Évaluation du potentiel terminologique de candidats termes

Patrick Drouin[1], Philippe Langlais[2]

[1] OLST / ÉCLECTIK, Département de linguistique et de traduction, Université de Montréal
C.P. 6128, Succursale Centre-Ville Montréal (Québec) H3C 3J7, Canada

patrick.drouin@umontreal.ca

[2] RALI, Département d'informatique et de recherche opérationnelle, Université de Montréal
C.P. 6128, Succursale Centre-Ville Montréal (Québec) H3C 3J7, Canada

felipe@iro.umontreal.ca

Abstract

On a daily basis, terminologists must go through specialized documents and identify terms related to a specific domain. Software solutions now exist in order to tackle this task automatically but the units gathered by such automated systems are not always terms. In this paper, we compare the potential of different statistical measures (including a n-gram language model) to discriminate terms from non terms in a list of candidate terms.

Résumé

Une des étapes du travail terminologique vise à identifier, dans un corpus spécialisé, les termes spécifiques d'un domaine. Les logiciels d'acquisition automatique de termes ont pour objectif de prendre en charge cette étape. Cependant, dans les listes d'unité terminologiques retenues par ces systèmes, de nombreuses unités ne sont pas des termes. Dans cet article, nous comparons l'apport de différentes mesures statistiques (incluant un modèle de langue n-gramme) utilisées dans le but de discerner au sein de listes produites par un extracteur de termes les termes des non-termes.

Mots-clés : terminologie, sélection de termes candidats, mesures statistiques.

1. Introduction

Dans le cadre de leurs tâches quotidiennes, les spécialistes procèdent à l'identification des unités terminologiques dans un corpus spécialisé sans obstacles majeurs. Ils parviennent ainsi à distinguer les unités qui ont un statut terminologique des autres unités. Cette étape est nommée dépouillement terminologique. La prise en charge de cette activité par des méthodes informatiques relève de l'acquisition automatique des termes, sous-domaine de la terminologie computationnelle et du traitement automatique de la langue.

L'automatisation du dépouillement n'est cependant pas chose facile. En effet, l'ordinateur parvient à dresser une liste, mais cette dernière contient certaines unités qui n'ont aucun intérêt terminologique. Puisque l'humain réussit là où l'ordinateur échoue partiellement, nous suggérons de pousser plus loin le processus d'acquisition automatique des termes et de cerner le potentiel terminologique (PT) à l'aide de diverses mesures statistiques. Dans cet article, nous comparons le potentiel de différentes mesures statistiques couramment employées à distinguer les termes des non-termes. Nous étudions également le recours à un modèle de langue pour mener à bien cette tâche.

L'expérimentation effectuée ici se démarque des travaux antérieurs par le fait que les performances sont mesurées par rapport à un dépouillement manuel d'un corpus par un

terminologue. L'utilisation d'un corpus étalon annoté manuellement illustre clairement les forces et les faiblesses de chacune des approches testées. Le recours à un modèle de langue pour l'évaluation du PT constitue également une contribution nouvelle.

Dans la section 2, nous présentons les travaux antérieurs. Nous décrivons ensuite (section 3) les corpus qui ont fait l'objet de l'analyse ainsi que le protocole que nous avons mis en place. En section 4, nous présentons les différentes mesures statistiques que nous avons comparées ; les résultats obtenus sont analysés en section 5.

2. Travaux antérieurs

Afin de permettre aux systèmes informatiques de distinguer les unités terminologiques des unités non terminologiques dans les listes de candidats termes (CT) identifiés par les systèmes d'acquisition automatique de termes, des chercheurs ont tenté de cerner le « **potentiel terminologique** » (***termhood*)** des CT. Ces études font souvent l'hypothèse que la structure de surface d'un CT et la réutilisation du matériau lexical au sein de divers CT sont des indices révélateurs du statut terminologique.

Nombreux sont les chercheurs à utiliser la structure de surface des termes sous la forme de patrons de formation syntagmatiques. Cette utilisation des structures à titre d'indice pour l'extraction des termes correspond au *unithood* décrit par Kageura et Umino (1996). En effet, les modes de formation de termes sont bien documentés et ont fait l'objet d'études et de descriptions dans le cadre de la terminologie (OLF 1979 ; Sager 1990, etc.).

Les recoupements de formes entre les CT retenus par les systèmes sont aussi exploités. L'utilisation de cet indice se fait habituellement en tenant compte de la fréquence d'inclusion d'un CT dans un autre (ex. : *connexion Internet*/***connexion Internet** à large bande*) (Drouin et Ladouceur 1994 ; Assadi et Bourigault 1996 ; Frantzi et Ananiadou 1996). Des liens sémantiques potentiels entre termes sont donc dépistés à partir de recoupements simples de chaînes de caractères. Une telle approche apporte de l'information intéressante sur la réutilisation du matériau lexical, mais elle ne permet cependant pas de se prononcer de façon catégorique quant au statut terminologique d'un CT.

