécialisé de dix traités internationaux. L'identification des contextes syntactico-sémantiques permet de vérifier si le signifié des termes est identique dans tous les schémas syntactico-sémantiques ou s'il s'agit de cas de polysémie.

La caractéristique traditionnelle de la monosémie des unités terminologiques d'un corpus spécialisé justifie le deuxième adjectif de notre étude, à savoir *sémantique*. L'objectif principal de notre étude sémantique est de vérifier si les unités lexicales de notre 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 traditionnelle, nous procédons à une étude de corpus à grande échelle. Il faut donc opérationnaliser la thèse monosémiste et la reformuler en une question 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[1] de la langue spécialisée (ou 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. L'idée que les unités lexicales, qui sont plus ou moins spécifiques, sont plus ou moins monosémiques, implique l'idée de gradation, c'est-à-dire l'idée de degré de spécificité et de degré de monosémie.

Pour répondre à cette question quantitative, nous envisageons une double analyse quantitative. D'une part, nous proposons de quantifier l'analyse sémantique en développant une mesure de monosémie, qui permet d'accorder une valeur numérique à chaque unité lexicale analysée, en fonction de son degré de monosémie. D'autre part, nous lui accordons une valeur numérique en fonction de son degré de spécificité. Ainsi, nous avançons l'hypothèse que les unités lexicales (les plus) spécifiques du corpus technique ne sont pas (les plus) monosémiques, contrairement à la thèse monosémiste traditionnelle. Le corpus technique analysé contient effectivement des mots polysémiques, 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 ». Afin d'étayer notre hypothèse, nous procédons à des analyses statistiques, qui permettent de vérifier la corrélation entre les données quantitatives de spécificité et de monosémie de toutes les unités lexicales spécifiques ou représentatives du corpus technique. De par son approche, notre étude vise donc à réconcilier la linguistique et la technique (notamment l'informatique et la statistique), parce

[1] Il est à noter que les unités grammaticales seront supprimées de la liste des unités spécifiques, qui ne comprendra que des unités lexicales.

qu'elle recourt à des techniques quantitatives pour mieux comprendre et expliquer des phénomènes linguistiques.

Le corpus technique d'analyse 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 et consiste en 4 sous-corpus, datant de 1996 à 2002 : 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). 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. Les deux sous-corpus des normes et directives et des guides et manuels sont plus normatifs et prescriptifs que les deux autres sous-corpus issus des revues électroniques et des fiches techniques, qui sont plus descriptifs. Les normes et directives et les fiches techniques s'adressent plutôt à des professionnels, tandis que les guides et manuels (et dans une certaine mesure les revues) sont plus didactiques et vulgarisants et visent un public d'étudiants et de semi-experts. 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.

Après cette explication méthodologique, nous préciserons les deux volets de la double analyse quantitative (section 2). Ensuite, nous présenterons les résultats des analyses statistiques pour le corpus technique entier, ainsi que les interprétations linguistiques qui en découlent (section 3). Dans la section 4, nous commenterons les analyses statistiques par classe lexicale et les interprétations linguistiques respectives. Nous terminerons cet article par les conclusions et les perspectives de recherche (section 5).

2. Une double analyse quantitative

2.1. Quantifier la spécificité des unités spécifiques

Le premier volet de la double analyse quantitative consiste à identifier les unités spécifiques du corpus technique, c'est-à-dire les « spécificités[2] » ou les « mots-clés » (ou *keywords*), et à déterminer leur degré de spécificité. En termes relatifs, les spécificités sont significativement plus fréquentes dans le corpus technique que dans un corpus de référence de langue générale. Afin de relever les spécificités, nous recourons à la méthode des mots-clés (*Keywords Method*) (Bertels, 2005), implémentée dans le logiciel *Abundantia Verborum Frequency List Tool*[3] et basée sur la statistique du LLR (*log likelihood ratio*) (log de vraisemblance) (Dunning, 1993). Cette méthode permet aussi de déterminer le degré de spécificité des unités spécifiques, grâce à la mesure statistique du LLR. Plus une unité linguistique est typique ou spécifique dans le corpus technique par rapport au corpus de référence de langue générale, plus son degré de spécificité sera élevé. Le degré de spécificité permet donc d'ordonner les spécificités, d'accorder un rang de spécificité[4] et de les situer sur un continuum de spécificité.

