Acquisition des termes simples fondée sur les pivots lexicaux spécialisés

Patrick Drouin

Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7
patrick.drouin@umontreal.ca
http://mapageweb.umontreal.ca/drouinp

Résumé

Le présent article décrit une technique d'acquisition automatique des termes reposant sur les spécificités lexicales des corpus techniques. Plus spécifiquement, nous nous intéressons à l'acquisition automatique des termes simples en langue anglaise et au gain en précision réalisé grâce à la méthodologie proposée. Nous donnons tout d'abord une description de la méthodologie utilisée pour l'acquisition des spécificités lexicales des corpus. Par la suite, nous nous proposons une stratégie d'acquisition automatique de termes qui exploite ces spécificités. Enfin, nous présentons les corpus utilisés dans le cadre de notre démarche ainsi que les résultats obtenus.

1. Introduction

Cet article décrit une technique d'acquisition automatique des termes reposant sur l'opposition du lexique de corpus de nature différente (Drouin 2002). Au cœur de la stratégie proposée se trouve la notion de vocabulaire spécifique. Un sous-ensemble de ce lexique, qui constitue ce que nous nommons les *pivots lexicaux spécialisés* (*PLS*), est utilisé pour l'acquisition des candidats-termes (CT). Plus spécifiquement, cet article s'intéresse à l'acquisition automatique des termes simples en langue anglaise et au gain en précision réalisé grâce à l'utilisation des PLS.

Nous présentons tout d'abord la méthodologie utilisée pour identifier les PLS, le processus d'acquisition des termes ainsi que les corpus utilisés. La seconde partie consiste en une analyse des résultats ; nous y décrivons l'étape de validation des candidats-termes et procédons à une évaluation de la précision et du rappel obtenus à l'aide du logiciel TermoStat, qui a été élaboré afin de tester l'approche par PLS. Le logiciel s'intègre au sein d'une démarche terminologique assistée par ordinateur et se veut un outil de dépouillement automatique de corpus techniques.

2. Pivots lexicaux spécialisés

Les travaux sur les spécificités lexicales des corpus sont issus des expérimentations de Muller (1977, 1979) sur le lexique. Ces travaux ont inspiré depuis de nombreuses recherches,

dont celles de Lafon (1980), de Lebart et Salem (1994) et de Camlong (1996). L'identification des particularités lexicales de sous-corpus permet d'observer une foule de phénomènes utiles à divers domaines dont la stylistique, la lexicologie, l'analyse du discours, etc. Ces travaux n'ont cependant pas été exploités dans le cadre d'une démarche terminologique visant l'acquisition automatique des termes ; c'est l'objectif que nous nous fixons dans le cadre de la présente recherche.

Les travaux sur les spécificités lexicales reposent sur l'opposition du lexique de sous-corpus en vue d'en faire ressortir les divergences lexicales. A notre avis, une comparaison des fréquences d'occurrences des mots entre deux corpus (un corpus dit de référence de type journalistique et un corpus dit d'analyse de type technique) permettra de faire ressortir les spécificités lexicales du *corpus d'analyse* (*CA*) et que ces dernières sont étroitement liées à la terminologie du corpus technique.

Pour fins de description des cas de figure possibles, on considère que le *corpus de référence* (*CR*) sert de point de départ pour les observations. Une comparaison des fréquences dans les deux corpus conduit à l'observation d'une fréquence supérieure, inférieure ou identique à la fréquence théorique projetée à partir du corpus de référence.

Jusqu'à maintenant, les travaux sur les spécificités ont été effectués sur des corpus statiques et l'identification des spécificités a uniquement été faite sur des parties (sous-corpus) de ces corpus. Notre recherche repose sur un corpus dynamique qui est mis en forme lors de l'analyse conduisant à l'acquisition des termes et qui n'existe que d'un point de vue informatique. Ce corpus, nommé *corpus global* (*CG*), est composé du CR et du CA. La dichotomie dans le type de discours est exploitée pour mettre en évidence les particularités lexicales du CA par rapport au CG.