La mesure du PT proposée par Frantzi (1997), la NC-Value, repose sur la prise en compte de certains mots (noms, verbes et adjectifs) avoisinant le CT en corpus. Cette technique est semblable à celle mise de l'avant par Nakagawa et Mori (1998) qui prennent en considération l'ensemble des unités lexicales simples utilisées au sein d'un CT complexe. Successeur de la C-Value (Frantzi et Ananiadou, 1996) à titre d'indicateur du PT, la SNC-Value (Maynard et Ananiadou, 2000) pousse un peu plus loin l'objectif visant à déterminer le potentiel terminologique. Cet indicateur fait intervenir des relations sémantiques dépistées dans des ressources terminologiques existantes (dictionnaires, thésaurus, etc.). Bien que le SNC-Value conduise à des résultats intéressants, le recours à des ressources terminologiques limite considérablement, de l'aveu même des auteures, l'application potentielle de cette technique. De plus, dans sa forme actuelle, cette dernière ne peut s'appliquer qu'aux CT complexes.

De nombreuses mesures statistiques ont été exploitées en vue de déterminer le PT des candidats recensés automatiquement. Ces mesures reposent principalement sur les concepts de fréquence et de répartition et elles sont généralement issues de travaux en repérage d'information ou en linguistique de corpus (Kageura et Umino, 1996). Les travaux entrepris dans ce cadre ont pour objet de mettre en lumière des phénomènes liés au comportement des CT observable sur le plan textuel au sein d'un corpus. Une telle technique n'est pas sans rappeler les travaux de Bourigault (1994) et de son exploitation de techniques endogènes pour

l'acquisition de la terminologie. Cette approche consiste à fournir au système un minimum d'information et à mettre en place des procédures d'apprentissage qui lui permettent d'acquérir des connaissances par lui-même suite à l'observation des particularités des CT en corpus. Le corpus est donc à la fois objet du traitement et source d'information. Daille *et al.* (1994) testent diverses mesures pour en venir à la conclusion que la fréquence constitue le meilleur indice du PT. Paslaru Bontas et Schlangen (2005) ont recours à un indice simple tiré de domaine de la recherche d'information, l'indice *tf.idf* qui repose sur la fréquence des CT et sur leur distribution au sein de corpus. Leurs travaux impliquent aussi la prise en considération d'éléments morphologiques (préfixes et suffixes).

Dans un autre ordre d'idée, Kit (2002) suggère une mise en opposition de corpus afin d'améliorer les résultats des divers tests de PT. Des auteurs se sont déjà intéressés à une approche impliquant la comparaison de corpus pour faire ressortir le PT des candidats termes : Ahmad (1994), Drouin (2002), Chung (2003), Gillam *et al.* (2005), Lemay et al. (2005), Drouin et Bae (2005). La fréquence des unités lexicales du corpus faisant l'objet du processus d'acquisition automatique des termes (*corpus d'analyse*, *CA*) est comparée à celle observée dans un c*orpus de référence* (*CR*). Ce dernier se veut généralement représentatif d'un usage non spécialisé. La comparaison s'opère grâce à divers indices statistiques qui ont pour but de quantifier la *déviation* de la fréquence observée entre les deux corpus et de mettre en évidence les formes qui adoptent un comportement spécifique dans le corpus spécialisé. Cette quantification de la spécificité correspond à une évaluation du PT des unités relevées.

3. Méthodologie

3.1. Corpus

3.1.1. Corpus de référence

Le corpus de référence est utilisé à titre de point de comparaison, il constitue ainsi, d'une certaine façon, la norme linguistique sur laquelle sera modelé le comportement des unités lexicales. Pour la présente recherche, nous utilisons un corpus de langue journalistique construit à partir de l'ensemble des articles publiés au cours de l'année 2002 dans le journal *Le Monde* qui totalise environ 30 millions d'occurrences, soit près de 1 180 000 phrases. Les articles du quotidien sont rédigés dans une langue non spécialisée et portent sur des sujets très diversifiés.

3.1.2. Corpus d'analyse

Le dépouillement terminologique est effectué sur le corpus d'analyse. Ce dernier a un contenu spécialisé tiré du site Web de la société KDS, spécialisée dans le domaine de la réservation en ligne. Le corpus possède donc des caractéristiques lexicales particulières attachées à la technologie Web et un vocabulaire spécialisé portant sur la réservation en ligne. Le défi consiste donc à relever les termes de ce domaine tout en sachant écarter les termes spécialisés relevant des autres domaines. Le corpus d'analyse est composé d'un peu plus de 43 000 occurrences.