[2] Nous adoptons le terme « spécificités » pour désigner les mots les plus représentatifs et les plus caractéristiques du corpus technique d'analyse, indépendamment de la méthode utilisée (Calcul des spécificités (Lafon, 1984) versus *Keywords Method*).

[3] Abundantia Verborum : http://wwwling.arts.kuleuven.be/genling/abundant/obtain.htm.

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

Les unités les plus représentatives ou les plus spécifiques sont généralement très fréquentes[5] dans le corpus technique (par exemple *machine, outil, usinage*, *broche*) et elles reflètent clairement la thématique du domaine. Après avoir supprimé les hapax, les mots grammaticaux et les noms propres, nous recensons 4 717 unités lexicales spécifiques dans notre corpus technique.

2.2. Quantifier la monosémie des unités spécifiques

Le deuxième volet de notre double analyse quantitative vise à quantifier l'analyse sémantique, afin de pouvoir déterminer le degré de monosémie des 4 717 spécificités 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). L'accès à la sémantique des cooccurrents d'un mot de base (ou d'une unité lexicale spécifique) se fait 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 (Martinez, 2000). Le degré de ressemblance 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. Nous déterminons le degré de monosémie des 4 717 spécificités à partir du degré de recoupement des cooccurrents de leurs cooccurrents, qui est calculé à partir d'une mesure de recoupement[6] (pour les détails : voir Bertels et al., 2006). Une fois obtenu, le degré de monosémie permet de situer les spécificités sur un continuum d'homogénéité sémantique (ou de monosémie) et d'accorder un rang de monosémie (cf. section 2.1).

3. Analyses pour le corpus technique entier

Comme nous l'avons signalé ci-dessus, le but de notre étude est de vérifier si les unités lexicales les plus spécifiques ou les plus représentatives du corpus technique sont effectivement les plus monosémiques ou les plus homogènes sémantiquement. Il s'agit donc de vérifier s'il existe une corrélation entre le continuum de spécificité et le continuum de monosémie. A cet effet, les données quantitatives de spécificité et de monosémie (cf.

[5] Par contre, les unités les plus fréquentes du corpus technique ne sont pas nécessairement des unités spécifiques (ou mots-clés). Ainsi, les unités grammaticales *de, le, à, pour*, etc. sont très fréquentes dans le corpus technique, mais elles sont également très fréquentes dans le corpus de référence de langue générale. Par conséquent, ces unités ne sont pas significativement plus fréquentes dans le corpus technique.

[6] Mesure permettant de calculer le degré de monosémie ou le degré de recoupement :

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

sections 2.1 et 2.2) sont soumises à une analyse de régression simple, qui permet d'étudier l'impact du rang de spécificité sur le rang de monosémie. Afin de connaître l'impact combiné de plusieurs variables susceptibles d'influer sur le rang de monosémie, nous procédons aussi à une analyse de régression multiple. Nous présenterons ci-dessous les résultats de ces analyses statistiques pour le corpus technique entier, ainsi que les interprétations linguistiques qui en découlent, dans le but de mieux situer et expliquer les interprétations linguistiques des analyses détaillées (cf. section 4).

