A la découverte de la polysémie des spécificités du français technique

Ann Bertels

ILT – K.U.Leuven
Dekenstraat 6 – B-3000 LEUVEN (Belgique)
ann.bertels@ilt.kuleuven.ac.be

Mots-clefs – Keywords

sémantique lexicale, langue spécialisée, spécificités, polysémie, cooccurrences

lexical semantics, language for specific purposes (LSP), keywords, polysemy, co-occurrences

Résumé – Abstract

Cet article décrit l'analyse sémantique des spécificités dans le domaine technique des machines-outils pour l'usinage des métaux. Le but de cette étude est de vérifier si et dans quelle mesure les spécificités dans ce domaine sont monosémiques ou polysémiques. Les spécificités (situées dans un continuum de spécificité) seront identifiées avec la *KeyWords Method* en comparant le corpus d'analyse à un corpus de référence. Elles feront ensuite l'objet d'une analyse sémantique automatisée à partir du recouvrement des cooccurrences des cooccurrences, afin d'établir le continuum de monosémie. Les travaux de recherche étant en cours, nous présenterons des résultats préliminaires de cette double analyse.

This article discusses a semantic analysis of pivotal terms (keywords) in the domain of machining terminology in French. Building on corpus data, the investigation attempts to find out whether, and to what extent, the keywords are polysemous. In order to identify the most typical words of the typicality continuum, the KeyWords Method will be used to compare the technical corpus with a reference corpus. The monosemy continuum will be implemented in terms of degree of overlap between the co-occurrents of the co-occurrents of the keywords. We present some preliminary results of work in progress.

1 Introduction et question de recherche

Cet article s'inscrit dans le cadre d'une thèse de doctorat sur la sémantique du vocabulaire spécifique d'un corpus de français technique. Comme le corpus d'analyse relève du domaine

technique des machines-outils pour l'usinage des métaux, l'analyse sémantique porte sur les spécificités[1] d'une langue spécialisée.

Dans la langue spécialisée, les besoins communicatifs requièrent plus de précision, ce que la terminologie traditionnelle définit comme l'univocité, la monoréférentialité et la monosémie des unités terminologiques de la langue spécialisée. La terminologie traditionnelle prescriptive et normative adopte une approche onomasiologique par domaine. Récemment, la monosémie et l'univocité de la langue spécialisée ont été remises en question par la Théorie Communicative de la Terminologie (Cabré, 1998, 2000), par la socioterminologie (Gaudin, 1993) et par la terminologie socio-cognitive (Temmerman, 1997). Les termes font partie intégrante de la langue naturelle, mais véhiculent des connaissances spécialisées (Lerat, 1995). Les partisans de la terminologie descriptive rejettent la dichotomie entre la langue générale et la langue spécialisée et adoptent une approche sémasiologique et linguistique, basée sur l'étude de corpus de textes spécialisés (Condamines & Rebeyrolles, 1997).

Pour quantifier la thèse monosémiste de la terminologie traditionnelle, nous nous proposons de la reformuler en une question de recherche opérationnelle et mesurable : « Y a-t-il une corrélation entre, d'une part, le continuum de spécificité et, d'autre part, le continuum de monosémie (continuum de sens) ? » L'hypothèse de recherche avancée pose que, contrairement à la thèse traditionnelle, les mots (les plus) spécifiques du corpus technique ne sont pas nécessairement (les plus) monosémiques. L'analyse se propose donc de vérifier la polysémie des mots du corpus technique d'analyse, p.ex. le mot *broche* (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques ». A cet effet, ces mots sont ordonnés en fonction de leur spécificité et situés sur une échelle de spécificité allant des mots les plus spécifiques aux mots les moins spécifiques, mais comprenant toujours des spécificités statistiquement significatives du corpus technique. Un deuxième classement situe les mêmes mots sur une échelle de monosémie, à partir d'une analyse des cooccurrences de deuxième ordre, c'est-à-dire les cooccurrences des cooccurrences. La question de recherche principale (corrélation entre le degré de spécificité et le degré de monosémie) sera complétée par des questions de recherche secondaires faisant intervenir les facteurs influant sur le degré de monosémie, notamment la fréquence et la classe lexicale. Une analyse de régression multiple permettra de vérifier l'impact des variables indépendantes (spécificité, fréquence, classe lexicale, etc.) sur le degré de monosémie.