Afin d'isoler les spécificités, nous utilisons une technique qui s'inspire de celle décrite par Muller (1977, 1979) ainsi que par Lebart et Salem (1988, 1994). Le calcul des spécificités met à notre disposition deux indices qui peuvent être utilisés pour l'analyse des résultats : la *valeur-test* et la *probabilité* ; nous reprenons ici la terminologie de Lebart et Salem (1994 : 317). Nous avons décidé d'adopter les valeurs-tests pour regrouper les formes plutôt que les probabilités puisque les valeurs calculées pour ces dernières sont beaucoup moins nuancées et notre analyse perdrait ainsi beaucoup de subtilité. En effet, l'examen d'une table des probabilités reliées aux valeurs-tests permet de constater que, pour des valeurs supérieures à 3,09, les probabilités sont inférieures à 0,001 (1/1 000) et que les écarts sont de plus en plus petits et difficiles à cerner. L'utilisation des valeurs-tests permet aussi d'avoir une meilleure idée des écarts de comportement des diverses formes et d'apporter des nuances là où les probabilités ne permettent pas de le faire.

Nous proposons une nouvelle notion, le *pivot lexical spécialisé* (*PLS*), qui correspond à un sous-ensemble du lexique isolé par le calcul des spécificités. Les spécificités qui nous intéressent sont celles dont la fréquence se démarque de façon positive dans CA, les formes surreprésentées.

Afin de ne retenir que les formes intéressantes et très significatives, nous nous attardons sur celles dont les valeurs-tests sont supérieures à 3,09, comme le suggèrent Lebart et Salem (1994 : 183). L'adoption de ce seuil nous permet d'assurer qu'il n'y a que 1 chance sur 1 000 que la fréquence observée dans CA soit due au hasard. À titre de comparaison, mentionnons qu'un seuil de 1,96 correspond à une probabilité de 5 % alors qu'un seuil de 2,33 correspond à une probabilité de 1 %.

Nos visées terminologiques nous amènent à imposer une contrainte supplémentaire aux unités identifiées à l'aide des critères précédents et nous limitons les PLS aux formes

nominales et adjectivales.

3. Acquisition automatique des termes

L'acquisition automatique des termes est prise en charge par le logiciel TermoStat. Ce dernier a été conçu spécifiquement pour tester la méthodologie d'acquisition des termes fondée sur les PLS. Le recensement de ces derniers correspond d'ailleurs la première étape d'analyse des corpus et il constitue le noyau central du processus d'acquisition.

Le point de départ de l'algorithme est l'utilisation d'une forme nominale tirée de la liste des PLS comme tête des CT. Comme l'ont démontré les travaux de Bourigault (1994) et de Assadi et Bourigault (1996), la productivité de la tête est un bon indice de la qualité des CT. À notre avis, l'utilisation de la fréquence de la tête comme critère décisif pour l'acquisition des termes ne saurait être suffisante en elle-même; c'est pourquoi nous proposons d'utiliser les résultats des tests statistiques qui mettent en évidence leur pertinence pour le corpus. Cette sélection des têtes permet ainsi d'éliminer dès le départ des CT potentiels qui se construiraient à partir de têtes fréquentes mais non pertinentes pour le corpus à l'étude.

Le processus d'acquisition repose sur le principe de frontière de terme tel que défini par Bourigault (1992). Notre conception des frontières de termes diffère légèrement de celle décrite dans les travaux de cet auteur. Notre conception des frontières de termes est plus lâche et les règles qui sont appliquées afin de déterminer ce qui constitue une frontière sont plus simples, car elles n'impliquent pas d'analyse syntaxique locale.

Les règles utilisées dans le cadre de notre démarche sont donc moins contraignantes. Par contre, une contrainte probabiliste vient s'ajouter : toutes les formes au sein d'un candidat-terme (CT) doivent faire partie de la liste des PLS. Cette nouvelle dimension s'ajoute à la définition de la frontière et des adjectifs et des substantifs peuvent ainsi être considérés comme des frontières. Cette utilisation des PLS à titre de noyau de l'acquisition automatique des termes constitue un des éléments novateurs de l'approche proposée. Les algorithmes d'acquisition mis en place prennent ainsi en considération la spécificité des unités lexicales pour le corpus qui fait l'objet de l'analyse.

