

Etude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique

Ann Bertels^{1,2} Dirk De Hertog² Kris Heylen²

(1) ILT, KU Leuven, Dekenstraat 6, B-3000 Leuven (Belgique)

(2) QLVL, KU Leuven, Faculty of Arts, Blijde-Inkomststraat 21, B-3000 Leuven (Belgique)

ann.bertels@ilt.kuleuven.be, dirk.dehertog@arts.kuleuven.be,

kris.heylen@arts.kuleuven.be

RESUME

Cet article présente les résultats d'une analyse sémantique quantitative des unités lexicales spécifiques dans un corpus technique, relevant du domaine des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les mots-clés du corpus technique sont monosémiques. A cet effet, nous procédons à une analyse statistique de régression simple, qui permet d'étudier la corrélation entre le rang de spécificité des mots-clés et leur rang de monosémie, mais qui soulève des problèmes statistiques et méthodologiques, notamment un biais de fréquence. Pour y remédier, nous adoptons une approche alternative pour le repérage des unités lexicales spécifiques, à savoir l'analyse des marqueurs lexicaux stables ou *Stable Lexical Marker Analysis* (SLMA). Nous discutons les résultats quantitatifs et statistiques de cette approche dans la perspective de la corrélation entre le rang de spécificité et le rang de monosémie.

ABSTRACT

Semantic analysis of keywords and stable lexical markers in a technical corpus

This article presents the results of a quantitative semantic analysis of typical lexical units in a specialised technical corpus of metalworking machinery in French. The study aims to find out whether and to what extent the keywords of the technical corpus are monosemous. A simple regression analysis, used to examine the correlation between typicality rank and monosemy rank of the keywords, points out some statistical and methodological problems, notably a frequency bias. In order to overcome these problems, we adopt an alternative approach for the identification of typical lexical units, called *Stable Lexical Marker Analysis* (SLMA). We discuss the quantitative and statistical results of this approach with respect to the correlation between typicality rank and monosemy rank.

MOTS-CLES : unités lexicales spécifiques, analyse des mots-clés, analyse des marqueurs lexicaux stables, sémantique quantitative, analyse de régression.

KEYWORDS : typical lexical units, Keyword Analysis, Stable Lexical Marker Analysis (SLMA), quantitative semantics, regression analysis.

1 Introduction

Cette communication s'inscrit dans le contexte d'une étude sémantique quantitative effectuée sur un corpus de textes techniques relevant du domaine technique spécialisé des machines-outils pour l'usinage des métaux. L'étude vise à vérifier si et dans quelle mesure les unités lexicales spécifiques du corpus technique sont monosémiques. Selon la

Terminologie traditionnelle, qui adopte une approche onomasiologique et prescriptive, la langue spécialisée se caractérise par la monosémie et l'univocité (Wüster, 1991). Les termes de la langue spécialisée sont idéalement monosémiques, tandis que la polysémie est réservée aux mots de la langue générale. Cela aboutit à une double dichotomie, qui oppose les termes de la langue spécialisée aux mots de la langue générale et la monosémie à la polysémie. Récemment, l'idéal de monosémie dans la langue spécialisée ainsi que la double dichotomie ont été remis en question par les partisans de la terminologie descriptive et linguistique, sémasiologique, distributionnelle et (con)textuelle (Bourigault et Slodzian, 1999 ; Cabré, 2000 ; Temmerman, 2000 ; Gaudin, 2003). Par ailleurs, des expérimentations ponctuelles menées sur des corpus spécialisés ont abouti à l'observation de cas de polysémie dans la langue spécialisée, même à l'intérieur d'un domaine spécialisé (Condamines et Rebeyrolle, 1997 ; Eriksen, 2002 ; Ferrari, 2002). Ces études sémantiques ponctuelles et qualitatives ainsi que les remises en question théoriques nous ont incités à évaluer la thèse monosémiste traditionnelle à grande échelle, dans un corpus de textes techniques, et à adopter une approche alternative, quantitative et scalaire.

Afin de procéder à une étude sémantique à grande échelle, à partir d'une analyse de corpus, il est nécessaire de reformuler la thèse monosémiste traditionnelle qualitative en une question de recherche quantitative. S'il est vrai que les unités lexicales d'un corpus spécialisé sont monosémiques, ce sera d'autant plus vrai pour les unités lexicales les plus spécifiques et les plus représentatives de ce corpus. Nous examinons donc si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques, comme le prétendent les partisans de la terminologie traditionnelle, ou s'il existe des unités lexicales spécifiques qui sont polysémiques, comme le suggèrent les partisans de la terminologie descriptive. A cet effet, nous procédons à une double analyse quantitative. Pour l'extraction des unités lexicales spécifiques et pour le calcul de leur degré de spécificité, nous recourons à l'analyse des mots-clés (*Keyword Analysis*) (Scott, 2006). Pour quantifier l'analyse sémantique, nous calculons le degré de monosémie des mots-clés à partir du degré de recoupement de leurs cooccurrents de deuxième ordre, par le biais d'une mesure de recoupement (Bertels *et al.*, 2010). Ces données quantitatives de spécificité et de monosémie mènent ensuite à une analyse statistique de régression simple. Elle consiste à étudier la corrélation entre le rang de spécificité et le rang de monosémie, pour ainsi répondre à la question de recherche quantitative. Les résultats de l'analyse statistique confirment les observations des études sémantiques antérieures et permettent de réfuter la thèse monosémiste traditionnelle, comme nous l'avons décrit précédemment (Bertels *et al.*, 2010). Dans cet article, nous discutons la pertinence de l'analyse des mots-clés pour l'identification des unités spécifiques et nous proposons une approche méthodologique alternative.

