Acquisition de la Terminologie et Définition des Tâches à Effectuer, Deux Principales Indissociables

Mathieu Roche

Acquisition de la terminologie et définition des tâches à effectuer, deux principes indissociables

Mathieu Roche

Équipe TAL, LIRMM, UMR 5506, Université Montpellier 2
mathieu.roche@lirmm.fr

Résumé

L'acquisition de la terminologie à partir de textes spécialisés est une étape essentielle. Elle est souvent indispensable pour extraire des connaissances permettant une meilleure compréhension de ces textes. Ainsi, les termes extraits peuvent représenter des traces utiles pour l'expert du domaine étudié. Afin de définir la pertinence de la terminologie extraite, cet article montre la nécessité de rigoureuseusement prendre en compte les tâches et les sous-tâches à effectuer avec les termes. En effet, d'une tâche à l'autre, les terminologies obtenues à partir d'un même corpus peuvent se révéler extrêmement différentes.

Mots clés

Fouille de textes, TAL, Terminologie

1. Introduction

Dans cet article, nous nous intéressons à l'étude des groupes de mots qui sont appelés des **collocations**. Nous nous intéressons plus particulièrement à l'extraction des collocations à partir de corpus (ensembles de textes homogènes). Comme nous le montrerons dans cet article, ces collocations peuvent représenter des traces qui seront utiles pour l'extraction de connaissances dans les textes. Ces connaissances extraites sont alors susceptibles d'apporter une meilleure compréhension des textes étudiés. Les collocations pertinentes sont des traces linguistiques de concepts. Cette définition a été introduite par Yves Kodratoff dans (Kodratoff, 2004). Dans cet article, nous montrerons comment les collocations peuvent ou non représenter de véritables traces linguistiques de concepts. Ainsi, nous montrerons que la pertinence d'une collocation devra être définie selon une tâche déterminée par l'expert.

Plus formellement, (Clas, 1994) donne deux propriétés définissant une collocation. Dans un premier temps, une collocation est définie comme un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe. Par exemple, "jour faste" est considéré comme une collocation car le sens global de ce groupe de mots peut être déduit des deux mots "jour" et "faste". En nous appuyant sur cette définition, l'expression "tirer son chapeau" n'est pas une collocation car son sens ne peut pas être déduit de chacun des mots. De telles formes sont appelées des **combinaisons figées**[1]. Une deuxième propriété est ajoutée par (Clas, 1994) pour définir une collocation. Le sens des mots qui composent la collocation doit être limité. Par exemple "acheter un chapeau" n'est pas une collocation car le sens de "acheter" et de "chapeau" n'est pas limité. En effet, de multiples objets, voire des personnes, peuvent être achetés. De tels groupes de mots sont appelés des **combinaisons libres**.

Dans nos travaux, nous considérons les combinaisons figées et libres comme des collocations contrairement à la définition de (Clas, 1994). Prenons l'exemple "first lady" issu d'un ensemble de textes étudiés au cours du challenge TREC'2004 (Text REtrieval Conference)[2]. Une telle collocation devrait être considérée comme une combinaison figée désignant la femme du chef de l'exécutif. Dans le cas des textes décrivant la campagne sénatoriale d'Hillary Clinton en 2000, cette combinaison est utile et pertinente pour caractériser Hillary Clinton. Cet aspect pragmatique, nous a motivé à considérer les combinaisons figées mais également libres comme des collocations.

Dans cette étude qui a pour but de définir de quelle manière définir la pertinence d'une collocation, nous nous appuierons sur deux corpus écrits en français :

- Corpus composé de Curriculum Vitæ fournis par la société VediorBis (120000 mots après divers prétraitements décrits dans (Roche, 2004)). Une des particularités de ce corpus tient au fait qu'il est composé de phrases très courtes avec de nombreuses énumérations.
- Corpus issu du domaine des Ressources Humaines (société PerformanSe) correspondant à des commentaires de tests de psychologie de 378 individus

[1]Les combinaisons figées non ambiguës peuvent alors être lexicalisées pour former des expressions lexicalisées.

[2]http://trec.nist.gov/. L'exemple donné est issu de la tâche Novelty du Challenge TREC'04. Cette tâche consiste à rechercher des phrases pertinentes et nouvelles à partir de textes journalistiques.

(600000 mots). Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique avec l'utilisation de tournures souvent littéraires.