En plus d'une modification du concept initial de frontière, nous procédons, d'une certaine façon, à son extension en prenant en compte de l'information telle que la ponctuation, les éléments de mise en page (cellule de tableau), etc. La conception de frontière exploitée par TermoStat repose donc à la fois sur une vision linguistique, probabiliste et textuelle du fonctionnement des termes en corpus.

Afin de contraindre la longueur des CT, nous imposons à l'algorithme une fenêtre de repérage. Cette dernière correspond au nombre de mots à gauche de la tête potentielle que le logiciel peut explorer. Les limites de l'expansion du terme n'étant pas déterminées à partir de critères logiques, on ne peut se fier à des critères purement syntaxiques afin de déterminer où se termine l'expansion d'un terme. Les travaux effectués sur l'acquisition automatique des termes, dont ceux de Justeson et Katz (1993) et de Nkwenti-Azeh (1994), nous ont amené à limiter la longueur des CT identifiés par TermoStat à six formes.

Une dernière contrainte est imposée à l'algorithme d'acquisition : la contrainte d'autonomie des CT. Cette dernière consiste à s'assurer que les CT atteignent un certain degré de fonctionnement linguistique autonome dans le corpus d'analyse. En comparant les CT recensés, les recoupements entre ces CT et leur fréquence respective, le logiciel TermoStat procède à un élagage de CT. La technique consiste à examiner chacun des candidats et à vérifier s'il est utilisé à titre de tête ou d'expansion d'un candidat plus long. Si

c'est le cas, la fréquence des deux CT sont comparées. Les candidats qui sont systématiquement inclus dans ces CT plus longs sont considérés comme des fragments de CT plus longs et sont éliminés de la liste finale présentée à l'utilisateur. Cette dernière contrainte entraîne donc une diminution du nombre de CT simples retenus par TermoStat.

4. Corpus

La première étape de traitement est la segmentation des corpus ;l 'algorithme de segmentation est fondé sur celui versé dans le domaine public par Robert MacIntyre de la University of Pennsylvania. Le corpus est ensuite étiqueté à l'aide de l'étiqueteur conçu par Éric Brill (Brill 1992) qui offre un bon niveau de fiabilité (Brill 1994 : 726 ; Habert *et al.* 1997 : 170). Le processus d'apprentissage a été laissé de côté dans le cadre de notre recherche et les textes ont été utilisés sans être revus.

Les corpus sont ensuite soumis à une étape de lemmatisation qui repose sur des heuristiques issues d'observations empiriques sur corpus. Au cours de cette étape, les ressources du corpus sont exploitées au maximum puisque l'information qu'il contient est utilisée pour prendre des décisions relatives au statut de formes tirées, elles aussi, du corpus. On rejoint ici la position de Brill (1994 et 1995) ainsi que de Bourigault et Gonzalez (1994) qui adopte une approche par apprentissage endogène. Un ensemble restreint de huit règles de lemmatisation a été identifié pour affiner rapidement les résultats de l'étiquetage. Ces règles simples nous permettent d'obtenir des résultats satisfaisants. Ainsi, une analyse des résultats sur un échantillon de 1 000 formes nominales prélevées au hasard conduit à une bonne lemmatisation dans 98,7 % des cas.

4.1. Corpus de référence

Le corpus de référence est composé d'articles tirés du quotidien *The Gazette* publié à Montréal entre mars 1989 et mai 1989. Le CR est constitué de 13 746 articles de journaux portant sur des sujets variés. Cette variété de sujets est importante et nécessaire à notre démarche puisqu'elle vient minimiser l'uniformité thématique du CR. Cet aspect hétérogène permet au corpus de référence de se distinguer du corpus d'analyse. En effet, ce dernier est plus uniforme du point de la thématique abordée. La taille totale du CR est d'environ 7 400 000 occurrences, qui correspondent à environ 82 700 formes différentes.