Ce corpus a été annoté manuellement par un terminologue, spécialiste du domaine, afin de mettre à notre disposition un corpus étalon. L'annotation a été effectuée dans le cadre d'un projet conjoint avec l'Office québécois de la langue française et la société AMEX Canada. L'objectif du dépouillement manuel était un recensement complet du vocabulaire lié au domaine de la réservation en ligne. Les consignes données au terminologue étaient de

recenser toutes les unités lexicales spécialisées relevant du domaine. Ainsi, les unités nominales, les adjectifs, les verbes et les adverbes ont été recensés.

Le format d'annotation utilisé est relativement simple et les unités lexicales considérées comme des termes par le terminologue sont encadrées des balises <TERM>… </TERM> dans le texte.

> Les <TERM>voyageurs</TERM> bénéficient ainsi de tous les <TERM>tarifs affaires</TERM> de la Deutsche Bahn et de sa nouvelle <TERM>fonction</TERM> de <TERM>e-ticket</TERM> (DB OnlineTicket) qui leur permet d'imprimer les <TERM>billets</TERM> eux-même jusqu'à une heure avant le <TERM>départ</TERM> du <TERM>train</TERM>.

Il est important de souligner que le dépouillement initial par l'humain du corpus d'analyse ne visait pas la réutilisation de ce corpus pour une démarche associée au traitement automatique de la langue. Ainsi, certaines unités terminologiques complexes imbriquées ont été annotées comme une seule unité terminologique ; les exemples qui suivent fournissent quelques exemples de ces cas potentiellement problématiques.

> … *automatiser l'<TERM>achat et la gestion des déplacements professionnels</TERM>. des*
> … *<TERM>actes administratifs et de management</TERM> et ...*
> … *aux <TERM>horaires, tarifs et disponibilités des trains</TERM> et de réserver …*

Le dépouillement du corpus par le terminologue a conduit à l'établissement d'une liste de référence contenant un total de 1382 unités terminologiques. Il est à noter que les termes de cette liste ne sont pas lemmatisés. L'annotation des occurrences en corpus conduit donc au recensement des variantes orthographiques *voyage d'affaire*, *voyage d'affaires*, *voyages d'affaire* et *voyages d'affaires* comme des termes différents.

3.2. Extraction des candidats termes

Afin d'établir une liste de candidats termes, le corpus d'analyse a été traité à l'aide du logiciel TermoStat (Drouin et Bae, 2005). Pour la présente recherche, les structures de surface de termes potentiels recensées par le logiciel sont les suivantes :

$nom
ex. : *alimentation, structure,* etc.

$nom(?: $nom)+
ex. : *mécanisme barillet*

$nom(?: $adj| $adjpar)+
ex. : *unité centrale spécialisée*

$nom(?: $adj| $adjpar)*(?: $prep $nom(?: $adj)*)+
ex.: *unité centrale de traitement, système conversationnel d'accès facile*

$nom(?: $adj| $adjpar)+ $coord(?: $adj| $adjpar)+
ex.: *ressources physiques ou matérielles, états distincts et exclusifs*

$nom(?: $adj| $adjpar)? $prep $nom $coord d. $nom(?: $adj| $adjpar)?
ex.: *touches d'insertion et de suppression, processus de lecture ou d'écriture*

La couverture de ces patrons de formation suffit généralement au recensement de la grande majorité des unités nominales du corpus d'analyse. Comme le laisse entrevoir les patrons de formation qui précèdent, le logiciel d'extraction ne prend pas ici en charge les adjectifs, les verbes et les adverbes isolés. Il y a donc une divergence entre les consignes de dépouillement manuel et les unités sélectionnées par le logiciel.

> … *(à la fois en <TERM>tarifs négociés et de dernière minute</TERM>).*

… aux <TERM>horaires, tarifs et disponibilités des trains</TERM> et de réserver …

Comme le démontrent les exemples qui précèdent, il est parfois difficile de prévoir les réalisations de surface des termes dans les corpus. Le terminologue, dans ces contextes précis, désirait relever les unités *tarifs négociés*, *tarifs de dernière minute*, *horaires des trains*, *tarifs des trains* et *disponibilités des trains*. Une telle stratégie de recensement des termes est très éloignée de ce que l'on peut espérer obtenir à l'aide d'un outil d'extraction de termes. De plus, il est fort à parier que l'inclusion de matrices permettant le recensement des unités des contextes précédents conduirait à une augmentation du bruit.

