Spécificités lexicales et acquisition de la terminologie

Patrick Drouin

OLST/ ÉCLECTIK, Université de Montréal, C. P. 6128,
Succursale Centre-ville, Montréal (Québec), H2S 2B6, Canada
patrick.drouin@umontreal.ca

Abstract

In this paper, we present a technique that uses corpus specific vocabulary in order to gain access to terminology. The method exploits a dynamically built heterogeneous corpus made of the merger of a technical and a non technical corpus so as to identify specialized vocabulary in the technical corpus. This highly specialized vocabulary (adjectives and nouns), is validated in order to establish its usefulness for terminology processing.

Résumé

Dans cet article, nous présentons une méthode exploitant le calcul de spécificités dans le cadre d'un processus d'identification de la terminologie. La méthodologie proposée repose sur la constitution dynamique d'un corpus hétérogène (technique/non technique) visant à faire ressortir la trace lexicale laissée dans le corpus technique par la terminologie d'un domaine. Cette trace est identifiée à l'aide du calcul des spécificités. Nous procédons à la validation de la pertinence d'un sous-ensemble des spécificités (nominales et adjectivales hautement spécifiques) afin de vérifier leur utilité dans le cadre du travail du terminologue.

Mots-clés : spécificités, terminologie, acquisition automatique de la terminologie, corpus, analyse de fréquence, langue de spécialité.

1. Introduction

Malgré les progrès récents effectués dans le domaine de l'acquisition automatique des termes (Bourigault *et al.*, 2001 ; Jacquemin, 2001), l'élaboration de dictionnaires spécialisés, bien qu'assistée par ordinateur, demeure une tâche ardue et essentiellement manuelle. Le processus de confection de ces dictionnaires repose encore principalement sur le dépouillement d'une masse de documents spécialisés portant sur un domaine du savoir humain. Le phénomène de plus en plus important de libre circulation des documents en format électronique fait en sorte que le terminologue s'attaque à des corpus de plus en plus volumineux. Les divers documents techniques qui composent ces corpus possèdent bien souvent de nombreuses caractéristiques qui les distinguent des documents non techniques (lexique, style, syntaxe, taille, public cible, etc.). La lecture d'un de ces documents par un non-spécialiste l'amène rapidement à constater la très forte présence de termes techniques.

Nous croyons que la trace lexicale laissée par la terminologie peut être exploitée afin de mettre en lumière les éléments terminologiques contenus dans un corpus de documents techniques. Afin de l'exploiter, nous proposons une méthodologie qui consiste à mettre en opposition le comportement des unités lexicales de corpus de niveaux de spécialisation différents. Pour y parvenir, nous utilisons le calcul des spécificités (Lafon, 1980). Nous ne proposons donc pas un nouvel indice en vue de la description de la spécificité des unités lexicales; nous exploitons plutôt un indice bien connu dans un cadre différent. L'objectif du présent article est de vérifier si l'application du calcul des spécificités à corpus hétérogènes permet

l'obtention de résultats satisfaisants et intéressants pour le travail terminologique.

La section 2 constitue un survol rapide des travaux sur les spécificités, de ceux portant sur des notions apparentées dans le cadre de la terminologie, ainsi que du travail de quelques auteurs qui se sont intéressés à la mise en opposition de corpus dans le but d'identifier la terminologie. La section 3 brosse un tableau de la méthodologie adoptée dans le cadre de nos travaux alors que la section 4 porte sur les résultats obtenus à l'aide du logiciel d'acquisition automatique de la terminologie *TermoStat*.