Lorsque l'on décrit les systèmes d'extraction de la terminologie, il est coutume de préciser que l'on souhaite extraire les collocations pertinentes. Cependant, une des difficultés majeures est de déterminer ce que nous entendons par "collocation pertinente". Ainsi, la section 2 définit la pertinence d'une collocation. Dans la suite de cet article, les collocations pertinentes seront appelées des "**termes**". Les sections 3 et 4 présenteront respectivement le type de collocations extraites pour les tâches de normalisation des textes et de construction d'une classification conceptuelle. Enfin, nous présenterons quelques perspectives à ce travail.

2. Qu'est ce qu'une collocation pertinente ?

Pour définir la pertinence d'une collocation ou plus généralement d'une trace présente dans les textes, il est nécessaire de déterminer la tâche que l'on souhaite effectuer avec celle-ci. Nous montrerons par exemple dans les sections suivantes qu'une collocation peut être pertinente pour une tâche de normalisation des textes mais inutile pour construire une classification conceptuelle.

Il est également important de préciser que nous pouvons ajouter un niveau supplémentaire relatif aux sous-tâches pour une tâche principale. À titre d'exemple, une classification conceptuelle peut être construite pour différents objectifs :

- Découvrir des règles d'association entre concepts présents dans les textes (Kodratoff *et al.*, 2003, Azé & Roche, 2003) ou entre les instances de concepts (Janetzko *et al.*, 2004).
- Extraire des informations en utilisant des patrons d'extraction grâce aux concepts (Freitag, 1998), etc.

Ces deux objectifs propres à la tâche principale d'acquisition d'une classification conceptuelle caractérisent plusieurs sous-tâches. La pertinence d'une collocation peut en être affectée. Par exemple, une collocation peut être adaptée pour une classification conceptuelle pour rechercher des informations dans les textes à base de patrons d'extraction. Pourtant, cette même collocation peut être jugée comme non pertinente pour découvrir des règles d'association entre concepts. Nous reviendrons en détail sur cet exemple dans la section 4. Ce fait est donc à prendre

en considération dans la définition des collocations pertinentes. Dorénavant, nous considérons que si des sous-tâches propres à une tâche principale sont identifiées alors l'expert devra définir la pertinence selon les sous-tâches à réaliser.

Plus généralement, nous montrons dans cet article qu'une suite de mots présente dans les textes qui respecte une structure grammaticale déterminée est une **trace pertinente** si elle est utile pour une tâche ou une sous-tâche à réaliser par l'expert. Dans les sections suivantes, nous discutons de l'utilisation des mesures statistiques pour deux tâches principales : la normalisation des textes et la classification conceptuelle. Ces mesures statistiques permettent de classer les collocations extraites. Nous allons expliquer de quelle manière ces mesures peuvent aider à privilégier les collocations propres à une tâche principale. En effet, nous estimons pouvoir identifier les mesures statistiques les plus adaptées pour une tâche principale particulière mais il nous semble difficile de déterminer les mesures classiques à utiliser pour les sous-tâches souvent plus fines.

3. Acquisition des termes pour la normalisation des textes

Dans la chaîne globale de fouille de textes (Kodratoff *et al.*, 2003, Mathiak & Eckstein, 2004), après l'acquisition du corpus, sa normalisation est la première tâche importante à effectuer. Ces normalisations consistent à éliminer le bruit présent dans les textes, à uniformiser le vocabulaire, etc. Nous donnons ci-dessous, quelques exemples de sous-tâches utiles pour la phase de normalisation :

- détection du bruit dans les textes : problèmes liés au nettoyage, aux fautes d'orthographe, etc.

- détection de collocations utiles pour la phase de constitution ou de mise à jour de lexiques de noms propres : noms de lieux, noms de sociétés, noms d'établissements, noms d'organisations (politiques, associatives, religieuses, etc.), couples "prénom" associé à un "nom de famille", etc.

Sur le corpus de CVs, les deux sous-tâches de normalisation qui sont le nettoyage (présence de bruit et de fautes d'orthographe) et la reconnaissance de collocations pouvant constituer un lexique de noms propres ont été considérées. Ces lexiques peuvent être utilisés de différentes façons pour la tâche de normalisation. Les collo-

cations d'un tel lexique de noms propres peuvent être utiles pour une phase de pré-terminologie. La pré-terminologie consiste à extraire des collocations particulières qui peuvent être considérées comme des mots à part entière. Ces collocations étant repérées, un trait d'union (ou un "blanc souligné" plus spécifiquement utilisé pour ce type d'entitées nommées) peut être placé entre chacun des mots composant ces collocations. Ceci permet alors de les considérer comme des mots à part entière lors des étapes suivant la phase de normalisation (étiquetage grammatical, analyse syntaxique, etc.). Bien entendu, ce type de collocation est souvent spécifique aux domaines étudiés.

