EXTRACTION AUTOMATIQUE D'UNITÉS LEXICALES COMPLEXES: UN ENJEU FONDAMENTAL POUR LA RECHERCHE DOCUMENTAIRE

Gaël Dias[***], Sylvie Guilloré[**], Jean-Claude Bassano[**] et José Gabriel Pereira Lopes[*]

Résumé – Abstract

Reconnaître les unités lexicales complexes des collections de textes toujours croissantes constitue un enjeu fondamental dans le cadre de la recherche documentaire. En effet, les unités polylexicales sont souvent moins ambiguës et plus motivées que les mots simples et permettent ainsi une meilleure approximation des thèmes abordés. L'identification des ces suites de mots permet soit d'indexer les textes avec une plus grande précision soit de guider l'utilisateur dans sa quête d'information. Dans cet article, nous présentons un analyseur statistique qui extrait à partir d'un texte brut un ensemble d'unités lexicales complexes contiguës et non-contiguës sans recourir aux méthodes d'amorçage ni à la définition de valeurs seuil globales.

The acquisition of multiword lexical units from large text collections is a fundamental issue in the context of Information Retrieval. Indeed, multiword lexical units embody meaningful sequences of words that are less ambiguous than single words and consequently allow to approximate more accurately the contents of texts. The identification of these sequences of words leads to improvements in the indexing process or allows to guide the user in his search for information. In this paper, we present a statistical analyser that extracts contiguous and non-contiguous multiword lexical units from raw texts, without using enticement techniques or defining global thresholds.

Mots Clés - Keywords

Unités lexicales complexes, Indexation, Mesures d'Association Normalisées

Multiword Lexical Units, Indexing, Normalised Association Measures

1. INTRODUCTION

Reconnaître les unités lexicales complexes des collections de textes toujours croissantes constitue un enjeu fondamental dans le cadre de la recherche documentaire. En effet, l'identification des termes complexes permet soit d'indexer les textes avec une plus

[*] Universidade Nova de Lisboa, FCT/DI, Quinta da Torre, 2825-114 Caparica, Portugal. email : {ddg, gpl}@di.fct.unl.pt

[**] Laboratoire d'Informatique Fondamentale d'Orléans, BP 6102 – 45061 Orléans Cedex 2, France. email : {dias, guillore, bassano}@lifo.univ-orleans.fr

grande précision soit de guider l'utilisateur dans sa quête d'information. Dans le premier cas, la selection de termes discriminants (ou descripteurs) pour représenter le contenu des textes est un problème critique. Idéalemement, les termes d'indexation devraient décrire directement les concepts présents dans les documents. Cependant, la plupart des systèmes de recherche d'information indexent les textes de la base documentaire à partir d'unités lexicales simples. Or, ces unités atomiques ne sont pas suffisamment spécifiques pour évoquer le contenu des textes. Afin d'améliorer la qualité de l'indexation, certains systèmes bénéficient de l'existence de thesauri prédéfinis. Dans ce cas, les descripteurs sont choisis parmi les éléments du thesaurus. Ainsi, les mécanismes pour retrouver les documents utilisent les liens directs entre le thesaurus et les textes, et quelquefois les liens de synonymie, d'hyperonymie et/ou d'hyponymie entre les éléments du thésaurus (Cf. Betts R. 1991). Malheureusement, de nombreux domaines ne disposent pas de thesauri spécialisés et parallèlement peu de projets incluent la construction automatique de thesauri (Cf. Grefenstette G. 1994). L'identification des unités lexicales complexes d'une base documentaire propose une alternative aux problèmes précédemment exprimés. En effet, les unités polylexicales sont souvent moins ambiguës et plus motivées que les unités lexicales simples et permettent ainsi une meilleure approximation des thèmes abordés. Ainsi, les termes complexes, préalablement extraits des documents de la base, sont utilisés pour représenter le contenu des textes sous la forme d'une liste de termes représentatifs de ce contenu (Cf. Evans D. 1993). Dans le deuxième cas, le système de recherche documentaire doit pouvoir guider l'utilisateur dans sa quête d'information et ainsi lui permettre de rafiner sa requête en lui proposant une liste de descripteurs qui s'apparient à sa requête initiale. L'utilisateur doit pouvoir ainsi "voyager" dans l'espace de la base documentaire à partir des différents concepts listés par le système. Le pouvoir de représentation des unités lexicales complexes joue un rôle primordial dans cette opération en proposant une meilleure discrimination des sujets abordés par les textes (Cf. Quaresma P. *et al* 1998).

Comme nous venons de le justifier, l'identification des unités lexicales complexes est un enjeu important pour la recherche documentaire. Cependant, l'ensemble des unités polylexicales est ouvert et à compléter. En particulier, une partie essentielle de la néologie lexicale dans les domaines techniques et scientifiques s'opère par le biais de séquences complexes. Par exemple, *World Wide Web*, *adresse IP* et *réseau TCP/IP* sont des termes complexes particulièrement récents dans le domaine de l'informatique qu'il est peu probable de trouver répertoriés dans des bases de connaissance lexicale. En effet, la création de banques terminologiques est un travail long et difficile. L'un des enjeux de la recherche en langage naturel est donc de fournir des outils permettant l'extraction automatique d'unités lexicales complexes à partir de corpora spécialisés. Ainsi, le problème du dépouillement terminologique a été abordé suivant trois axes distincts. La première approche utilise des techniques structurelles fondées sur l'analyse syntaxique de l'énoncé (Cf. David S. & Plante P. 1990; Bourigault D. 1996). Malheureusement, cette démarche nécessite des connaissances approfondies de la langue qui freinent son application à de nouveaux domaines (Cf. Habert B. 1997) ou à de nouvelles langues. Afin de faire face aux problèmes mis en évidence par l'approche précédente, un grand nombre de chercheurs proposent des techniques hybrides qui associent modèles statistiques et filtres syntaxiques (Cf. Enguehard C. 1993; Justeson J. 1993; Daille B. 1995; Herviou-Picard M.L. 1996; Feldman R. 1998). Cependant, ces systèmes n'autorisent que l'extraction de termes complexes de type nominal. De plus, leur application à de nouvelles langues nécessite la redéfinition des filtres syntaxiques. Finalement, la troisième approche propose des techniques statistiques et numériques qui décèlent les associations préférentielles présentes dans le corpus (Cf. Church K.W. & Hanks P. 1990; Dunning T. 1993; Smadja F. 1993; Shimohata S. 1997). Cette dernière méthode se différencie des approches précédentes de part sa flexibilité d'adaptation à de nouveaux domaines et/ou à de nouvelles langues ainsi que par l'ensemble non-restreint des termes complexes extraits. Cependant, les systèmes statistiques proposés exhibent deux inconvénients majeurs. Premièrement, ils recourent à la définition de valeurs seuil globales qui sont sujettes à erreurs. Deuxièmement, ils ne mesurent que les associations textuelles binaires et doivent ainsi utiliser des techniques d'amorçage pour l'extraction d'associations n-aires (Cf. Habert B. & Jacquemin C. 1993).

Dans cet article, nous présentons un analyseur statistique qui extrait à partir d'un texte brut[1], un ensemble d'unités lexicales complexes contiguës[2] et non-contiguës[3] sans recourir aux méthodes d'amorçage ni à la définition de valeurs seuil globales. Ainsi, nous conjugons une nouvelle mesure d'association fondée sur le concept d'expectative normalisée, l'Expectative Mutuelle (Cf. Dias G. *et al* 1999(a)), avec un nouveau processus d'extraction basé sur un algorithme de maxima locaux, le LocalMaxs (Cf. Silva J. *et al* 1999). En particulier, l'Expectative Mutuelle permet de caractériser la structure des termes complexes sans se limiter aux associations binaires[4] et le LocalMaxs permet d'éviter la définition de valeurs seuil globales. Les résultats obtenus par l'analyseur appliqué à des collections de textes de différents domaines et de différentes langues montrent l'extraction de termes de base ainsi que l'acquisition de termes obtenus par composition et modification. Finalement, nous illustrons comment l'ensemble des unités lexicales complexes extraites à partir d'une base documentaire du domaine législatif portugais peut être intégré dans le moteur de recherche développé à l'Université Nouvelle de Lisbonne afin de guider l'utilisateur dans sa quête d'information.

2. ÉTUDE DES SYSTÈMES EXISTANTS

Comme nous l'avons mentionné dans l'introduction, trois axes de recherche ont été préférablement adoptés pour l'acquisition d'unités lexicales complexes à partir de corpora spécialisés. Dans cette partie, nous illustrons chacune de ces approches par l'étude d'un système particulier. Nous nous attacherons à présenter les problèmes posés par chacune des méthodes afin de motiver le caractère novateur de notre analyseur.

D. Bourigault (1993) propose une technique d'analyse syntaxique locale par patron de surface pour le repérage de termes complexes dans un texte. Cette technique a été développée dans le cadre du logiciel LEXTER. Le principe de base est de découper le texte en repérant des frontières potentielles entre lesquelles il est possible d'isoler des syntagmes nominaux. Ainsi les verbes, les pronoms, les conjonctions et les séquences prépositions+article indéfini (entre autres) constituent des frontières qui permettent de détacher "en négatif" les unités lexicales complexes de type nominal. Les résultats obtenus par LEXTER révèlent cependant un taux de bruit important dû à la complexité des phrases qui font l'objet de nombreuses combinaisons syntaxiques. D'autre part, LEXTER n'autorise que l'extraction de termes complexes qui dérivent de la composition nominale. Finalement, LEXTER recourt à un grand nombre de contraintes linguistiques par le biais d'heuristiques qui limitent son champ d'application au français. Afin d'illustrer les limitations de ce système, nous présentons un ensemble d'unités lexicales complexes que les contraintes imposées ne permettent pas d'extraire: *l'offre et la demande*, *transmission via FTP anonyme*, *monter une partition*, *consommateurs d'alcool et de drogues*, *prendre une cuillère à soupe*.