3.3. Mesures de performance

Les résultats obtenus à partir des mesures statistiques sont évalués à l'aide des concepts de *précision*, de *rappel* et de *f-mesure*. La première mesure, la précision, correspond au nombre de termes valides identifiés dans l'ensemble des CT recensés par le système. Le rappel constitue la proportion de termes valides identifiés par rapport à l'ensemble des termes annotés manuellement dans le texte original. La f-mesure est la moyenne harmonique de ces deux taux.

Pour cette expérimentation, la précision globale et le rappel global sont déterminés par les performances du système d'acquisition automatique des termes puisque toutes les mesures statistiques décrites dans la section 4 travaillent sur une liste identique de candidats termes. La précision moyenne obtenue par le logiciel est de 17,40 % (867/4984) alors que le rappel est de 62, 74% (867/1382). Comme nous l'avons mentionné à la section précédente, les divergences dans les objectifs de dépouillement du terminologue et celles imposées au logiciel sont à l'origine des performances obtenues.

Dans le cadre de la présente recherche, nous nous intéressons plus particulièrement au potentiel des mesures statistiques à séparer les termes des non-termes dans la liste des candidats termes. Afin d'y parvenir, nous avons sélectionné des mesures qui permettent de comparer le comportement des CT dans des corpus de niveaux de spécialisation différents et de quantifier la déviation observée à l'aide d'un score. Ce score, que nous associons au potentiel terminologique des CT, est ensuite utilisé pour trier la liste des CT générée par TermoStat de façon à regrouper en tête de liste ce que les mesures considèrent comme des termes potentiels.

La section suivante présente les mesures statistiques utilisées pour quantifier le potentiel terminologique. Les performances obtenues sont ensuite représentées sous forme de graphiques dans la section 5 où nous nous intéressons plus particulièrement aux caractéristiques des termes effectivement validés par les mesures se trouvant dans la première moitié des listes de CT (437 termes au total). Cette analyse a pour but de mettre en évidence les forces de chacune des mesures statistiques.

4. Mesures statistiques

Afin de comparer les fréquences des CT dans les corpus de référence (CR) et corpus d'analyse (CA), nous utilisons la table de contingence suivante qui illustre les divers scénarios possibles. Ces données seront utilisées pour tester le potentiel terminologique des CT à l'aide des mesures décrites dans les paragraphes qui suivent.

Corpus	**CR**	**CA**	**Total**
Fréquence CT	*a*	*b*	*a+b*
Fréquences autres mots	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*N=a+b+c+d*

Tableau 1 : Tableau de contingence pour représentation des fréquences des unités

4.1.1. Fréquence

La première mesure testée est directement observable dans les corpus, il s'agit de la fréquence brute. L'intérêt de cet indice pour l'évaluation du PT des candidats termes a été démontré par Daille et al. (1994). Les observations de ces auteurs confirment l'intuition des terminologues selon laquelle la fréquence représente un bon critère pour l'identification des termes dans les corpus spécialisés (OLF 1979). Pour la présente expérimentation, nous nous contentons d'évaluer la qualité de l'évaluation du PT par la fréquence brute sans avoir recours à la fréquence relative ou encore à un ratio de fréquence tel que proposé par Chung (2003) ou Gillam (2005). La liste obtenue du logiciel TermoStat est donc tout simplement triée en ordre décroissant de fréquence.

4.1.2. Spécificité

Le calcul de spécificité a été proposé par Lafon (1980) afin de cerner le vocabulaire spécifique à un sous-corpus par rapport à l'ensemble d'un corpus.

$$\log P(X=b) = \log (a+b)! + \log (N-(a+b))! + \log (b+d)! + \log (N-(b+d))! - \log N! - \log b! - \log a! - \log b! - \log (N-(a+b+d))!$$

Cette approche permet de comparer le comportement des unités lexicales en fonction de critères variables. Nous adaptons légèrement la démarche en fusionnant le corpus de référence et le corpus d'analyse afin de vérifier si le lexique de ce dernier se comporte comme le lexique du premier. Le calcul des spécificités conduit à l'obtention d'un score qui facilite le classement des CT les uns par rapport aux autres. Nous faisons l'hypothèse que les unités les plus spécifiques sont de meilleurs candidats termes et la liste est triée en ordre décroissant de spécificité.

4.1.3. X^2

Le test du X^2 a été utilisé par Rayson *et al.* (1997) pour l'analyse des conversations au sein du British National Corpus. Il a aussi été exploité par Kilgariff (2001) pour évaluer l'homogénéité des corpus.

$$X^2 = N(ad-bc)^2/((a+b)(c+d)(a+c)(b+d))$$

Nous l'utilisons ici tout simplement pour comparer les fréquences d'occurrence de CT. Les unités qui se démarquent le plus se verront attribuer une valeur plus élevée; la liste obtenue à l'aide du X^2 sera donc triée en ordre décroissant.