2 Corpus technique d'analyse et corpus de référence

Le corpus technique d'analyse est constitué de textes techniques du domaine des machines-outils pour l'usinage des métaux et comprend environ 1.760.000 mots. Le corpus a été étiqueté et lemmatisé par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1998 à 2001 : revues techniques électroniques (800.000) et fiches techniques (300.000) trouvées sur Internet, normes ISO et directives (300.000) et quatre manuels numérisés (360.000). Les textes se situent à différents niveaux de normalisation et de vulgarisation, s'adressant tant à des professionnels (revues et fiches) qu'à des étudiants (manuels). Afin de pouvoir déterminer

[1] Nous adoptons le terme « spécificités » pour désigner les mots les plus spécifiques et caractéristiques du corpus d'analyse, indépendamment de la méthode utilisée (calcul des spécificités vs. KeyWords Method).

les spécificités du corpus technique, il est complété par un corpus de référence de langue générale. Celui-ci est constitué d'articles journalistiques du journal Le Monde (janvier – septembre 1998). Il a également été lemmatisé et comprend environ 15.300.000 mots.

Les fichiers générés par Cordial se composent de trois colonnes, avec un mot par ligne: (1) la forme fléchie ou forme graphique, (2) le lemme ou forme canonique et (3) le code Cordial, comparable à un POS-tag (Part-Of-Speech) indiquant la classe lexicale. Dans les fichiers texte, nous avons corrigé quelques fautes de frappe. Les fichiers lemmatisés ont également fait l'objet d'un nettoyage, à savoir quelques regroupements (p.ex. lemmes avec et sans point *Fig./Fig* et lemmes avec et sans trait d'union) et la correction des erreurs de lemmatisation (p.ex. *machines-outils* sous le lemme *machine-outil*). Ces opérations de nettoyage ont été effectuées, tant pour le corpus d'analyse technique que pour le corpus de référence.

3 Approche méthodologique : spécificités et polysémie

Comme la recherche porte sur la question de savoir s'il y a une corrélation entre le continuum de spécificité et le continuum de monosémie, la réponse et l'analyse linguistique qui en découle requièrent une approche méthodologique double. Il faut d'une part le calcul des spécificités et d'autre part une mesure pour déterminer le degré de monosémie. Ces spécificités se situent, non seulement au niveau des unités simples, p.ex. *fraisage, commande*, mais également au niveau des unités polylexicales, p.ex. *commande numérique*.

3.1 Spécificités

La recherche en langue spécialisée prend généralement comme point de départ l'identification des spécificités, c'est-à-dire des mots spécifiques qui caractérisent le corpus spécialisé et qui le différencient d'un corpus de langue générale. Les spécificités ne sont pas les mots les plus fréquents de ce corpus, mais les mots les plus représentatifs. Du point de vue relatif, ces mots figurent de façon significative plus fréquemment dans le corpus de langue spécialisée que dans un corpus de langue générale. Afin de déterminer les spécificités, les fréquences dans le corpus spécialisé sont comparées aux fréquences dans un corpus de référence, compte tenu de la taille des deux corpus, ce qui revient à comparer la fréquence observée (corpus d'analyse) à la fréquence attendue (corpus de référence). S'il y a une différence entre la fréquence observée et la fréquence attendue, il faut vérifier si elle est statistiquement significative. A cet effet, deux méthodologies sont désormais disponibles : le calcul des spécificités (Lafon, 1984) implémenté dans le logiciel Lexico3[2], outils de statistique textuelle, et la *KeyWords Method* des logiciels WordSmith Tools[3] et Abundantia Verborum[4] (Speelman, 1997). Les deux méthodologies 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.