Motivée par la réduction du bruit du processus d'acquisition exhibé par LEXTER, B. Daille (1995) propose un système hybride qui associe filtres linguistiques et statistiques lexicales. Dans un premier temps, le système ACABIT identifie les "termes de base" du corpus qui s'apparient à l'une des structures morpho-syntaxiques suivantes: Nom+Adjectif, Nom+Nom, Nom+à(Det)+Nom, Nom+de(Det)+Nom et enfin Nom+Prep+Nom. Dans un deuxième temps, les candidats termes sont classés par ordre décroissant de pertinence terminologique par l'application du coefficient de vraisemblance (Cf. Dunning T. 1993). Les résultats obtenus montrent une amélioration sensible des performances. Cependant, plusieurs caractéristiques des termes complexes ne sont pas abordées. D'une part, ACABIT

---

[1] Le texte n'est ni lémmatisé, ni étiqueté morpho-syntaxiquement, ni épuré à l'aide de listes de "stop-words".

[2] Séquences inninterrompues de mots.

[3] Séquences de mots interrompues par un espace ou plus qui représentent les occurrences de mots généralement synonymes.

[4] Nous répondons ainsi à la question posée par B. Habert et C. Jacquemin (1993:40).

ne permet pas l'extraction de termes complexes qui ne dérivent pas de la composition nominale. D'autre part, la définition du coefficient de vraisemblance ne permet de mesurer que les associations textuelles binaires. Cette contrainte a deux effets principaux. D'abord, comme il n'est pas possible de mesurer les associations pour les termes contenant plus de deux mots pleins, les termes complexes obtenus par surcomposition ou modification ne peuvent être classés. Ensuite, toutes les informations présentes dans le texte ne sont pas utilisées du fait que seuls les mots pleins comptent pour le classement des termes complexes. Afin d'illustrer les limitations d'ACABIT, nous présentons un ensemble d'unités lexicales complexes que le système n'est pas à même d'identifier: *home page du projet de documentation Linux*, *membre permanent du Conseil de Sécurité*, *des droits et des obligations*, *exporter une variable*, *agents thérapeutiques dérivés de ressources naturelles*, *Ministre des Affaires Étrangères*.

Parallèlement aux techniques linguistiques et hybrides, F. Smadja (1993) propose un système purement statistique pour l'extraction de termes complexes : XTRACT. Dans une première étape, XTRACT extrait l'ensemble des associations binaires dont le z-score est supérieur à une valeur seuil globale déterminée par l'utilisateur. Dans une deuxième étape, XTRACT examine le contexte immédiat de chaque association binaire retenue et analyse les distributions des mots coocurrents. Il identifie une association ternaire si la probabilité d'occurrence d'un mot donné dépasse un certain seuil. Finalement, le processus est itératif et termine lorsque plus aucune unité lexicale complexe n'est identifiée. XTRACT se distingue des deux autres méthodes précédemment citées par sa flexibilité d'utilisation. En effet, il ne dépend pas (du moins dans sa première partie) de la définition de contraintes morphologiques ou syntaxiques et peut être ainsi appliqué à tout type de texte (domaine ou langue). Cependant, XTRACT montre deux problèmes majeurs qui sont l'utilisation de méthodes d'amorçage et la définition de valeurs seuil globales. En effet, les associations binaires retenues pour la deuxième étape du système sont choisies sur la base d'une valeur seuil globale qui ne peut être définie qu'à partir d'expériences et qui par conséquent est sujette à un taux d'erreur non négligeable. D'autres part, le résultat des méthodes d'amorçage dépend des associations binaires retenues lors de la première étape du processus. Par exemple, l'identification du terme complexe *Traité de Maastricht*, dépend de l'identification préalable de l'association binaire *Traité de*. Or, le degré d'association entre *Traité* et *de* est généralement sous-évalué par les mesures statistiques du fait de la forte fréquence de la préposition *de* dans l'ensemble du corpus. Ainsi, l'unité complexe *Traité de Maastricht* ne sera probablement pas identifiée du fait de la non sélection du bigram *Traité de* durant l'etape initiale. Comparativement au français et au portugais, Il est de noter que ce problème est moins contraignant pour l'anglais qui fait fort usage de la composition nominale par juxtaposition. Par exemple, *Traité de Maastricht* est traduit par *Maastricht Treaty*. Afin de résoudre les problèmes présentés ci-dessus, nous présentons un analyseur statistique qui extrait à partir d'un texte brut un ensemble d'unités lexicales complexes grâce à l'Expectative Mutuelle et le LocalMaxs. En particulier, l'Expectative Mutuelle permet de caractériser la structure des termes complexes sans se limiter aux associations binaires et le LocalMaxs permet d'éviter la définition de valeurs seuil globales.

3. SPÉCIFICATION DES UNITÉS LEXICALES COMPLEXES

B. Habert et C. Jacquemin (1993) soulignent l'absence de critères purement linguistiques permettant de délimiter l'ensemble des unités lexicales complexes et espèrent que les analyses automatiques viennent compléter les indices fragiles fournis par les linguistes. Dans cette partie, nous proposons une spécification probabiliste des unités lexicales complexes basée sur la définition de collocation.

Plusieurs auteurs ont proposés dans le cadre de travaux théoriques et appliqués, différentes définitions du concept de collocation (Cf. Hausmann F. 1979; Cowie A. 1981; Benson M. 1989; Smadja F. 1993). De notre point de vue, M. Benson (1989) propose l'une des définitions les plus compréhensives : "*une collocation ... est une combinaison arbitraire et recurrente de mots*". Cependant, cette définition n'est pas suffisante dans le cadre des analyses statistiques puisqu'elle ne suggère que la fréquence comme facteur décisif pour l'extraction des collocations. Ainsi, F. Smadja (1993) introduit la notion essentielle de

plausibilité et définit une collocation comme étant : " *... une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard et qui correspondent à une utilisation arbitraire*". Malheureusement, les méthodes quantitatives n'assurent pas le caractère arbitraire de l'utilisation des mots qui se trouvent ensemble dans une combinaison. En effet, suivant la théorie cognitive, nous devons considérer toute suite de mots qui se réfère à un concept ou à un objet du domaine comme une unité lexicale complexe. Or, pour une grande partie de ces cas, le sens global de la combinaison n'est pas entièrement opaque. À l'extrême, le sens global de l'unité lexicale peut même être calculé par compositionalité. Par exemple, l'unité lexicale complexe *accord salarial* ne correspond pas une utilisation arbitraire des mots qui la forment. En effet, son sens peut raisonablement être approché par la compositon des sens de ses constituants. En opposition, l'unité polylexicale *coup de coeur* correspond à une utilisation arbitraire des mots qui la composent et son sens est complètement opaque. Par conséquent, nous définissons le concept d'unité lexicale complexe à partir de la spécification statistique des collocations en omettant leur caractère arbitraire. Ainsi, une unité lexicale complexe est *une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard*.

D'après leur définition, les unités lexicales complexes peuvent être divisées en trois structures différentes. Premièrement, les unités lexicales complexes peuvent être contiguës. Dans ce cas, ce sont des suites de mots ininterrompues telles que *huiles essentielles*, *disquette de boot*, *Président de la République*. Elles peuvent aussi être non-contiguës. Leur structure est alors représentée par une séquence de mots interrompue par un espace ou plus qui représentent un ensemble de mots interchangeables et généralement synonymes. Par exemple, *transport de _____ dangereuses* est une unité lexicale complexe où l'espace (i.e. "_____") peut être rempli par l'une des deux unités lexicales simples suivantes: *substances* ou *matières*. Finalement, les unités polylexicales peuvent être flexibles. Dans ce cas, les constituants de l'unité ne sont pas contraints par des positions fixes. Par exemple, *mettre à disposition* est une unité lexicale complexe flexible puisqu'elle peut être rencontrée dans le texte suivant des positions variables comme dans *mettre rapidement à disposition* ou *mettre plusieurs ressources importantes à disposition*.

4. ARCHITECTURE GLOBALE DU SYSTÈME

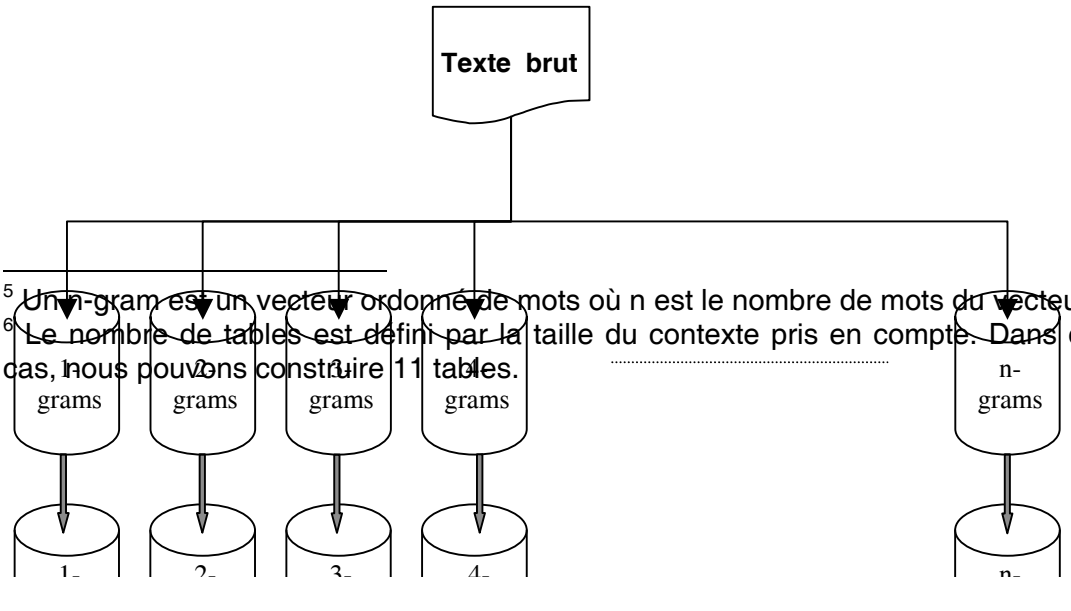
Après avoir défini le concept d'unité lexicale complexe du point de vue statistique et montré les différentes structures qui le caractérisent, nous présentons l'architecture globale de notre analyseur. Comme nous l'avons défini antérieurement, l'analyseur a été élaboré autour de l'idée principale de garantir une flexibilité maximum d'utilisation. Ainsi, il est possible d'extraire, à partir d'un texte brut, un ensemble d'unités lexicales complexes de toute taille et ceci indépendemment du domaine ou de la langue de l'enoncé.