4.1.4 Log-likelihood

Proposé par Dunning (1993), le log-likelihood a été utilisé, entre autres, par Rayson et Garside (2000) pour la comparaison de corpus (et non des unités lexicales directement). Il est calculé de la façon suivante :

$$E_1 = ((a+c)(a+b))/((a+c)(b+d))$$
$$E_2 = ((b+d)(a+b))/((a+c)(b+d))$$
$$LL = 2*((a*ln(a/E_1)) + (b*ln(b/E_2)))$$

Tout comme pour les deux mesures qui précèdent, le log-likelihood conduit à l'obtention d'un poids. Plus l'écart entre la fréquence relative observée dans le corpus d'analyse et celle que l'on pourrait prédire à partir du corpus de référence est important, plus le score de vraisemblance est grand. Ainsi, un CT qui obtient un poids élevé est potentiellement plus intéressant d'un point de vue terminologique qu'un CT ayant une valeur plus basse. La liste sera donc encore une fois triée en ordre décroissant des valeurs de log-likelihood.

4.1.5 Modèle de langue

Le modèle construit est un modèle trigramme lissé par une variante de la méthode décrite initialement dans (Kneser & Ney, 1995). L'idée principale de cette méthode est de retirer de chaque *n*-gramme vu au moins une fois dans le corpus d'entraînement, un compte fractionnaire (un *discount*) fixe D. Les comptes ainsi retenus sont alors redistribués sur un modèle d'ordre inférieur via la combinaison linéaire suivante :

$$p_{KN}(w_i \mid w_{i-2}w_{i-1}) = \frac{\max\left\{c\left(w_{i-2}^{i}\right) - D, 0\right\}}{\sum_{w_i} c\left(w_{i-2}^{i} - 2\right)} + \frac{D \times \left|\left\{w_i : c\left(w_{i-2}^{i} - 2\right) > 0\right\}\right|}{\sum_{w_i} c\left(w_{i-2}^{i} - 2\right)} p_{KN}\left(w_i \mid w_{i-1}\right)$$

où c(•) indique la fréquence de • en corpus. Le modèle d'ordre inférieur (modèle bigramme) prend la forme qui suit, dont l'idée se résume ainsi : la probabilité d'un bigramme dépend du nombre de contextes gauches immédiats dans lesquels il est observé en corpus. Ce modèle pouvant produire une probabilité nulle (dans le cas où le bigramme $w_{i-1}w_{i}$ n'a jamais été vu en corpus) il est à son tour lissé par la même méthode qui vient d'être décrite (la récurrence s'arrête avec un modèle uniforme).

$$p_{KN}(w_i \mid w_{i-1}) = \frac{\left|\{w_{i-2} : \mathrm{c}\left(w_{i-2}^{i}\right) > 0\}\right|}{\sum_{w_i}\left|\{w_{i-2} : \mathrm{c}\left(w_{i-2}^{i}\right) > 0\}\right|}$$

Pour entraîner notre modèle, nous avons utilisé la boîte à outils SRILM (Stolcke, 2002) qui implémente une variante de cette technique décrite dans (Chen & Goodman, 1998).

Nous avons noté chaque candidat terme w_1^n avec ce modèle selon l'équation :

$$p(w_1^n) = \left(\log p_{KN}(w_1) + \delta(n \geq 2)\left(\log p_{KN}(w_2 \mid w_1) + \sum_{i=3}^{n} \log p_{KN}(w_i \mid w_{i-2}, w_{i-1})\right)\right) / n$$

$$où \quad \delta(x) = \begin{cases} 1 si\ x\ est\ vrai \\ 0 si\ x\ est\ faux \end{cases}$$

Le modèle de langue entraîné sur le corpus de référence est utilisé pour attribuer une probabilité aux CT retenus par le système d'acquisition automatique de termes. Une séquence bien notée par ce modèle respecte la distribution des mots dans le corpus de référence. par conséquent, des valeurs faibles de probabilité sont susceptibles d'indiquer des structures propres au corpus d'analyse, donc des termes. La liste des CT est ainsi triée en ordre croissant de probabilité.