L'analyse de régression linéaire simple pour les 4 717 spécificités est hautement significative (valeur $p < 2.2e\text{-}16$). La variation du rang de spécificité permet d'expliquer 51,57% (R^2) de la variation du rang de monosémie, pour un coefficient de corrélation Pearson de -0,72. Les résultats statistiques permettent donc de remettre en question la thèse monosémiste traditionnelle, puisqu'ils démontrent une corrélation négative entre le rang de spécificité et le rang de monosémie. Il s'avère que les unités les plus spécifiques ou les plus représentatives du corpus technique ne sont pas les plus monosémiques, mais au contraire, ce sont les plus hétérogènes sémantiquement, par exemple *machine, pièce, tour*. En plus, les unités les moins représentatives ou spécifiques sont les plus monosémiques (*rationnellement, télédiagnostic*), à quelques exceptions près, comme *service* et *objet*. Cependant, la corrélation négative observée entre le rang de spécificité et le rang de monosémie n'est pas tout à fait linéaire et soulève un problème d'hétéroscédasticité. Certaines spécificités sont effectivement plus polysémiques qu'on n'aurait cru en tenant compte de leur rang de spécificité (par exemple *service, objet*). Par contre, d'autres spécificités sont plus monosémiques qu'on n'aurait cru en tenant compte de leur rang de spécificité (par exemple *autocalibrage, hydrauliquement*). Afin de résoudre ce problème d'hétéroscédasticité, nous recourons d'abord aux solutions techniques les plus courantes, à savoir des transformations logarithmiques et polynomiales, une analyse de régression simple pondérée et une analyse de régression non linéaire. Ces solutions techniques permettent de résoudre le problème et elles confirment notre hypothèse que les unités les plus spécifiques du corpus technique ne sont pas les plus monosémiques. Cependant, les solutions techniques s'avèrent difficiles à interpréter du point de vue linguistique.

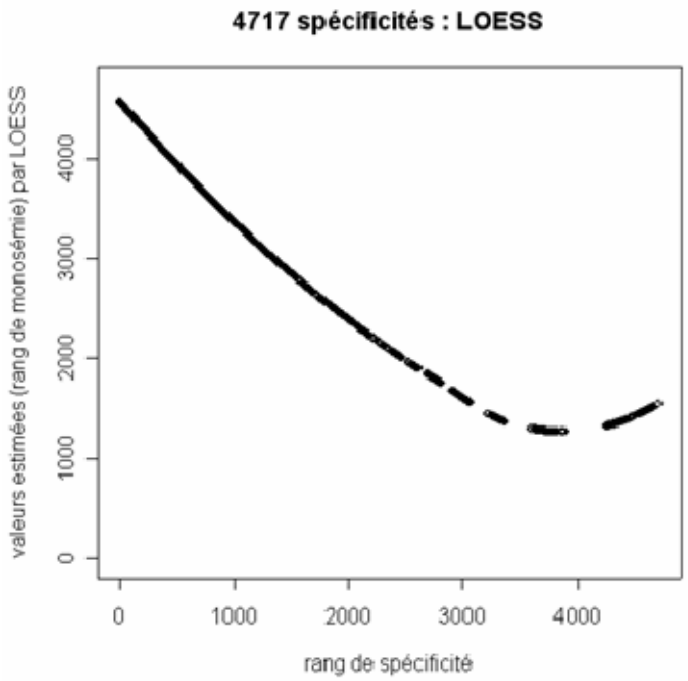

Figure 1 : Visualisation de l'analyse de régression non linéaire : LOESS

La visualisation de la régression non linéaire (LOESS) (cf. figure 1) indique que la tendance négative ne s'applique pas à toutes les spécificités, puisque les mots les moins spécifiques, qui se situent dans la partie inférieure droite de la visualisation, échappent à la tendance générale. Par conséquent, nous optons pour la solution d'exclusion d'un sous-ensemble de mots, dans le but de découvrir l'origine de l'hétéroscédasticité et de trouver une solution opérationnelle et interprétable du point de vue linguistique. Le meilleur critère d'exclusion permettant de résoudre l'hétéroscédasticité pour le sous-ensemble restant est le critère de la fréquence générale. En effet, les mots les plus fréquents dans le corpus de référence de langue générale échappent en partie à la corrélation négative entre le rang de spécificité et le rang de monosémie.