La plupart des études proposées dans le cadre de l'acquisition d'unités lexicales complexes reposent sur l'analyse de textes étiquetés morpho-syntaxiquement. Les termes complexes sont ainsi extraits à partir de règles ou de patrons syntaxiques qui sont suportés par l'information linguistique préalablement incorporée dans les textes. Cependant, les unités lexicales complexes témoignent souvent d'une grande cohésion illustrée par des contraintes de figement telles que le blocage des paradigmes synonymiques ou le blocage des propriétés transformationnelles. G. Gross (1996) met par exemple l'accent sur le fait qu'un mot peut se trouver dans une construction avec son sens habituel sans pour autant pouvoir jouir de toutes les transformations habituelles de cette construction. J. Justeson (1993) corrobore cette constatation et affirme que plus une suite de mots est figée (i.e. moins elle accepte les transformations morphologiques et syntaxiques), plus il est probable que cette suite soit une unité lexicale complexe. En nous basant sur la théorie du figement, nous croyons que les unités polylexicales sont des suites de mots suffisamment figées pour proposer qu'elles soient extraites à partir de la seule information générale présente dans les textes et que l'introduction d'informations linguistiques n'est pas nécessaire dans le processus d'acquisition. L'analyseur reçoit ainsi en entrée un texte brut qui n'est ni lemmatisé, ni étiqueté morpho-syntaxiquement, ni épuré à partir de listes de "stop-words".

Cette décision est bien entendu sujette à débat. En effet, G. Grefenstette (1998) montre que le dosage efficace entre méthodes linguistiques et statistiques bénéficie une grande partie des applications du langage naturel. M. Sussna (1993) postule même que la description la plus riche est nécessairement la plus appropriée. Cependant, cela ne va pas de soi. L'introduction d'informations linguistiques dans les textes implique également l'addition de contraintes qui ne sont pas originellement comprises dans les énoncés. Par exemple, il n'est pas certain que la lémmatisation systématique (Cf. Church K.W. 1995) ou la morphologie dérivationnelle, avec notamment le regroupement des mots appartenant à la même famille dérivationnelle (ou *stemming*) (Cf. Gaussier É. *et al.* 1997), améliorent les performances du processus d'acquisition. Par ailleurs, les traitements linguistiques sont lourds et parfois imprécis, comme dans le cas des étiqueteurs morpho-syntaxiques. Finalement, l'épurage des textes à partir de listes de "stop-words" a été jusqu'ici accepté sans préoccupation par la communauté scientifique. Il faudrait pourtant évaluer la perte d'information due à l'utilisation exclusive des mots pleins pour le calcul des associations textuelles (Cf. Enguehard C. 1993; Daille B. 1995). En effet, cette approche définit les mots fonctionnels comme étant dénués de sens terminologique. Cela est loin d'être sûr et pose de nombreux problèmes. Ainsi, notre préoccupation a été de n'utiliser que l'information présente dans les textes et toute l'information présente dans les textes. Nous sommes bien entendu convaincus de la nécessité de l'apport linguistique dans le processus d'acquisition terminologique. En effet, les travaux statistiques basés sur l'étude des corpora permettent d'identifier les termes complexes utilisés dans leur contexte immédiat et n'autorisent pas de façon irréfutable à les cataloguer en tant que structures terminologiques. Ainsi, une certaine partie des termes doit être traitée linguistiquement pour peupler les banques terminologiques. L'objectif des analyseurs statistiques est alors de proposer le plus grand nombre possible de termes complexes terminologiquement valides afin de diminuer le traitement linguistique qui comme nous l'avons vu, implique un certain nombre de contraintes qu'il reste à évaluer de façon décisive.

L'architecture globale de notre analyseur s'articule autour de quatre étapes séquentielles qui sont présentées dans la Figure (1). Premièrement, l'analyseur transforme le texte brut reçu en entrée en un ensemble de tables de n-grams. Dans les deux étapes suivantes, l'analyseur calcule respectivement la fréquence puis l'Expectative Mutuelle de chaque n-gram. Finalement, la dernière étape de l'analyseur consiste à appliquer l'algorithme de sélection LocalMaxs sur l'ensemble des n-grams associés à leur mesure d'association. Nous exemplifions dans le paragraphe suivant le processus de transformation du texte brut en tables de n-grams.

De nombreux travaux lexicographiques (Cf. Sinclair J. 1974 ; Mason O. 1997) montrent que la plupart des relations lexicales associent des mots qui sont séparés les uns des autres par au plus cinq autres mots. Or, les unités lexicales complexes sont des relations lexicales spécifiques. Ainsi, une unité lexicale complexe peut être définie structurellement comme étant un n-gram[5] spécifique (contigu ou non) calculé à partir d'un mot pivot dans un contexte immédiat de 5 mots à gauche et 5 mots à droite du mot pivot. Par conséquent, le texte est traité séquentiellement et chaque mot étant tour à tour un mot pivot, tous les n-grams contigus et non-contigus sont calculés et stockés dans différentes tables[6].



[5] Un n-gram est un vecteur ordonné de mots où n est le nombre de mots du vecteur.
[6] Le nombre de tables est défini par la taille du contexte pris en compte. Dans ce cas, nous pouvons construire 11 tables.

Figure 1. Architecture Globale

Par exemple, si on considère la phrase (1) comme l'énoncé reçu en entrée et *Traité* (=w_1) le mot pivot en étude, alors la Figure (2) exhibe respectivement un 3-gram non-contigu (i.e. un vecteur ordonné de trois mots interrompus) et un 3-gram contigu (i.e. un vecteur ordonné de trois mots ininterrompus).

(1) "*Après de nombreuses négociations le Traité de Maastricht a été ratifié par tous les Etats -membres.*"

En effet, chaque 3-gram correspond à une combinaison spécifique qui contient 3 mots dont le mot pivot et qui est calculée dans le contexte immédiat de 5 mots à droite et 5 mots à gauche du mot pivot. Ainsi, le calcul de tous les 3-grams à partir de ce contexte ne peut dépasser à gauche le mot *Après* et à droite le mot *ratifié* pour l'étude du mot pivot *Traité*.

w_1	$position_{12}$[7]	w_2	$position_{13}$	w_3
Traité	-2	*négociations*	+5	*ratifié*
Traité	+1	*de*	+2	*Maastricht*

Figure 2. Exemples de 3-grams

En ce qui concerne la notation des n-grams non-contigus, chaque interruption du n-gram peut être identifiée par un espace (i.e. "_____") qui représente l'ensemble des occurrences du corpus qui remplissent la position libre (Expression (2)) ou par l'explicitation des positions (Expression (3)).

[7] Les $position_{12}$ et $position_{13}$ sont respectivement les distances entre w_1 and w_2 et entre w_1 and w_3. Le signe "+" ("-") est utillisé pour les mots à droite (gauche) of du mot pivot.

(2) *négociations* _____ *Traité* _____ _____ _____ _____ *ratifié*

(3) [*Traité* -2 *négociations* +5 *ratifié*]

En ce qui concerne la notation des n-grams contigus, chaque n-gram peut être représenté par la suite des ses constituants comme ils apparaissent dans le corpus (Expression (4)) ou par l'explicitation des positions (Expression (5)).

(4) *Traité de Maastricht*

(5) [*Traité* +1 *de* +2 *Maastricht*]

Un n-gram générique sera identifié par un vecteur (Expression (6)) dans lequel p_{1i} dénote la distance qui sépare le mot w_i du mot w_1, pour i=2,...,n.

(6) [w_1 p_{12} w_2 p_{13} w_3 ... p_{1i} w_i ... p_{1n} w_n]

Nous définissons l'Expectative Mutuelle et le LocalMaxs dans les deux parties suivantes de cet article.

5. LA MESURE D'EXPECTATIVE MUTUELLE

Les modèles statistiques, numériques ou probabilistes proposés dans la littérature (Cf. Salem A. 1987; Church K.W. & Hanks P. 1990; Gale W. 1991; Smadja F. 1993; Dunning T. 1993; Smadja F. 1996; Shimohata S. 1997) ne sont définis que pour les bigrams (i.e. n-grams avec n=2) et se restreignent généralement à l'acquisition d'associations textuelles binaires. Pour les associations de plus de deux unités lexicales simples, l'acquisition requiert un travail complémentaire où les paires d'association acquises initialement jouent le rôle d'amorce (Cf. Habert B. & Jacquemin C. 1993). Afin d'éviter le recours aux techniques d'amorçage qui comme nous l'avons vu précedemment posent un certain nombre de problèmes, nous présentons dans cet article le concept de normalisation qui permet d'évaluer le degré de cohésion de tout n-gram, pour tout $n \geq 2$. Parallèlement, la plupart des modèles statistiques, numériques ou probabilistes sont sensibles à l'occurrence d'unités lexicales atomiques fréquentes et par conséquent sous-évaluent généralement les associations entre constituants. Ainsi, B. Daille (1995) et C. Enguehard (1993) ne considèrent que les occurrences des mots pleins pour évaluer les forces de cohésion et évitent l'intégration des fragments fonctionnels souvent fréquents dans l'application des mesures statistiques. Malheuresement, cette caractéristique devient évidente lors de la normalisation et aboutit à la définition de valeurs de cohésion "incohérentes". Par conséquent, nous introduisons une nouvelle mesure d'association, l'Expectative Mutuelle (EM) (Cf. Dias G. *et al* 1999(a)), basée sur la notion d'Expectative Normalisée (EN).