5. Résultats

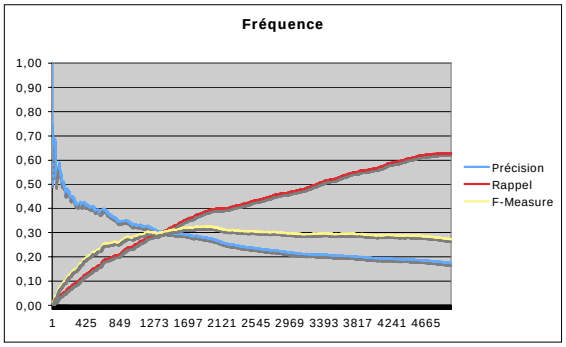
Dans les graphiques 1-5, l'axe des abscisses représente le rang attribué aux CT de 1 à 4984, qui est le nombre total de CT identifiés par le logiciel d'extraction. L'axe des ordonnées représente la précision, le rappel et la f-mesure sur une échelle de 0 à 1.

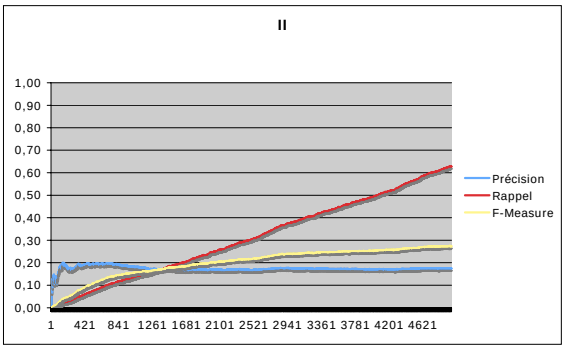
L'observation de ces graphiques permet d'observer une similarité marquée entre les courbes générées par les mesures. En effet, les courbes obtenues à partir du calcul des spécificités (*Spec*) et celles de la fréquence brute (*Freq*) sont semblables, tout comme celles générées à partir du calcul du X^2 (*X2*), du log-likelihood (*ll*) et du modèle de langue (*LogProb*).

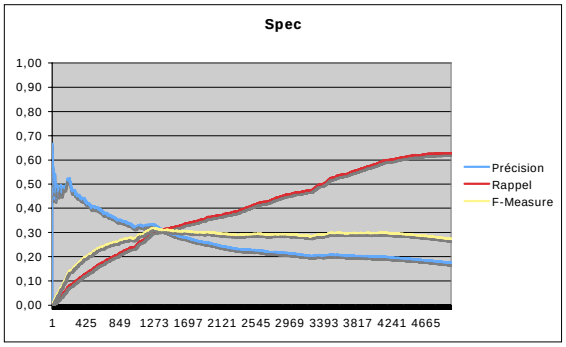
Nos résultats confirment les observations de Daille *et al.* (1994) au sujet de la fréquence. En effet, on observe que la meilleure performance est obtenue à l'aide de la fréquence qui concentre un maximum de véritables termes en tête de la liste des CT. La spécificité conduit aussi à des résultats intéressants, bien que légèrement moins élevés que la fréquence.

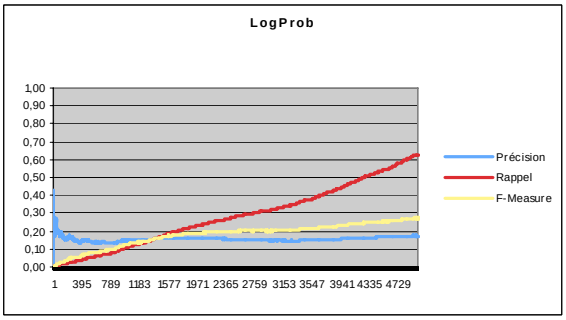
L'union de tous les termes classés dans la première moitié des listes permet de recenser 754 termes. De ce total, 109 termes sont communs aux 5 mesures, 248 sont communs à 4 mesures, 433 à 3 mesures et 618 termes sont communs à 2 mesures. Les cinq tests statistiques ne semblent donc pas s'entendre sur les candidats qui doivent apparaître en tête de liste et une stratégie ne relevant que l'intersection des listes sur la première partie conduirait à l'identification d'un ensemble limité de 109 termes.

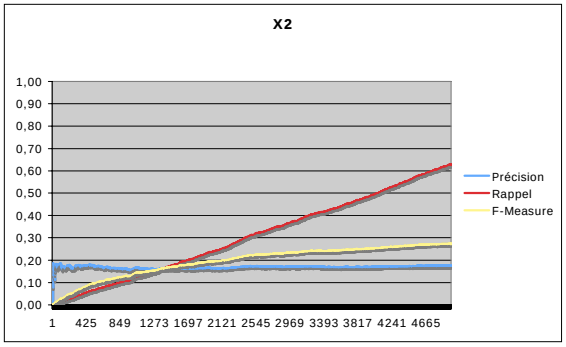
Le recoupement des données entre les listes générées par les indicateurs de PT permet d'ailleurs de confirmer les observations faites sur les courbes. En effet la plus grande intersection se situe entre les listes du *log-likelihood* et du X^2 qui contient 357 termes. La seule autre intersection entre les listes qui se rapproche de ce nombre est celle entre la fréquence et la spécificité avec 332 termes. La plus petite intersection est constituée des éléments relevés par le modèle de langue et la spécificité qui contient 208 termes.











