Etudier la sémantique des termes techniques : des théories à la pratique

Ann Bertels

ILT et QLVL, Université de Leuven (Belgique)
ann.bertels@ilt.kuleuven.be

Résumé : L'objectif de cet article est de montrer comment on peut étudier la sémantique des termes techniques et comment on peut passer des théories sémantiques de la terminologie à la pratique, c'est-à-dire à l'analyse de corpus. Nous procédons à une analyse sémantique de plusieurs milliers de mots spécifiques d'un corpus de textes techniques. A cet effet, nous adoptons une méthodologie qui repose sur une double approche quantitative et scalaire, d'une part pour l'identification et l'extraction des mots spécifiques et d'autre part pour leur analyse sémantique. Nous présentons non seulement les résultats de l'analyse globale des mots spécifiques du corpus technique, mais également ceux de l'analyse détaillée de la classe lexicale des substantifs et de quelques sous-groupes de mots spécifiques.

Mots-clés : Terminologie prescriptive, terminologie descriptive, corpus spécialisés, sémantique quantitative, unités simples, unités polylexicales.

1 Introduction

Cet article s'appuie sur les résultats d'une étude effectuée sur un corpus de textes techniques relevant du domaine spécialisé des machines-outils pour l'usinage des métaux. L'étude vise non seulement à analyser la sémantique des mots techniques, mais également et surtout à développer une méthodologie permettant d'analyser simultanément des milliers de mots du corpus technique. Ce corpus de 1,7 million d'occurrences a été étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1996 à 2002 : des revues électroniques, des fiches techniques, des normes ISO et des manuels. Le corpus de référence de langue générale (15,3 millions d'occurrences lemmatisées) est constitué d'articles du journal *Le Monde* (1998).

Lorsqu'on recourt à un corpus technique ou à un corpus d'un domaine spécialisé, il est indispensable de se pencher sur les particularités de la langue spécialisée et dès lors sur les théories sémantiques de la terminologie. La théorie traditionnelle de la terminologie, qui adopte une approche onomasiologique et prescriptive, préconise la monosémie et l'univocité dans la langue spécialisée. De ce fait, les partisans de la terminologie traditionnelle sont parfois qualifiés de « monosémistes ». Les théories descriptives et linguistiques, par contre, plus récentes et basées sur l'analyse de la réalité langagière, remettent en question cet idéal de monosémie, après avoir observé des cas de polysémie dans la langue spécialisée.

Face à ces deux théories sémantiques de la terminologie, nous proposons de passer à la pratique et de procéder à une analyse sémantique à grande échelle. Le but 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[1]. Si l'on veut évaluer la thèse monosémiste de la théorie traditionnelle, en s'appuyant sur un corpus spécialisé et donc en ayant recours aux outils de la linguistique de corpus, on est confronté à plusieurs défis.

Tout d'abord, la thèse monosémiste doit être opérationnalisée et reformulée en une question opérationnelle et mesurable. Le premier défi consiste donc à formuler une question de recherche quantitative. S'il est vrai que les unités lexicales de la langue spécialisée et donc du 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 de notre corpus technique sont effectivement les plus monosémiques. Cette question de recherche quantitative soulève deux autres questions : quelles sont ces unités lexicales spécifiques et comment procéder à leur analyse sémantique ? Par conséquent, le deuxième défi à relever consiste à identifier les unités les plus spécifiques du corpus technique et à déterminer à quel point ces unités sont spécifiques ou représentatives du corpus technique. Ensuite, le troisième défi consiste à quantifier et à automatiser l'analyse sémantique, parce qu'il est impossible d'analyser manuellement tous les contextes d'apparition de toutes les occurrences de plusieurs milliers d'unités spécifiques. Finalement, ces données quantitatives font l'objet d'une analyse statistique de régression, permettant d'étudier leur corrélation. Par conséquent, le quatrième et dernier défi réside dans l'interdisciplinarité. De par son approche, notre étude vise à réconcilier la linguistique (l'analyse sémantique), la terminologie (l'étude du corpus technique), l'informatique (la quantification et l'automatisation de la recherche) et la statistique (l'analyse de régression).

Dans cet article, nous présentons d'abord les principes des théories sémantiques de la terminologie (section 2) et ensuite la méthodologie de notre recherche pratique (section 3). Finalement, nous discutons les résultats de l'analyse statistique de base et de quelques sous-groupes intéressants du point de vue terminologique (section 4). Nous expliquons tant les résultats quantitatifs et statistiques que les interprétations linguistiques.