5.1. L'expectative normalisée

Nous définissons l'expectative normalisée (EN) existant entre n mots comme étant l'expectative moyenne de voir apparaître un mot dans une position donnée sachant que les (n-1) autres mots apparaissent dans le texte contraints par leurs positions. Par exemple, l'expectative normalisée du 3-gram [*Traité* +1 *de* +2 *Maastricht*] doit représenter l'expectative moyenne de voir apparaître "*Traité*" sachant que "*de Maastricht*" conditionne son apparition, mais aussi de voir apparaître la préposition "*de*" entre "*Traité*" et "*Maastricht*" et finalement de voir apparaître "*Maastricht*" sachant que "*Traité de*" contraint son occurrence. Nous illustrons cette situation dans la Figure (3) dans laquelle chaque ligne correspond à une expectative particulière.

L'idée de base de l'expectative normalisée est d'évaluer le coût de la perte d'un mot dans un n-gram. Ainsi, plus une suite de mots est figée et témoigne d'une forte cohésion, moins cette séquence accepte la perte d'un de ses constituants et plus la valeur de l'expectative normalisée doit être élevée. Le concept sous-jascent à l'expectative normalisée est celui de la normalisation de la probabilité conditionnelle. À ce stade, nous voulons alerter le lecteur du fait que la probabilité conditionnelle est une mesure assymétrique. Or, comme le processus de normalisation définit une mesure symétrique, nous préférons utiliser le terme d'expectative normalisée plutôt que celui de probabilité conditionnelle normalisée qui peut porter à confusion.

Dans un premier temps, nous présentons la définition de la probabilité conditionnelle (Équation (1)). La probabilité conditionnelle mesure l'expectative d'apparition de l'événement X=x sachant que l'événement Y=y conditionne son apparition. p(X=x,Y=y) représente la fonction de densité conjointe entre deux variables aléatoires X, Y et p(Y=y) est la fonction de densité marginale de la variable Y. Ainsi, à chaque expectative spécifique correspond une probabilité conditionnelle.

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

Équation 1. Probabilité conditionnelle

D'où, le 3-gram *Traité de Maastricht* implique la définition de trois probabilités conditionnelles qui sont clairement identifiées dans la Figure (3).

Expectative d'apparaître le mot	Sachant l'occurrence de
Traité	[_____ +1 *de* +2 *Maastricht*]
de	[*Traité* +1 _____ +2 *Maastricht*]
Maastricht	[*Traité* +1 *de* +2 _____]

Figure 3. Expectatives possibles pour un 3-gram

Par conséquent, l'expectative normalisée doit caractériser en un seul tenant l'ensemble de toutes les probabilités conditionnelles impliquées par l'extraction d'un mot dans un n-gram. Afin de rendre possible cette caractérisation, nous proposons une normalisation de la probabilité conditionnelle en introduisant le concept d'Unité Moyenne d'Expectative.

Considérons le n-gram $[w_1\ p_{12}\ w_2\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$. Ce n-gram est équivalent au n-gram $[w_1\ p_{12}\ w_2\ p_{23}\ w_3\ ...\ p_{2i}\ w_i\ ...\ p_{2n}\ w_n]$ où $p_{2i} = p_{1i} - p_{12}$ pour i=3,...,n et p_{2i} dénote la distance qui sépare le mot w_i du mot w_2. Cette transformation est nécessaire car nous voulons considérer un n-gram comme la composition de n pseudo-2-grams obtenus à partir du n-gram en lui retirant un mot à la fois. Cette opération est à l'origine de la définition de n événemements que nous représentons dans la Figure (4) dans laquelle l'espace (i.e. "_____") correspond au mot extrait du n-gram.

(*n-1*)-gram résultant	mot extrait
$[\ \_\_\_\_\_\ w_2\ \ p_{23}\ \ w_3\ ...\ p_{2i}\ \ w_i\ ...\ p_{2n}\ \ w_n]$	w_1
$[w_1\ \_\_\_\_\_\ \ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$	w_2
...	...
$[w_1\ p_{12}\ w_2\ p_{13}\ w_3 ... p_{1(i-1)}\ w_{(i-1)}\ \_\_\_\_\_\ p_{1(i+1)}\ w_{(i+1)} ... p_{1n}\ w_n]$	w_i
...	...
$[w_1\ p_{12}\ w_2\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1(n-1)}\ w_{(n-1)}\ \_\_\_\_\_\ ]$	w_n

Figure 4. n pseudo-2-grams contenus dans un n-grams

Comme nous l'avons mentionné ci-dessus, la normalisation a pour objectif de réunir en une seule mesure d'association les n probabilités conditionnelles qui résultent de la décomposition d'un n-gram en n pseudo-2-grams et qui sont définies dans les équations (2) et (3).

$$p(w_1 p_{12} \mid [\ w_2\ \ p_{23}\ \ w_3 ... p_{2i}\ \ w_i ... p_{2n}\ \ w_n]) = \frac{p([\ w_1 p_{12} w_2\ \ p_{23}\ \ w_3 ... p_{2i}\ \ w_i ... p_{2n}\ \ w_n])}{p([w_2\ \ p_{23}\ \ w_3 ... p_{2i}\ \ w_i ... p_{2n}\ \ w_n])}$$

Équation 2. Extraction du premier mot du n-gram.

$$\forall i, i = 2..n, \quad p(w_i p_{1i} \mid [\, w_1 \ldots p_{1(i-1)}\ w_{(i-1)}\ p_{1(i+1)}\ w_{(i+1)} \ldots p_{1n}\ w_n]) = \frac{p([\, w_1 p_{12}\ w_2\ \ p_{13}\ \ w_3 \ldots p_{1i}\ \ w_i \ldots p_{1n}\ \ w_n])}{p([w_1 \ldots p_{1(i-1)}\ w_{(i-1)}\ p_{1(i+1)}\ w_{(i+1)} \ldots p_{1n}\ w_n])}$$

Équation 3. Extraction des mots autres que le premier du n-gram.

Dans le cadre de la normalisation de la probabilité conditionnelle, nous introduisons la notion d'Unité Moyenne d'Expectative (UME). En effet, l'analyse de l'ensemble des probabilités conditionnelles illustrées dans les équations (2) et (3) montre que les numérateurs restent inchangés d'une probabilité à l'autre et que seuls les dénominateurs changent. Par conséquent, l'introduction de la normalisation des dénominateurs (i.e. l'UME) dans la définition de la probabilité conditionnelle (Équation (1)) propose une solution élégante pour la normalisation de la probabilité conditionnelle. Ainsi, l'UME est définie comme étant la moyenne arithmétique de tous les dénominateurs des équations (2) et (3) et réunit en un événement moyen et unique tous les événements particuliers impliqués par chaque probabilité. Rigoureusement, l'UME d'un n-gram donné est la moyenne arithmétique de toutes les probabilités conjointes des n (n-1)-grams contenus dans le n-gram et est définie par l'équation (4). L'accent circomflexe "^" correspond à une convention fréquemment utilisée en Algèbre qui consiste à écrire un "^" au-dessus du terme omis d'une suite indexée de 2 à n.

$$UME([w_1 p_{12}\ w_2 \ldots p_{1i}\ \ w_i \ldots p_{1n}\ \ w_n]) = \frac{1}{n}\left(p([w_2 \ldots p_{2i}\ \ w_i \ldots p_{2n}\ w_n]) + \sum_{i=2}^{n} p\left(\left[w_1 \ldots \hat{p}_{1i}\ \ \hat{w}_i \ldots p_{1n}\ w_n \right]\right)\right)$$

Équation 4. Définition de l'Unité Moyenne d'Expectative d'un n-gram

Finalement, l'expectative normalisée d'un n-gram est introduite comme étant une probabilité conditionnelle "juste" qui utilise le concept de l'Unité Moyenne d'Expectative et est défini par l'équation (5).

$$EN([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = \frac{p([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])}{UME([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])}$$

Équation 5. Définition de l'Expectative Normalisée d'un n-gram

5.2. L'expectative mutuelle

J. Justeson (1993) et B. Daille (1995) ont montré dans leurs études que l'un des critères les plus importants pour l'identification des unités lexicales

complexes est la fréquence. G. Gross (1996) corrobore cette opinion et affirme que la compréhension d'une unité polilexicale est un processus itératif, étant nécessaire qu'une unité soit prononcée plus d'une fois pour que sa compréhension soit possible. Finalement, de par définition, les unités lexicales complexes sont des suites de mots récurrentes. Or, l'expectative normalisée mesure le degré de cohésion qui lie les constituants d'un n-gram mais ne rend pas compte de l'hypothèse formulée précédemment. Certains de cette assomption, nous déduisons qu'entre deux n-grams ayant la même expectative normalisée, il est plus probable que le plus fréquent des deux corresponde à une unité lexicale complexe. Ainsi, à partir de l'expectative normalisée et de la fréquence relative nous définissons l'expectative mutuelle dans l'équation (6).

$$EM([w_1 ... p_{1i}\ w_i\ ...\ p_{1n}\ w_n]) = p([w_1 ... p_{1i}\ w_i\ ...\ p_{1n}\ w_n]) \times EN([w_1 ... p_{1i}\ w_i\ ...\ p_{1n}\ w_n])$$

Équation 6. Définition de l'Expectative Mutuelle d'un n-gram

L'expectative mutuelle permet donc de mesurer le degré de cohésion de tout n-gram sans être limité aux associations binaires. Ainsi, il est possible de classer n'importe quel n-gram (i.e. $n \geq 2$) suivant son degré de pertinence. À partir de l'ensemble des n-grams associés à leur valeur d'expectative mutuelle, nous présentons la dernière étape de notre analyseur qui applique l'agorithme LocalMaxs pour l'extraction des unités lexicales complexes.