Cependant, d'un corpus à l'autre et d'un lexique de noms propres à l'autre, leur utilisation peut différer. Notons ci-dessous des exemples liés à une utilisation différente des lexiques :

- *Utilisation différente d'un même lexique selon les corpus.*
  Par exemple, sur un corpus d'articles de journaux (Soboroff & Harman, 2003), les collocations issues d'un lexique de prénoms associés à un nom de famille doivent être considérées comme de la pré-terminologie. Ce type d'information est en effet essentiel dans un corpus journalistique. Sur un corpus de CVs, de telles informations sont inutiles et inopportunes dans un souci d'anonymat de certaines informations textuelles. Il est donc nécessaire de les supprimer. Cet exemple montre donc qu'un même lexique peut être utilisé de manière différente selon les corpus.

- *Utilisation différente des lexiques sur un même corpus.*
  Par exemple, l'utilisation de lexiques des noms de sociétés peut être utile dans une phase de pré-terminologie sur un corpus de CVs. Au contraire, sur ce même corpus de CVs, un lexique de prénoms associés à un nom de famille est utilisé afin de supprimer ces informations. Cet exemple montre donc que, sur un même corpus, les lexiques peuvent être utilisés de manière différente.

Plusieurs expérimentations sur le corpus de CVs non normalisé ont été effectuées afin de comparer deux mesures statistiques. Pour cela, nous nous sommes appuyés sur la relation Nom-Nom du corpus de CVs qui présente le nombre le plus important de collocations (voir tableau 1). Pour extraire ces collocations, une étape préalable est effectuée qui consiste à étiqueter grammaticalement les mots des

textes avec l'étiqueteur de Brill (Brill, 1994). Ceci permet alors d'extraire aisément les collocations respectant des patrons précis de type Nom-Nom, Nom-Préposition-Nom, Nom-Adjectif, Adjectif-Nom, etc. Notons que l'approche d'extraction de la terminologie que nous utilisons suit un processus itératif détaillé dans (Roche, 2004) afin de construire des termes complexes (composés de nombreux mots). Par exemple, si à la première itération, le terme "système complexe" de type Nom-Adjectif est extrait, à la deuxième itération, nous pouvons extraire le terme "système complexe naturel" (exemple issu du titre des thèmes des Journées de Rochebrune).

Collocations	**Nombre total**	**Collocations**	**Nombre total**
Nom-Prép-Nom	5340	Nom-Adjectif	2904
Nom-Nom	9394	Adjectif-Nom	878

Table 1: Nombre total de collocations sur le corpus de CVs non normalisé.

Nous montrons, dans la suite, de quelle manière les mesures statistiques peuvent être utilisées pour identifier ce type de collocations pouvant représenter des traces utiles pour l'étape de normalisation. Afin de classer les collocations extraites, nous nous appuyons sur deux mesures classiques du domaine : l'Information Mutuelle (Church & Hanks, 1990) et le Rapport de Vraisemblance (Dunning, 1993). Ces mesures sont décrites dans (Roche, 2004). Comme cela est montré dans (Daille, 1994, Roche, 2004), le Rapport de Vraisemblance privilégie les collocations les plus fréquentes contrairement à l'Information Mutuelle qui place en tête les collocations les plus rares.

Dans le tableau 2, nous présentons l'analyse manuelle des 100 premières collocations Nom-Nom en utilisant ces deux mesures statistiques. Dans ce tableau, les expérimentations réalisées avec l'ensemble des collocations de type Nom-Nom extraites sans appliquer d'élagage sont présentées. Nous avons associé manuellement chacune des 100 premières collocations binaires aux différentes sous-tâches propres à la normalisation. Ce tableau montre que certaines mesures statistiques sont plus ou moins adaptées pour la tâche de normalisation. Ainsi, près des deux tiers (65%) des premières collocations extraites avec l'Information Mutuelle sont utiles pour la tâche globale de normalisation (voir tableau 2). A contrario, avec le Rapport de Vraisemblance, moins d'un tiers (28%) sont utiles pour la phase de nor-

malisation. Précisons également que l'Information Mutuelle est particulièrement efficace pour constituer ou enrichir des lexiques composés des couples "prénom nom" contrairement au Rapport de Vraisemblance. Les travaux de (Thanopoulos *et al.*, 2002) ont montré que, d'une manière générale, l'Information Mutuelle permettait d'extraire des entités nommées.