Il est à noter que cet article ne se situe pas dans le domaine de l'extraction de termes. Cette discipline se caractérise par un classement catégoriel (termes vs non-termes), alors que nous visons à étudier la spécificité dans une perspective graduelle. Nous n'avons pas l'intention d'extraire la terminologie du domaine technique des machines-outils pour l'usinage des métaux. Nous cherchons avant tout à répondre à notre question de recherche, soulevée par le corpus de langue spécialisée, qui consiste à étudier la corrélation entre la spécificité et la monosémie.

Dans cet article, nous expliquons d'abord la méthodologie et les résultats de l'étude sémantique des mots-clés du corpus technique (section 2), ainsi que les problèmes statistiques et méthodologiques (section 3). Ensuite, nous présentons une autre approche pour l'identification des unités lexicales spécifiques et pour le calcul de leur degré de spécificité, à savoir l'analyse des marqueurs lexicaux stables ou *Stable Lexical Marker Analysis* (section 4). Nous discutons finalement les résultats quantitatifs et statistiques de cette approche dans la perspective de la corrélation entre le rang de spécificité et le rang de monosémie (section 5).

2 Etude sémantique des mots-clés du corpus technique

Le corpus technique constitué dans le cadre de cette étude relève du domaine spécialisé des machines-outils pour l'usinage des métaux et il comprend 1,7 million d'occurrences. Il a été étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1996 à 2002, à savoir des revues électroniques (800.000 occurrences), des fiches techniques (300.000), des normes ISO et directives (300.000) et 4 manuels numérisés (360.000). Le corpus de référence de langue générale compte 15,3 millions d'occurrences lemmatisées et il est constitué d'articles du journal *Le Monde* de la même période (1998).

2.1 Identification des unités lexicales spécifiques

Pour le repérage des unités lexicales spécifiques, plusieurs approches méthodologiques sont envisageables. Elles permettent de générer une liste d'unités spécifiques, pourvues d'une indication de leur degré de spécificité. Les différences les plus importantes résident dans la méthodologie et les mesures statistiques sous-jacentes. La méthodologie du calcul des spécificités (Lafon, 1984 ; Labbé et Labbé, 2001) est basée sur la distribution hypergéométrique et sur le test statistique de Fisher Exact¹. Elle est implémentée notamment dans Lexico3², Hyperbase³ et TermoStat⁴. Du point de vue méthodologique, le calcul des spécificités procède par comparaison partie-tout. Une section d'un corpus est comparée au corpus entier dans le but d'identifier les mots spécifiques de la section par rapport au corpus entier. Le calcul des spécificités utilise seulement des informations appartenant au domaine en question. Le résultat est un coefficient de spécificité. Plus élevé ce coefficient, plus faible sera la probabilité de la fréquence observée (par rapport au corpus entier) et plus spécifique sera le mot. L'analyse des mots-clés (*Keyword Analysis*) (Scott, 2006) se caractérise par une approche contrastive, qui consiste à comparer les fréquences relatives des mots dans un corpus spécialisé à celles dans un corpus de référence de langue générale. Un mot est « clé » ou spécifique dans le corpus spécialisé si sa fréquence relative dans ce corpus est plus élevée que sa fréquence relative

¹ Le test statistique de Fisher Exact est généralement utilisé pour des données de taille modeste, des corpus peu volumineux et des fréquences plutôt faibles ($n < 20$).

² SYLED – CLAZT, Paris3 : <http://www.tal.univ-paris3.fr/lexico/>. [consulté le 20/01/2012].

³ Hyperbase : <http://ancilla.unice.fr/~brunet/pub/hyperbase.html>. [consulté le 20/01/2012].

⁴ TermoStat : http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html. [consulté le 20/01/2012]. Dans TermoStat le corpus de référence et le corpus d'analyse sont fusionnés en un corpus virtuel, pour vérifier si le lexique du corpus d'analyse se comporte comme celui du corpus de référence.

dans le corpus de référence et si la différence de fréquence est statistiquement significative. Une mesure statistique, comme le rapport de vraisemblance (*Log-Likelihood Ratio* ou LLR ou G^2) (Dunning, 1993), permet de décider s'il s'agit d'un mot-clé du corpus spécialisé. L'analyse des mots-clés est implémentée notamment dans WordSmith⁵, AntConc⁶, TermoStat et AV Frequency List Tool⁷.

Pour identifier le vocabulaire spécifique de notre corpus technique, nous adoptons l'analyse des mots-clés, parce qu'elle compare deux corpus différents. Nous recourons à la mesure statistique du log de vraisemblance (LLR), qui permet de conduire à des possibilités de classement précis et donc à des degrés de spécificité avec une granularité aussi fine que possible. Plusieurs études antérieures ont validé la mesure du LLR et démontré la pertinence de l'analyse des mots-clés pour relever les unités spécifiques d'un domaine particulier (Paquot *et al.*, 2009 ; Kwary, 2011). Nous nous limitons dans cette étude au niveau des unités simples, comme *fraisage*, *commande*, *machine*. Des recherches futures devront certainement porter sur l'étude des unités polylexicales, telles que *machine à fraiser*, *commande numérique*, puisque la plupart des unités terminologiques d'un domaine spécialisé se situent au niveau des unités complexes⁸. L'analyse des mots-clés est effectuée dans AV Frequency List Tool, à partir de deux listes de fréquence des lemmes des deux corpus, réalisées à l'aide de scripts en Python. Le logiciel AV génère une liste de lemmes spécifiques du corpus technique et indique leur degré de spécificité, à savoir la valeur du LLR (*keyness*), et une valeur p associée. Nous relevons 4717 mots-clés ($p < 0,05$), après suppression des mots grammaticaux, des noms propres et des hapax. Le degré de spécificité ou de *keyness* permet de situer ces 4717 mots-clés sur un continuum de spécificité et de leur accorder un rang de spécificité.