6. LE LOCALMAXS

Dans le cadre du processus d'acquisition des unités lexicales complexes, la plupart des approches statistiques et hybrides se basent sur la définition de valeurs seuil globales (Cf. Church K.W. & Hanks P. 1990; Smadja F. 1993; Dunning T. 1993; Daille B. 1995; Smadja F. 1996; Shimohata S. 1997). Les seuils (i.e. valeurs limites globales qui permettent de définir si une association textuelle est d'intérêt ou non) font l'objet d'un ajustement qui est crucial pour la réussite des expériences statistiques présentées dans la litérature. Il s'agit d'un compromis entre des valeurs (de fréquence ou de mesure d'association) assez permissives pour que la collecte soit importante (taux de rappel) et des valeurs pas trop généreuses pour que le résultat soit précis (taux de précision). Malheureusement, cette approche se révèle peu fiable et peu flexible. En effet, les résultats dépendent des expériences et sont par conséquent sujets à erreur. De plus, suivant que la longueur, le type, le domaine et la langue du corpus ainsi que la mesure d'association utilisée changent, il est nécessaire de réajuster les valeurs des seuils. Ainsi, nous introduisons un nouvel algorithme, le LocalMaxs (Cf. Silva F. *et al* 1999), qui ne dépend d'aucun seuil pré-établi ou mesuré par expérimentation et qui élit tout n-gram dont le degré d'association correspondant est un maximum local.

Le LocalMaxs elit les unités lexicales complexes à partir de l'ensemble des n-grams associés à leur mesure d'association, en s'appuyant sur deux

hypothèses. D'une part, les mesures d'association montrent que plus une suite de mots est figée et cohésive, plus sa valeur de mesure d'association est forte[8]. D'autre part, les unités lexicales complexes témoignent d'une forte cohésion dans leur contexte immédiat et sont par conséquent localement motivées. Ainsi, un n-gram est une unité lexicale complexe si le degré d'association qui lie ses n constituants est supérieur aux mesures d'association de tous ses sous-groupes de (n-1) mots (i.e. toutes les suites de (n-1) mots contenues dans le n-gram) et strictement supérieur aux mesures d'association de tous ses super-groupes de (n+1) mots (i.e. toutes les suites de (n+1) mots qui contiennent le n-gram) . Le LocalMaxs analyse ainsi les fluctuations locales des mesures d'associations dans le contexte immédiat des n-grams et ne recourt donc pas à la définition de valeurs seuil globales. Le LocalMaxs base son processus d'extraction sur l'analyse de valeurs seuils locales motivées par le contexte immédiat de chaque n-gram.

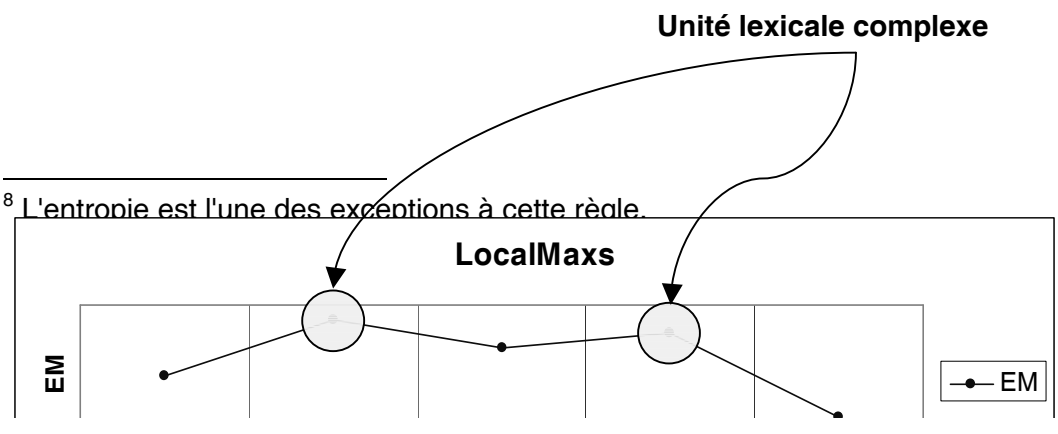
Le LocalMaxs est formellement définit dans la Figure (5). Soient, une mesure d'association, *assoc*, un n-gram, *W*, l'ensemble de tous les (n-1)-grams contenus dans *W*, Ω_{n-1}, l'ensemble de tous les (n+1)-grams contenant *W*, Ω_{n+1} et une fonction *taille* qui retourne la longueur d'un n-gram *W* donné en argument, alors:

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$

W est une unité lexicale complexe si $(taille(W)=2 \wedge assoc(W) > y) \vee (assoc(W) \geq x \wedge assoc(W) > y)$

Figure 5. Algorithme du LocalMaxs

Afin que le lecteur se familiarise avec cet algorithme, nous exemplifions son fonctionnement sur le 3-gram [*Traité* +1 *de* +2 *Maastricht*]. D'après le LocalMaxs, le 3-gram [*Traité* +1 *de* +2 *Maastricht*] peut être considéré comme une unité lexicale complexe si tous ses sous-groupes ont une mesure d'association inférieure à la sienne et si tous ses super-groupes (comme définis ci-dessus) ont une mesure d'association strictement inférieure à la sienne. Or, les valeurs d'expectative mutuelle des 2-grams [*Traité* +1 *de*], [*Traité* +2 *Maastricht*] et [*de* +1 *Maastricht*] sont, dans le pire des cas égales, mais jamais supérieures à celle du 3-gram. En effet, la perte d'un des constituants du 3-gram diminue évidemment son intégrité et sa cohésion. Dans le cas des super-groupes contenant le 3-gram, comme il n'existe aucun mot qui puisse renforcé la cohésion de *Traité de Maastricht*, tous les 4-grams qui contiennent le 3-gram ont une mesure d'expectative mutuelle strictement inférieure à celle de [*Traité* +1 *de* +2 *Maastricht*]. Ainsi, le 3-gram est considéré par le système comme étant une unité lexicale complexe.



[8] L'entropie est l'une des exceptions à cette règle.

Figure 6. Exemple d'extraction par juxtaposition

Entre autres, le LocalMaxs montre deux propriétés intéressantes. D'une part, il peut être appliqué à toute mesure d'association qui respecte la première hypothèse formulée dans le deuxième paragraphe (i.e. plus une suite de mots est figée et cohésive, plus sa valeur de mesure d'association est forte). Ainsi, nous avons mené plusieurs expériences, appliquant tour à tour au LocalMaxs les mesures d'association normalisées suivantes: l'Information Mutuelle spécifique (Cf. Church K.W. & Hanks P. 1990), le coefficient Dice (Cf. Smadja F. 1996), le ϕ^2 (Cf. Gale W. 1990) et le coefficient de vraisemblance (Cf. Dunning T. 1993)[9]. D'autre part, le LocalMaxs permet l'extraction de termes complexes obtenus par composition. En effet, comme le LocalMaxs extrait les unités lexicales complexes à partir de l'analyse des contextes immédiats des n-grams, l'algorithme peut identifier des associations entre constituants qui sont eux-mêmes extraits comme étant des associations pertinentes. Par exemple, le LocalMaxs élira l'unité lexicale complexe *Ministre de l'Intérieur Jean-Pierre Chevènement* à partir des deux unités complexes également extraites *Ministre de l'Intérieur* et *Jean-Pierre Chevènement*. Cela est illustré dans la Figure (6). En effet, la valeur de l'expectative mutuelle du pentagram (ou 5-gram) *Ministre de l'Intérieur Jean-Pierre Chevènement* correspond à un maximum local. Nous expliquons les principes de base qui permettent l'extraction de cette unité juxtaposée. La valeur de l'expectative mutuelle associée à *Ministre de l'Intérieur* est strictement supérieure à celle associée à *Ministre de l'Intérieur Jean-Pierre* puisqu'il existe plusieurs ministres de l'intérieur dans le cadre de l'Union Européenne. Ainsi, l'apport d'un mot dans le 3-gram ne renforce pas son degré de cohésion. Bien au contraire, il introduit un facteur de flexibilité. Par contre, l'apport du nom *Chevènement* à l'unité *Ministre de l'Intérieur Jean-Pierre*, vient renforcer les liens qui associent tous les constituants en relation. Ainsi, l'expectative mutuelle du pentagram *Ministre de l'Intérieur Jean-Pierre Chevènement* est plus forte que celle du tetragram (ou 4-gram) *Ministre de l'Intérieur Jean-Pierre* et parallèlement plus forte que celle de l'hexagram (ou 6-gram) *Ministre de l'Intérieur Jean-Pierre Chevènement rentre*. Finalement, le LocalMaxs élit les unités lexicales complexes suivantes: *Ministre de l'Intérieur*, *Jean-Pierre Chevènement* et *Ministre de l'Intérieur Jean-Pierre Chevènement*. Cette propriété est à réhausser comparativement aux sytèmes statistiques qui associent méthodes d'amorçage et valeurs seuil globales. En effet, ces derniers n'autorisent pas l'extraction d'unités lexicales complexes juxtaposées puisque le processus de sélection termine pour une suite de mots lorsque celle-ci a été élue par le système.

Finalement, le LocalMaxs propose une solution robuste et flexible au problème de la sélection des unités lexicales complexes parmi l'ensemble des n-grams associés à leur mesure d'association. En effet, il ne dépend pas de valeurs seuil globales et permet l'extraction de termes complexes obtenus par composition.

[9] Les coefficients de Cramer et de Pearson ont été également testés (Cf. Bhattacharyya G. & Johnson R. 1977).