Sous-tâches et leur description		**IM**		**RV**	
Nettoyage	Bruit	5%	5 %	1%	1 %
	Fautes d'orthographe	0%		0%	
Constitution ou enrichissement de lexiques de noms propres	noms de lieux	5 %	60 %	4 %	27 %
	noms de sociétés, d'établissements, d'organisations	25 %		20 %	
	association prénoms/noms	30 %		3 %	

Table 2: Pourcentage des 100 premières collocations Nom-Nom (première itération) vérifiant des sous-tâches de la phase de normalisation. Ces collocations sont classées avec deux mesures : l'Information Mutuelle (IM) et le Rapport de Vraisemblance (RV). Expérimentations à partir du corpus de CVs non normalisé sans appliquer d'élagage pour la relation Nom-Nom.

Dans cette section, les expérimentations présentées sont relatives au corpus de CVs. Notons que sur d'autres corpus, l'Information Mutuelle extrait également des collocations pertinentes pour la constitution de lexiques spécifiques qui sont utiles pour l'étape de normalisation. Cependant, ces lexiques peuvent être de natures différentes. À titre d'exemple, un lexique constitué de "termes littéraires" (expressions linguistiques) peut être construit à partir du corpus des Ressources Humaines. En effet, l'auteur de ce corpus utilise un vocabulaire caractéristique composé de nombreux termes littéraires qui sont extrêment bien classés en utilisant l'Information Mutuelle : par exemple, "*statu quo*", "*voeux pieux*", "*carte blanche*", "*bâton rompus*", "*coudées franches*", etc. De telles collocations sont des combinaisons figées.

Après avoir mis en valeur les avantages de l'Information Mutuelle pour la tâche principale de normalisation, nous allons comparer le Rapport de Vraisemblance et l'Information Mutuelle pour une autre tâche principale : l'acquisition des termes pour la classification conceptuelle du domaine.

4. Acquisition des termes pour la classification conceptuelle

Pour construire une classification conceptuelle, les collocations évoquant des concepts du domaine sont extraites puis regroupées. Le tableau 3 présente des exemples de collocations associées à des concepts à partir des deux corpus en français étudiés. Les concepts sont définis par les experts selon les différents objectifs qu'ils souhaitent effectuer avec la classification conceptuelle. Nous détaillerons des exemples d'objectifs que les experts peuvent se fixer dans la suite de cet article.

CURRICULUM VITÆ		RESSOURCES HUMAINES	
Collocations	**Concepts**	**Collocations**	**Concepts**
aide comptable	Activité Gestion	besoin d'information	Communication
gestion administrative	Activité Gestion	capacité d'écoute	Communication
chef de service	Activité Encadrement	contexte professionnel	Environnement
direction générale	Activité Encadrement	lieu de travail	Environnement
employé libre service	Activité Commerce	sentiment de malaise	Stress
assistant marketing	Activité Commerce	tension permanente	Stress

Table 3: Extrait des classifications conceptuelles.

Afin de valider les collocations extraites, plusieurs catégories de pertinence (ou de non pertinence) ont été identifiées :

- $1^{\textbf{ère}}$ **catégorie :** La collocation est pertinente pour la classification conceptuelle (exemple du corpus de CVs : "*baccalauréat littéraire*")
- $2^{\textbf{ème}}$ **catégorie :** La collocation est pertinente mais très spécifique et pas nécessairement pertinente pour le domaine (exemple du corpus de CVs : "*écosystème marin*")
- $3^{\textbf{ème}}$ **catégorie :** La collocation est pertinente mais très générale et pas nécessairement pertinente (exemple du corpus de CVs : "*situation actuelle*")
- $4^{\textbf{ème}}$ **catégorie :** La collocation est non pertinente (exemple du corpus de CVs : "*jour quotidienne*")
- $5^{\textbf{ème}}$ **catégorie :** L'expert ne peut pas juger si la collocation est pertinente (exemple du corpus de CVs : "*master franchisé*").

Comme précisé précédemment, les collocations représentant des traces linguistiques de concepts doivent être pertinentes selon une sous-tâche à réaliser. Par

exemple, les collocations qui sont des instances de concepts peuvent être utilisées pour découvrir des règles d'association entre concepts présents dans les textes. Ceci permet alors de déterminer la force des associations qui peut exister entre les concepts. Les concepts peuvent également être utilisés pour construire des patrons d'extraction utiles pour la recherche d'informations.