2.2 Quantification de l'analyse sémantique

Les 4717 mots-clés font ensuite l'objet d'une analyse sémantique quantitative et automatisée. A cet effet, nous recourons à l'analyse des cooccurrences (Grossmann *et al.*, 2003 ; Condamines 2005 ; Blumenthal et Hausmann, 2006 ; Mayaffre 2008), parce qu'elle permet de quantifier et d'objectiver la monosémie en l'implémentant en termes d'homogénéité sémantique (Habert *et al.*, 2005). Une unité lexicale monosémique

⁵ WordSmith Tools : <http://www.lexically.net/wordsmith/>. [consulté le 20/01/2012].

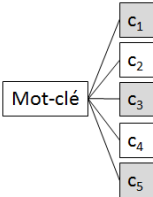
⁶ AntConc : <http://www.antlab.sci.waseda.ac.jp/software.html>. [consulté le 20/01/2012].

⁷ AV Frequency List Tool : <http://www.ling.arts.kuleuven.be/av-tools/>. [consulté le 20/01/2012].

⁸ Plusieurs outils d'extraction terminologique, tels que LEXTER (Bourigault *et al.* 2001), permettent de repérer les unités polylexicales. Toutefois celles-ci posent problème lors du calcul des spécificités. Pour l'instant, il n'est pas possible de déterminer le degré de spécificité des unités complexes de façon fiable et statistiquement significative, notamment parce que la plupart d'entre elles sont absentes dans un corpus de référence de langue générale. Par ailleurs, les techniques d'extraction automatique de termes s'appuient souvent sur un algorithme hybride avec une composante syntaxique importante, c'est-à-dire des structures syntaxiques récurrentes (Lemay *et al.*, 2005). Ainsi, plusieurs variables concourent au repérage des unités terminologiques complexes plutôt qu'une seule. Or, l'analyse de régression à laquelle nous procédons pour étudier la corrélation entre les données de spécificité et de monosémie, requiert une seule variable linguistique, c'est-à-dire un critère de spécificité clair et précis.

apparaît dans des contextes plutôt homogènes sémantiquement. Par contre, une unité lexicale polysémique se caractérise par des cooccurents plus hétérogènes sémantiquement, qui appartiennent à des champs sémantiques différents (Véronis, 2003 ; Habert *et al.*, 2004). L'accès à la sémantique des cooccurents de premier ordre (ou *c*) peut se faire à partir de leurs cooccurents, c'est-à-dire les cooccurents de deuxième ordre (ou *cc*). Ceux-ci se caractérisent principalement par des relations paradigmatiques avec le mot de base (hyponymes, hyperonymes, synonymes, antonymes) et dès lors ils sont intéressants pour caractériser sémantiquement le mot de base. Ils ont permis entre autres de mettre en évidence des relations de synonymie (Martinez, 2000).

Si les cooccurents de premier ordre d'un mot de base, en l'occurrence un mot-clé, partagent beaucoup de cooccurents de deuxième ordre, ces derniers se recoupent formellement, ce qui constitue une indication de l'homogénéité sémantique des cooccurents de premier ordre et, dès lors, du mot de base. Le degré de monosémie d'un mot-clé pourra donc être déterminé en fonction du degré de recouvrement formel de ses cooccurents de deuxième ordre. Une représentation schématique (Cf. figure 1) fait intervenir un mot-clé, ses 5 *c* différents et tous leurs *c* (10 *cc* différents et 26 *cc* au total). Ce schéma permettra d'expliquer le poids de chaque *cc* pour le recouvrement global. Un *cc* partagé par tous les *c* (p.ex. *cc*₃), figure 5 fois dans la liste des *cc*, constituée de 5 blocs de *cc* (un bloc par *c*). Le *cc* figurant 5 fois aura donc un poids maximal de 5/5. Par contre, un *cc* qui figure dans un seul bloc (p.ex. *cc*₂ ou *cc*₄) est un *cc* isolé avec un poids minimal de 1/5. De même, le poids d'un *cc* qui figure 2 fois dans la liste des *cc* ou dans 2 blocs (p.ex. *cc*₅) équivaut à 2/5. Ainsi, on pourra calculer facilement le poids de chaque *cc* dans la liste des 26 *cc*. Le poids de chaque *cc* correspond au rapport entre la fréquence du *cc* dans la liste des *cc* et le nombre de *c*. Pour connaître le recouvrement global, calculé à partir du recouvrement de tous les *cc*, on fera d'abord la somme des poids individuels (donc 26 répétitions du calcul précédent). Ce résultat sera divisé par 26 (nombre total de *cc* dans la liste), parce que chaque *cc* contribue pour 1/26 au recouvrement global calculé pour le mot-clé. Le résultat final se situe toujours entre 0 et 1 et représente le degré de recouvrement moyen pour un mot-clé, c'est-à-dire son degré d'homogénéité sémantique.



| | <i>cc</i> ₁ | <i>cc</i> ₂ | <i>cc</i> ₃ | <i>cc</i> ₄ | <i>cc</i> ₅ | <i>cc</i> ₆ | <i>cc</i> ₇ | <i>cc</i> ₈ | <i>cc</i> ₉ | <i>cc</i> ₁₀ |
|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-------------------------|
| <i>c</i> ₁ | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| <i>c</i> ₂ | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>c</i> ₃ | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| <i>c</i> ₄ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| <i>c</i> ₅ | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |

FIGURE 1 – Schéma : mot-clé + cooccurents de premier et de deuxième ordre.

