# L'extraction des termes complexes : une approche modulaire semi-automatique

Ismaïl Biskri[1,2], Jean-Guy Meunier[1], Sylvain Joyal[2]

[1]LANCI – UQAM – C.P. 8888, Succ. Centre-Ville – Montréal – Québec, H3C 3P8 – Canada
[2]DMI – UQTR – C.P. 500 – Trois-Rivières – Québec, G9A 5H7 – Canada
ismail_biskri@uqtr.ca
meunier.jean-guy@uqam.ca

## Abstract

Complex terms extraction systems have achieved good rates of success in the last decade. However, all these systems do not take in account users points of view, perspective, knowledge and subjectivity. Many researchers reject this fact. They argue that multiplicity of points of view leads to more than only one kind of results. In our paper we present a semi-automatic method and software tool for complex term identification. Our approach is hybrid in that it combines numeric (Bayesian approach + N-grams of words) and linguistic filters. The software tool (ESATEC) is different from other term identification tools in that it is *by design* semi-automatic: i.e. it is interactive and constantly under the user's control. The software supports the knowledge engineer's work, the (corpus) domain's expert, or the linguist, by helping them do their job more efficiently. We justify this semi-automatic approach by the need to have a more flexible and customisable tool to perform certain term identification tasks. We don't want impose on users a pseudo-standardised vision of the world. This work can be useful in terminology, indexation, information retrieval, etc.

## Résumé

Durant la dernière décennie, plusieurs outils d'extractions et de repérage de termes complexes ont été mis au point. Certains de ces outils ont même été considérés comme relativement bons. Toutefois, tous ces systèmes avaient un « handicap » commun : ils ne tenaient pas compte du point de vue, de la perspective, de la connaissance et de la subjectivité de l'utilisateur. Ce que plusieurs chercheurs rejettent. Ils affirment en effet que toute opération d'interprétation ne mène pas nécessairement vers un seul type de résultat mais bien plusieurs étant donnée la multiplicité de points de vue au moment de l'interprétation. Dans notre article nous présentons une méthode ainsi qu'un outil semi-automatiques pour le repérage de termes complexes. Notre approche est hybride. Elle combine des filtres numériques (approche bayésienne + N-grams de mots) et linguistiques. L'outil mis au point, en l'occurrence ESATEC, est interactif et sous le contrôle constant de l'usager. Il assiste l'ingénieur des connaissances, l'expert du domaine ou le linguiste dans leur tâche. Nous justifions l'approche semi-automatique par un besoin d'outils flexibles et personnalisables. Nous refusons d'imposer à un usager une vision standardisée du monde. Notre travail peut être utile en terminologie, en indexation, en recherche d'information, etc.

**Mots-clés :** Termes complexes, approche semi-automatique, multilinguisme, apprentissage, n-grams.

## 1. Introduction

La langue écrite ou parlée, la traduction, le résumé, la gestion documentaire ou de l'information et bien sûr la terminologie et dans la dernière décennie l'ontologie, un repérage complet et adéquat des termes complexes dans un corpus traitant d'un domaine spécifique est considéré comme un pré-traitement des plus importants pour l'obtention de résultats d'une meilleure qualité (Strzalkowski, 1999). Dans un passé très récent, un certain nombre d'outils pour le repérage de termes ont été développés et proposés à la littérature scientifique. Ces outils acceptent comme input un texte ou corpus, généralement, pré-traité (étiqueté par exemple). Ils

produisent de façon automatique une liste de candidats termes soit au moyen d'une approche statistique (bayésienne par exemple) soit au moyen d'une approche linguistique. Les approches statistiques peuvent être multilingues. Elles sont cependant bruitées (Remaki et Meunier, 2000 ; Smadja, 1993). Les approches linguistiques sont moins bruitées, mais ne peuvent toutefois rendre compte de corpus multilingues ou certains néologismes dans des domaines spécifiques. Ces dernières approches semblent plus adaptées à des textes stéréotypés (Frath *et al.*, 2000).