2 Théories sémantiques de la terminologie

Différentes approches et théories sémantiques ont marqué l'histoire de la terminologie. Ce que l'on appelle généralement la « Théorie classique de la terminologie » ou la « Théorie traditionnelle de la terminologie » et dès lors la

[1] Il est à noter que quelques dictionnaires spécialisés font bel et bien état d'une pluralité de sens pour certains termes spécialisés. Notre étude confirme donc, au travers d'une analyse de corpus, les attestations de polysémie dans les dictionnaires. Nous avons procédé à une approche méthodologique quantitative et scalaire approfondie, puisque notre étude a été réalisée dans le cadre d'une thèse de doctorat.

« terminologie traditionnelle », renvoie à la Théorie générale de terminologie (TGT), conçue par Eugen Wüster dans les années 1930. Ingénieur autrichien et spécialiste des vocabulaires spécialisés, il est préoccupé par la précision de la communication spécialisée. La terminologie traditionnelle wüsterienne se caractérise par une approche prescriptive, conceptuelle et onomasiologique : elle part de l'identification et de l'établissement des concepts dans un champ de connaissances particulier pour en fixer les dénominations standardisées correspondantes. Elle privilégie le principe de la bi-univocité (*Eineindeutigkeit*) : chaque concept est désigné par un terme et chaque terme dénomme un concept (Wüster, 1931 et 1991). La bi-univocité implique que la polysémie, l'homonymie et la synonymie sont évitées ou limitées. En conclusion, la terminologie traditionnelle préconise pour les termes de la langue spécialisée la monoréférentialité et la monosémie. La polysémie est réservée aux mots de la langue générale. La terminologie traditionnelle adopte donc une approche dichotomique : elle oppose les termes (de la langue spécialisée) aux mots (de la langue générale) comme elle oppose la monosémie à la polysémie.

Récemment, c'est-à-dire depuis l'essor de la linguistique de corpus et depuis la constitution de corpus électroniques spécialisés, les principes de la terminologie traditionnelle ont fait l'objet d'une révision fondamentale. L'approche prescriptive, conceptuelle et onomasiologique est remise en question par les adeptes d'une approche descriptive, linguistique et sémasiologique, basée sur l'étude de textes spécialisés. Citons à titre d'exemple la Théorie Communicative de la Terminologie (Cabré, 2000) et l'approche linguistique de Condamines & Rebeyrolle (1997), la Socioterminologie (Gaudin, 2003) et son approche discursive, la Terminologie socio-cognitive (Temmerman, 2000), basée sur la théorie du prototype, et l'approche textuelle de Bourigault & Slodzian (1999). Kocourek (1991) soutient également l'idée d'une approche descriptiviste et textuelle permettant l'étude du contexte (linguistique) des termes. Les partisans de la terminologie descriptive remettent en question l'idéal de monosémie dans la langue spécialisée, ainsi que la double dichotomie. En raison des phénomènes de (dé)terminologisation et de nomadisation, les termes circulent et voyagent d'un domaine à un autre. Ils s'enrichissent et s'appauvrissent, tout en gardant un noyau de sens commun (Delavigne & Bouveret, 1999).

Des expérimentations récentes menées sur des corpus spécialisés, dans la perspective distributionnelle et contextuelle de la terminologie descriptive, ont abouti à l'observation de cas de polysémie dans la langue spécialisée, même à l'intérieur d'un domaine spécialisé (Arnzt & Picht, 1989 ; Condamines & Rebeyrolle, 1997 ; Temmerman, 2000 ; Eriksen, 2002 ; Ferrari, 2002). Ces travaux antérieurs étudient, comme nous, la polysémie dans un corpus spécialisé, mais se limitent à l'analyse de quelques mots seulement. Ainsi, Condamines & 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. 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. Malgré leur champ d'étude limité, ces

études sémantiques ponctuelles fournissent des indications concrètes sur la présence de polysémie dans la langue spécialisée. Les résultats convaincants de ces études et les récentes remises en question théoriques nous incitent à effectuer une étude sémantique à grande échelle dans un corpus d'un domaine spécialisé.