7. RÉSULTATS

L'objectif de cette partie est dans un premier temps de justifier l'introduction de l'expectative mutuelle dans le cadre des modèles probabilistes pour l'acquisition d'associations textuelles et dans un deuxième temps, d'illustrer l'efficacité de l'analyseur en exhibant un certain nombre de termes complexes extraits à partir de corpora de différentes langues et de différents domaines.

7.1. Analyse comparative

Afin de pouvoir évaluer les résultats obtenus avec l'expectative mutuelle, nous avons testé le LocalMaxs avec quatre autres modèles mathématiques couramment utilisés dans la littérature: l'Information Mutuelle spécifique (Cf. Church K.W. & Hanks P. 1990), le coefficient Dice (Cf. Smadja F. 1996), le ϕ^2 (Cf. Gale W. 1990) et le coefficient de vraisemblance (Cf. Dunning T. 1993). Nous rappelons leurs formules pour les associations binaires dans les équations (7), (8), (9) et (10), où N et f correspondent respectivement au nombre de mots dans le corpus et à la fonction "fréquence" qui retourne le nombre d'occurrences d'un n-gram.

$$\mathrm{MI}\ (\lceil w_1 p_{12}\ w_2 \rceil) = \log_2\left(\frac{N * f(\lceil w_1 p_{12}\ w_2 \rceil)}{f(\lceil w_1 \rceil) * f(\lceil w_2 \rceil)}\right)$$

Équation 7. Information Mutuelle Spécifique

$$\mathrm{Dice}\ (\lceil w_1 p_{12}\ w_2 \rceil) = \frac{2.f(\lceil w_1 p_{12}\ w_2 \rceil)}{f(\lceil w_1 \rceil) + f(\lceil w_2 \rceil)}$$

Équation 8. Coefficient Dice

$$\Phi^2\ (\lceil w_1 p_{12}\ w_2 \rceil) = \frac{(N\ * f(\lceil w_1 p_{12}\ w_2 \rceil) - f(\lceil w_1 \rceil)\ * f(\lceil w_2 \rceil))^2}{f(\lceil w_1 \rceil)\ * (N - f(\lceil w_1 \rceil))\ * f(\lceil w_2 \rceil)\ * (N - f(\lceil w_2 \rceil))}$$

Équation 9. le ϕ^2

$$\mathrm{Loglike}\ (\lceil w_1 p_{12}\ w_2 \rceil) = -2\log\lambda\ =$$
$$2 * \left(\log\theta_1^{s1}(1-\theta_1)^{n1-s1} + \log\theta_2^{s2}(1-\theta_2)^{n2-s2} - \log\theta^{s1}(1-\theta)^{n1-s1} - \log\theta^{s2}(1-\theta)^{n2-s2}\right)$$

avec

$$\begin{cases} s1 = f([w_1 p_{12} w_2]) \\ n1 = f([w_1]) \\ \theta_1 = \dfrac{s1}{n1} \end{cases} \quad \text{et} \quad \begin{cases} s2 = f([w_2]) - f([w_1 p_{12} w_2]) \\ n2 = N - f([w_1]) \\ \theta_2 = \dfrac{s2}{n2} \end{cases} \quad \text{et} \quad \theta = \frac{f([w_2])}{N}$$

Équation 10. Coefficient de Vraisemblance

Comme nous l'avons mentionné antérieurement, les quatre mesures d'association proposées par Church, Smadja, Gale et Dunning ne sont définies que pour le cas des 2-grams. Par conséquent, afin d'intégrer chaque modèle au LocalMaxs, il a fallu préalablement procéder à leur normalisation[10]. Finalement, l'évaluation du processus d'extraction a été réalisée à partir d'un corpus en français d'environ 300000 mots qui a été extrait d'une base documentaire multilingue recensant un vaste ensemble de débats du Parlement Européen traduits en plusieurs langues de l'Union Européenne[11].

Les résultats montrent que l'expectative mutuelle appliquée au LocalMaxs permet un saut qualitatif significatif comparativement aux autres modèles (Cf. Dias G. *et al.* 1999(b)). Nous nous attarderons dans cette partie sur l'une des caractéristiques les plus marquantes que chaque mesure partage à l'exception de l'expectative mutuelle. En effet, les modèles proposés par Church, Smadja, Gale et Dunning se montrent particulièrement sensibles à l'occurrence d'unités lexicales atomiques fréquentes dans les n-grams et aboutissent à la définition de valeurs de cohésion "incohérentes". Par exemple, le coefficient Dice et le ϕ^2 elisent le bigram *turcs* _____ *kurdes* comme étant l'expression la plus discriminante du corpus. Or, la probabilité que la conjonction "*et*" apparaisse entre "*turcs*" et "*kurdes*" vaut un. En fait, comme le 3-gram [*turcs* +1 *et* +2 *kurdes*] reçoit une valeur d'association inférieure à celle du 2-gram [*turcs* +2 *kurdes*], le LocalMaxs élit préférablement l'association binaire. En effet, la fréquence élevée de la conjonction "*et*" dans le corpus sous-estime le degré de cohésion du 3-gram et aboutit à une mesure "incohérente". Par opposition, l'analyseur utilisé avec l'expectative mutuelle élit l'unité lexicale complexe la plus longue et la plus fréquente qui englobe les deux mots "*turcs*" et "*kurdes*". Dans le but d'exemplifier le résultat obtenu à partir de l'expectative mutuelle, nous présentons le produit du concordanceur (Figure (7)) lorsque les mots "*turcs*" et "*kurdes*" sont recherchés dans le texte en étant séparés l'un de l'autre par un seul mot. Le résultat montre clairement que lorsque "*turcs*" et "*kurdes*" se

[10] Plusieurs normalisations ont été appliquées et nous continuons à tester de nouvelles possibilités. Quoiqu'il en soit, pour chaque mesure, la meilleure des normalisations (i.e c'est-à-dire la normalisation qui produit le plus grand nombre d'unités précises) a été prise en compte pour effectuer l'évaluation.

[11] Cette base documentaire a été acquise auprès de l'European Language Resources Association (ELRA) - http://www.icp.grenet.fr/ELRA/home.html

trouvent ensemble, le nom "*réfugiés*", l'adjectif "*politiques*" et la conjonction "*et*" apparaissent également pour former l'unité lexicale complexe "*réfugiés politiques turcs et kurdes*".

de la faim sept	*réfugiés politiques turcs et kurdes*	*en Grèce Sept réf*
de la faim des	*réfugiés politiques turcs et kurdes*	*en protestation cont*
entants des sept	*réfugiés politiques turcs et kurdes*	*qui font la grève de*
n Grèce les sept	*réfugiés politiques turcs et kurdes*	*qui sont détenus en*
maines que sept	*réfugiés politiques turcs et kurdes*	*tenus en isolement*

Figure 7. Concordances de *turcs _____ kurdes*

Par conséquent, l'analyseur associé à l'expectative mutuelle extrait l'unité polylexicale "*réfugiés politiques turcs et kurdes*". Pareillement au Dice et au ϕ^2, l'Information Mutuelle spécifique et le coefficient de vraisemblance montrent le même problème. Par exemple, l'Information Mutuelle spécifique élit le bigram *surface _____ globe* plutôt que d'extraire l'unité lexicale complexe "*changement climatique à la surface du globe*" qui corresopnd à l'unité la plus longue et la plus fréquente contenant les deux mots "*surface*" et "*globe*"[12]. De la même façon, le coefficient de vraisemblance extrait le trigram *communauté _____ ses Etats* plutôt que la suite "*la Communauté et ses Etats membres*" qui est classée par l'expectative mutuelle comme l'une des unités lexicales complexes les plus représentatives du corpus. Contrairement aux quatre autres modèles statistiques présentés, l'expectative mutuelle exhibe une cohérence totale des résultats. Ainsi, chaque fois qu'une unité lexicale complexe non-contiguë est élue, chaque interruption présente dans le n-gram correspond à l'occurrence d'au moins deux mots différents du corpus. Par exemple, l'expectative mutuelle associée au LocalMaxs élit l'unité non-contiguë "*membres des _____ d'inspection*" où l'espace correspond aux occurrences de "*missions*" et "*services*".

7.2. Analyse qualitative

Grâce à sa flexibilité, l'analyseur a pu être testé sur des corpora de différentes tailles, domaines et langues (Cf. Dias G. *et al* 1999(c)). Dans tous les cas de figure, il a permis l'extraction directe de termes de base, de termes obtenus par composition et de termes construits par modification. Dans ce cadre, l'analyseur assure un taux de précision de 70%[13]. Nous avons même montré lors d'une étude multi-domaine que la précision peut atteindre les 80% si un simple post-traitement est appliqué à l'ensemble des n-grams extraits par le LocalMaxs (i.e. exclusion de l'ensemble des associations textuelles communes à plusieurs domaines). Cependant, comme le rappellent B. Habert et C.

[12] L'unité "*surface du globe*" peut être considérée comme une unité lexicale complexe. Cependant, ses constituants ne se trouvent jamais ensemble dans le corpus sans que l'autre unité polylexicale "*changement climatique*" s'y trouve aussi (et inversement). Par conséquent, il n'existe aucune base statistique pour préférer l'expression "*surface du globe*" à celle de "*changement climatique à la surface du globe*".

[13] Nous n'avons retenu que les unités polylexicales contenant au plus une interruption. En effet, les résultats montrent que plus le nombre d'interruptions est important et plus l'information lexicographique est disperse.

Jacquemin (1993), les systèmes statistiques identifient également des associations textuelles terminologiquement non valides. Notre analyseur n'échappe pas à cette règle. En effet, l'analyse détaillée des résultats exhibe l'extraction de locutions adjectivales, adverbiales, prépositives et conjonctives ainsi que l'identification de déterminants composés et quelques fragments syntaxiques. Dans le cadre de la recherche documentaire, nous nous attacherons à exemplifier les résultats obtenus par l'analyseur pour l'extraction des termes complexes. Suivant cet objectif, nous structurons notre étude autour de la classification des termes complexes en trois catégories: les termes de base, les termes obtenus par composition et les termes obtenus par modification.