La plupart des méthodes d'extraction de termes complexes préconisent l'utilisation d'un filtre linguistique pour le repérage de termes. Ce filtre utilise des patrons de termes comme ceux montrés dans Daille (1994) et Sta (1998). Dans une seconde étape, et ce pour réduire le bruit, elles utilisent des filtres statistiques ou de nature syntaxique voir sémantique comme dans Bourigault (1996) et Condamines et Rebeyrolle (2001). Les raisons qui sont données dans Daille (1994) pour justifier ce choix sont multiples :

- La perte de termes modifiés par un adverbe ou un adjectif.
- L'utilisation de filtres statistiques avant celle de filtres linguistiques induit beaucoup de bruits.
- La fréquence des termes est parfois erronée, particulièrement quand il n'y a pas au préalable une opération de lemmatisation.
- Les méthodes statistiques sont sensibles à la taille du corpus. Plus le corpus est grand plus ces méthodes donnent de meilleurs résultats.

Toutefois, malgré ces raisons, certains auteurs continuent à privilégier les méthodes statistiques, avec l'introduction de filtres linguistiques simples pour apporter des corrections aux bruits (par exemple MANTEX (Frath *et al.*, 2000)). Ces auteurs affirment que les approches linguistiques cachent pour la plupart des problèmes complexes voire majeurs.

- L'étiquetage des termes. Certains auteurs évaluent à 40 minutes le travail nécessaire pour corriger 1000 mots étiquetés.
- La lemmatisation. Le problème devant être surmonté est la nature polysémique de la langue. Le sens des mots varie très souvent en fonction du contexte dans lequel il est utilisé.
- La structure du terme. La structure syntaxique du terme est généralement considérée de type « syntagme nominal ». Pourtant cette règle souffre des exceptions et n'est pas vraie pour tous les domaines.

Le lecteur pourrait vouloir lire certains travaux relatifs au repérage des termes complexes. Nous lui conseillons, parmi d'autres, les références suivantes : Smadja (1993), Ananiadou (1994), Collier *et al.* (1998), Dagan et Church (1994), Frantzi (1997) et Lauriston (1995).

La majorité des outils disponibles dédiés à l'extraction des termes complexes ont une propriété commune : ils sont automatiques et n'interagissent que très peu avec l'usager. L'emphase mise par les programmeurs de ces outils sur leur aspect automatique cache, généralement, des pré- ou post-traitements (manuels ou pas) non-triviaux, en particulier : l'étiquetage du corpus, la lemmatisation et l'évaluation des termes candidats. Mais ce qui est particulièrement problématique est le dépouillement des résultats eux mêmes, particulièrement si on considère leur taux de rappel et de précision. La méthode et l'outil que nous présentons dans cet article permettent le repérage d'expressions récurrentes (termes) à partir d'un corpus. Toutefois, notre approche est très différente des autres :

- Par son design. Il est interactif et permet le contrôle permanent de l'usager.
- L'outil a des capacités d'apprentissage. L'identification des termes complexes est fondée sur un ensemble de termes préalablement validé par l'usager.

Le logiciel semi-automatique que nous avons développé assiste l'ingénieur des connaissances, l'expert du domaine (traité dans un corpus), ou le linguiste dans leur tâche. Dans l'esprit de nos précédents travaux (Meunier et Biskri, 2003), nous pensons que l'intervention humaine est incontournable dès lors qu'il s'agit de traitement des langues naturelles pour des résultats de haute qualité. L'identification des termes complexes n'en est qu'une modalité. La traduction ou le résumé sont d'autres exemples d'applications où le traitement automatique réalise des performances relativement pauvres comparées à des standards humains. La communauté de linguistique computationnelle ainsi que celle de l'intelligence artificielle semblent partagées en deux groupes : d'une part ceux dont l'objectif est la complète automaticité qui écarte toute intervention, d'autre part ceux dont l'objectif est d'assister intelligemment des humains dans des tâches qui ne peuvent être faites ou contrôlées que par des humains. Notre travail est définitivement représentatif du second groupe. Il y a une autre raison importante pour maintenir le contrôle humain : permettre une analyse qui tienne compte de la perspective, de la subjectivité et des connaissances du domaine de l'usager. En d'autres termes, plusieurs usagers utilisant le même outil peuvent obtenir plusieurs résultats différents : c'est ce que nous appelons une approche flexible pour le repérage des termes complexes. Les mêmes termes complexes ne sont pas nécessairement similaires par exemple en médecine et en anthropologie. C'est la raison pour laquelle la compétence d'un expert est importante. Un autre aspect que nous avons pris en considération dans notre travail a trait à l'apprentissage. Celui utilisé dans notre système est relativement simple. Il est par contre un ajout à même de favoriser le point de vue de l'usager. Notre système de repérage de termes améliore la qualité des résultats en se basant sur un ensemble de termes préalablement validés par l'usager. Cet ensemble de termes représente en soi un patrimoine qui permet d'améliorer les performances du logiciel. Enfin, le système est de conception modulaire. Chaque module (fonction) est indépendant des autres. Seul l'usager (étant donnés ses besoins) peut décider quel module exécuter. Ce genre de représentation de LATAO (Lecture et Analyse de Textes Assistées par Ordinateur) est inspiré d'un projet plus général : SATIM (Système de l'Analyse et du Traitement de l'Information Multidimensionnelle) (Meunier et Biskri, 2003).