3 Passer à la pratique : une double analyse quantitative

Comme nous l'avons annoncé ci-dessus, la question de recherche principale est celle de savoir si les unités lexicales les plus spécifiques de notre corpus technique sont effectivement les plus monosémiques. Cette question constitue non seulement une solution alternative à l'approche dichotomique de la terminologie traditionnelle, elle permet en plus de quantifier l'analyse et donc de l'automatiser. Nous adoptons une approche scalaire pour étudier la sémantique des unités lexicales, classées des plus spécifiques aux moins spécifiques. A présent, nous nous limitons aux unités simples[2] ou monolexicales (p.ex. *usinage, découpe*) ; les unités polylexicales (p.ex. *machine à fraiser*) feront l'objet d'une étude ultérieure[3]. La question de recherche requiert une double analyse quantitative et automatisée, qui mène à une analyse statistique et qui permet non seulement d'étudier un nombre important d'unités lexicales du corpus technique, mais aussi d'aboutir à des résultats quantitatifs et objectifs.

3.1 Déterminer les unités spécifiques et leur degré de spécificité

Dans un premier temps, nous identifions toutes les unités spécifiques du corpus technique ou « mots-clés » (*keywords*), et nous déterminons leur degré de spécificité. Les unités spécifiques ne sont pas les unités les plus fréquentes du corpus technique, mais ce sont les unités les plus représentatives. En termes relatifs, elles sont significativement plus fréquentes dans le corpus technique que dans le corpus de référence de langue générale. Pour identifier et extraire unités spécifiques simples, nous recourons à la méthode des mots-clés (*Keywords Method*)[4] (Bertels, 2005), implémentée dans le logiciel *Abundantia Verborum Frequency List Tool*[5] ou dans

---

[2] Nous considérons comme unités simples tous les lemmes (ou formes canoniques) des unités typographiques, telles qu'elles sont identifiées par l'analyseur Cordial, donc non seulement les unités linguistiques séparées par des blancs, mais aussi les unités linguistiques avec trait d'union ou apostrophe (par exemple *machine-outil*).

[3] Même s'il existe des outils d'extraction terminologique qui permettent de repérer les unités polylexicales (Bourigault et al., 2001), ces unités complexes posent problème lors du calcul des spécificités. Pour l'instant, il n'est guère possible de déterminer le degré de spécificité des unités complexes de façon fiable et statistiquement significative.

[4] On peut aussi envisager de recourir au calcul des spécificités (Lafon, 1984), implémenté notamment dans le logiciel Lexico3 et basé sur la distribution hypergéométrique. Les deux méthodes aboutissent grosso modo à des résultats similaires, à savoir une liste de mots spécifiques pourvus d'une mesure statistique indiquant le degré de spécificité. Les différences les plus importantes résident dans la méthodologie et la statistique sous-jacentes.

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

WordSmith[6] et basée sur la mesure statistique du LLR[7] (*log likelihood ratio*) (log de vraisemblance) (Dunning, 1993). La mesure statistique peut être considérée comme le degré de spécificité des unités spécifiques, parce qu'elle indique à quel point ces unités sont spécifiques du corpus technique par rapport au corpus de référence. Plus une unité est spécifique ou représentative, plus son degré de spécificité sera élevé. Le degré de spécificité permet de classer les unités spécifiques, d'accorder un rang de spécificité et de les situer sur un continuum de spécificité. Les unités les plus spécifiques sont généralement très fréquentes[8] dans le corpus technique (par exemple *machine, outil, usinage*) et elles reflètent clairement la thématique du domaine. Après avoir supprimé les hapax, les mots grammaticaux et les noms propres, nous recensons 4717 unités lexicales spécifiques dans notre corpus technique.

3.2 Quantifier et automatiser l'analyse sémantique

Dans un deuxième temps, les 4717 unités lexicales spécifiques font l'objet d'une analyse sémantique. Un corpus électronique de textes spécialisés offre une information indispensable pour l'analyse sémantique, à savoir le contexte linguistique, mais l'exploitation efficace de grandes quantités de textes requiert une approche quantitative et automatisée.