Ces 1 507 mots exclus sont des mots généraux, tels que *service, objet, commercial*, qui se caractérisent par une polysémie à la fois générale et technique : leurs (divers) sens généraux se retrouvent aussi dans le corpus technique. Ces mots produisent un effet perturbateur par rapport à la tendance générale de corrélation négative et échappent à une prédiction de leur rang de monosémie à partir de leur rang de spécificité, du fait qu'ils sont de toutes façons plutôt polysémiques ou hétérogènes sémantiquement, quel que soit leur rang de spécificité. Les 3 210 spécificités restantes sont très peu fréquentes ou même absentes du corpus de langue générale. Elles se caractérisent par l'homoscédasticité et par une bonne corrélation linéaire négative entre le rang de spécificité et le rang de monosémie (R^2 de 60,35% pour un coefficient de corrélation Pearson de -0,78). Parmi ces 3 210 spécificités plutôt techniques, les mots les plus spécifiques et représentatifs sont plutôt hétérogènes sémantiquement, par exemple *usinage* et *broche*. Par contre, les mots les moins spécifiques sont plutôt homogènes sémantiquement, par exemple *adhésif* et *présentoir*. Les résultats de l'analyse de régression

simple pour ces 3 210 spécificités plutôt techniques conduisent donc également à une remise en question quantitative de la thèse monosémiste traditionnelle[7].

Les résultats de l'analyse de régression multiple confirment les résultats de l'analyse de régression simple et apportent des précisions grâce à l'intégration de toutes les variables indépendantes susceptibles d'influencer le rang de monosémie. Les variables indépendantes significatives expliquent 80,65% de la variation du rang de monosémie et la variable indépendante la plus significative est celle du rang de fréquence dans le corpus technique. Les autres variables significatives sont le rang de spécificité, la longueur (en nombre de caractères) et le nombre de classes lexicales. Le rang de fréquence technique et le rang de spécificité ont un rapport de corrélation négative avec le rang de monosémie. Plus les mots sont fréquents dans le corpus technique et plus ils sont spécifiques (représentatifs), moins ils sont sémantiquement homogènes. Cette observation corrobore certainement la conclusion de l'analyse de régression simple pour le rang de spécificité. La longueur se caractérise aussi par un rapport de corrélation négative : plus les mots sont longs, plus ils sont sémantiquement homogènes (rangs de monosémie près de 1). Finalement, nous observons un léger impact du nombre de classes lexicales : un mot qui appartient à plusieurs classes lexicales à la fois est plus hétérogène sémantiquement. Notons à ce sujet que l'appartenance à plusieurs classes lexicales pourrait s'interpréter comme un cas d'homonymie. Toutefois, jusqu'à présent, notre mesure de monosémie ne permet pas encore de faire la distinction entre la polysémie et l'homonymie, ni entre la polysémie et le vague (ou la sous-détermination).

4. Analyses détaillées par classe lexicale

Les analyses de régression simple par classe lexicale permettent d'étudier la corrélation entre le rang de spécificité et le rang de monosémie pour les substantifs, les adjectifs, les verbes et les adverbes. Le but de ces analyses détaillées est de vérifier si les conclusions des analyses pour le corpus technique entier s'appliquent aussi à des sous-ensembles des 4 717 spécificités. Pour les quatre classes lexicales de la liste des 4 717 spécificités, les analyses montrent une corrélation négative entre le rang de spécificité et le rang de monosémie (cf. tableau 1). Par conséquent, les mots les plus spécifiques d'une classe lexicale sont les plus hétérogènes sémantiquement à l'intérieur de cette classe lexicale. Il s'avère que la classe lexicale des substantifs représente le mieux cette corrélation négative (R^2 de 54,75%) et corrobore, dès lors, le mieux le pouvoir explicatif du rang de spécificité. Cette constatation renforce la remise en question de la thèse monosémiste, d'autant plus que les substantifs sont généralement bien représentés dans les textes techniques. Les adverbes illustrent moins bien la corrélation négative (R^2 de 38,31%).

[7] Il convient de vérifier, dans des recherches futures, si et à quel point la « monosémie » des monosémistes traditionnels correspond exactement à notre mesure de monosémie, qui implémente la monosémie en termes d'homogénéité sémantique.