## 2. Méthodologie

La méthode utilisée est hybride. Elle combine un calcul statistique bayésien avec des filtres numériques et linguistiques (nous présentons ces filtres à la section 3). La plupart de nos filtres sont « computationnellement » peu coûteux et faciles à opérer. Ils sont également facilement adaptables au traitement d'autres langues que le français et l'anglais.

Un terme complexe est considéré comme un n-gram de mots. On définit le n-gram de mots par une suite de deux mots (bi-gram), de trois mots (tri-gram) ou des fois de quatre mots (quadri-gram), voire de cinq mots (5-gram), etc. La probabilité qu'un n-gram de mots soit admis comme terme dépend de la probabilité du dernier mot de la chaîne étant donnés les mots qui le précèdent. Pour ce faire la formule générale exprimée en terme de probabilités conditionnelles, pour la reconnaissance de termes complexes est donnée dans la figure 1 ('Prob' : pour Probabilité, 'W' : pour Mot, et 'Π' : pour la multiplication).

Équation bayésienne générale :

$$\text{Prob}(W_{1\ldots n}) = \Pi_{1\ldots k}\ \text{Prob}(W_k \mid W_{1\ldots k-1})$$

Équation bayésienne pour bi-grams :

$$\text{Prob}(W_{1\ldots n}) \approx \Pi_{1\ldots k}\ \text{Prob}(W_k \mid W_{k-1})$$

*Figure 1. Formules bayésiennes*

Notre texte en entrée est un simple fichier texte qui n'est ni étiqueté ni lemmatisé. La seule information dont a besoin notre système est celle contenue dans les listes : liste de mots fonctionnels, liste de verbes, liste d'adverbes, etc. – ce type d'informations est dépendant de la langue mais indépendant du domaine. Le calcul bayésien détermine la probabilité des séquences ordonnées de mots dans le corpus. Les n-grams correspondent à des séquences de n mots qui peuvent correspondre à des termes complexes. En pratique, la valeur du n fréquemment utilisée est 2 (bi-grams), 3 (tri-grams), 4 (quadri-grams). Plus la probabilité d'un n-gram particulier est haute plus l'usager aura tendance à considérer ce n-gram comme terme complexe. Ainsi, la probabilité bayésienne agit comme un indicateur pour décider si un candidat terme doit être considéré comme valide ou non. Ces probabilités peuvent être perçues comme des approximations d'un phénomène linguistique complexe. Elles vont induire un taux d'erreur assez élevé, en particulier, un taux de précision assez bas. Ce qui a comme contrainte de rendre le processus de décision de l'usager plus compliqué et plus exigent en temps. Comme nous l'avons montré dans d'autres publications (Biskri *et al.*, 2003 ; Biskri et Delisle, 1999 ; Biskri et Meunier, 1998), une combinaison hybride de modèles statistiques et linguistiques peut influer positivement sur l'approche numérique pure en améliorant la granularité de leur output et ainsi leur valeur utile pour l'usager. La même idée est utilisée ici : la combinaison d'un calcul bayésien simple avec des filtres numériques et linguistiques flexibles pour l'élimination du bruit induit par le modèle bayésien de base. Un autre calcul numérique est également proposé pour permettre une phase d'apprentissage et détecter les termes déjà rencontrés et validés par l'usager.

Notre logiciel développé appelé ESATEC (pour Extraction Semi-Automatique de TErmes Complexes) est implémenté en Visual C++ et interagit avec l'usager au moyen d'une interface usager graphique. L'outil prend en entrée un corpus textuel sous format ascii, duquel il extrait le lexique. Il construit alors une matrice de collocation à partir de laquelle sont calculées les probabilités bayésiennes associées aux n-grams de mots pris dans le corpus. Dans la section suivante, nous montrons comment des filtres numériques et linguistiques sont utilisés pour améliorer la qualité des résultats.