---

[2] Lexico3 : SYLED – CLA2T, Paris3 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

[3] WordSmith Tools version 3 : http://www.lexically.net/wordsmith/ et http://www.oup.com

[4] Abundantia Verborum : http://wwwling.arts.kuleuven.ac.be/genling/abundant/obtain/

Premièrement, le calcul des spécificités (Lafon, 1984 et Labbé & Labbé, 1991) compare une section d'un corpus au corpus entier afin d'identifier le vocabulaire spécifique de cette section. La comparaison *partie-tout* permet de décider si la fréquence relative d'un mot dans la section est supérieure à ce qui serait attendu, en fonction de la fréquence relative dans le corpus entier. L'analyse statistique sous-jacente au calcul des spécificités utilise le test de Fisher Exact, basé sur les probabilités exactes de la distribution hypergéométrique. Fisher Exact est généralement utilisé pour un ensemble de données de taille modeste ($n \leq 20$). En outre, les factorielles de la formule pour le calcul de la probabilité de la fréquence observée (Lafon, 1984) mèneraient à des nombres astronomiques, si la formule était appliquée à un corpus de quelques milliers, voire des millions de mots. Pour remédier à ce problème, Lafon propose d'utiliser des logarithmes. Dès lors, le résultat du calcul *x* est à interpréter comme l'exposant de la base 10, d'où résulte la probabilité 10^x. Dans Lexico3, ce sont les exposants (résultats de la formule du calcul hypergéométrique) qui figurent dans la colonne du coefficient de spécificité. Les spécificités positives indiquent un suremploi dans la section analysée, tandis que les spécificités négatives signalent un sous-emploi. Le calcul des spécificités est surtout utilisé par la communauté francophone (Cf. Zimina, 2004 et Drouin, 2004).

La deuxième méthodologie permettant d'identifier les mots les plus spécifiques est surtout utilisée par des utilisateurs du logiciel WordSmith–KeyWords (Berber-Sardinha, 1999). Elle est couramment appelée *KeyWords Method* ou méthode des mots-clefs. Les fréquences dans le corpus spécialisé sont comparées aux fréquences dans un corpus de référence de langue générale, compte tenu de la taille des deux corpus, ce qui permet d'identifier les mots significativement plus fréquents dans le corpus spécialisé. Il s'agit donc de la comparaison de deux corpus différents, et non d'une comparaison *partie-tout*. La deuxième différence entre les deux méthodologies réside dans la statistique sous-jacente. La *KeyWords Method* se sert du ratio du log de vraisemblance (*log likelihood ratio*) (Dunning,1993). Cette statistique de test n'est pas basée sur des probabilités exactes et par conséquent, elle s'applique facilement à des corpus plutôt volumineux. Le ratio du log de vraisemblance sera d'autant plus élevé que le mot est plus fréquent dans le corpus spécialisé par rapport au corpus de référence, indiquant dès lors son degré de spécificité. La valeur p correspondante permet de supprimer les spécificités statistiquement non significatives ($p \leq 0.05$). En plus, le tri des spécificités en fonction de la statistique de test (ratio du log de vraisemblance) permet de les classer par ordre de spécificité décroissante et par conséquent, de les situer dans un continuum de spécificité.

3.2 Polysémie

Les spécificités relevées dans le corpus technique d'analyse, font ensuite l'objet d'une analyse sémantique. Pour chaque spécificité, on déterminera le degré de monosémie, dans le but de vérifier si les mots les plus spécifiques sont en effet (les plus) monosémiques et si les mots moins spécifiques ont plus tendance à être polysémiques. Il s'agira donc d'objectiver et de quantifier l'analyse sémantique, en ayant recours aux cooccurrences. Selon Schütze (1998), Véronis (2003) et d'autres, les cooccurrences permettent de distinguer les différents usages et sens des mots. Audibert (2003) recourt même aux cooccurrences comme critères de désambiguïsation sémantique automatique. Dans Véronis (2003), les cooccurrences d'un mot, à partir d'un grand corpus, sont regroupées suivant leur similarité ou dissimilarité (en fonction de leur co-fréquence) pour identifier les différents sens du mot.