L'analyse des cooccurrences est effectuée, de façon récurrente, dans une fenêtre d'observation (ou *span*) de 5 mots à gauche et 5 mots à droite, sans informations de position ni d'orientation. Cette fenêtre apporte suffisamment d'informations sémantiques pertinentes, sans introduire trop de bruit, et permet un traitement informatique efficace.

Les cooccurrents de premier et de deuxième ordre sont considérés au niveau des formes graphiques, ce qui permet de faire la distinction entre, par exemple, *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage ») et dès lors de tenir compte de la différence sémantique. La mesure d'association utilisée pour déterminer les cooccurrents statistiquement pertinents est la mesure statistique du LLR (log du rapport de vraisemblance). Le seuil de significativité très sévère (valeur $p < 0,0001$) permet de relever uniquement les cooccurrents de premier et de deuxième ordre sémantiquement pertinents. La mesure de recouplement est concrétisée à l'aide de scripts en Python pour calculer le degré de recouplement des 4717 mots-clés à partir du recouplement formel de leurs cooccurrents de deuxième ordre. Ce degré de recouplement ou d'homogénéité sémantique permet de situer les 4717 mots-clés sur un continuum d'homogénéité sémantique ou de monosémie et permet de leur accorder un rang de monosémie.

Comme nous ne disposons pas de listes de sens préétablis, ni d'autres mesures sémantiques comparables, nous avons procédé à une validation manuelle de la mesure de recouplement à partir de l'analyse manuelle des cooccurrents les plus pertinents, ainsi qu'à une validation externe au moyen de dictionnaires. Les résultats de ces validations confirment les résultats de notre mesure de recouplement pour un échantillon de 50 mots-clés. Il est à noter que des recherches supplémentaires s'imposent pour examiner la relation précise entre, d'une part, notre mesure de monosémie, implémentant la monosémie en termes d'homogénéité sémantique, et, d'autre part, ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure, dans le but opérationnel de développer un critère mesurable. Sans recherches supplémentaires, il serait impossible d'affirmer que les degrés de monosémie calculés correspondent parfaitement à ce que les terminologues traditionnels considèrent comme monosémie ou polysémie. Notons toutefois que ces derniers omettent de fournir des critères opérationnels à ce sujet.

2.3 Corrélation entre la spécificité et la monosémie

Pour répondre à la question de recherche et pour examiner la corrélation entre le continuum de spécificité et le continuum de monosémie (ou d'homogénéité sémantique), les données quantitatives de spécificité et de monosémie sont soumises à une analyse statistique de régression linéaire 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 le coefficient de détermination ou le pourcentage de variation expliquée R^2 . Il représente le pourcentage de la variation du rang de monosémie que l'on pourra expliquer ou prédire à partir de la variation du rang de spécificité des 4717 mots-clés.

Les résultats de l'analyse de régression simple permettent d'infirmar la thèse monosémiste traditionnelle, car ils montrent une corrélation négative (coefficient de corrélation Pearson de -0,72) et un pourcentage de variation expliquée R^2 de 51,57% (valeur $p < 2,2e^{-16}$). Il s'avère donc que les mots-clés les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais, au contraire, les plus hétérogènes sémantiquement (p.ex. *machine*, *pièce*, *tour*). En plus, les mots-clés les moins spécifiques du corpus technique sont les plus homogènes sémantiquement (par exemple *rationnellement*, *télédiagnostic*), à quelques exceptions près (*service* et *objet*). En effet, la

visualisation ci-dessous (Cf. figure 2) montre que la droite de régression s'incline vers le bas. Parmi les mots-clés spécifiques qui sont 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 (tels que *tour*), ainsi que des mots vagues (comme *usage*) dont le sens sous-déterminé est précisé par le contexte linguistique.

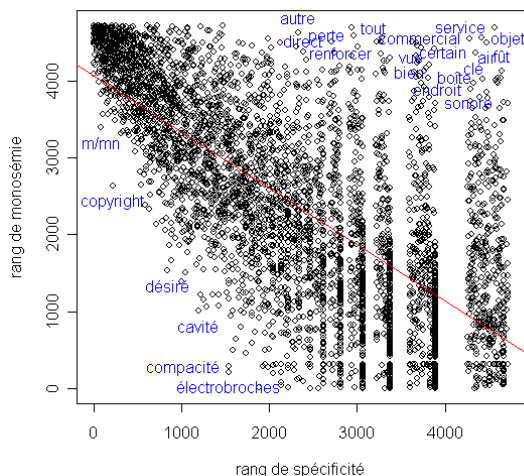


FIGURE 2 – Visualisation de l'analyse de régression.

3 Problèmes statistiques et méthodologiques

3.1 Hétéroscédasticité : mots généraux

La visualisation ci-dessus (Cf. figure 2) montre que la corrélation négative ne s'applique pas à tous les mots-clés et qu'elle n'est pas tout à fait linéaire. Le test statistique de Goldfeld-Quandt soulève effectivement un problème statistique d'hétéroscédasticité (statistique F du GQ-test 2,07), ce qui veut dire que les variances des erreurs ne sont pas constantes. En effet, certaines observations se caractérisent par un résidu important, c'est-à-dire par une grande différence entre leur valeur observée et la valeur estimée par le modèle de régression, par exemple *service*, *objet*, *commercial*. L'écart (ou le résidu) entre leur valeur observée (visualisée par la petite boule) et leur valeur estimée située sur la droite de régression est très important, ce qui donne lieu à une erreur importante lors de la prédiction de leur rang de monosémie à partir de leur rang de spécificité. Ces mots se situent principalement dans la partie supérieure droite, c'est-à-dire parmi les mots-clés les moins spécifiques. Ils sont plus polysémiques qu'on n'aurait cru en prenant en considération leur rang de spécificité.