## 3. Repérage des termes complexes dans ses différentes étapes

À l'étape initiale, une fois que le lexique est extrait, l'usager sélectionne les mots qui l'intéressent (il peut sélectionner l'ensemble du lexique aussi). Ces derniers peuvent être spécifiques à un domaine d'intérêt propre à l'usager. Ils forment ainsi un ensemble de mots pôles servant à déterminer uniquement les termes complexes jugés proches du domaine d'expertise de l'usager. Par exemple, dans « acide sulfurique » et « acide chlorhydrique », le mot « acide » est un mot pôle. L'usager doit aussi spécifier la taille en nombre de mots des termes complexes.

Dans la seconde étape, l'outil s'intéresse au repérage de candidat termes complexes à proprement parler en utilisant l'information donnée par l'usager. À cette étape, tout ce que nous avons est une liste de candidat termes qui sera soumise à un ensemble de filtres semi-automa-

tiques de nature numérique et/ou linguistique. Ces filtres sont indépendants les uns des autres. L'ordre dans lequel ils sont présentés n'est pas significatif. L'ordre dans lequel ils sont appliqués est par contre lui significatif. Il dépend en ce sens du choix de l'usager.

Nous donnons une brève description de ces filtres ci-après (voir également la figure 2).

Le premier filtre élimine les candidats termes qui ont une probabilité inférieure à un certain seuil donné par l'usager. Ce seuil peut faire l'objet d'une expérimentation. Il peut être possible que sa valeur soit établie après un certain nombre de tests. Le logiciel propose une valeur particulière de seuil à l'usager. Celle-ci représente la moyenne des valeurs de probabilité de l'ensemble des candidats termes. L'usager peut bien entendu prendre en considération ce seuil ou non.

Le deuxième filtre élimine les candidats termes qui commencent ou se terminent par un mot fonctionnel, un verbe, un adverbe, etc. c'est ici que l'indépendance au domaine de l'information mentionnée à la section 2 est nécessaire. L'usager peut appliquer la totalité ou juste une partie de ce filtre. Ce type de filtre a déjà été utilisé dans d'autres travaux en l'occurrence le projet Lexter (Bourigault, 1996). Il permet de déterminer les frontières d'un groupe nominal.

Le troisième filtre élimine les candidats termes qui commencent ou se terminent avec des mots spécifiques choisis par l'usager dans le lexique. Dans ce cas là, ce sont les connaissances de l'usager du domaine du corpus qui sont utiles pour éviter du bruit non productif.