4.2. Corpus d'analyse

Les documents qui font l'objet du dépouillement terminologique ont été mis à notre disposition par le groupe *Optical Networks* de la société *Nortel Networks*. Les documents trois ont été rédigés au cours de l'année 2000. Afin de les désigner, la notation suivante est utilisée : CA_1, CA_2 et CA_3.

Il est difficile de classifier catégoriquement un document comme relevant d'un seul domaine de l'activité humaine. La nature multidisciplinaire de ce domaine conduit à l'inclusion de concepts venus du domaine des télécommunications, de la physique optique, de l'informatique, etc. Par contre, pour simplifier la discussion, nous adoptons le point de vue de Kageura (1999 : 29-30), qui considère que la classification en domaines, bien que naïve, est extrêmement utile. Ainsi, nous considérons que le corpus d'analyse relève du domaine des télécommunications.

Corpus	Nombre d'occurrences	Nombre de formes

CA_1	11 947	1 207
CA_2	28 583	2 066
CA_3	8 676	1 053

La taille des documents analysés varie considérablement. Cette variation permet de tester la stabilité des algorithmes dans diverses conditions. On pourrait cependant s'objecter à l'utilisation de documents de petite taille. Nous croyons que la taille d'un corpus doit être déterminée en fonction des objectifs de travail. Étant que notre démarche s'insère dans le cadre d'une entreprise, la taille des corpus doit correspondre à un échantillon représentatif traité par les terminologues en situation de travail. Notre objectif impose donc une restriction significative sur le corpus et la taille du corpus d'analyse est ainsi dictée par des critères externes. Jusqu'où pouvons nous pousser cette logique et où devons-nous placer le seuil quant à la taille des corpus? La constitution d'un corpus est avant tout un exercice d'équilibre qui s'insère dans un ensemble de contraintes qui doivent être soupesées afin de déterminer une taille qui répond aux objectifs visés.

5. Analyse des résultats

Comme nous l'avons mentionné en introduction, nous nous intéressons dans cet article, à la performance de l'algorithme d'acquisition pour les termes simples. Étant donnée les contraintes imposées lors de l'acquisition, ces derniers correspondent à un sous-ensemble des PLS nominaux, ce sont les PLS autonomes ou les têtes productives. La présente section décrit le processus de validation des résultats ; les performances de TermoStat sont ensuite évaluées à l'aide des indices bien connus que sont la précision et le rappel.

3.5.1. Validation des résultats

Afin d'être à même d'évaluer les performances de TermoStat, les CT sont soumis à un processus de validation en deux étapes. Les CT sont d'abord validés automatiquement à l'aide d'une banque de terminologie spécialisée en télécommunications mise à notre disposition par la société *Nortel Networks.*. La liste de termes mise à la disposition de TermoStat par la banque de terminologie est une liste dite à plat, c'est-à-dire que cette liste ne comporte pas d'information sémantique permettant de confirmer hors de tout doute que le CT relevé correspond au terme de la banque de terminologie. Les CT sont donc comparés, sur le plan de la graphie, avec la liste des termes de la banque de terminologie. Cette démarche s'inspire de celle décrite dans Daille (1994 : 123-124). Étant donné l'étroitesse du domaine et la corrélation entre les documents du corpus d'analyse et la banque de terminologie consultée, nous croyons qu'il s'agit d'une source fiable pour assurer la qualité des décisions prises par TermoStat.

Afin d'assurer la validation des CT recensés qui n'apparaissent pas dans la banque de terminologie, nous avons eu recours à trois terminologues experts du domaine des télécommunications. Les experts avaient à leur disposition une interface comportant la liste des CT ainsi que l'ensemble des occurrences de ces CT en contexte.

3.5.2. Précision

La variation de précision entraînée par le recours à une approche reposant sur les PLS ne peut être évaluée que dans la mesure où les résultats sont comparés aux résultats d'un logiciel qui utilise une technique différente. Cherchant à minimiser les sources d'influences possibles sur les CT retenus, le logiciel TermoStat a été modifié afin de procéder à l'acquisition automatique des termes sans la contrainte des PLS. Cette version du logiciel utilise donc toute unité nominale comme une tête de terme potentielle.