Ce sont majoritairement des mots généraux, très fréquents dans le corpus de référence et dès lors peu spécifiques dans le corpus technique, en dépit de leur fréquence élevée dans le corpus technique. Pour ces mots, qui se trouvent dans la zone marginale de spécificité (valeur p légèrement inférieure à 0,05), le modèle de régression n'est pas une bonne prédiction de leur rang de monosémie à partir de leur rang de spécificité. 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 sont plutôt hétérogènes sémantiquement, quel que soit leur rang de spécificité.

3.2 Multicollinéarité : fréquence technique

Le deuxième problème est soulevé par une analyse statistique de régression multiple, qui évalue l'impact combiné et simultané de plusieurs variables indépendantes sur la variable dépendante. Parfois, deux ou plusieurs variables indépendantes sont corrélées les unes avec les autres. Elles expliquent en grande partie la même variation de la variable dépendante, ce que l'on qualifie de problème de multicollinéarité⁹.

Pour les 4717 mots-clés, nous observons un problème de multicollinéarité pour trois variables, à savoir le log du LLR, le rang de spécificité et le rang de fréquence technique. En effet, il y a une corrélation (trop) importante (0,87) entre la valeur du LLR, utilisée pour identifier et classer les mots-clés, et la fréquence technique. Par ailleurs, la mesure statistique du LLR est trop sensible à la fréquence technique, car pour les fréquences techniques (très) élevées, elle gonfle la valeur du LLR et dès lors le degré de spécificité. Il s'ensuit que les mots très fréquents dans le corpus technique ont un degré de spécificité relativement plus élevé que les mots moyennement ou faiblement fréquents. Par conséquent, certains mots très fréquents se situent à tort parmi les mots les plus spécifiques. Bien entendu, dans l'analyse de régression simple, nous considérons le rang de spécificité, qui permet tout de même d'effacer les différences trop importantes en termes de degrés de spécificité. Notons également que la fréquence technique élevée de certains mots s'explique par leur fréquence très élevée dans une des parties du corpus, en dépit de leur fréquence plutôt normale dans les autres parties. Ce biais de fréquence local est souvent causé par un biais de sujet (*topical bias*). En effet, le calcul du degré de spécificité compare la fréquence relative dans le corpus technique entier (de 1,7 million de mots) à la fréquence dans le corpus général entier (de 15,3 millions de mots), sans tenir compte de la dispersion des mots à travers les corpus. Or, la prise en considération de la dispersion des mots s'avère importante lors de l'extraction des mots-clés (Paquot *et al.*, 2009). Comme notre corpus technique consiste en 4 sous-corpus (revues, fiches techniques, normes et manuels), cette hétérogénéité des sources aura probablement un impact sur la dispersion et la spécificité des unités lexicales spécifiques.

En conclusion, deux problèmes se posent. D'une part, il y a trop de mots généraux parmi les mots-clés et ils entraînent un effet perturbateur et de ce fait un problème statistique d'hétéroscédasticité. D'autre part, la mesure statistique du LLR est trop sensible à la fréquence technique élevée et elle souffre d'un biais de sujet.

⁹ Valeurs VIF dans l'analyse de régression multiple : log du LLR (VIF 36,26), rang de spécificité (VIF 26,32) et rang de fréquence technique (VIF 14,72).

4 Solutions

Pour remédier à ces problèmes, nous proposons d'adopter une méthode alternative, qui permet de prendre en considération également la dispersion des mots à travers les corpus. Ainsi, on évite qu'un mot soit spécifique du corpus technique à cause de sa surreprésentation dans une seule partie. Or, la dispersion ne permet pas de résoudre tout le problème de la sensibilité à la fréquence. Par conséquent, nous adoptons également une autre mesure statistique, capable de refléter de façon plus fiable les unités lexicales spécifiques du corpus technique et leur degré de spécificité.

4.1 Stable Lexical Marker Analysis (SLMA)

La nouvelle méthode, appelée *Stable Lexical Marker Analysis* (SLMA) ou analyse des marqueurs lexicaux stables, a été développée dans le domaine de la linguistique variationnelle (Speelman *et al.*, 2006). Le but était d'identifier les variantes lexicales régionales typiques ou les « marqueurs lexicaux stables » des différences régionales entre le néerlandais utilisé aux Pays-Bas et en Flandre (Belgique) (Speelman *et al.*, 2008). La méthode s'applique aussi à l'extraction d'unités terminologiques, par exemple dans le domaine juridique de la législation financière (De Hertog *et al.*, 2010). La SLMA compare deux corpus à partir de leurs listes de fréquence et permet ainsi d'identifier les différences lexicales consistantes et stables entre les corpus. Elle s'inspire de la méthode des mots-clés de Scott (2006), en ce qu'elle consiste à comparer des listes de fréquence d'un corpus d'analyse à des listes de fréquence d'un corpus de référence. Toutefois, au lieu de comparer une liste de fréquence d'analyse à une liste de fréquence de référence, elle compare plusieurs fois de telles listes de fréquence. Elle fait donc intervenir de multiples tests d'hypothèse pour ainsi rendre compte de la dispersion.