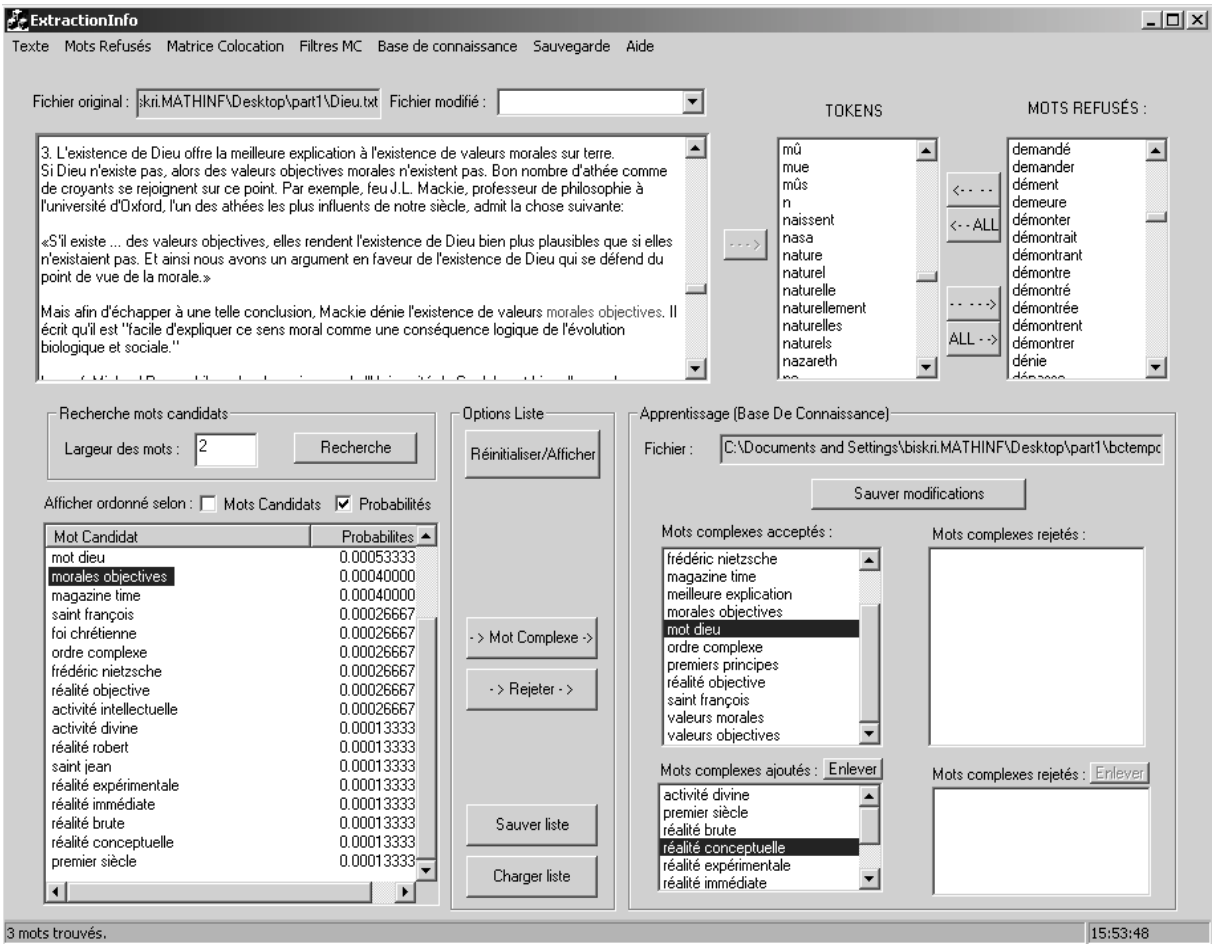


*Figure 2. Interface usager et fonctionnalité*

D'autres filtres peuvent être intégrés à la plate-forme, sans que cela modifie quoique ce soit au niveau de celle-ci. Certains filtres sont d'ailleurs actuellement en cours de développement. En particulier, un filtre qui consiste en l'application d'une analyse syntaxique pour éliminer l'ensemble des candidats termes qui ne sont pas des groupes nominaux (Biskri *et al.*, 2003). Des filtres similaires ont été montrés dans la littérature (Daille, 1994). Ils utilisent des patrons de termes complexes pour le français comme : Nom « de » (déterminant) Nom ; Nom « à » (déterminant) Nom ; etc. Notre analyse est basée sur le modèle de la Grammaire Catégorielle Combinatoire Applicative (Biskri et Desclés, 1997) qui a l'avantage d'être dans son essence indépendante de la langue et se prête donc, via quelques aménagements mineures à être multilingues.

## 4. Apprentissage

Les filtres présentés dans la section précédente nous permettent d'éliminer du bruit dans la liste des candidats termes. Ceci est en soi une « bonne nouvelle » qui permet d'améliorer le taux de précision. Toutefois, cette amélioration induit des fois une « mauvaise nouvelle » : baisse du taux de rappel. Dans le sens de cette remarque contraignante, il sera établi que le filtrage du bruit sera tributaire de la règle générale suivante : *les candidats termes ne peuvent être éliminés par un filtre s'ils ont été préalablement validés par un usager et ainsi stockés dans une table de termes complexes*. Autrement dit, les termes complexes appris par le système sont utilisés dans une ultime vérification. Aussi, plus il y a de corpus traités par le logiciel, plus la liste des termes complexes valides utilisée pour la phase d'apprentissage s'allonge et par conséquent plus la dernière vérification devient discriminante. Ce qui est intéressant avec cette dernière règle est que le taux de rappel s'en voit amélioré sans pour autant affecter le taux de précision. Cette règle peut être plus générale *: un candidat terme ne peut être éliminé par un filtre s'il dérive d'un terme complexe valide*. Par exemple, si la liste des termes valides contient « acide chlorhydrique » alors même si un candidat terme comme « acide sulfurique » apparaît avec une faible probabilité (donc susceptible d'être supprimé par le premier filtre) il ne sera pas éliminé. Une simple fonction, qui stipule que le découpage en n-grams de caractères[1] des deux termes en l'occurrence *acide sulfurique* et *acide chlorhydrique* sont similaires étant donné un certain seuil, en est garante.