Le tableau qui suit illustre la variation de précision pour les trois corpus à l'étude. Les pourcentages correspondent à la variation de précision lorsque les PLS sont utilisés en comparaison de l'approche sans PLS. Ainsi, pour une fréquence d'occurrence minimale de 2, on note une augmentation de la précision de 7,35 % dans le cas du corpus CA_2. Comme le démontre le tableau, l'approche par PLS conduit, pour le repérage des termes simples, à une augmentation intéressante de la précision pour des fréquences minimales égales ou inférieures à 5. On note un apport positif de l'approche par PLS sur l'ensemble de la gamme des fréquences pour CA_2. Cette excellente performance est en partie due au fait que le corpus comporte une série de formes simples semblables à des acronymes, qui servent à désigner des caractéristiques techniques et qui ne se retrouvent pas dans les autres documents. Sans l'apparition de ces formes, les performances du prototype seraient comparables sur l'ensemble des documents du corpus d'analyse.

Fréq. min.	CA_1	CA_2	CA_3
1	+6,83 %	+16,21 %	+5,78 %
2	+4,65 %	+7,35 %	+6,30 %
3	+2,68 %	+4,86 %	+3,27 %
4	+1,20 %	+4,70 %	+1,54 %
5	+0,29 %	+4,59 %	+1,10 %
6	-0,31 %	+3,34 %	-0,13 %
7	-0,20 %	+1,74 %	0,00 %
8	-0,08 %	+0,69 %	0,00 %
9	-0,03 %	+0,11 %	0,00 %
10	-0,04 %	+0,30 %	0,00 %

On peut affirmer que pour l'acquisition des termes simples, l'apport des PLS se situe principalement au niveau des basses fréquences. Ces derniers permettent de recenser des formes qui passent habituellement inaperçues aux yeux des terminologues. En effet, l'identification de ces termes moins fréquents représente un défi pour le terminologue dans des corpus volumineux. L'approche par PLS facilite donc le recensement de ces formes, souvent laissées de côté par l'humain ou par les techniques d'acquisition automatique de termes exploitant un seuil minimal de fréquence.

Labbé et Labbé (2001) ont démontré que la fiabilité du calcul des spécificités diminue lorsque la fréquence des événements considérés est basse. Nos propres expérimentations ont aussi permis de mettre en évidence une certaine instabilité des résultats de basse fréquence lorsque divers corpus de référence sont utilisés. Malgré ces constatations, une évaluation qualitative des données, plutôt que purement quantitative, permet de démontrer leur utilité dans le cadre d'une démarche terminologique. L'évaluation de la pertinence des termes

simples recensés par TermoStat le démontre bien. Le tableau qui suit donne un aperçu du niveau de précision atteint par le logiciel TermoStat lors de l'acquisition des termes simples. Ces valeurs correspondent aux variations indiquées dans le tableau précédent mais laissent transparaître le degré de qualité des résultats obtenus.

Fréq. min.	CA_1	CA_2	CA_3
1	91,10 %	85,11 %	82,38 %
2	93,08 %	84,44 %	87,74 %
3	95,34 %	85,77 %	91,89 %
4	95,12 %	87,39 %	91,01 %
5	94,37 %	87,30 %	91,89 %
6	93,85 %	87,21 %	91,80 %
7	95,45 %	86,93 %	91,38 %
8	96,08 %	87,94 %	93,75 %
9	97,56 %	87,88 %	95,00 %
10	97,18 %	86,60 %	93,94 %

La fréquence absolue est généralement considérée comme un indice permettant de déterminer la validité des CT. Cependant, une démarche qui prend aussi en considération la probabilité d'occurrence des formes semble conduire à de meilleurs résultats pour les basses fréquences. Afin de maximiser la précision, on peut donc envisager d'avoir recours à une approche par PLS pour l'acquisition des CT peu fréquents et à une approche utilisant un simple seuil de fréquence pour les fréquences les plus élevées.