2. Travaux antérieurs

Dans le cadre de la terminologie traditionnelle et de la terminotique, les études ayant recours à la mise en opposition de corpus sont relativement récentes et peu diffusées. Les approches utilisées dans le cadre de la terminologie se fondent essentiellement sur des analyses de fréquence. Ahmad *et al.* (1994) et Chung (2003) proposent des approches mettant en opposition des corpus dans le but d'identifier le vocabulaire propre à la langue médicale anglaise. Pour sa part, Nelson (2000) s'intéresse à l'identification du vocabulaire du monde des affaires. Il ne s'agit pas à proprement parler d'une étude terminologique et les travaux de cet auteur se rapprochent plus de ceux de Phal (1971) et de Huizong (1986) en ce qu'ils cherchent plutôt à décrire les caractéristiques du vocabulaire d'un ou plusieurs secteurs.

La majorité des études citées précédemment opposent, d'un point de vue de la fréquence d'occurrence, un corpus restreint à un domaine particulier du savoir à un corpus dit « général », composé de documents non liés à un domaine spécifique. Ces études portent sur des documents n'ayant pas l'objet de traitements linguistiques (étiquetage grammatical, lemmatisation, etc.). L'enrichissement des documents peut conduire à des analyses plus fines permettant de distinguer les formes graphiquement identiques, mais possédant des catégories grammaticales différentes. La lemmatisation, pour sa part, permet de regrouper les occurrences d'une forme et d'obtenir une meilleure idée de son importance dans les corpus.

Même si elles exploitent essentiellement la fréquence d'occurrence des formes d'un corpus, les techniques utilisées jusqu'à maintenant sont difficilement comparables. De plus, la variété des seuils de pertinence sélectionnés ne facilite pas la tâche en vue d'une étude comparative. À titre d'exemple, Chung (2003) considère que les formes qui apparaissent 50 fois plus souvent dans son corpus de médecine que dans son corpus général sont dignes d'intérêt. À notre avis, une approche probabiliste, s'éloignant des observations fondées sur les fréquences brutes d'occurrence, pourrait être plus facilement adaptable d'un domaine à un autre et d'une démarche à une autre. Même si l'approche proposée par Lafon (1980) a suscité de l'intérêt dans le cadre du travail en lexicographie (Leselbaum et Labbé 2002 ; Zimina 2002), à notre connaissance, aucune étude n'a envisagé d'avoir recours au calcul des spécificités en terminologie. Nous proposons, dans cet article, une méthode ayant pour but d'évaluer l'utilité des spécificités pour ce type de travail.

3. Méthodologie

3.1. *Description des corpus*

La démarche proposée requiert la constitution de deux corpus, un corpus de référence (CR) et un corpus d'analyse (CA). Afin de tester la stabilité de l'approche, nous reproduirons les tests sur trois corpus d'analyse nommés CA_1, CA_2 et CA_3. Tous les corpus analysés sont en anglais.

La taille totale du corpus de référence est d'environ 7 400 000 occurrences, qui correspondent à approximativement 82 700 formes différentes. Le CR est un corpus non technique composé de 13 746 articles de journaux portant sur des sujets variés tirés du quotidien montréalais The Gazette et publiés entre mars 1989 et mai 1989. Cette diversité de thèmes traités est importante et nécessaire à notre démarche puisqu'elle vient minimiser l'uniformité thématique du CR. On ne peut, bien sûr, s'assurer entièrement qu'un corpus journalistique ne comporte aucune thématique dominante. En effet, les articles qui composent le quotidien sont nécessairement liés à l'actualité et ainsi, à de grandes thématiques sociales. On pourrait aussi envisager de constituer un corpus plus équilibré à partir d'échantillons provenant de documents tirés domaines différents et de documents plus généraux.

Les corpus d'analyse sont de nature technique et correspondent à un seul document. Le CA_1 comporte 11 947 occurrences (1 207 formes), le CA_2 28 583 occurrences (2 066 formes) et le dernier corpus d'analyse (CA_3) est composé de 8 676 occurrences (1 053 formes). Ces corpus présentent donc un éventail de tailles qui rendra possible la validation de la méthodologie sur des ensembles textuels qui possèdent des caractéristiques différentes. Les corpus utilisés pour les expérimentations sont petits, mais nous croyons que leur taille peut être déterminée en fonction des objectifs de travail. Ainsi, dans le cas des documents qui composent les corpus d'analyse, leur taille doit correspondre à un échantillon représentatif traité par les terminologues en situation de travail. Cet objectif, adopté au début de nos travaux, impose une restriction considérable sur le corpus. La taille des corpus d'analyse est donc avant tout dictée par des critères externes.