Plus concrètement, pour un n = 3, le découpage en tri-grams de caractères de « acide sulfurique » et « acide chlorhydrique » donnent respectivement la suite des séquences de trois caractères suivantes : (aci, cid, ide, de , e s, su, sul, ulf, lfu, fur, uri, riq, iqu, que) et (aci, cid, ide, de , e c, ch, chl, hlo, lor, orh, rhy, hyd, ydr, dri, iqu, que). Les deux termes complexes sont alors considérés comme similaires s'ils partagent un certain nombre de tri-grams. Ce nombre devant être supérieur à un certain seuil établi par l'usager.

## 5. Évaluation[2]

Parce que notre approche est semi-automatique, la comparer à d'autres approches automatiques serait maladroit. Nous devons dès lors considérer notre évaluation sous un angle et une

[1] Pour rappel la notion de n-grams de caractères a été utilisée dans plusieurs travaux sur le traitement des textes oraux. Plus récemment cette notion a été au centre de l'intérêt de Greffenstette (1995) pour l'identification de la langue, et de Damashek (1995) pour le traitement du texte écrit. Ces deux chercheurs ont particulièrement prouvé que la notion de n-grams n'induisait pas de perte d'information. Des applications récentes sur l'indexation (Mayfield et MacNamee, 1998) sur l'hypertextualisation multilingue (Haleb et Lelu, 1998) et la désambiguïsation lexicale (Biskri et Delisle, 2001) confirment cela.

[2] Par souci d'homogénéiser nos évaluations mais aussi de les alléger, nous avons volontairement choisi de nous limiter aux repérages des termes complexes formés de deux mots.

perspective différents. Pour ce faire, nous avons pris un article scientifique de 20 pages, écrit en français, traitant de traitement automatique des langues.

Deux premiers résultats fondamentaux sont constatés :

- Le premier filtre qui consiste à supprimer tous les candidats termes dont la probabilité est inférieure à un certain seuil fixé (arbitrairement) par l'usager induit une baisse du taux de rappel.
- Le deuxième et troisième filtre contribuent quant à eux à améliorer le taux de précision.

Pour arriver à ce constat plusieurs évaluations sur le même texte ont été effectuées. Une première fixait un seuil de probabilité pour l'admissibilité des candidats termes de 0,001 alors qu'une seconde le fixait à 0,0005. On a pu ainsi récupérer dans le premier cas 92 candidats termes sur lesquels seuls 10 étaient acceptables, un taux de précision relatif donc de 11 %. Dans le deuxième cas, 391 candidats termes ont pu être récupérés parmi lesquels 37 étaient valables, un taux de précision relatif de 10 %. Outre le taux de précision qui est très bas dans les deux évaluations, il semble que le taux de rappel (qui reste relatif dans le cas de notre évaluation) est inversement proportionnel au seuil de probabilité. Il augmente quand le seuil de probabilité diminue. Selon nos chiffres ce taux a augmenté de 73 % en diminuant le seuil de probabilité d'acceptabilité des candidats termes de 0,001 à 0,0005.

Deux autres évaluations complémentaires aux deux premières ont été réalisées. Dans ces deux évaluations il s'agissait d'appliquer le deuxième filtre sur les listes de candidats termes obtenus après l'utilisation du premier filtre avec des probabilités d'admissibilité identiques à celles utilisées dans la première et la deuxième évaluation. Ainsi dans la troisième évaluation on applique le premier filtre avec un seuil d'acceptabilité de 0,001 puis le deuxième filtre. On obtient une liste de 10 candidats termes sur lesquels 8 sont effectivement des termes complexes. On remarque un taux de précision relatif de 80 %. Pour ce qui est de la quatrième évaluation, les mêmes deux filtres sont appliqués avec cependant un seuil d'acceptabilité de 0,0005 pour le premier filtre. Il est obtenu une liste de 41 candidats termes parmi lesquels 33 sont valides. Un taux de précision relatif de 80 % également. Il semble que le taux de précision augmente dans les deux évaluations du fait de l'utilisation du deuxième filtre. Toutefois, le taux de rappel diminue légèrement. Ainsi, dans la troisième évaluation le taux de rappel relatif diminue de 20 % alors que dans la quatrième évaluation il diminue de 10 %. Cette diminution n'est en rien dramatique. Il est possible de la corriger avec la phase d'apprentissage.