3.5.3. Rappel

L'évaluation du rappel obtenu par un logiciel sous-entend la disponibilité d'une mesure étalon, de documents dépouillés pouvant servir de point de référence. Étant donné l'absence de ce point de comparaison pour la présente recherche, nous adoptons le dépouillement effectué par le prototype d'acquisition sans contrainte comme point de référence.

Le point de départ du processus d'acquisition automatique des termes, pour l'algorithme évoluant sans la contrainte des PLS, est composé de l'ensemble des substantifs des documents. Ce bassin est donc très large et le logiciel ne tente pas de déterminer quelles formes sont plus ou moins représentatives du document. Les seules formes qui sont éliminées au cours du processus d'acquisition sont celles qui ne parviennent pas à satisfaire la contrainte d'autonomie. Le tableau suivant présente le rappel obtenu par l'algorithme des PLS pour chaque seuil de fréquence minimale. Il est important de garder à l'esprit que tous les CT retenus par l'approche qui n'exploite pas les PLS ne sont pas valides. Par exemple, pour une fréquence minimale de 4, TermoStat parvient à recenser 91,76 % des CT identifiés dans la contrainte des PLS dans CA_1. Cependant, ces derniers ne sont pertinents que dans 93,92 % des cas (95,15 % - 1,20 %).

Fréq. min.	CA_1	CA_2	CA_3
1	80,18 %	84,73 %	78,24 %

2	88,00 %	84,98 %	86,08 %
3	91,09 %	86,80 %	93,58 %
4	91,76 %	90,23 %	95,29 %
5	93,71 %	93,22 %	98,55 %
6	94,57 %	96,15 %	98,25 %
7	95,45 %	96,38 %	100 %
8	98,00 %	95,38 %	100 %
9	98,77 %	95,08 %	100 %
10	98,57 %	96,36 %	100 %

Comme le montre le tableau qui précède, le recours aux PLS entraîne une diminution du nombre de termes simples identifiés. Cette diminution est en relation directe avec la nature des PLS. En effet, la sélection de certaines formes au sein du bassin de formes nominales ne peut conduire qu'à un sous-ensemble de ces mêmes formes une fois les contraintes appliquées. De plus, c'est sur le plan des unités simples que les contraintes des PLS se manifestent le plus puisqu'elles entraînent une élimination immédiate de certains CT.

Cette élimination est cependant en accord avec l'objectif initial de notre démarche qui consiste à favoriser la précision par rapport au rappel. Les exemples suivants sont tirés de la listes des termes identifiés par l'approche sans contrainte et qui ne font pas partie de la liste générée par la version régulière du logiciel TermoStat : *building, bus, canada, care, case, division, end, exchange, field, incident, manager, need, page, performance, point, state, store.* On retrouve, parmi les termes non retenus, des unités nominales ayant une fréquence élevée au sein du CA mais dont la fréquence était prévisible.

Lorsqu'on s'intéresse de plus près à cette liste, on constate que certaines formes sont en relation directe avec le domaine des télécommunications. Il semble donc que la spécificité de ces formes ne soit pas lexicale mais sémantique. L'approche par PLS ne permet cependant pas de bien cerner cette spécificité sémantique (opposition mot – terme) et les formes sont ainsi retranchées de la liste des termes simples et le rappel s'en trouve ainsi diminué.

6. Conclusion et perspectives de recherche

Nous avons proposé une technique, fondée sur le calcul des spécificités, permettant d'isoler un ensemble de formes terminologiquement intéressantes au sein d'un corpus spécialisé : les pivots lexicaux spécialisés. Ces derniers sont isolés à l'aide d'une comparaison des fréquences d'occurrence des formes dans deux corpus possédant des caractéristiques différentes.

L'intégration des PLS dans le processus d'acquisition automatique des termes implanté à l'aide du logiciel TermoStat permet d'atteindre une précision moyenne minimale de 86,2 % lors de l'acquisition des termes simples. La méthode est donc bien adaptée à cette démarche. Elle met aussi en évidence le fait que malgré certains doutes de la validité des données d'un point de vue quantitatif, elles peuvent s'avérer forts utiles d'un point de vue qualitatif.