7.2.1. Termes de base

Un terme de base correspond à un n-gram contigu candidat qui n'est constitué d'aucun autre n-gram élu. Ainsi, un terme de base n'est ni restreint par sa longueur ni par sa catégorie syntaxique. Nous présentons dans la Figure (8) quelques exemples de termes de base qui ont été extraits à partir du corpus en français du Parlement Européen.

EM	Fréquence	Terme de base
0.000105067	21	*Parlement européen*
0.000105067	2	*José Saramago*
0.000100499	28	*Coopération politique européenne*
0.000100499	2	*Conseil de sécurité*
0.000100499	2	*droit de vote*
0.000100499	2	*statuant à l'unanimité*
0.000102732	4	*prix européen de littérature*
0.000102732	4	*chômage de longue durée*
0.000102732	4	*Comité économique et social*
0.000102732	2	*ministères des Länder allemands*
0.002322371	2	*l'offre et la demande*
0.000128415	2	*Petites et moyennes entreprises*
0.000102732	2	*faire don de ses organes*
0.000103191	4	*« Semaine européenne de l'entreprise »*
0.000128415	2	*changement climatique à la surface du globe*

Figure 8. Termes de base

7.2.2. Termes complexes obtenus par composition

Un terme complexe obtenu par composition correspond à un terme construit à partir d'un ou plusieurs termes de base. Cette catégorie englobe entre autres les termes complexes construits par juxtaposition, substitution, postposition de modifieurs et coordination, selon la classification proposée par B. Daille (1995)[14]. Nous présentons dans la Figure (9) quelques exemples de termes complexes obtenus par composition.[15]

EM	F	Termes obtenus par composition
0.000102	3	[*conseil municipal*] *de Torfaen*

[14] Nous ne suivons pas cette classification car B. Daille utilise des patrons syntaxiques prédéfinis pour définir chacune de ces catégories. Or, nous ne disposons pas de telles informations. De plus, les resultats mis en évidence par notre analyseur exhibent un plus grand nombre de phénomènes linguistiques que ceux proposés par B. Daille.

[15] Les termes de base constituants sont identifés entre crochets.

0.000102	4	*la compétence des* [*États membres*]
0.000128	2	*Le* [*Conseil européen*] *de Lisbonne*
0.000102	2	[*Journal officiel*] *des* [*Communautés européennes*]
0.000102	2	*règlements régissant les* [*Fonds structurels*]
0.000102	2	*l'exposition aux* [*rayonnements non ionisants*]
0.000102	2	[*secteur communautaire*] *des* [*fibres synthétiques*]
0.000102	2	[*autorités compétentes*] *des* [*États membres*]
0.000128	2	*le dialogue entre* [*partenaires sociaux*]
0.000128	3	*la situation des* [*droits de l'homme*]
0.000128	3	*les émissions de* [*dioxyde de carbone*]
0.000128	2	[*Convention d'application*] *de* [*l'Accord de Schengen*]
0.000128	2	[*recommandation du Conseil*] *du 8 novembre 1988*
0.000128	2	[*mission de surveillance*] de la [*Communauté européenne*]
0.000128	2	[*Chefs d'État*] des [*pays de la Communauté*]

Figure 9. Termes obtenus par composition

7.2.3. Termes complexes obtenus par modification

Un terme obtenu par modification est un n-gram non-contigu contenant exactement une et une seule interruption[16]. En effet, l'insertion de modifieurs à l'intérieur d'un terme implique l'introduction d'un facteur de flexibilité qui doit correspondre à une interruption. De fait, il est possible que plusieurs occurrences modificatrices s'insèrent à l'intérieur d'un terme complexe. Par exemple, les deux modifieurs "*européen*" et "*mondial*" peuvent s'intégrer dans le terme de base "*Conseil des télécommunications*" pour donner "*Conseil européen des télécommunications*" ou "*Conseil mondial des télécommunications*". Dans le cadre de notre analyseur, l'étude des n-grams non-contigus permet la représentation des termes obtenus par modification par l'identification de l'élément modifieur sous la forme d'un saut dans la suite des mots. Ainsi, l'analyseur élira le terme complexe non-contigu *Conseil _____ des télécommunications* qui représente le concept incarné par les deux formes modifiées et dans lequel l'interruption identifie l'élément flexible du terme complexe. Les résultats obtenus montrent que l'élément modifieur ne correspond pas forcément à un adjectif ou à un adverbe comme le propose B. Daille mais peut s'apparier à un éventail de formes morpho-syntaxiques (Voir Figure (10)). En effet, un grand nombre de termes non-contigus identifient par le biais de l'interruption leur propre processus de figement. Par exemple, l'unité "*passation _____ marchés*" accepte les deux modifieurs "*de*" et "*des*" qui mettent en évidence la transformation de la préposition contractée "*des*" en la préposition "*de*" pour former un terme complexe figé.

EM	F	Termes obtenus par modification	Modifieurs
0.000105	9	*passation _____ marchés*	*de* *des*
0.000105	2	*contrôles _____ frontières*	*de* *des*
2.708e-05	2	*transport de _____ dangereuses*	*matières* *substances*
2.708e-05	2	*la sélection _____ candidats*	*de* *des*
2.708e-05	3	*publier des _____ nationaux*	*programmes* *plans*
2.708e-05	4	*l'article _____ du règlement*	*6* *32* *premier*

[16] Les termes caractérisés par plusieurs interruptions exhibent des phénomènes lexicographiquement valides mais terminologiquement moins fiables.

0.000102	2	*proposition de _____ du Conseil*	*Directive Réglement*
0.000102	2	*la _____ des énergies renouvelables*	*promotion divulgation*
0.000128	2	*l'égalité de _____ entre hommes et femmes*	*rémunération traitement*

Figure 10. Termes obtenus par modification

Les termes obtenus par modification sont un atout non négligeable dans le cadre de la recherche documentaire, tant au niveau de l'indexation comme au niveau de l'aide à l'utilisateur. En effet, un terme non-contigu représente un concept général qui peut être spécialisé en appariant son interruption à l'un de ses modifieurs. Ainsi, il est possible de généraliser une requête en proposant comme terme d'indexation[17] son concept sous-jascent. Par exemple, si un utilisateur désire accéder à l'information relative aux "*transports de substances dangereuses*", il est souhaitable que le système de recherche documentaire puisse lui proposer les textes concernant les "*transports de matières dangereuses*". Ainsi, l'utilisation du terme non-contigu "*transports de _____ dangereuses*" comme terme d'indexation propose une solution élégante pour la généralisation de la requête, permettant le recouvrement de textes en relation directe. Les termes obtenus par modification sont par conséquent primordiaux pour une meilleure flexibilité des systèmes de recherche documentaire.

Par ailleurs, les termes non-contigus permettent l'identification d'hapaxes discriminants qui possibilitent le recouvrement d'un plus vaste ensemble de textes pertinents. En effet, un terme obtenu par modification correspond à un ensemble de termes spécialisés qui peuvent se manisfester une et une seule fois dans la base documentaire. Par exemple, l'unité lexicale complexe "*proposition de _____ du Conseil*" correspond aux deux hapaxes "*proposition de Directive du Conseil*" et "*proposition de Réglement du Conseil*". Les termes complexes non-contigus utilisés comme termes d'indexation permettent ainsi l'identification de termes représentatifs du domaine indépendamment de leur fréquence. Ainsi, l'un des principaux problèmes des analyses statistiques peut être raisonablement minimisé par l'extraction d'unités lexicales complexes non-contiguës: le problème de la "fréquence un". En effet, l'extraction statistique d'unités lexicales complexes n'est possible que pour les suites de mots qui apparaissent au minimum deux fois dans le corpus. Bien que notre analyseur n'échappe pas à cette contrainte, il propose une amélioration sensible au problème en autorisant l'élection de termes complexes non-contigus.

8. CONCLUSION ET TRAVAIL FUTUR

Comme nous l'avons motivée tout au long de cet article, la reconnaissance des unités lexicales complexes des collections de textes toujours croissantes constitue un enjeu fondamental dans le cadre de la recherche documentaire. En effet, les unités polylexicales sont souvent moins ambiguës et plus motivées que les mots simples et permettent ainsi une meilleure approximation des thèmes abordés. Dans cet article, nous avons proposé un analyseur statistique qui extrait à partir d'un texte brut un ensemble d'unités lexicales complexes contiguës et non-contiguës sans recourir aux méthodes d'amorçage ni à la définition de valeurs seuil globales. Ainsi, nous avons conjugué une nouvelle mesure d'association fondée sur le concept d'expectative normalisée, l'Expectative Mutuelle, avec un nouveau processus d'extraction basé sur un algorithme de maxima locaux, le LocalMaxs. En particulier, l'Expectative Mutuelle permet de caractériser la structure des termes complexes sans se limiter aux associations binaires et le LocalMaxs permet d'éviter la définition de valeurs seuil globales dans le processus d'acquisition. Par ailleurs, l'introduction de la

[17] Un terme d'indexation est soit une clé d'indexation (indexation des documents) soit un terme qui est proposé à l'utilisateur dans sa quête d'information (aide à l'utilisateur).

mesure d'Expectative Mutuelle a permis une amélioration sensible dans le cadre de l'extraction d'unités lexicales complexes, comparativement aux modèles mathématiques couramment mentionnés dans la littérature: l'Information Mutuelle spécifique, le coefficient Dice, le ϕ^2 et le coefficient de vraisemblance. Finalement, les résultats obtenus par l'analyseur appliqué à des collections de textes de différents domaines, tailles et langues ont mis en évidence l'extraction de termes de base ainsi que l'acquisition de termes obtenus par composition et modification avec un taux de précision significatif.