Pour déterminer à quel point les unités lexicales spécifiques sont monosémiques, nous recourons à l'analyse des cooccurrences (Grossmann & Tutin, 2003 ; Condamines, 2005 ; Blumenthal & 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 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 cooccurrents de premier

---

[6] WordSmith : http://www.lexically.net/wordsmith/

[7] Il est à noter qu'il existe d'autres mesures pour tester la dépendance de deux variables (p.ex. chi-carré, information mutuelle, score Z). Toutefois, les estimations de l'information mutuelle, qui sont basées directement sur les fréquences, ont tendance à surestimer la significativité des mots de faible fréquence. Les valeurs du test du chi-carré (χ^2) de Pearson et celles du score Z (basées sur la distribution normale) ne sont pas fiables pour des fréquences attendues inférieures à 5 ou même à 10 (Müller, 1992a ; Rayson & Garside, 2000). Pour remédier aux problèmes de l'analyse des mots peu fréquents dans les corpus linguistiques, Dunning (1993) propose la mesure statistique du log de vraisemblance (*log likelihood ratio* ou LLR). Celle-ci permet la comparaison directe de la significativité de mots plus fréquents et de mots moins fréquents en raison de son meilleur comportement asymptotique (approximatif). Par conséquent, la significativité des mots rares est plus fiable. Il convient de signaler tout de même que la mesure du LLR s'avère sensible à la fréquence lors de l'analyse de mots extrêmement fréquents.

[8] 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 comme elles sont également très fréquentes dans le corpus de référence, ces unités ne sont pas « spécifiques ».

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 ou d'homogénéité sémantique des 4717 unités lexicales spécifiques en fonction du degré de recoupement des cooccurrents de leurs cooccurrents. Celui-ci est calculé à partir d'une mesure de recoupement ou de monosémie (1). La formule est basée sur le recoupement formel des cooccurrents des cooccurrents (*cc*), tenant compte de la fréquence d'un *cc* dans la liste des *cc* (fq *cc*), du nombre total de *c* et du nombre total de *cc*. Un *cc* sera d'autant plus important pour le recoupement 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 (ou *c*).

$$\sum_{cc} \frac{\text{fq } cc}{\text{nbr total } c \cdot \text{nbr total } cc} \qquad (1)$$

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 de la formule le nombre de *c* qui ont ce *cc* en commun (fq *cc*), par exemple 5, et d'inclure dans le dénominateur le nombre total de *c*, par exemple 7. Le recoupement est donc exprimé par la fraction 5/7. En exprimant pour chaque *cc* le recoupement par la fraction *nombre de c avec le cc (*ou fq *cc)* divisé par *nombre total de c*, le résultat se situe toujours entre 0 (pas ou peu de recoupement) et 1 (recoupement important ou parfait) et sera dès lors facilement interprétable. Nous considérons les *c* et *cc* au niveau des formes graphiques et non pas au niveau des lemmes (formes canoniques), ce qui permet de faire la distinction entre, par exemple, *pièce usinée* et *pièce à usiner*.

La mesure respecte un seuil de significativité très sévère (à savoir une valeur $p<0,0001$), afin de ne relever que les cooccurrents les plus fortement associés et donc sémantiquement les plus pertinents. Le résultat de l'analyse sémantique quantitative, à savoir le degré de recoupement ou le degré d'homogénéité sémantique, permet de situer les 4717 unités lexicales spécifiques sur un continuum d'homogénéité sémantique (ou de monosémie) et de leur accorder un rang de monosémie, à l'instar du rang de spécificité (Cf. section 3.1).

Il est à noter que des recherches supplémentaires s'imposent pour examiner la relation précise entre notre mesure de monosémie, implémentant la monosémie comme homogénéité sémantique, et ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure dans le but de développer un critère opérationnalisable et mesurable. Sans recherche supplémentaire, il serait impossible d'affirmer que notre mesure de monosémie et les degrés de monosémie calculés correspondent parfaitement à ce que les terminologues traditionnels considèrent comme monosémie ou polysémie.

4 Discussion des résultats

4.1 Analyse des 4717 unités lexicales spécifiques

Rappelons que le but de notre étude est de vérifier si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques, c'est-à-dire les plus homogènes sémantiquement. Cela revient donc à vérifier s'il existe une corrélation entre le continuum de spécificité et le continuum de monosémie (ou d'homogénéité sémantique). Pour y arriver, nous soumettons les données quantitatives de spécificité et d'homogénéité sémantique à une analyse statistique de régression simple. Celle-ci permet d'étudier l'impact d'une variable indépendante ou explicative (ici : le rang de spécificité) sur la variable dépendante ou expliquée (ici : le rang de monosémie). Le résultat de cette analyse est un pourcentage de variation expliquée R^2, qui 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 d'un ensemble de données, en l'occurrence la liste des 4717 unités spécifiques. Le résultat comprend aussi une valeur p, indiquant la significativité statistique et donc la fiabilité de la capacité prédictive du modèle.