En somme ce qui ressort de ces quatre évaluations se résume en deux points :

- Le seuil de probabilité utilisé dans le premier filtre pour la validation des candidats termes doit être le plus bas possible pour garantir un haut taux de rappel.
- L'application des deuxième et troisième filtres contribuera à élever le taux de précision. Ces mêmes filtres pouvant diminuer légèrement le taux de rappel, l'utilisation de l'apprentissage permettra d'y remédier.

Ces évaluations comme on peut le constater ne tiennent pas compte encore de la phase d'apprentissage. Leur seul objectif était de montrer qu'une intrication de modules divers dans une chaîne de traitement pouvaient rendre compte de l'extraction des termes complexes et qu'on pouvait mesurer l'impact de chaque module en terme de rappel relatif et de précision relative.

Une évaluation supplémentaire plus complète s'imposait donc. Dans cette dernière évaluation nous avons soumis un texte philosophique de 20 pages à notre analyseur, tout en prenant aussi en entrée une « base de connaissances philosophiques »[3] contenant des termes complexes préalablement validés. Il en ressort que plusieurs termes complexes valides qui auraient du être éliminés par les filtres précédents (étant donnée leur nature) sont récupérés du fait de l'apprentissage du système étant donnée la base de connaissance. Aussi, le taux de rappel a pu augmenter de près de 50 %. Toutefois, le taux de précision a baissé de 7%. Cette baisse reste quand même mineure vu l'importante augmentation du taux de rappel.

Ces derniers résultats restent malgré tout bien relatifs. Ils dépendent du contenu de la base de connaissance. Or celle augmentera de taille au fur et à mesure des traitements avant de se stabiliser. Ceci aura tendance à favoriser le taux de rappel peu importe le seuil de probabilité choisi pour l'application du premier filtre. Ceci étant, notons

De cette dernière évaluation, il en ressort les termes complexes de la figure 3. Nous avons choisi les 19 premiers termes complexes par commodité pour ne pas surcharger cet article.

| |
|---|
| Meilleure explication ; Big Bang ; valeurs morales ; premiers principes ; valeurs objectives ; morales objectives ; Saint-François ; foi chrétienne ; ordre complexe ; Frédéric Nietzsche ; réalité objective ; activité intellectuelle ; activité divine ; Saint-Jean ; réalité expérimentale ; réalité immédiate ; réalité brute ; réalité conceptuelle ; premier siècle. |

*Figure 3. Résultats d'une évaluation sur un texte philosophique*

Bien entendu, cette liste reflète nos maigres connaissances en philosophie. Une expertise en philosophie aurait certainement engendré un résultat produit au moyen de ESATEC différent. C'est d'ailleurs ce que nous voulons : que la perspective et les connaissances de l'usager influent sur le résultat.

D'autres évaluations, principalement qualitatives, que nous réservons à nos prochaines publications ont été réalisées, en particulier sur l'anglais. Des résultats presque semblable ont été obtenus. L'aspect multilingue de notre approche a pu aussi être vérifiée.

## 6. Conclusion

Nous avons présenté dans cet article une méthode ainsi qu'un outil semi-automatiques pour le repérage des termes complexes. Notre approche est différente de la plupart des autres outils du fait que son design est semi-automatique. Nous justifions ce choix par un besoin d'avoir un outil plus flexible et surtout plus « personnalisable ». Plus concrètement, nous voulons, dans certaines situations et dans une certaine mesure, permettre que la perspective, que les connaissances ainsi que la subjectivité de l'usager influencent le résultat. Que l'usager lise le texte pour la première fois ou pas, que plusieurs usagers considèrent le même texte, nous pensons que permettre à un même outil d'arriver à plusieurs résultats possibles représentatifs de leur état d'esprit au moment de l'analyse est d'un grand intérêt. Manifestement, cette flexibilité n'est pas possible avec la plupart des outils de repérage de termes complexes puisque la majorité produit des résultats uniques.

[3] Nous pouvons considérer autant de bases de connaissances que de domaines d'application ou d'utilisateurs. Par ailleurs nous évoquons ici une base de connaissances philosophiques. Nous précisons qu'elle a été construite pour les besoins de l'évaluation sans aucune assistance d'une expertise en philosophie.

Un autre aspect est relatif aux capacités d'apprentissages de l'outil : il peut influencer le résultat de façon significative. Comme présenté à l'évaluation, l'apprentissage peut améliorer le taux de rappel sans affecter grandement le taux de précision. En outre, la base de connaissance fait office de patrimoine que peuvent se partager les membres d'une communauté scientifique ou professionnelle quelconque.