Bien qu'il soit difficile de classifier catégoriquement un document comme relevant d'un seul domaine de l'activité humaine, nous considérons que les corpus d'analyse traitent du domaine des télécommunications. La nature multidisciplinaire de ce domaine conduit cependant à l'inclusion de concepts venus de différents domaines. Même si celui des télécommunications sert de dénominateur commun aux corpus d'analyse, le corpus CA_1 traite de l'interface de programmation de composantes informatiques. Pour leur part, les corpus CA_2 et CA_3 abordent d'un sujet plus étroit au sein du domaine des télécommunications, celui de la structure physique des réseaux de fibres optiques et de leurs composantes. Le corpus CA_1 s'adresse à des informaticiens qui conçoivent des applications destinées aux composantes présentées dans les documents CA_2 et CA_3. Ces derniers sont rédigés pour des intervenants du domaine des télécommunications ayant une bonne connaissance de la structure physique des réseaux de fibres optiques. Leur public cible est principalement constitué d'architectes de réseaux, d'installateurs, de réparateurs, d'ingénieurs, de testeurs, d'administrateurs de réseaux, etc. Les documents décrivent les possibilités, les caractéristiques, l'entretien, l'utilisation et l'installation des éléments d'un tel réseau.

3.2. *Préparation de corpus*

La première étape de traitement est une segmentation des corpus. L'algorithme de segmentation utilisé est fondé sur celui placé dans le domaine public par Robert MacIntyre de la *University of Pennsylvania*. Le corpus est ensuite étiqueté, sans entraînement préalable, à l'aide de l'étiqueteur conçu par Éric Brill (1992).

Les corpus sont soumis à une étape de lemmatisation heuristique, reposant sur des observations empiriques sur corpus. L'algorithme de lemmatisation consiste à identifier une forme nominale (couple forme/partie du discours ; ex. : *matrices/NNS*), à vérifier si elle comporte un suffixe potentiellement pluriel (ex. : *-ices*), à en retrancher une partie (ex. : -ces), à ajouter un suffixe singulier (ex. : -x), et à rechercher un couple correspondant (ex. : *matrix/NN*) dans la

liste des couples identifiés dans le corpus. Si un couple correspondant est repéré, on considère que la lemmatisation a été effectuée avec succès.

Règle	Suffixe	Retranché	Ajouté	Long. min.	Exemple
1	*-ices*	*-ces*	*-x*	5	*matrices / matrix*
2	*-ives*	*-ves*	*-fe*	5	*knives / knife*
3	*-sses*	*-es*		5	*accesses / access*
4	*-ches*	*-es*		5	*switches / switch*
5	*-eet*	*-eet*	*-oot*	4	*feet / foot*
6	*-ies*	*-ies*	*-y*	4	*possibilities / possibility*
7	*-i*	*-I*	*-us*	4	*stimuli / stimulus*
8	*-s*	*-s*		4	*cars / car*

Tableau 1. Règles de lemmatisation

Nous rejoignons ici la position de Brill (1994) ainsi que de Bourigault et Gonzalez (1994) qui adoptent une approche par apprentissage endogène exploitant le contenu d'un corpus afin de déduire des informations relatives aux autres éléments du corpus. Les règles de lemmatisation présentées dans le tableau 1 nous permettent d'obtenir des résultats satisfaisants. Une analyse d'un échantillon de 1 000 formes nominales prélevées au hasard conduit à une bonne lemmatisation dans 98,7 % des cas.