En effet, le corpus spécialisé est subdivisé en plusieurs partitions (disons n partitions), tout comme le corpus de référence (m partitions). Pour chaque partition des deux corpus, on établit une liste de fréquence. Il y a donc $n*m$ listes de fréquence. Ensuite, chaque partition du corpus spécialisé A (p.ex. A_1, A_2, \dots, A_n) est comparée à chaque partition du corpus de référence B (p.ex. B_1, B_2, \dots, B_m), par le biais de leur liste de fréquence, ce qui revient à $n*m$ comparaisons de partitions. Chaque comparaison par paire de partitions permet de générer une liste de mots-clés spécifiques de la partition spécialisée (LLR et valeur $p < 0,05$). Les mots qui sont spécifiques dans la plupart de ces comparaisons (au maximum $n*m$) sont qualifiés de « marqueurs lexicaux stables », parce qu'ils sont stables et consistants à travers le corpus spécialisé entier. Le nombre de comparaisons significatives par paire de partitions (qualifié de SLM) est une première indication du degré de spécificité. Ces unités lexicales sont spécifiques (globalement relativement plus fréquentes dans le corpus spécialisé) et stables (avec une dispersion uniforme à travers le corpus spécialisé). Le découpage des corpus en partitions peut se réaliser à l'aide de scripts en Python, tout comme les multiples comparaisons des listes de fréquence.

4.2 Odds Ratio

La mesure statistique du log Odds Ratio (log OR), permet d'obtenir une indication de spécificité à granularité plus fine que le nombre de comparaisons significatives par paire de partitions (SLM). Le log OR permet également de prendre en considération la réelle

importance de la différence de fréquence d'un mot dans les deux (partitions de) corpus, ce que l'on qualifie de *effect size*. Le log OR fait intervenir la fréquence relative du mot, ainsi que celle de tous les autres mots, ce qui évite de gonfler le résultat pour les fréquences élevées (Cf. LLR). Pour un mot w_k donné, on calcule ainsi le score SMEA (*Stable Marker Effect size Analysis*), c'est-à-dire $SMEA(w_k, A, B)$, en calculant le log OR pour chaque comparaison significative, dans un corpus spécialisé A (n partitions) et un corpus de référence B (m partitions). La somme est divisée par le nombre total de comparaisons de partitions.

$$SMEA(w_k, A, B) = \frac{1}{n * m} \sum_{i=1}^n \sum_{j=1}^m (\log(\frac{F_{w_k}^{A^i} / F_{-w_k}^{A^i}}{F_{w_k}^{B^j} / F_{-w_k}^{B^j}})) * S(F_{w_k}^{A^i}, F_{-w_k}^{A^i}, F_{w_k}^{B^j}, F_{-w_k}^{B^j})$$

avec $F_{w_k}^{A^i}$ la fréquence du mot w_k dans la partition i du corpus A et $F_{-w_k}^{A^i}$ la fréquence de tous les mots autres que w_k (et de même pour les fréquences du corpus B). $S()$ est une fonction booléenne qui égale 1 si la distribution des mots est significativement différente dans les corpus A et B ; sinon elle égale 0. Si le nombre de comparaisons significatives est plus élevé, donc si le mot spécifique est mieux dispersé, le score SMEA sera plus élevé. Le score SMEA est une indication de la spécificité du mot ainsi que de sa dispersion, mais elle échappe au gonflement pour les mots très fréquents. Sa granularité très fine permet de classer les marqueurs lexicaux stables et de déterminer leur nouveau rang de spécificité, appelé rang de SMEA. De Hertog et al. (2010) ont démontré la fiabilité de cette approche par l'extraction de candidats-termes à partir d'un corpus de textes juridiques européens et par leur validation contre la base de données terminologique officielle des services européens.

5 Etude sémantique des marqueurs lexicaux stables du corpus technique

5.1 Identification des marqueurs lexicaux stables

Pour déterminer les marqueurs lexicaux stables du corpus technique, celui-ci est subdivisé en 5 partitions, c'est-à-dire une partition par sous-corpus, pour les normes, les fiches et les manuels (entre 300.000 et 360.000 occurrences) et 2 partitions de 400.000 occurrences pour le sous-corpus des revues. Ces 5 partitions sont de taille comparable et raisonnable et respectent l'ordre des mots et les frontières des sous-corpus thématiques et stylistiques. Le corpus de référence de langue générale est réparti en 36 partitions de taille similaire à celle des partitions techniques (environ 400.000 occurrences). L'analyse de la SLMA est effectuée sur les lemmes au lieu des formes fléchies, à l'instar de l'analyse des mots-clés dans AV Frequency List Tool (Cf. section 2.1). Après l'extraction et avant l'interprétation, la liste des marqueurs lexicaux stables repérés subit le même traitement que la liste des mots-clés, à savoir la suppression des mots grammaticaux, des noms propres et des hapax. Dans le corpus technique, nous recensons ainsi 3479 marqueurs lexicaux stables, statistiquement significatifs ($p < 0,05$), dont 3123 formes (ou presque 90%) figurent aussi dans la liste des 4717 mots-clés.