	Coefficient de corr. Pearson	Analyse de régression simple (R^2)	Analyse de régression multiple (R^2) et variables indépendantes significatives[8] par ordre décroissant de pertinence	
4 717 spécif.	-0,72	51,57%	80,65%	rvfq1 ; rvspec ; long ; *nbr_claslex*
Adj. (1 083)	-0,70	49,48%	77,18%	rvfq1 ; *écart* ; *nbr_claslex*
Adv. (141)	-0,62	38,31%	70,13%	rvfq1 ; *log_LLR* ; long
Subst. (2 923)	-0,74	54,75%	81,95%	rvfq1 ; long ; rvspec ; *nbr_claslex*
Verbes (541)	-0,67	45,20%	82,30%	rvfq1 ; rvspec ; long

Tableau 1 : Résultats des analyses statistiques détaillées par classe lexicale

Les résultats des analyses de régression multiple pour les quatre classes lexicales sont convergents : le rang de fréquence dans le corpus technique (rvfq1) est la variable indépendante la plus significative et se caractérise partout par une corrélation négative avec le rang de monosémie. Par conséquent, les spécificités réparties par classe lexicale les plus fréquentes dans le corpus technique sont les moins monosémiques et, dès lors, les plus hétérogènes sémantiquement. Inversement, les spécificités les moins fréquentes dans le corpus technique sont les plus monosémiques, ce qui confirme les observations que nous avons faites à partir des analyses pour les 4 717 spécificités. Les observations concernant la longueur et le nombre de classes lexicales se voient confirmées également, même si ce n'est que pour certaines classes lexicales. En effet, la longueur se caractérise par une corrélation négative avec le rang de monosémie. Les adverbes et les substantifs les plus longs, tels que *perpendiculairement* et *affûteuse-rectifieuse*, sont les plus monosémiques alors que les adverbes et les substantifs les moins longs (*plus, bien, axe, air*) sont les moins monosémiques. La variable du nombre de classes lexicales s'avère significative uniquement pour les adjectifs et pour les substantifs. En effet, ce sont principalement ces deux classes lexicales qui sont impliquées dans les étiquettes à plusieurs classes lexicales. La corrélation positive est confirmée : les adjectifs et les substantifs qui appartiennent en même temps à une autre classe lexicale (respectivement celle des substantifs et des adjectifs) ont des rangs de monosémie plus élevés (près de 4 717) et sont plus hétérogènes sémantiquement. En l'occurrence, ils sont homonymiques, par exemple *technique, manuel, standard*. Le rang de spécificité ou sa variante « log_LLR » figurent aussi parmi les variables significatives, ce qui confirme les résultats des analyses de régression simple.

Avant d'interpréter les résultats des analyses détaillées, nous tenons à rappeler que les analyses statistiques pour le corpus technique entier (cf. section 3) permettent d'isoler un sous-ensemble de 1 507 spécificités fréquentes dans le corpus de référence de langue générale. Elles sont responsables du problème de l'hétéroscédasticité et entraînent un effet perturbateur pour l'ensemble des 4 717 spécificités, dans la mesure où elles échappent à la tendance de corrélation négative entre le rang de spécificité et le rang de monosémie. Dans le

[8] Les variables qui se caractérisent par une corrélation positive sont indiquées en italique. Il est à noter que les variables « écart » (écart entre le rang de fréquence générale et le rang de fréquence technique) et « log_LLR » représentent la spécificité dans les analyses de régression multiple où le problème de multicolinéarité requiert la suppression du rang de spécificité.

but de vérifier si les spécificités les plus générales par classe lexicale ont le même effet perturbateur et dans le but d'interpréter les résultats des analyses détaillées, nous procédons à une explication quantitative et linguistique.