3.3. *Identification des spécificités*

Nous avons procédé à l'identification des spécificités à l'aide du logiciel d'acquisition automatique des termes TermoStat (Drouin, 2003). La première étape de traitement du corpus par le logiciel en vue de l'extraction des termes est l'identification des spécificités. Ces dernières sont ensuite utilisées à titre de pivots pour la recherche de termes spécifiques au CA. Certaines contraintes sont cependant appliquées au cours du processus d'acquisition et le logiciel ne relève que les spécificités positives nominales et adjectivales. Cette contrainte, bien que très importante, se justifie par la vocation ultime du logiciel et par la nature des termes en langue anglaise qui sont très majoritairement constitués de substantifs et d'adjectifs.

En vue de l'identification des spécificités, TermoStat procède à la constitution dynamique d'un corpus global hétérogène en fusionnant virtuellement le corpus de référence et le corpus d'analyse. Il s'agit ici d'une utilisation inhabituelle du calcul des spécificités qui porte généralement sur un sous-corpus dans le but d'identifier ses spécificités par rapport au corpus d'où il est issu. Nous introduisons donc volontairement un document (le CA) à titre de sous-corpus au sein d'un corpus relativement uniforme (le CR) afin de vérifier dans quelle mesure le comportement des unités lexicales du CA se démarque de ce que l'on observe dans le corpus de référence. Nous avons implémenté au sein du logiciel le calcul de spécificités par approximation normale du calcul hypergéométrique décrit dans Lebart et Salem (1994 : 182).

Du sous-ensemble des spécificités identifiées (nominales et adjectivales), nous restreignons à nouveau le bassin de spécificités et ne retenons que les formes dont la valeur-test est supérieure ou égale à 3,09. Ce seuil minimal nous permet de retenir les formes pour lesquelles nous avons moins d'une chance sur 1 000 d'observer une fréquence égale ou supérieure à celle constatée au sein dans le corpus d'analyse. Il s'agit donc de formes qui sont très forte-

ment représentées au sein du CA et qui, selon nous, devraient être en relation directe avec la trace lexicale laissée par la terminologie dans ce corpus. Aucune contrainte de fréquence minimale n'est imposée aux formes retenues.

4. Résultats

4.1. *Processus de validation*

Notre objectif est de déterminer si les résultats obtenus à l'aide du calcul des spécificités peuvent être utilisés par le terminologue et, si c'est le cas, dans quelle mesure ils peuvent l'être. Afin de valider les données issues de l'acquisition des spécificités, nous avons recours à une banque de terminologie et à des terminologues spécialistes du domaine des télécommunications. La banque de terminologie nous a été fournie par la société *Nortel Networks*, qui a aussi mis à notre disposition les corpus d'analyse. La banque de terminologie comporte essentiellement de la terminologie du domaine des télécommunications. La validation à l'aide de la banque de terminologie consiste en une comparaison *à plat* des listes de spécificités construites à partir des documents qui composent le CA et de la liste extraite de la banque de terminologie. Les formes spécifiques qui sont présentes au sein de la banque de terminologie sont considérées comme pertinentes alors que les autres sont ensuite soumises à une validation humaine afin de juger de leur pertinence. La comparaison purement orthographique effectuée possède des avantages et des inconvénients, mais l'étroitesse du domaine et l'origine commune des documents du CA et de la banque de terminologie nous laissent penser qu'il s'agit d'une approche suffisamment fiable.

Les consignes données aux terminologues pour la validation des formes spécifiques sont simples et ils doivent se limiter à évaluer deux aspects : la pertinence de la forme pour le corpus d'analyse et sa pertinence pour le domaine des télécommunications. Ainsi, si la forme est utilisée dans le domaine des télécommunications ou si elle est représentative du contenu du document, elle est alors considérée comme valide. Ces consignes étendent le champ de validité d'une spécificité à l'ensemble d'un domaine et non seulement au corpus. Certaines unités lexicales pourraient en effet paraître banales au sein d'un corpus, mais elles n'en demeurent pas moins essentielles du point de vue de la terminologie d'un domaine. En effet, le calcul des spécificités ne peut être utilisé que pour déterminer la pertinence d'une forme par rapport à un corpus particulier tiré, dans le cadre de notre démarche, d'un domaine d'activité plus ou moins spécifique.