5.2 Marqueurs lexicaux stables versus mots-clés

Le tableau ci-dessous (Cf. table 1) montre que les mots-clés les plus fréquents et les plus spécifiques comme *machine*, *outil*, *pièce*, visualisés dans la colonne de droite aux rangs de spécificité (LLR) 1, 2 et 4 respectivement, ne figurent pas parmi les marqueurs lexicaux stables les plus spécifiques, visualisés à gauche du tableau. En effet, ils se retrouvent respectivement aux rangs de SMEA 33, 55 et 170. On observe également que la prise en considération de la dispersion relègue certaines unités lexicales, très fréquentes dans les revues (p.ex. *Fig*) à un rang moins spécifique (37 au lieu de 9). Les vrais termes, qui sont spécifiques du domaine, occupent des rangs plus spécifiques (*usinage*, *broche*, *copeau*, *fraisage*, *serrage*, ...). Ensuite, les mots généraux, qui ont des emplois généraux et techniques, occupent à juste titre des rangs un peu moins spécifiques (*machine*, *outil*, *pièce*, ...). Enfin, les mots généraux peu fréquents et à peine spécifiques (i.e. la queue de la liste des 4717 mots-clés) ne se retrouvent pas parmi les 3479 marqueurs lexicaux stables. Il s'avère que la fréquence technique moyenne des 4717 mots-clés est plus faible (140,77) que celle des 3479 marqueurs lexicaux stables (182,16). Par ailleurs, la corrélation entre la fréquence dans le corpus technique et la valeur de SMEA, qui indique le degré de spécificité, est moins problématique dans la liste des marqueurs lexicaux stables (0,32) que dans la liste des 4717 mots-clés, où la corrélation entre la fréquence technique et la valeur du LLR était trop importante (0,87).

| | lemme | SMEA | SLM | fréq.tech. | | mots-clés |
|----|---------------|------------|-----|------------|----|-----------|
| 1 | usinage | 85,699726 | 180 | 6720 | 1 | machine |
| 2 | broche | 74,8200697 | 180 | 2893 | 2 | outil |
| 3 | copeau | 73,7392965 | 180 | 2557 | 3 | usinage |
| 4 | fraisage | 68,364653 | 180 | 1873 | 4 | pièce |
| 5 | usiner | 68,3239216 | 180 | 1577 | 5 | mm |
| 6 | machine-outil | 67,6419261 | 180 | 1005 | 6 | vitesse |
| 7 | serrage | 66,3394778 | 180 | 939 | 7 | coupe |
| 8 | perçage | 62,8188634 | 180 | 846 | 8 | broche |
| 9 | fraise | 62,0265842 | 180 | 1571 | 9 | Fig |
| 10 | meule | 61,8557297 | 180 | 776 | 10 | axe |

TABLE 1 – Top 10 des marqueurs lexicaux stables du corpus technique (à gauche), par rapport au top 10 des 4717 mots-clés (à droite).

5.3 Corrélation entre la spécificité et la monosémie

Pour étudier la corrélation entre le rang de spécificité (rang de SMEA) et le rang de monosémie des 3479 marqueurs lexicaux stables, nous procédons à une analyse statistique de régression simple. Elle montre une corrélation négative entre le rang de

spécificité et le rang de monosémie (-0,49). Il s'avère donc que les marqueurs lexicaux les plus stables et les plus spécifiques ne sont pas les plus monosémiques. Toutefois, la corrélation (-0,49) est moins convaincante que celle pour les 4717 mots-clés (-0,72). Le pourcentage de variation expliquée R^2 de 23,87% (valeur $p < 2,2e^{-16}$) est également moins convaincant que celui pour les 4717 mots-clés (51,57% et valeur $p < 2,2e^{-16}$). Ces résultats moins concluants pour les marqueurs lexicaux stables s'expliquent principalement par l'absence des mots généraux peu fréquents et très peu spécifiques, qui sont très monosémiques, et par le fait que les mots les plus fréquents occupent, à juste titre, des rangs moins spécifiques. En raison de leur fréquence plus élevée dans le corpus technique, ces derniers ont plus de chances d'être polysémiques et/ou de constituer la tête d'unités polylexicales, où ils sont désambiguïsés par les autres composants (par exemple *machine à fraiser*, *machine à rainurer*).

Notons que le test de Goldfeld-Quandt soulève aussi un problème d'hétéroscédasticité (statistique F du GQ-test 1,37), mais moins important que pour les 4717 mots-clés (2,07). Le problème de l'hétéroscédasticité est donc résolu en partie, mais suggère la présence d'une variable supplémentaire, cachée jusqu'à présent, qui prédit peut-être une partie de la variation du rang de monosémie. Cette variable pourrait être liée au fait que les mots spécifiques constituent la tête d'unités polylexicales dans le corpus technique. Des recherches futures permettront de vérifier si elle permet d'expliquer l'hétéroscédasticité et dans quelle mesure.

6 Conclusion

Dans cet article, nous avons étudié les unités lexicales spécifiques d'un corpus technique relevant du domaine spécialisé restreint des machines-outils pour l'usinage des métaux. Nous nous sommes tout particulièrement intéressés à la corrélation entre le rang de spécificité et le rang de monosémie de ces unités spécifiques.

Une double analyse quantitative a permis de générer une liste de 4717 mots-clés, avec un degré de spécificité et un degré de monosémie ou d'homogénéité sémantique. Ces données quantitatives ont permis de classer les 4717 mots-clés dans un continuum de spécificité et dans un continuum de monosémie afin d'examiner la corrélation entre le rang de spécificité et le rang de monosémie par le biais d'une analyse statistique de régression simple. Nous avons observé une corrélation négative, qui indique que les unités lexicales les plus spécifiques du corpus technique, relevées avec la méthodologie de l'analyse des mots-clés, ne sont pas les plus homogènes sémantiquement, au contraire. Cette observation a permis de remettre en cause la thèse monosémiste traditionnelle. La méthode alternative de l'analyse des marqueurs lexicaux stables (*Stable Lexical Marker Analysis* ou SLMA) a permis de remédier aux problèmes statistiques et méthodologiques d'hétéroscédasticité et de multicollinéarité, en prenant en considération la dispersion et en utilisant une autre mesure statistique. Elle a généré une liste de 3479 marqueurs lexicaux stables avec un nouveau rang de spécificité (rang de SMEA). Les résultats de l'analyse de régression simple confirment la corrélation négative entre le nouveau rang de spécificité (rang de SMEA) et le rang de monosémie des marqueurs lexicaux stables, bien qu'elle soit moins forte. Ces premières expérimentations montrent donc que l'analyse des marqueurs lexicaux stables constitue une alternative valable pour l'analyse des mots-clés.