Dans la première des sous-tâche (découvrir des règles d'association entre concepts), les concepts utilisés doivent être précis afin de déterminer des associations éventuelles. Ce travail a des similarités avec les approches de (Srikant & Agrawal, 1997) qui consistent à utiliser une taxonomie pour généraliser des règles d'association extraites. Dans nos travaux présentés dans (Kodratoff *et al.*, 2003, Azé & Roche, 2003) et dans la thèse de Jérôme Azé (Azé, 2003), les règles d'associations découvertes sont de la forme $concept_1 ... concept_{n-1} \rightarrow concept_n$ où n est le nombre de concepts impliqués dans les règles d'association extraites. Le détail de l'algorithme est présenté dans (Azé, 2003). À titre d'exemple, nous donnons une règle d'association extraite à partir du corpus des Ressources Humaines : "Stress" $\rightarrow$ "Environnement". Cette règle signifie que le stress s'exerce par l'intermédiaire de l'environnement. Cet exemple montre que les règles d'association permettent une meilleure compréhension des corpus étudiés. L'extraction des règles d'association s'effectue avec des concepts très précis qui intéressent l'expert. Ainsi, en reprenant les différentes catégories évoquées au début de cette section, les collocations pertinentes afin de découvrir des règles d'association entre concepts sont les collocations de la catégorie 1. Les collocations issues des catégories 2, 3 et 4 sont jugées comme non pertinentes. Enfin, les collocations de la catégorie 5 ne sont pas prises en considération car l'expert n'a pas été apte à les valider. Cette dernière catégorie correspond en fait à des collocations qui sont ambiguës ou qui n'ont pas pu être évaluées par méconnaissance partielle du domaine.

Définissons la seconde sous-tâche évoquée qui peut être associée à la même tâche principale (classification conceptuelle). Cette sous-tâche consiste à déterminer des concepts dans le but de construire des patrons d'extraction. Les patrons d'extraction prenant en compte des noms de personnes peuvent être utiles dans la recherche d'informations. Par exemple, le concept de *Nom de Personnes* permet d'appliquer un patron de type "Concept_Nom_de_Personnes suivi du verbe "*s'intéresser à*" afin d'extraire dans les textes tous les centres d'intérêt des personnes. Ainsi, l'utilisation des collocations spécifiques ou générales (catégories 2 et

3) permet de construire des patrons d'extraction couvrant davantage les textes et être donc plus efficace pour la recherche d'informations. Pour cette sous-tâche, les collocations pertinentes sont donc les collocations des catégories 1, 2 et 3 et les collocations non pertinentes sont celles issues de la catégorie 4. Ainsi, plus généralement, les collocations de la catégorie 4 ne peuvent être considérées comme de véritables traces.

	Information Mutuelle (IM)		
Catégories	Nb de collocations	Découvrir des règles d'association	Construire des patrons d'extraction
1. pertinent	6	*positif :* 6	*positif :* 83
2. spécifique	77	*négatif :* 83	
3. général	0		
4. non pertinent	6		*négatif :* 6
5. indécis	11	11	11
	Rapport de Vraisemblance (RV)		
Catégories	Nb de collocations	Découvrir des règles d'association	Construire des patrons d'extraction
1. pertinent	52	*positif :* 52	*positif :* 88
2. spécifique	30	*négatif :* 45	
3. général	6		
4. non pertinent	9		*négatif :* 9
5. indécis	3	3	3

Table 4: Corpus des CVs (non normalisés). Évaluation des collocations pour deux sous-tâches : établir une classification pour découvrir des règles d'association entre concepts et pour construire des patrons d'extraction. Les 100 premières collocations de type Nom-Nom classées avec l'Information Mutuelle et le Rapport de Vraisemblance sans application d'élagage sont évaluées.

Dans la suite, nous allons examiner les 100 premières collocations de type Nom-Nom du corpus de CVs classées avec l'Information Mutuelle et le Rapport de Vraisemblance. Ces collocations, analysées pour la tâche principale de classification conceptuelle, correspondent aux mêmes jeux de données que l'étude sur la normalisation (section 3). Les résultats sont présentés dans le tableau 4. Dans un premier temps, nous remarquons que les collocations pertinentes sont beaucoup plus nombreuses lorsqu'elles sont extraites avec le Rapport de Vraisemblance. L'Information Mutuelle extrait davantage de collocations spécifiques. Ceci corrobore les résultats présentés dans la section précédente qui montraient que l'Informa-

tion Mutuelle était une mesure particulièrement bien adaptée pour extraire des collocations afin de constituer ou enrichir des lexiques spécifiques.