L'explication quantitative, qui apporte une solution au problème de l'hétéroscédasticité pour les 4 717 spécificités, s'applique aussi aux résultats des analyses par classe lexicale. Si un sous-ensemble de spécificités comprend plus de spécificités générales (c'est-à-dire fréquentes dans le corpus de langue générale), celui-ci se prête moins bien à une analyse de régression simple. En effet, pour les unités lexicales les plus générales, le modèle de régression n'est guère satisfaisant, parce qu'il donne lieu à l'hétéroscédasticité et/ou à des pourcentages de variation expliquée R^2 plutôt faibles. Ainsi, les adverbes sont relativement plus fréquents dans le corpus général (la moyenne de leur fréquence générale absolue est plus élevée) et ils sont mieux représentés dans le sous-ensemble perturbant des 1 507 spécificités (5%) que dans la liste entière des 4 717 spécificités (3%). Les substantifs, par contre, sont plus nombreux dans la liste entière (62%) que dans le sous-ensemble perturbant (51%). En plus, ils sont relativement moins fréquents dans le corpus de référence de langue générale.

L'explication quantitative s'accompagne d'une explication linguistique, en termes de caractéristiques syntaxiques et collocationnelles, différentes selon la classe lexicale. En effet, le mécanisme collocationnel des adverbes est moins puissant que celui des substantifs ou des adjectifs, par exemple. Les substantifs sont désambiguïsés par des adjectifs qualificatifs, par des déterminants et par des verbes, avec lesquels ils ont des relations collocationnelles très fortes. Par conséquent, les substantifs ont relativement plus de cooccurrents stables et statistiquement très significatifs. Les adjectifs et les verbes en particulier, forment souvent de vraies collocations avec les substantifs, par exemple *avance technologique* (« progression »), *augmenter l'avance (d'un outil)* (« la vitesse »), *usiner une pièce*. Par contre, le mécanisme désambiguïsateur et collocationnel des adverbes est généralement moins clair : l'applicabilité de l'analyse des cooccurrences est donc plus restreinte pour les adverbes, dans la mesure où ceux-ci ont peu de cooccurrents stables ou statistiquement très significatifs. Le pouvoir désambiguïsateur de leurs cooccurrents, moins nombreux et moins forts, a par conséquent un impact considérable sur les rangs de monosémie et dès lors, sur les résultats de l'analyse de régression simple. Cette observation nous amène à envisager une mise au point de notre mesure de monosémie, basée sur l'homogénéité sémantique. Il serait en effet intéressant d'intégrer des informations syntaxiques ou de considérer une mesure différente par classe lexicale.

5. Conclusions et perspectives

Notre étude a permis d'apporter des réponses quantitatives et linguistiques à des questions sémantiques, grâce à des analyses quantitatives et statistiques. A son tour, elle soulève de nouvelles questions, notamment en ce qui concerne les unités polylexicales et la mesure de monosémie. Les résultats des analyses statistiques ont permis d'ébranler la thèse monosémiste traditionnelle. En effet, plus les unités lexicales sont spécifiques et représentatives dans le corpus technique, plus elles sont hétérogènes sémantiquement. Les analyses statistiques détaillées par classe lexicale ont confirmé cette conclusion.

La poursuite de nos travaux passe inévitablement par les unités polylexicales, étant donné que la plupart des unités lexicales spécifiques d'un corpus technique se situent à ce niveau. Nous aimerions aussi enrichir la mesure de monosémie en y intégrant plus d'informations linguistiques et syntaxiques. Finalement, nous projetons de compléter la mesure de

monosémie par des analyses statistiques multivariées de regroupement (*cluster analysis*), dans le but d'affiner les résultats et les interprétations.

Références

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

Bertels A. (2005). A la découverte de la polysémie des spécificités du français technique. In Hernandez N., Jardino M. and Pitel G., editors, *Actes de TALN-RECITAL 2005 (12e Conférence sur le Traitement Automatique des Langues Naturelles)*, pages 575-584.

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

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

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

Condamines A (editor). (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.

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

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

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

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. *Actes de JADT 2000 (5es Journées internationales d'Analyse statistique des Données Textuelles)*, pages 78-84.

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 (10e Conférence sur le Traitement Automatique des Langues Naturelles)*, pages 265-274.

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