Afin de déterminer le degré de monosémie des spécificités, nous proposons d'aller plus loin et d'étudier les cooccurrences de deuxième ordre. Les cooccurrences des cooccurrences permettent de trouver des synonymes d'un mot, selon Martinez (2000). Pour le mot *mesures*, il trouve les cooccurrents *nouvelles*, *prises*, etc. qui cooccurrent à leur tour avec *décisions*. D'après Denhière & Lemaire (2003), les cooccurrences de deuxième ordre et même d'ordre supérieur déterminent le degré d'association de deux mots M1 et M2, même si ces deux mots ne figurent jamais ensemble. Si les cooccurrences M1-M3 et M2-M3 sont suffisamment fortes, on considère que M1 et M2 sont associés et des cooccurrents d'ordre 2. Il est également possible d'extraire automatiquement les sens des mots à partir d'un réseau de cooccurrences lexicales de deuxième ordre, comme l'explique Ferret (2004). La connectivité des cooccurrents formant un sens est plus importante que leur connectivité avec les autres cooccurrents définissant les autres sens de ce mot, la mesure de cohésion étant l'information mutuelle normalisée (Ferret, 2004).

Les cooccurrents de deuxième ordre étant des critères désambiguïsateurs puissants, ils seront très précieux lors de l'analyse sémantique des spécificités. En effet, le degré de recouvrement des cooccurrences de deuxième ordre sera un indice important du degré de monosémie du mot de base. Pour étudier le caractère monosémique ou polysémique d'une unité linguistique, on vérifie si les contextes peuvent être considérés comme sémantiquement homogènes ou non (Condamines & Rebeyrolles, 1997). L'accès à la sémantique des cooccurrences pourra se faire (automatiquement) par le biais des cooccurrences de deuxième ordre. Le degré de recouvrement de ces cooccurrences de deuxième ordre indiquera à quel point les cooccurrences de premier ordre (contextes du mot de base) sont sémantiquement homogènes.

- Si les cooccurrents des cooccurrents (« cc » ou cooccurrents de deuxième ordre) sont formellement très différents et se recouvrent très peu, les différents cooccurrents (« c » ou cooccurrents de premier ordre) seront sémantiquement plus diversifiés, une structure formelle de cooccurrence différente indiquant un sens différent. Les cooccurrents sémantiquement diversifiés appartenant à plusieurs champs sémantiques, la spécificité aura moins de chances d'être monosémique.

- Et inversement, plus les cooccurrences des cooccurrences se recouvrent, plus les cooccurrents sont sémantiquement homogènes. Le degré de ressemblance ou similarité lexicale des cooccurrents d'un mot étant proportionnel au degré de monosémie de ce mot, un fort recouvrement des cooccurrents de deuxième ordre signale un degré de monosémie plus important.

4 Premiers résultats de la recherche : spécificités et polysémie

4.1 Spécificités

La liste des spécificités du corpus technique d'analyse (1,7 million de mots) est générée avec les logiciels Abundantia Verborum et AV Frequency List Tool. Une expérimentation sur un échantillon restreint du corpus technique sur les trois logiciels disponibles pour le calcul des spécificités montre des résultats comparables en termes de spécificités relevées. Pour garantir la comparaison des résultats, le corpus technique a été incorporé dans le corpus de référence dans le logiciel Lexico3, procédant par comparaison *partie-tout*. Force est de constater que la

même procédure d'incorporation pour le corpus entier (corpus technique de 1,7 million et corpus de référence de 15,3 millions) n'aboutit pas aux résultats escomptés. Même si la liste de fréquence du grand corpus entier s'affiche, l'étape suivante du calcul des spécificités échoue en raison de la taille trop importante du corpus.

Appliquée au corpus technique lemmatisé, la *KeyWords Method* produit une liste d'environ 13.000 spécificités pour le corpus technique ($p \leq 0.05$). A l'aide des codes Cordial, une liste de mots grammaticaux (450) et une liste de noms propres (7200) sont générées, permettant de les supprimer. Les opérations de filtrage et de nettoyage génèrent une liste de spécificités techniques définitive de 7240 mots (lemmes), Cf. Figure 1, pour un aperçu des 25 mots les plus spécifiques. Ces 7240 mots feront l'objet de l'analyse sémantique automatisée pour établir le degré de monosémie. La première colonne (Cf. Figure 1) contient les lemmes spécifiques, les deuxième et troisième colonnes donnent la fréquence absolue dans le corpus technique (FREQ_ABS1) et dans le corpus de référence (FREQ_ABS 2). Dans la colonne 4, on voit la statistique de test LLR indiquant le degré de spécificité et dans la colonne 5, le complément de la valeur p correspondante (1-p). Les colonnes 6 et 7 affichent les fréquences relatives (multipliées par 10000) pour les deux corpus et la dernière colonne informe sur le type de spécificité (1 pour une spécificité positive et –1 pour une spécificité négative). Il est à remarquer que cette liste de spécificités contient aussi des mots de la langue générale (p.ex. *type, permettre*), spécifiques de ce corpus technique. Ce ne sont pas des termes, mais ils sont maintenus car nous nous proposons de comparer leur degré de monosémie à celui des termes (dans le corpus technique) et à leur degré de monosémie dans un corpus de langue générale.