Les résultats statistiques permettent d'infirmer la thèse monosémiste traditionnelle, parce qu'ils montrent une corrélation négative (R^2 de 51,57% et coefficient de corrélation Pearson de -0,72). Il s'avère donc que les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais, au contraire, les plus hétérogènes sémantiquement (par exemple *machine, pièce, tour*). En plus, les unités lexicales les moins spécifiques du corpus technique sont les plus homogènes sémantiquement (par exemple *rationnellement, télédiagnostic*)[9], à quelques exceptions près (*service* et *objet*). La visualisation ci-dessous (Cf. figure 1) confirme les résultats statistiques, puisque la droite de régression s'incline vers le bas. Parmi les unités lexicales hétérogènes sémantiquement, nous retrouvons effectivement des unités polysémiques, telles que *découpe*, dont les sens « action de découper » et « résultat de la découpe » se caractérisent par une relation métonymique. Nous recensons également des homonymes (*tour*) et des mots vagues (comme *usinage*), dont le sens sous-déterminé est précisé par le contexte linguistique.

Les résultats quantitatifs de l'analyse de régression et leur visualisation permettent donc d'infirmer la thèse monosémiste traditionnelle et de confirmer les observations des études de corpus antérieures (Cf. section 2), limitées à l'analyse sémantique de quelques mots seulement. Or, pour interpréter correctement les résultats de l'analyse statistique, il est indispensable de tenir compte des particularités de la mesure sous-jacente et de considérer la monosémie en termes d'homogénéité sémantique.

[9] Il convient de signaler qu'une analyse de régression multiple a permis d'observer une corrélation négative (statistiquement significative) entre la longueur des unités lexicales spécifiques (simples), mesurée en nombre de caractères, et leur rang de monosémie. Plus les mots sont longs, plus ils sont monosémiques ou homogènes sémantiquement; plus les mots sont courts, plus ils sont polysémiques ou hétérogènes sémantiquement.