Figures 1-5 : Courbes précision/rappel/f-mesure pour l'ensemble des indices

En plus des recoupements entre les éléments retenus par les diverses mesures, il est intéressant d'évaluer la contribution originale de chacune. Le tableau 2 illustre le nombre de termes originaux proposés par les mesures. Les contributions sont relativement équivalentes à l'exception de la fréquence brute qui ne contient aucun terme n'ayant pas été proposé par les autres indices dans la première moitié de la liste. Malgré la bonne performance de la fréquence brute lors du tri des données ayant pour objectif de mettre en lumière le caractère terminologique des CT, elle ne permet pas d'augmenter le rappel obtenu à partir des autres mesures sur la première moitié de la liste des CT.

ll	***X^2***	***Freq***	***Spec***	***LogProb***
14	37	0	41	44

Tableau 2 : Nombre de termes uniques par mesure

Mots	Total	Intersection	ll	X^2	Freq	Spec	LogProb
1	249	64	0	18	0	0	17
2	200	18	2	18	0	9	15
3	241	26	7	1	0	14	12
4	48	0	2	0	0	5	0
5	80	1	5	0	0	5	0
6	23	0	0	0	0	2	0
7	20	0	0	0	0	4	0
8	2	0	0	0	0	1	0
9	4	0	0	0	0	1	0

Tableau 3 : Longueur en nombre de mots des termes recensés

L'observation des listes de termes relevés permet de distinguer certaines caractéristiques « physiques » qui varient d'une liste à une autre. Le tableau 3 présente la longueur, évaluée en nombre de mots, de tous les termes recensés, de l'intersection entre les 5 mesures et de la contribution unique de chaque indice.

Le modèle de langue (LogProb) favorise les termes les plus courts, tout comme le X^2. Cette contribution à la liste est importante d'un point de vue terminologique puisque les termes simples (composés d'un seul mot) sont bien souvent laissés de côté par les systèmes d'extraction automatique. Tous les indices permettant de valider leur intérêt terminologique sont donc à retenir et exploiter. La tendance qu'a le modèle de langue à proposer des unités plus courtes pourrait être attribuable au fait qu'il repose sur des trigrammes. Une analyse des termes proposés par cet indice montre que ce sont de termes plus répandus dans la langue courante (*touristes, système de gestion, tarifs aériens, mode de communication, factures*, etc.) que ceux proposés par les autres indices. Il s'agit peut-être là d'un effet de l'entraînement du modèle sur du texte journalistique. Il s'agit d'un phénomène intéressant dans la mesure où ces unités lexicales relèvent tout de même du domaine et qu'ils doivent être repérés. D'un point de vue terminologique, on peut envisager de l'utiliser pour la description ou le dépistage de la déterminologisation ou de la banalisation lexicale. L'inverse est aussi vrai pour les cas de spécialisation du lexique comme *intuitivité, protocole*.

La présence d'unités lexicales spécialisées constituées de plusieurs mots dans un texte de langue spécialisée n'est pas un phénomène rare alors que ces unités sont plus rares en langue non spécialisée. Cette mise en opposition, sur laquelle repose l'approche décrite dans cet article, semble être bien exploitée par la mesure de spécificité qui recense des termes plus longs que les autres indices. Une telle propension est très intéressante pour l'extraction des termes dans les domaines techniques où les unités lexicales ont souvent tendance à être plus complexes.

6. Conclusion

Ces premières expérimentations avec divers indices sur un texte annoté par un humain nous permettent d'avoir une bonne idée de la contribution de mesures statistiques. L'utilisation de plusieurs indices n'est pas vaine puisqu'elle permet d'augmenter le rappel.

Les mesures testées nous ont permis, dans un premier temps, de confirmer l'intérêt de la fréquence pour l'identification de termes au sein d'un ensemble de non-termes. D'abord documenté de façon intuitive par les langagiers dans le cadre de leurs travaux, l'utilité de cet indice pour l'extraction automatique des termes avait été documentée par Daille et al. (1994). La mise en parallèle de la fréquence avec divers indices statistiques nous a cependant permis de constater que cette dernière n'apporte pas de contribution importante au rappel.