LEMME	FREQ_ABS1	FREQ_ABS2	LLR	1-P	FREQ_REL1	FREQ_REL2	SPEC_POS
machine	12671	1052	50521,91	1	74,51	0,71	1
outil	8306	918	32037,72	1	48,84	0,62	1
usinage	6720	8	30468,41	1	39,52	0,01	1
pièce	7556	2219	24407,46	1	44,43	1,50	1
mm	5490	191	23357,57	1	32,28	0,13	1
vitesse	5283	900	19108,78	1	31,07	0,61	1
coupe	6730	4153	17063,37	1	39,58	2,80	1
broche	2893	12	13010,42	1	17,01	0,01	1
Fig	2680	0	12194,00	1	15,76	0,00	1
axe	3206	420	12079,16	1	18,85	0,28	1
copeau	2557	0	11634,18	1	15,04	0,00	1
plaquette	2407	35	10592,46	1	14,15	0,02	1
diamètre	2415	95	10200,09	1	14,20	0,06	1
commande	2751	850	8765,71	1	16,18	0,57	1
acier	2252	277	8558,49	1	13,24	0,19	1
fraisage	1873	0	8521,34	1	11,01	0,00	1
arête	1870	29	8213,91	1	11,00	0,02	1
précision	2263	541	7663,01	1	13,31	0,36	1
usiner	1577	11	7045,52	1	9,27	0,01	1
surface	2258	758	7037,02	1	13,28	0,51	1
type	2820	1830	6994,07	1	16,58	1,23	1
système	4052	5165	6915,85	1	23,83	3,48	1
fraise	1571	45	6745,88	1	9,24	0,03	1
gamme	1860	545	6006,35	1	10,94	0,37	1
permettre	4883	9504	5848,03	1	28,71	6,41	1

Figure 1 : Les 25 mots les plus spécifiques du corpus technique d'analyse

4.2 Polysémie

Le degré de monosémie (ou inversement de polysémie) dépend du degré de recouvrement des cooccurrents des cooccurrents. Afin d'établir les listes des cooccurrences pertinentes à deux niveaux et de générer une base de données qui sera interrogée pour le calcul automatique du degré de recouvrement, nous avons recours à un algorithme de scripts Python[5].

Pour déterminer le degré de monosémie, nous proposons une formule (Cf. Figure 2), basée sur le recouvrement des cooccurrents des cooccurrents (cc), en tenant compte (1) de la fréquence d'un cc dans la liste des cc (= nombre de cooccurrents (c) apparaissant avec ce cc), (2) du nombre total de c et (3) du nombre total de cc. La mesure d'association utilisée pour déterminer les cooccurrences pertinentes est la statistique LLR (log de vraisemblance). Nous ne prenons en considération que les cooccurrences statistiquement significatives ($p \leq 0.05$).

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

Figure 2 : Formule de recouvrement des cooccurrents des cooccurrents

Dans un premier temps, nous dressons une liste des cooccurrences pertinentes à partir des fichiers lemmatisés du corpus technique, contenant le collocatif, la base[6] et leur co-fréquence. Deux autres fichiers sont dérivés de ces informations et contiennent respectivement les bases et les collocatifs et leurs fréquences. Toutes ces informations permettront de générer une base de données avec des informations statistiques, à savoir la statistique de test LLR et la valeur p. En fait, deux listes de cooccurrences avec leur base de données correspondante sont ainsi dressées : une première liste avec les spécificités comme base et leurs cooccurrents comme collocatif et une deuxième liste de cooccurrences avec les cooccurrents de la première liste comme base et leurs cooccurrents (d'ordre 2) comme collocatif.