Un dernier aspect important réside dans la méthodologie particulière qu'un utilisateur peut adopter en utilisant le logiciel. Par exemple, il pourrait préférer traiter d'abord (en partie) un corpus standard afin d'extraire une liste standard de termes complexes qui devenant disponible peut servir à l'identification suivante de termes appliquée au même corpus ou à un corpus différent. En fait, il s'avère que l'ingénieur de la langue travaille de cette manière : il établit de nouvelles connaissances sur les connaissances déjà établies.

## Références

Ananiadou S. (1994). A Methodology for Automatic Term Recognition. In *Proceedings of COLING'94* : 1034-1038.

Biskri I. et Delisle S. (2001). Les n-grams de caractères pour l'aide à l'extraction de connaissances dans des bases de données textuelles multilingues. In *Actes de TALN 2001* : 93-102.

Biskri I., Meunier J.-G., Joyal S. et Gayton F. (2003). Extraction of the complex terms : the contribution of categorial grammars. In *Proceedings of PACLING'03* : 109-114.

Biskri I. et Delisle S. (1999). Un modèle hybride pour le textual data mining - un mariage de raison entre le numérique le linguistique. In *Actes de TALN 1999* : 55-64.

Biskri I. et Meunier J.-G. (1998). Vers un modèle Hybride pour le traitement de l'information lexicale dans les bases de données textuelles. In *Actes des JADT 1998.*

Biskri I. et Desclés J.-P. (1997). Applicative and Combinatory Categorial Grammar (from syntax to functional semantics). In Mitkov R et Nicolov N. (Eds), *Recent Advances in Natural Language Processing.* John Benjamins Publishing Company.

Bourigault D. (1996). Conception et exploitation d'un logiciel de termes : problèmes théoriques et méthodologiques. In *IVème journées scientifiques du réseau thématique — AUPELF-UREF* : 137-146.

Collier N., Hirakawa H. et Kumano A. (1998). Machine Translation vs Dictionary Term Translation – a Comparison for English-Japanese News Article Alignment. In *Proceedings of COLING-ACL'98* : 263-267.

Condamines A. et Rebeyrolle J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to Terminological Knowledge Base (CTKB). In Bourigault D., Jacquemin C. et L'Homme M.-C. (Eds), *Recent Advances in Computational Terminology*. John Benjamins Publishing Company.

Dagan I. et Church K. (1994). Termight : Identifying and Translating Technical Terminology. In *Proceedings of CANLP-ACL'94* : 34-40.

Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of CSSALW'94* : 29-36.

Damashek M. (1995). Gauging Similarity with n-Grams : Language-Independent Categorization of Text. *Science*, vol. (267) : 843-848.

Frantzi K.T. (1997). Incorporating context information for the extraction of terms. In *Proceedings of EACL'97* : 501-503.

Frath P., Oueslati R. et Rousselot F. (2000). Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques. In Charlet J., Zacklad M., Kassel G. et Bourigault D. (Eds), *Ingénierie des connaissances : Évolutions récentes et nouveaux défis.* Eyrolles.

Greffenstette G. (1995). Comparing Two Language Identification Schemes. In *Actes des JADT* 1995 : 85-96.

Halleb M. et Lelu A. (1998). Hypertextualisation automatique multilingue à partir des fréquences de n-grammes. In *Actes des JADT 1998*.

Lauriston A. (1995). Criteria for Measuring Term Recognition. In *Proceedings of EACL'95* : 17-22.

Lelu A., Halleb M. et Delprat B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes. In *Actes des JADT 1998.*

Mayfield J. et McNamee P. (1998). Indexing Using both n-Grams and Words. In *Proceedings of TREC7'98* : 419-424.

Meunier J.G. et Biskri I. (2003). SATIM : une plate-forme modulaire pour la construction de chaînes d'analyse de textes assistée par ordinateur. In Arnould J.C. et Blum C. (Eds), *L'édition électronique : état des lieux*. Presses de l'Université de Rouen.

Remaki L. et Meunier J.-G. (2000). Un modèle HMM pour la détection des mots composés dans un corpus textuel. In *Actes des JADT 2000.*

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, vol.(19/1) : 143-178

Sta J.D. (1998). Automatic acquisition of terminological relations from a corpus for query expansion. In *Proceedings of ACM-SIGIR'98 (21st annual international conference on Research and development in information retrieval)* : 371-372.

Strzalkowski T. (1999). *Natural Language Information Retrieval*. Kluwer Academic Publishers.