Nous avons pu observer des contributions individuelles importantes de la part des indices statistiques. En effet, pendant que certains privilégient des termes plus courts, d'autres soulignent l'intérêt de termes composés de plusieurs mots. Un autre élément important mis en lumière par la comparaison est le fait que les termes identifiés possèdent des degrés de spécialisation variables. Il serait envisageable d'optimiser l'utilisation des indices et les listes de résultats obtenus en fonction de ces affinités. C'est l'idée sous-jacente à l'étude décrite par Patry et Langlais (2005) et qui repose sur l'exploitation d'un algorithme d'apprentissage automatique.

Malgré l'intérêt des résultats obtenus à partir des techniques quantitatives, il serait probablement intéressant de s'intéresser aux propriétés des termes dans une perspective plus linguistique qui viendrait confirmer l'intérêt d'un candidat terme tel qu'établi par des mesures statistiques.

Références

Ahmad, K. *et al.* (1994). What's in a Term ? The semi-automatic Extraction of Terms from Text, dans *Translation Studies. An Inter-discipline*, Amsterdam/Philadephia ; John Benjamins : 267-278.

Assadi, H. et Bourigault D. (1996). Acquisition et modélisation des connaissances à partir de textes : outils informatiques et éléments méthodologiques, dans *Actes RFIA'96 :* 505-514.

Bourigault, D. (1994). *Un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*. thèse de doctorat, École des Hautes Études en Sciences Sociales, 352 p.

Chen, S.F. et Goodman J. (1998). *Empirical study of Smoothing Techniques for Language Modeling.* rapport interne, Center for Research in Computing Technology, Harvard University, Cambridge.

Chung, T. M. (2003). A Corpus Comparison Approach for Terminology Extraction. Dans *Terminology*, 9(2) : 221.246.

Daille *et. al.* (1994). Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th conference on Computational linguistics* : 515-521.

Drouin, P. (à paraître). Termhood experiments ; quantifying the relevance of candidate terms. Dans *The 15th European Symposium on Language for Special Purposes (LSP-2005)*, Bergame, University of Bergamo, à paraître.

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

Drouin, P. et Bae H. S. (2005). Korean Term Extraction in the Medical Domain by Corpus Comparison. Dans *Terminology and Knowledge Engineering (TKE-2005)*, Copenhague, Copenhagen Business School : 349-361.

Drouin, P. et Ladouceur J. (1994). L'identification automatique de descripteurs complexes dans des textes de spécialité. Dans *Proceedings of the Workshop on Compound Nouns ; Multilingual Aspects of Nominal Composition*, Genève, ISSCO : 18-28.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Dans *Computational Linguistics*, 19(1) : 61-74.

Frantzi, K. *et al.* (2000). Automatic recognition of multi-word terms : the C-value/NC-value method. Dans *International Journal on Digital Libraries*, 3(2) : 115-130

Frantzi, K. T. et Ananiadou S. (1997). Automatic Term Recognition Using Contextual Cues. Dans *Proceedings of the 3rd DELOS Workshop*, 8 p.

Gillam, L. Tariq M.et Ahmad K. (2005). Terminology and the Construction of Ontology. Dans *Terminology*, 11(1) : 55-81.

Kageura, K. (1999). Theories "of" terminology : A quest for a framework for the study of term formation. Dans *Terminology*, 5(1) : 21-40.

Kageura, K. (1995). Toward the theoretical study of terms : A sketch from the linguistic viewpoint. Dans *Terminology*, 2(2) : 239-258.

Kneser, R. et Ney H. (1995). Improving backing-off for m-gram language modelling. *IEEE International Conference on Acoustics Speech and Signal Processing*, vol 1 : 181-184.

Kit, C. (2002). Corpus tools for retrieving and deriving termhood evidence. Dans *5th East Asia Forum of Terminology* : 69-80.

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

Maynard, D.G. and Ananiadou, S. (2000). Identifying Terms by Their Family and Friends. dans *COLING 2000*.

Nakagawa, H. et Mori T. (2003). Automatic term recognition based on statistics of compound nouns and their components. Dans *Terminology*, (9)2 : 221-246.

Olf (1979). *Table ronde sur les problèmes du découpage du terme*. Montréal, Office de la langue française, 214 p.

Paslaru B.E. et Schlangen D. (2005). Ontology Engineering for the Semantic Annotation of Medical Data. Dans *4th Workshop on Web Semantics - DEXA2005*.

Patry, A. et Langlais P. (2005). Corpus-Based Terminology Extraction. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Danemark, August 17-18, 2005 : 313-321

Sager, J.C. (1990). *A Practical Course in Terminology Processing.* Amsterdam/Philadelphie, John Benjamins Publishing Company, 254 p.

Stolcke, A. (2002). Srilm – an extensible language model toolkit. *International Conference on Spoken Language Processing*, Denver, Colorado.