Les paramètres modifiables sont le type de cooccurrent à relever (lemme ou forme fléchie) et la fenêtre d'observation. Nous optons pour une fenêtre de [-5,+5], 5 mots à gauche et 5 mots à droite, parce qu'elle apporte assez d'information sémantique, sans qu'il y ait trop de bruit et qu'elle permet un traitement informatique efficace. Au premier niveau d'analyse de la spécificité comme base, la base de la cooccurrence est nécessairement relevée sous forme lemmatisée, afin de pouvoir rattacher les informations sémantiques (degré de monosémie) aux informations de spécificité (liste de spécificités). Pour le collocatif, la forme fléchie s'impose, en raison des informations sémantiques plus riches qu'elle véhicule (Cf. différence entre *pièce à usiner* et *pièce usinée*). Comme ce collocatif est la base du deuxième niveau d'analyse, la forme fléchie s'impose à ce deuxième niveau tant pour la base que pour le collocatif.

Ces deux bases de données sont fusionnées en une grande base de données, interrogée pour l'analyse du recouvrement des cooccurrents de deuxième ordre. A cet effet, la fonction Python de l'algorithme prévoit les paramètres suivants : la base (spécificité à analyser), le

---

5 http://www.python.org/

6 la base (anglais : *node*) étant le mot étudié et le collocatif (anglais : *collocate*) étant un de ses cooccurrents

seuil de signification pour les cooccurrents de premier niveau (p.ex. 0.95 pour p≤0.05), le seuil pour les cooccurrents de deuxième niveau et la base de données. Il y a plus de recouvrement, si plus de cooccurrents (c) partagent le même cc, ce qui signifie un poids plus lourd pour ce cc (score près de 1). Un cc moins/pas partagé indique donc peu/pas de recouvrement (score près de 0).

La figure 3 ci-dessous montre les premiers résultats pour le corpus technique (1.7 million) et pour un seuil de signification des c et cc de p≤0.0001. Parmi les 25 mots les plus spécifiques du corpus technique entier, les 2 mots les plus spécifiques se caractérisent par le degré de monosémie le moins élevé (rangs 25 et 24), indiquant peu de recouvrement des cooccurrents d'ordre 2. Les mots en gras sont les moins monosémiques de cet échantillon, ce qui semble contredire la thèse de la corrélation[7] entre le degré de spécificité et le degré de monosémie.

LEMME	FREQ_ABS1	FREQ_ABS2	LLR	DEGRE DE MONOSEMIE	RANG DE MONOSEMIE
machine	12671	1052	50521,91	**0,0231**	25
outil	8306	918	32037,72	**0,0240**	24
usinage	6720	8	30468,41	0,0349	12
pièce	7556	2219	24407,46	0,0310	18
mm	5490	191	23357,57	0,0534	1
vitesse	5283	900	19108,78	0,0402	6
coupe	6730	4153	17063,37	0,0370	10
broche	2893	12	13010,42	0,0394	7
Fig	2680	0	12194,00	0,0483	3
axe	3206	420	12079,16	0,0340	13
copeau	2557	0	11634,18	**0,0299**	20
plaquette	2407	35	10592,46	**0,0282**	22
diamètre	2415	95	10200,09	0,0444	4
commande	2751	850	8765,71	0,0317	17
acier	2252	277	8558,49	**0,0282**	21
fraisage	1873	0	8521,34	0,0350	11
arête	1870	29	8213,91	0,0386	8
précision	2263	541	7663,01	0,0491	2
usiner	1577	11	7045,52	0,0406	5
surface	2258	758	7037,02	0,0321	15
type	2820	1830	6994,07	0,0372	9
système	4052	5165	6915,85	**0,0280**	23
fraise	1571	45	6745,88	0,0319	16
gamme	1860	545	6006,35	0,0324	14
permettre	4883	9504	5848,03	0,0303	19

Figure 3 : Degré et rang de monosémie des 25 mots les plus spécifiques (p ≤ 0.0001)

Pour les 100 mots les plus spécifiques, une première analyse de régression simple fait intervenir le degré de monosémie comme variable dépendante et le degré de spécificité comme variable indépendante ou prédictive. Elle montre qu'il y a une très faible corrélation négative (p=0.03) et que le degré de spécificité n'explique que 3.5% de la variation du degré