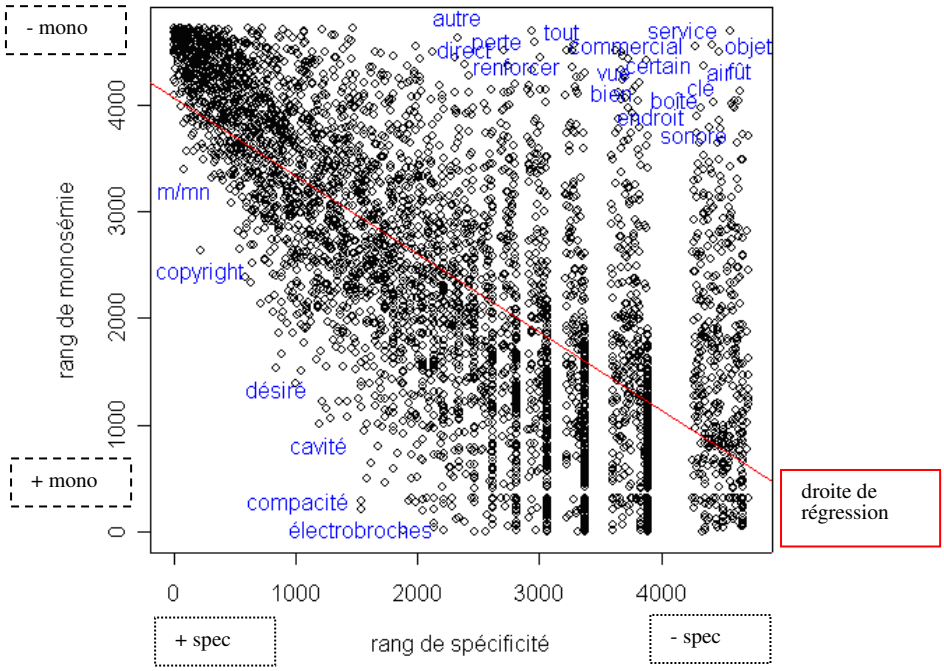

Fig. 1 – Visualisation de l'analyse de régression

La visualisation ci-dessus montre que la corrélation négative n'est pas tout à fait linéaire et qu'il y a un problème d'hétéroscédasticité. En effet, la corrélation négative ne s'applique pas à toutes les unités spécifiques, puisque certaines unités se situent très loin de la droite de régression des valeurs estimées. Elles se situent dans la partie supérieure droite, c'est-à-dire parmi les unités les moins spécifiques. Ces unités sont plus polysémiques qu'on n'aurait cru en tenant compte de leur rang de spécificité (par exemple *service, objet*). Ce sont majoritairement des mots généraux, très fréquents dans le corpus de référence de langue générale et dès lors peu spécifiques dans le corpus technique, en dépit de leur fréquence élevée dans le corpus technique. Ces mots sont hétérogènes sémantiquement et se caractérisent par une polysémie à la fois générale et technique : leurs (divers) sens généraux se retrouvent aussi dans le corpus technique. Ils 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.

4.2 Analyse détaillée de quelques sous-ensembles

Dans le but d'approfondir les résultats de l'analyse précédente, nous procédons à l'analyse de quelques sous-ensembles des 4717 unités spécifiques. Nous nous intéressons d'abord à la classe lexicale des substantifs. Ensuite, nous analysons le sous-ensemble des substantifs déverbaux, ainsi que celui des mots avec trait d'union.

L'analyse de régression pour les 2923 substantifs spécifiques (62% des 4717 unités lexicales spécifiques) montre également une corrélation négative entre le rang de spécificité de ces substantifs et leur rang de monosémie (R^2 de 54,75%). Cela signifie que les substantifs les plus spécifiques du corpus technique sont les plus hétérogènes sémantiquement. Compte tenu de la surabondance des substantifs dans les textes spécialisés (L'Homme 2004 ; Kocourek 1991), cette constatation confirme les résultats de l'analyse précédente et renforce la remise en question de la thèse monosémiste. Rappelons que dans l'analyse précédente, nous avions identifié un sous-ensemble d'unités fréquentes dans le corpus de référence, qui entraînaient un effet perturbateur pour l'ensemble des 4717 unités spécifiques, dans la mesure où elles échappaient à la tendance de corrélation négative. Afin de vérifier si les substantifs les plus généraux ont le même effet perturbateur et dans le but d'interpréter les résultats, nous procédons à une explication quantitative et linguistique.

D'abord, nous constatons que les substantifs sont plus nombreux dans la liste entière des 4717 unités spécifiques (62%) que dans le sous-ensemble perturbant (51%). En plus, les substantifs sont relativement moins fréquents dans le corpus de référence de langue générale ; la moyenne de leur fréquence générale absolue est moins élevée. A titre de comparaison, les adverbes (lexicaux) se caractérisent par un R^2 plus faible de 38,31%, donc ils se prêtent moins bien à la corrélation négative. Les adverbes sont relativement plus fréquents dans le corpus général et ils sont mieux représentés dans le sous-ensemble perturbant (5%) que dans la liste entière (3%). L'explication quantitative s'accompagne d'une explication linguistique en termes de caractéristiques syntaxiques et collocationnelles, différentes selon la classe lexicale, parce que le mécanisme collocationnel des adverbes est moins puissant que celui des substantifs. Les substantifs sont principalement désambiguïsés par des adjectifs qualificatifs, avec lesquels ils ont des relations collocationnelles très fortes. Par conséquent, ils ont relativement plus de cooccurrents stables et statistiquement très significatifs. Les adjectifs et les verbes forment souvent des collocations avec les substantifs, par exemple *avance technologique* (« progression »), *augmenter l'avance (d'un outil)* (« la vitesse »). Par contre, le mécanisme collocationnel des adverbes est généralement moins clair. L'applicabilité de la mesure de recoupement, basée sur l'analyse des cooccurrences, est donc plus restreinte pour les adverbes.

Passons finalement à l'analyse de quelques sous-ensembles d'unités lexicales spécifiques. Il se trouve que les substantifs déverbaux (en *–ion*, en *–age* et en *–ment*) affichent de meilleurs résultats (R^2 de 59%) que le groupe entier des substantifs spécifiques. Ils sont en moyenne moins fréquents dans le corpus général que les substantifs spécifiques pris ensemble. Cela confirme la conclusion formulée ci-dessus: un sous-ensemble d'unités spécifiques qui comprend moins d'unités fréquentes dans le corpus général, corrobore mieux le pouvoir explicatif du modèle de régression.