Du fait de sa flexibilité, notre analyseur va être associé au moteur de recherche développé par l'Université Nouvelle de Lisbonne dans le cadre d'un projet de fouille de textes dans une base documentaire du domaine législatif portugais (Quaresma P. *et al* 1998). Ainsi, nous proposons une solution intégrée au problème de la recherche documentaire. En effet, afin de guider l'utilisateur dans sa quête d'information, le moteur de recherche de l'Université Nouvelle de Lisbonne regroupe les textes recouvrés suivant l'ensemble des descripteurs[18] impliqués par la requête. Ainsi, chaque document pertinent est classé suivant un descripteur. Par exemple, si un utilisateur recherche l'ensemble des informations concernant le "*crime*" dans la base textuelle, le système regroupe les documents selon s'ils concernent les crimes politiques, les crimes militaires ou encore les crimes internationaux comme illustré dans la Figure (11).

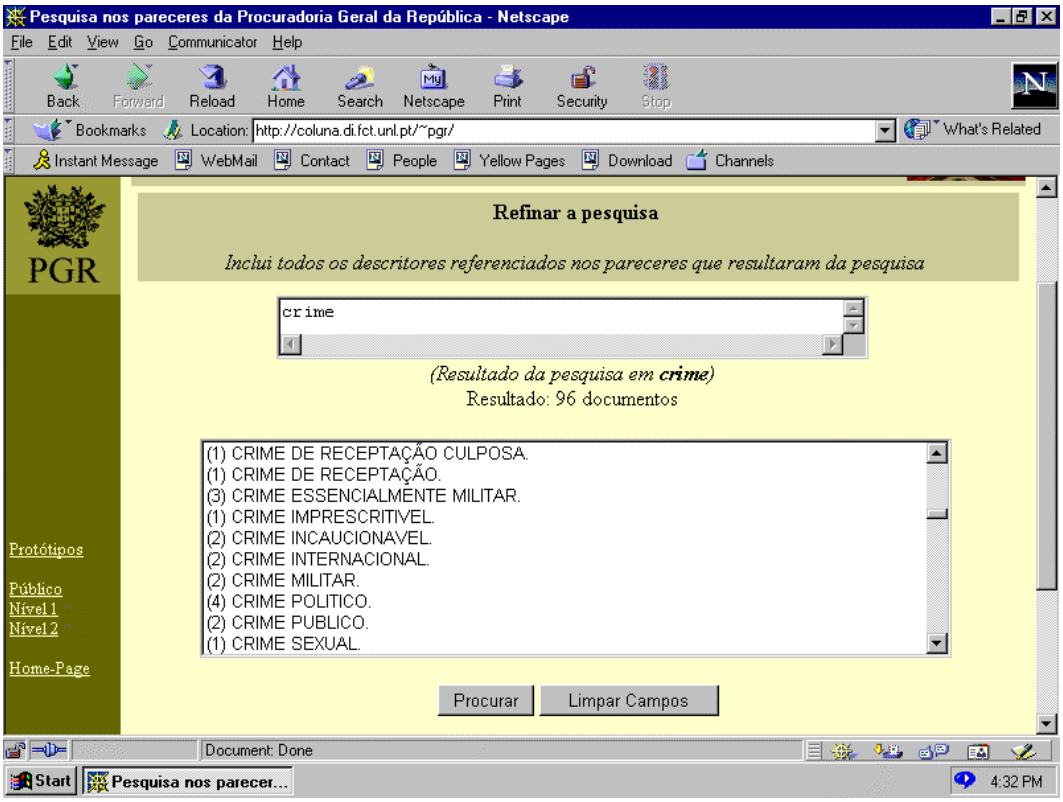


Figure 11. Structure proposée des documents pertinents

L'utilisateur peut ensuite décider de son domaine d'intérêt. Malheureusement, les descripteurs sont recensés dans un thesaurus prédéfini qu'il est difficile de maintenir du fait de l'accroissement rapide et constant des termes complexes spécialisés. Afin de suivre

[18] Les descripteurs sont recensés dans un thesaurus spécialisé pré-défini.

l'évolution du jargon associé au domaine législatif portugais, nous travaillons actuellement sur l'intégration de notre analyseur dans ce système de recherche. Notre objectif est de construire un système qui nous permette de calculer l'ensemble des unités lexicales complexes à partir de tous les textes (existants et futurs) de la base documentaire afin d'actualiser en continu l'ensemble des descripteurs du domaine.

RÉFÉRENCES

BHATTACHARYYA, Gouri.; Johnson, Richard. (1977): *Statistical Concepts and Methods*, New York, John Wiley & Sons.

BENSON, M. (1989): "The Structure of the Collocational Dictionary" in International Journal of Lexicography.

BETTS, R.; Marrable D. (1991): "Free Text vs controlled vocabulary, retrieval precision and recall over large databases", Online Inf. 91, London, 153-165.

BOURIGAULT, Didier (1996): "Lexter, a Natural Language Processing Tool for Terminology Extraction", Actes 7th EURALEX International Congress.

CHURCH, K.W.; Hanks P. (1990): "Word Association Norms Mutual Information and Lexicography" in Computational Linguistics, Vol. 16 (1), 23-29.

CHURCH, K.W. (1995): "One term or two?" in SIGIR, Seattle, EU, 310-318.

COWIE A. (1981): "The Treatment of Collocations and Idioms in Learners' Dictionaries" in Applied Linguistics, Vol. 11, 223-23.

DAILLE, Béatrice (1995): "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology" in The balancing act combining symbolic and statistical approaches to language, MIT Press.

DAVID, Sophie; Plante, Pierre (1990): "Termino Version 1.0", Rapport de recherche du Centre d'Analyse de Textes par Ordinateur, Université du Québec, Montréal.

DIAS, Gaël; Guilloré, Sylvie; Lopes, Gabriel (1999a): "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora", Actes Traitement Automatique des Langues Naturelles. Institut d'Etudes Scientifiques, Cargèse, France.

DIAS, Gaël; Guilloré, Sylvie; Lopes, Gabriel (1999b): "Mutual Expectation: a Measure for Multiword Lexical Unit Extraction", Actes VExTAL Venezia per il Trattamento Automatico delle Lingue, Universitá Cá Foscari, Venezia.

DIAS, Gaël; Guilloré, Sylvie; Lopes, Gabriel (1999c): "Multilingual Aspects of Multiword Lexical Units", Actes Workshop Language Technologies – Multilingual Aspects, Faculty of Art. Ljubljana, Slovenia.

DUNNING, Ted (1993): "Accurate Methods for the Statistics of Surprise and Coincidence" in Association for Computational Linguistics, Vol. 19-1.

ENGUEHARD, Chantal (1993): "Acquisition de Terminologie à partir de Gros Corpus", Actes Informatique & Langue Naturelle, 373-384.

EVANS, David; Lefferts Robert (1993): "Design and Evaluation of the CLARIT-TREC-2 System", *TREC93*, 137-150.

FELDMAN, R. (1998): "Text Mining at the Term Level", Actes PKDD'98, Lecture Notes in AI 1510, Springer Verlag.

GAUSSIER, É.;Grefenstette G.; Schulze, M. (1997): "Traitements du Langage naturel et recherche d'information: quelques expériences sur le français" in FRANCIL'97, 9-14.

GALE, William; Church K. (1991): "Concordances for Parallel Texts", Actes Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora, Oxford.

GREFENSTETTE, Gregory. (1994): *Explorations In Automatic Thesaurus Discovery*, Boston/Dordrecht/London, Kluwer Academic Publishers.

GREFENSTETTE, Gregory. (1998): *Cross-Language Information Retrieval*, Kluwer Editions.

GROSS, Gaston. (1996): *Les expressions figées en français*, Paris, Ophrys.

HABERT, Benoît; Jacquemin, C. (1993): "Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques" in Traitement Automatique des Langues (vol.34-nº2), Association pour le Traitement Automatique des langues, France.

HABERT, Benoît.; Nazarenko, Adeline.; Salem André. (1997): *Les linguistiques du Corpus*, Paris, Armand Colin.

HAUSMANN, F. (1979): "Un dictionnaire des collocations est-il possible?" in Travaux de linguistique et de littérature, Vol. 17, 187-195.

HERVIOU-PICARD, Marie-Luce (1996): "Construction de terminologies: une chaîne de traitement supportée par un atelier intégrant outils linguistiques et statistiques", Technical Report 96NO00018, EDF-DER.

JUSTESON, John (1993): "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text", IBM Research Report, RC 18906 (82591) 5/18/93.

MASON, Oliver (1997): "The Weight of Words: an Investigation of Lexical Gravity", Actes PALC'97.

QUARESMA, Paulo; Rodrigues, Irene; Lopes, Gabriel (1998): "PGR Project: The Portuguese Attorney General Decisions on the Web" in The Law in the Information Society, Instituto per la documentazione giuridica del CNR, C. Ciampi, E. Marinai (ed.), Florence, Italy.

SALEM, André (1987), *La pratique des segments répétés*. Klincksieck. Paris.

SILVA, Joaquim; Dias, Gaël; Guilloré, Sylvie; Lopes, Gabriel (1999): "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units" Actes 9th Portuguese Conference in Artificial Intelligence, Springer-Verlag.

SINCLAIR, J. (1974), *English Lexical Collocations: A study in computational linguistics*, Singapore, reprinted as chapter 2 of Foley, J. A. (ed). (1996), J. M. Sinclair on Lexis and Lexicography, Uni Press.

SHIMOHATA, S. (1997): "Retrieving Collocations by Co-occurrences and Word Order Constraints", Actes ACL-EACL'97, 476-481.

SMADJA, Frank (1993): "Retrieving Collocations From Text: XTRACT" in Computational Linguistics, Vol. 19 (1), 143-177.

SMADJA, Frank (1996): "Translating Collocations for Bilingual Lexicons: A Statistical Approach" in Association for Computational Linguistics, Vol. 22 (1).

SUSSNA, M. (1993): "Word sense disambiguation for free-text indexing using a massive semantic network", Actes Second International Conference on Information and Knowledge Management, ACM, 67-74.