4.2. *Présentation des résultats*

Le tableau 2 dresse la liste triée en ordre décroissant de valeur-test des 15 premières spécificités pour chacun des trois corpus d'analyse. On remarque que les abréviations sont très nombreuses (*OC*, *OPC*, *ID*, *SDH*, etc.), mais on y trouve aussi des formes pleines et en apparence moins spécifiques (*interface, parameter, amplifier*, etc.) .

Pour sa part, le tableau 3 présente la précision du processus d'acquisition des spécificités, telle qu'elle a été évaluée par l'équipe de terminologues, pour les trois documents qui composent le corpus d'analyse. La bonne performance obtenue doit être interprétée en contexte et en fonction des consignes données lors de l'étape de validation. Ces formes sont, selon les spécialistes, représentatives du corpus ou du domaine des télécommunications.

CA_1		CA_2		CA_3	
Forme	Valeur-test	Forme	Valeur-test	Forme	Valeur-test
interface	192.67	*amplifier*	283.15	*optical (n.)*[i]	257.33
oc	181.02	*optera*	234.11	*opc*	245.26
parameter	173.90	*module*	230.06	*optical (adj.)*	245.00
threshold	167.34	*haul*	211.91	*mor*	226.43
id	158.98	*dwdm*	167.35	*optera*	199.27
ne	150.23	*nm*	164.44	*sdh*	177.78
pm	140.18	*fiber*	163.03	*sonet*	170.81
objectid	137.23	*osc*	162.63	*osc*	169.11
pmbb	137.23	*long*	160.07	*amplifier*	160.26
dn	135.14	*wavelength*	159.82	*input*	154.82
ttp	130.85	*grid*	153.21	*span*	151.46
stm	122.32	*shelf*	150.70	*network*	145.16
invokeid	121.82	*band*	150.39	*haul*	145.15
equipmentid	114.58	*am2*	141.56	*orl*	142.35
attributeid	112.06	*optical*	141.28	*output*	135.32

Tableau 2. Présentation des 15 premières spécificités pour les 3 corpus d'analyse

	CA_1	CA_2	CA_3
Spécificités pertinentes	444	810	273
Spécificités non pertinentes	84	131	101
Précision	84,1 %	86,1 %	73,0 %

Tableau 3. Évaluation de la pertinence des spécificités pour les CA

Tel que nous l'avons mentionné auparavant, les spécificités dont la valeur-test était inférieure à 3,09 n'ont pas fait l'objet d'une validation par les terminologues. On retrouve, dans cette liste, des formes comme *time, rate, process* dans le CA_1, *house, loss, exchange* dans le CA_2 ou encore *point, building, state* et *manager* dans le CA_3. Ces formes sont typiques des documents liés au domaine des télécommunications, mais la mise en opposition des fréquences observées dans les corpus ne leur permet pas de se distinguer suffisamment dans le CA. Dans tous les cas, il s'agit de formes polysémiques ayant un sens non technique (mot) et un sens relevant du domaine des télécommunications (terme). La spécificité de ces formes est donc sémantique et non purement lexicale. Le calcul des spécificités, sans étiquetage sémantique des formes, ne permet malheureusement pas d'identifier cette particularité.

[i] La forme *optical* apparaît deux fois dans cette liste à titre de substantif et d'adjectif. Il s'agit ici d'une erreur d'étiquetage attribuable aux outils informatiques utilisés.