Références

- BERTELS, A., SPEELMAN, D. et GEERAERTS, D. (2010). La corrélation entre la spécificité et la sémantique dans un corpus spécialisé. In *Revue de Sémantique et de Pragmatique* n°27, pages 79–102.
- BHREATHNACH, U. et DE BARRA CUSACK, F., éditeurs (2010). *TKE 2010 : Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, Fiontar. Dublin City University.
- BLUMENTHAL, P. et HAUSMANN, F.J., éditeurs (2006). *Collocations, corpus, dictionnaires. Langue française*, n° 150.
- BOURIGAULT D., JACQUEMIN, C. et L'HOMME, M.-C., éditeurs (2001). *Recent advances in computational terminology*, Amsterdam/Philadelphia. John Benjamins Publishing Company.
- BOURIGAULT, D. et SLODZIAN, M. (1999). Pour une terminologie textuelle. In *Terminologies Nouvelles* n°19, pages 29–32.
- CABRE, M.T. (2000). Terminologie et linguistique : la théorie des portes. In *Terminologies Nouvelles* n°21, pages 10–15.
- CONDAMINES, A. et REBEYROLLE, J. (1997). Point de vue en langue spécialisée. In *Meta*, n°42(1), pages 174–184.
- CONDAMINES, A., éditeur (2005). *Sémantique et corpus*, Paris. Hermès-Science.
- DE HERTOOG, D., HEYLEN, K., SPEELMAN, D. et KOCKAERT, H. (2010). A variational linguistics approach to term extraction. In (Bhreatnach et de Barra Cusack, 2010), pages 229–248.
- DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* n°19(1), pages 61–74.
- ERIKSEN, L. 2002. Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder : Zur Terminologie der ‚Sache‘ im Deutschen. In *Hermes – Journal of Linguistics* n°28, pages 211–222.
- FERRARI, L. (2002). Un caso de polisemia en el discurso jurídico? In *Terminology* n°8(2), pages 221–244.
- GAUDIN, F. (2003). *Socioterminologie : une approche sociolinguistique de la terminologie*. Bruxelles. Duculot.
- GROSSMANN, F. et TUTIN, A., éditeurs (2003). *Les collocations, analyse et traitement, Travaux et Recherches en linguistique appliquée*, Série E, vol. 1.
- HABERT, B., ILLOUZ, G., FOLCH, H. (2004). Dégroupier les sens : pourquoi ? comment ? In *Actes des JADT 2004 (Journées internationales d'analyse statistique des données textuelles)*, Louvain-la-Neuve, pages 565–576.
- HABERT, B., ILLOUZ, G., FOLCH, H. (2005). Des décalages de distribution aux divergences d'acception, In (Condamines, 2005), pages 277–318.

- JUCKER, A., SCHREIER, D. et HUNDT, M. éditeurs (2009). *Corpora: Pragmatics and Discourse*, Amsterdam. Rodopi.
- KRISTIANSEN, G. et DIRVEN, R., éditeurs (2008). *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, Berlin/New York. Mouton de Gruyter.
- KWARY, D.A. (2011). A hybrid method for determining technical vocabulary. *In System* n°39(2), pages 175-185.
- KWARY, D.A. (2011). A hybrid method for determining technical vocabulary. *In System*, n°39(2), pages 175-185.
- LABBE, C. et LABBE, D. (2001). Que mesure la spécificité du vocabulaire ? *In Lexicometrica* n°3.
- LAFON, P. (1984). *Dépouillements et statistiques en lexicométrie*, Genève-Paris. Slatkine-Champion.
- LEMAY, C., L'HOMME, M.C. et DROUIN, P. (2005). Two methods for extracting specific single-word terms from specialized corpora. Experimentation and evaluation. *In International Journal of Corpus Linguistics*, n°10(2), pages 227-255.
- MARTINEZ, W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *In Actes des JADT 2000 (Journées internationales d'analyse statistique des données textuelles)*, Lausanne, pages 78-84.
- MAYAFFRE, D. (2008), Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence, *In Actes des JADT 2008 (Journées internationales d'analyse statistique des données textuelles)*, Lyon, pages 811-822.
- PAQUOT, M. et BESTGEN, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction », *In (Jucker et al., 2009)*, pages 247-269
- SCOTT, M. et TRIBBLE, C. (2006). *Textual Patterns. Key words and corpus analysis in language education*. Studies in Corpus Linguistics, vol. 22. Amsterdam. Benjamins.
- SPEELMAN, D., GRONDELAERS, S. et GEERAERTS, D. (2006). A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. *In (Wilson et al., 2006)*, pages 195-202.
- SPEELMAN, D., GRONDELAERS, S. et GEERAERTS, D. (2008). Variation in the choice of adjectives in the two main national varieties of Dutch. *In (Kristiansen et Dirven, 2008)*, pages 205-233.
- 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(Traitement automatique des langues naturelles)*, Batz-sur-Mer, pages 265-274.
- WILSON, A., ARCHER, D. et RAYSON, P., éditeurs (2006). *Corpus Linguistics around the World*, Amsterdam. Rodopi.
- WÜSTER, E. (1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, (3. Aufl.), Bonn. Romanistischer Verlag.