La seconde remarque à prendre en compte tient au fait que selon une sous-tâche à réaliser (par exemple, construire une classification conceptuelle pour découvrir des règles d'association entre concepts ou pour construire des patrons d'extraction) la qualité des collocations extraites diffère. Dans le tableau 4, les collocations utiles (resp. inutiles) pour chacune de ces sous-tâches sont appelées des exemples positifs (resp. négatifs). Une telle différence est significative avec l'Information Mutuelle. Bien que moins flagrantes, les différences restent importantes entre le nombre de collocations utiles et inutiles classées avec le Rapport de Vraisemblance pour chacune des sous-tâches (voir tableau 4).

5. Conclusion et Perspectives

Cet article montre que pour définir la pertinence d'une collocation, il est nécessaire de déterminer la tâche que l'expert souhaite effectuer avec la terminologie. Nous avons validé nos résultats à partir d'un corpus de CVs. Nous avons montré que selon les tâches et les sous-tâches à réaliser, certaines mesures statistiques étaient plus ou moins bien adaptées pour classer les collocations. Une perspective importante à ce travail consiste à étudier la qualité de la terminologie extraite en définissant différentes autres tâches et en utilisant d'autres mesures statistiques.

Plus généralement, une trace dans les textes qui peut être définie comme un groupe de mots ayant une structure grammaticale déterminée est réellement pertinente si elle est utile pour la tâche que l'expert souhaite réaliser. La plupart des systèmes d'extraction de la terminologie (et plus généralement des systèmes de Traitement Automatique du Langage) s'appuient seulement sur les textes en les traitant avec des outils statistiques et/ou linguistiques pour obtenir un résultat pertinent. Seulement, pour obtenir des traces pertinentes en choisissant les paramètres ou les méthodologies les plus adaptés, il est indispensable de prendre en considération l'utilisation souhaitée des traces obtenues. Cette utilisation peut d'ailleurs être plus ou moins indépendante des textes étudiés.

Remerciements : Je remercie Yves Kodratoff, Violaine Prince et les relecteurs pour leurs conseils ainsi que Serge Baquedano (société PerformanSe) pour la constitution d'un des corpus étudiés dans cet article.

Références

AZÉ J. & ROCHE M. (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *Revue RIA-ECA numéro spécial EGC03*, **17**, 283–294.

AZÉ J. (2003). *Extraction de Connaissances dans des Données Numériques et Textuelles*. PhD thesis, Université de Paris 11.

BRILL E. (1994). Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, p. 722–727.

CHURCH K. W. & HANKS P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, p. 22–29.

CLAS A. (1994). Collocations et langues de spécialité. *Meta*, **39**(4), 576–580.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

DUNNING T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

FREITAG D. (1998). Toward general-purpose learning for information extraction. In *Proceedings of the Annual Meeting of the ACL*, p. 404–408: Morgan Kaufmann Publishers.

JANETZKO D., CHERFI H., KENNKE R., NAPOLI A. & TOUSSAINT Y. (2004). Knowledge-based selection of association rules for text mining. In *Proceedings of ECAI'04, IOS Press, Valencia, Spain*, p. 485–489.

KODRATOFF Y. (2004). Induction extensionnelle : définition et application l'acquisition de concepts à partir de textes. *Revue RNTI E2, numéro spécial EGC'04*, **1**, 247–252.

KODRATOFF Y., AZÉ J., ROCHE M. & MATTE-TAILLIEZ O. (2003). Des textes aux associations entre les concepts qu'ils contiennent. *Numéro spécial de la revue RNTI "Entreposage et Fouille de données"*, **1**, 171–182.

MATHIAK B. & ECKSTEIN S. (2004). Five steps to text mining in biomedical literature. In *Proceedings of « Data Mining and Text Mining for Bioinformatics » workshop of ECML/PKDD Conference*, p. 44–49.

ROCHE M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. PhD thesis, Université de Paris 11.

SOBOROFF I. & HARMAN D. (2003). Overview of the trec 2003 novelty track. In *NIST Special Publication: SP 500-255 The Twelfth Text Retrieval Conference (TREC 2003)*.

SRIKANT R. & AGRAWAL R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, **13**(2–3), 161–180.

THANOPOULOS A., FAKOTAKIS N. & KOKKIANAKIS G. (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02)*, volume 2, p. 620–625.