Corpus	Fréquence<=5	Fréquence <=10	Fréquence >10
CA_1	332	430	98
CA_2	664	748	193
CA_3	247	300	74

Tableau 4. Répartition générale des spécificités en fonction de la fréquence

Labbé et Labbé (2001) ont démontré que la fiabilité du calcul des spécificités diminue lorsque la fréquence des événements considérés est basse. Le Tableau 4 donne un aperçu rapide de la répartition des spécificités en fonction de la fréquence dans les corpus traités. On remarque que la majorité des formes recensées ont une fréquence inférieure à 5 (63 % pour CA_1, 71 % pour CA_2 et 66 % pour CA_3). Il est intéressant de noter que notre évaluation par des terminologues de l'intérêt des données, de nature qualitative plutôt que quantitative, met en évidence leur utilité dans le cadre d'une démarche terminologique. L'importance accordée ici à la précision des résultats ne permet pas d'évaluer les performances de l'approche en ce qui concerne le rappel. Il serait intéressant de mesurer, du point de vue du terminologue, l'impact du silence. Cette problématique, beaucoup plus difficile à manipuler puisqu'elle nécessite un dépouillement systématique des corpus d'analyse, devra être abordée dans des études subséquentes.

5. Conclusion

Nous avons présenté une méthode exploitant le calcul de spécificités dans le cadre d'un processus d'identification de la terminologie. Les spécificités sont utilisées comme point de départ pour l'acquisition automatique de la terminologie. Nous avions comme objectif de mesurer la pertinence des spécificités par rapport à un corpus et à un domaine technique. La méthodologie proposée repose sur la constitution dynamique d'un corpus hétérogène visant à faire ressortir la trace lexicale laissée dans un corpus technique par la terminologie. Cette trace lexicale est identifiée à l'aide du calcul des spécificités.

Nous avons procédé à la validation de la pertinence d'un sous-ensemble des spécificités (nominales et adjectivales hautement spécifiques) et il en ressort que, malgré les limites d'une telle approche purement lexicale, ces dernières sont jugées comme utile pour le travail du terminologue. Étant donné les bonnes performances obtenues avec le calcul de spécificités en opposant des corpus de nature différente, nous envisageons de poursuivre nos travaux sur l'acquisition automatique des termes en utilisant les spécificités à titre de pivots pour le recensement des termes.

Références

Ahmad K., Davies A., Fulford H. et Rogers M. (1994). What's in a Term? The semi-automatic Extraction of Terms from Text. In Snell-Hornby, dans Pochhacker M.F. et Kaindl K. (Eds), *Translation Studies. An Interdiscipline.* John Benjamins.

Bourigault D., Jacquemin C. et L'Homme M.C. (Eds) (2001). *Recent Advances in Computational Terminology*. John Benjamins.

Bourigault D. et Gonzalez I. (1994). Acquisition automatique des termes complexes en français et en anglais, approche comparative. In *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition* : 29-43.

Brill E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of*

the 12th National Conference on Artificial Intelligence (AAAI-94) : 722-727.

Brill E. (1992). A Simple Rule-based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied -Natural Language Processing ANLP-1992* : 152-155.

Chung T. M. (2003). A Corpus Comparison Approach for Terminology Extraction. *Terminology*, vol. (9/2). A paraître.

Drouin P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology,* vol. (9/1) : 99-115.

Huizong Y. (1986). A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts. *Literary and Linguistic Computing,* vol. (1/2) : 93-103.

Jacquemin C. (2001). *Spotting and Discovering Terms through Natural Language Processing Techniques*. MIT Press.

Labbé C. et Labbé D. (2001). Que mesure la spécificité du vocabulaire? *Lexicometria*, vol. (3).

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, vol. (1) : 128-165.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Leselbaum J. et Labbé D. (2002). Lexicographie assistée par ordinateur. Signification de « banque » dans le vocabulaire économique In *Actes des JADT 2002* : 447-456.

Nelson M. (2000). *A Corpus-based Study of Business English and Business English Teaching Materials*. Unpublished PhD Thesis, University of Manchester.

Phal A. (1971). *Vocabulaire général d'orientation scientifique*. Crédif.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. *Lexicometria*, n° spécial.