Malgré la contrainte qui veut que les formes considérées comme des termes simples soient des PLS, le logiciel permet d'identifier au minimum 78,24 % des CT relevés sans cette

contrainte. L'imposition d'une telle contrainte n'a donc pas un impact important sur la couverture de l'algorithme si on prend en considération le gain en précision observé. L'évaluation du rappel laisse cependant entrevoir que certaines formes, qui devraient être identifiées comme spécifiques au corpus, sont laissées de côté puisque leur spécificité n'est pas de nature lexicale mais sémantique. Le calcul de la spécificité, utilisé pour l'identification des PLS, ne permet pas de les recenser.

La poursuite de nos travaux sur les PLS nous semble invariablement passer par la prise en charge corpus d'analyse plus volumineux; ceux utilisés dans le cadre de nos expérimentations étant de taille relativement modeste. Une diversification, d'un point de vue des domaines abordés au sein des corpus d'analyse, pourrait aussi conduire à des observations intéressantes sur le comportement des PLS.

Il serait aussi important de procéder à des analyses à l'aide d'un corpus de référence moins homogène. Bien que le CR utilisé dans le cadre de la présente thèse traite d'une foule de sujets différents, il ne relève qu'un seul type de discours, le style journalistique. On peut envisager qu'un corpus de référence plus varié, d'un point de vue du style, pourrait avoir une certaine influence sur les PLS identifiés dans un corpus technique.

Remerciements

Nous remercions Andy Lauriston, Lynne Leslie et Tricia Morgan pour le temps qu'ils ont consacré à la validation des résultats. Nous remercions également Marie-Claude L'Homme pour ses commentaires sur une version préliminaire de ce texte.

Références

ASSADI, H. et D. BOURIGAULT (1996), « Acquisition et modélisation des connaissances à partir de textes : outils informatiques et éléments méthodologiques », dans *Actes du 10ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, pp. 505-514.

BOURIGAULT, D. (1992), « Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases », dans *Proceedings of the Fourteenth International Conference on Computational Linguistics* (COLING 92), pp. 977-981.

BOURIGAULT, D. (1994), « Extraction et structuration automatique de terminologie pour l'aide à l'acquisition des connaissances à partir de textes », dans *Actes du 9ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'94)*, pp. 397-408.

BOURIGAULT, D. et I. GONZALEZ (1994), « Acquisition automatique des termes complexes en français et en anglais, approche comparative », dans *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, pp. 29-43.

BRILL, E. (1992), « A simple rule-based part-of-speech tagger », dans *Proceedings, 3rd Conference on Applied Natural Language Processing (ANPL '92)*, pp. 152-155.

BRILL, E. (1994), « Some Advances in Transformation-Based Part of Speech Tagging », dans *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pp. 722-727.

CAMLONG, A. (1996), *Méthode d'analyse lexicale textuelle et discursive*, Paris : Orphrys.

DAILLE, B. (1994), *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, thèse de doctorat, Paris : Université de Paris 7.

DROUIN, P. (2002), *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*, thèse de doctorat, Montréal : Université de Montréal.

JUSTESON, J. et S. KATZ (1993), *Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906*, IBM Research Division, tire à part.

LABBÉ, C et D. LABBÉ (2001), « Que mesure la spécificité du vocabulaire ? », dans *Lexicometria*, n° 3, tiré à part.

LAFON, P. (1980), « Sur la variabilité de la fréquence des formes dans un corpus », dans *MOTS*, n° 1, pp. 128-165.

LEBART, L. et A. SALEM (1994), *Statistique textuelle*, Paris : Dunod.

MULLER, C. (1977), *Principes et méthodes de statistique lexicale*, Paris: Hachette.

MULLER, C. (1979), *Langue française et linguistique quantitative : recueils d'articles*, Genève : Slatkine.

NKWENTI-AZEH, B. (1994). « Positional and combinational characteristic of terms : consequences for corpus-based terminography », dans *Terminology*, n° 1, vol. 1, pp. 61-95.