Nous avons aussi procédé à une analyse détaillée pour les unités spécifiques avec trait d'union (-) ou barre oblique (/), catégorisées par Cordial comme une seule unité lexicale, par exemple *t/min*. Ce sous-ensemble de 429 « mots composés » comprend presque seulement des mots absents du corpus de référence, à quelques exceptions près (*sous-traitance, technico-commercial*). La corrélation négative entre le rang de spécificité et le rang de monosémie (rangs de 1 à 4717) se maintient pour ces mots

très typiques de la langue spécialisée (R^2 de 61%). Toutefois, à l'intérieur du sous-ensemble des mots composés (rangs de 1 à 429), on observe une chute importante du pourcentage de variation expliquée (R^2 de 47%). Bien que la corrélation négative soit statistiquement significative et donc fiable, le nouveau rang de spécificité (de 1 à 429) permet moins bien de prédire le rang de monosémie. Ces mots avec trait d'union ou barre oblique sont apparentés aux unités polylexicales, en raison de leur caractère composé, qui facilite une certaine désambiguïsation. Par conséquent, nous pourrions déjà avancer l'hypothèse que les unités polylexicales se prêteront moins bien à une corrélation négative entre le rang de monosémie et le rang de spécificité. Des recherches futures permettront de le vérifier, ou non, fondées sur de nouvelles analyses statistiques de régression.

5 Conclusion

En s'appuyant sur l'analyse quantitative et automatisée d'un corpus technique, notre étude sémantique a permis d'ébranler la thèse monosémiste traditionnelle. En effet, plus les unités lexicales sont spécifiques dans le corpus technique, plus elles sont hétérogènes sémantiquement. La poursuite de nos travaux passe inévitablement par les unités polylexicales, qui constituent une partie importante des unités lexicales d'un corpus spécialisé. A ce sujet, nous envisageons de dissocier les deux composants des unités du dernier sous-ensemble discuté ci-dessus et de considérer le premier composant comme « mot de base » et le deuxième composant comme cooccurrent, dont les c seront considérés comme cc du mot de base. Cette dissociation nous permettra de vérifier l'effet de notre mesure de recoupement pour les mots composés avec trait d'union ou barre oblique et de faire ainsi un premier pas vers l'analyse des unités polylexicales.

Références

ARNTZ R. & 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 *Actes de TALN-RECITAL 2005.* p. 575-584.

BLUMENTHAL P. & HAUSMANN F.J. EDS (2006). Collocations, corpus, dictionnaires. *Langue française* 150.

BOURIGAULT D, JACQUEMIN C. & L'HOMME M.-C. (2001). *Recent advances in computational terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

BOURIGAULT D. & SLODZIAN M. (1999). Pour une terminologie textuelle. *Terminologies Nouvelles.* 19, p. 29-32.

CABRE M.T. (2000). Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles.* 2, p. 10-15.

CONDAMINES A. & REBEYROLLE J. (1997). Point de vue en langue spécialisée. *Meta*. 42-1, p. 174-184.

CONDAMINES A. ED. (2005). *Sémantique et corpus*. Paris : Hermes-Science.

DELAVIGNE V. & BOUVERET M. (1999). *Sémantique des termes spécialisés*. Rouen : Publications de l'Université de Rouen.

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

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

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

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

GROSSMANN F. & TUTIN A. EDS. (2003). Les collocations, analyse et traitement. *Travaux et Recherches en linguistique appliquée, Série E,* 1.

HABERT B., ILLOUZ G. & FOLCH H. (2004). Dégrouper les sens : pourquoi ? comment ? In *Actes de JADT 2004.* p. 565-576.

HABERT B., ILLOUZ G. & FOLCH H. (2005). Des décalages de distribution aux divergences d'acception. In A. CONDAMINES ED. *Sémantique et corpus*. p. 277-318. Paris.

KOCOUREK R. (1991). *La langue française de la technique et de la science*. Wiesbaden : Brandstetter Verlag.

L'HOMME M.C. (2004). *La terminologie : principes et techniques.* Montréal : Les presses de l'Université de Montréal.

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.* p. 78-84.

MÜLLER C. (1992). *Initiation aux méthodes de la statistique linguistique* (réimp. de l'édition de 1968). Paris : Champion.

RAYSON P. & GARSIDE R. (2000). Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)* : 1-6.

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.* p. 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.