[7] Nous ne recourons pas à une simple mesure de corrélation, en raison de l'effet d'interférence attendu des autres facteurs (fréquence, classe lexicale, etc.).

de monosémie. Une analyse de régression multiple, mesurant l'impact de plusieurs variables indépendantes (spécificité, fréquence, classe lexicale, nombre de classes lexicales et longueur), indique comme facteurs significatifs le nombre de classes lexicales (p=0.008), la classe lexicale (p=0.03) et la fréquence (p=0.03) pour une variation totale expliquée de 12% (p=0.004). Parmi les 100 mots les plus spécifiques du corpus technique, les mots les plus polysémiques, affichant le degré de monosémie le plus bas, se caractérisent par une fréquence absolue élevée et par leur appartenance à deux ou plusieurs classes lexicales, principalement les classes 'nom' et 'adjectif'.

5 Conclusion et perspectives

Pour étudier la sémantique des spécificités dans le domaine technique des machines-outils pour l'usinage des métaux, nous avons eu recours à une double analyse. D'une part, la *KeyWords Method* a permis de dresser la liste des spécificités du corpus d'analyse, ordonnées par ordre de spécificité décroissante. D'autre part, la formule pour le recouvrement des cooccurrences des cooccurrences a permis d'accéder à la sémantique des spécificités, en évaluant le degré de monosémie. L'analyse détaillée des résultats de recherche nous apprendra pour un nombre important de mots techniques, s'il y a une corrélation entre leur degré de spécificité et leur degré de monosémie, en fonction d'une série de facteurs, notamment la classe lexicale et la fréquence.

La formule pour le recouvrement des cooccurrents de deuxième ordre et pour le degré de monosémie sera soumise à plusieurs expérimentations en fonction de plusieurs paramètres, afin de mettre au point le calcul du degré de monosémie. Une fois recueillies pour le corpus technique entier, les données sur le degré de monosémie permettront de situer les spécificités dans un continuum de monosémie (continuum de sens). Des analyses statistiques dans R[8] permettront ensuite de vérifier la corrélation entre le continuum de spécificité et le continuum de monosémie et de procéder à une analyse linguistique détaillée en fonction des variables (linguistiques) indépendantes.

Nous envisageons cette double analyse du degré de spécificité et du degré de monosémie pour les collocations spécifiques également, étant donné que les termes se situent souvent au niveau des unités polylexicales. Nous nous proposons aussi de procéder à une validation manuelle de la formule déterminant le degré de monosémie à l'aide de l'analyse des collocations et cooccurrences relevées.

Références

Audibert L. (2003), Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences, Actes de *TALN 2003*, 35-44.

Berber Sardinha A. (1999), Word sets, keywords and text contents : An investigation of text topic on the computer, *DELTA*, 15-1, 141-149.

[8] http://www.r-project.org/

Cabré M.T. (1998), *La terminologie. Théorie, méthode et applications*, Ottawa, Les Presses de l'Université.

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

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

Drouin P. (2004), Spécificités lexicales et acquisition de la terminologie, Actes de *JADT 2004*, 345-352.

Denhière G., Lemaire B. (2003), Modélisation des effets contextuels par l'analyse de la sémantique latente. Actes de *EPIQUE 2003,* http://www.upmf-grenoble.fr/sciedu/blemaire/epique03.pdf

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

Ferret O. (2004), Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales, Actes de *TALN 2004*, http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Ferret.pdf

Gaudin F. (1993), *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles*, Rouen, Publications de l'Université de Rouen.

Labbé C., Labbé D. (2001), Que mesure la spécificité du vocabulaire?, *Lexicometrica*, 3, http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3/specificite2001.PDF

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

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

Martinez W. (2000), Mise en évidence de rapports synonymiques par la méthode des cooccurrences, Actes de *JADT 2000*, 78-84.

Schütze H. (1998), Automatic Word Sense Discrimination, *Computational Linguistics*, 24-1, 97-123.

Speelman D. (1997), *Abundantia verborum : a computer tool for carrying out corpus-based linguistic case studies*, PhD Thesis, K.U.Leuven.

Temmerman R. (1997), Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology, *Hermes*, 18, 51-90.

Véronis J. (2003), Cartographie lexicale pour la recherche d'informations, Actes de *TALN 2003*, 265-274.