# Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation

Mathieu Roche, Violaine Prince

# Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation

Mathieu Roche, Violaine Prince

{mroche, prince}@lirmm.fr

Equipe TAL, LIRMM, CNRS, Université Montpellier 2, France

## Abstract

The paper shows that the extraction of nominal syntagms offers a number of candidates the status of collocation. Syntactically, these structures are compatible with Mel'čuk's definition of *base* and *collocate,* if one of the elements has a governing role. Semantically and pragmatically, the status of collocation must be validated by a form of *relevance*. However, this relevance is completely dependent on the task to be realized. We evaluated this assumption on corpora of different natures and different styles.

## Résumé

L'article présenté montre que l'extraction de syntagmes nominaux offre un ensemble de candidats au statut de collocation. Syntaxiquement, ces structures sont a priori compatibles avec la définition en *base* et *collocatif* de Mel'čuk si un des éléments joue un rôle de gouverneur. Sémantiquement et pragmatiquement, le statut de collocation doit être validé par une forme de *pertinence*. Or cette dernière est totalement dépendante de la tâche à réaliser. Nous avons évalué cette hypothèse sur des corpus de nature et de styles différents.

**Mots-clés :** collocations, terminologie.

## 1. Introduction

Dans cet article, nous nous intéressons à l'extraction de **collocations** à partir de corpus par des méthodes automatiques ou semi-automatiques. Si linguistiquement, les collocations peuvent être décrites comme des *associations syntagmatiques binaires, restreintes, semi-figées et fortement dépendantes du contexte d'utilisation* (Grossmann et Tutin, 2003), sur un plan plus cognitif, des collocations dites **pertinentes** sont des traces linguistiques de concepts dans le texte (Kodratoff, 2004) et peuvent apporter des éléments appropriés à l'extraction de connaissances et à une meilleure compréhension du texte concerné.

Nous nous proposons ici, après avoir retenu une définition large de la collocation, d'étudier la manière de définir sa pertinence, en fonction de la tâche considérée. De nombreux travaux se sont attelés à repérer des collocations (Heid et Frebott, 1991), à les extraire de documents (Smadja, 1993) pour la construction de bases terminologiques (Daille, 1994) ou pour une meilleure compréhension et modélisation de l'information textuelle (Claveau et L'Homme 2006 ; Orliac 2006). La majorité des travaux a relié à juste titre le problème des collocations avec celui de la terminologie. Les campagnes d'évaluation de logiciels et de techniques de recherche d'information et de questions/réponses comme TREC (Text Retrieval Conference) montrent que le problème de détection et d'évaluation de la collocation transcende le cadre particulier de la construction terminologique. Des collocations apparaissent dans des textes, sont datées et contextualisées pour disparaître ensuite, et leur influence sur l'extraction du

fragment pertinent est sans doute prépondérante tout en étant circonscrite. Par conséquent, la question qui se pose est la suivante : la collocation extraite est-elle pertinente : **1)** pour la question considérée **2)** par rapport à la terminologie en vigueur ou détectée (se pose alors la question du genre textuel et du domaine de spécialité) **3)** par rapport aux traitements ultérieurs comme la traduction (Claveau et Zweigenbaum, 2005) ou la classification (Yang, 1999) ?

La section 3 définit et argumente la pertinence d'une collocation en fonction de critères à la fois linguistiques et mesurables. Dans la suite de cet article, les collocations pertinentes seront appelées des « **termes** ». Les sections 4 et 5 présenteront respectivement le type de collocations extraites pour les tâches de normalisation des textes et de construction d'une classification conceptuelle. Pour ce faire, nous nous appuierons sur deux corpus en Français :

- Une collection de Curriculum Vitae fournis par la société VediorBis (120 000 mots après divers prétraitements décrits dans (Roche, 2004)). Ce corpus est composé de phrases très courtes avec de nombreuses énumérations.
- Un corpus issu du domaine des Ressources Humaines (société PerformanSe) correspondant à des commentaires de tests de psychologie de 378 individus (600 000 mots). Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique avec l'utilisation de tournures souvent littéraires.

Enfin, la conclusion permettra de résumer le travail réalisé et d'ouvrir la voie à quelques perspectives d'extension de cette étude de faisabilité.

## 2. Caractérisation générale de la collocation

(Clas, 1994) donne deux propriétés principales pour définir une collocation. Premièrement, une collocation est considérée comme : *un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe*. Par exemple, « jour faste » est vu 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].

La deuxième propriété définie par l'auteur (Clas, 1994) est que *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 achetées. De tels groupes de mots sont appelés des **combinaisons libres**.

Le problème soulevé par une telle définition est le suivant :

- Jusqu'à quel point une expression peut être figée ?
- Inversement, n'y-a-t-il pas des combinaisons dites libres qui, dans certains contextes deviennent fortement récurrentes ?

Prenons l'exemple de « first lady » issu d'un ensemble de textes étudiés au cours du challenge TREC'2004 - tâche Novelty (Soboroff et Harman, 2005). Cette dernière consiste à rechercher des phrases pertinentes et nouvelles à partir de textes journalistiques. Une telle expression devrait être considérée comme une combinaison figée désignant la femme du chef de

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

l'exécutif (l'expression française correspondante de *première dame* joue le même rôle). Or, le degré de figement est relatif et passablement circonstanciel. Par ailleurs, d'autres éléments peuvent se combiner aussi bien avec *premier* (comme *premier ministre*, *première épouse*, *premier tour*, etc) qu'avec *dame* en seconde position (*grande dame*, etc).

Dans le cas des textes décrivant la campagne sénatoriale d'Hillary Clinton en 2000, cette combinaison est utile et pertinente pour caractériser Hillary Clinton elle-même et non pas sa fonction temporaire. Cet aspect fortement contextuel nous a motivé à considérer la notion de collocation au sens large, telle que proposée par (Grossmann et Tutin, 2003) plutôt que par (Clas, 1994).

Une définition plus formelle, due à Mel'čuk (Mel'čuk et al., 1984-1999) est donnée ci-dessous. L'expression AB formée des lexies A, la **base**, et B, le **collocatif**, est appelée collocation si et seulement si :

- Le sens de A est inclus dans le sens de AB
- A est sélectionné par le locuteur de façon régulière et non contrainte
- B est sélectionné en fonction de A et du sens de AB à exprimer

Par rapport à cette définition, l'exemple « first lady » s'analyse très bien : « lady » est la base, et « first » le collocatif : la structuration en *[Adjectif]-[Nom]* place le nom en gouverneur, et lui affecte naturellement le rôle de base. L'adjectif étant un modifieur, le sens de la base est partiellement conservé. La sélection par le locuteur est dépendante du contexte (de qui on parle et où), du thème (ici la politique américaine). Nous utiliserons cette définition pour affirmer si un groupe extrait comme *candidat à la collocation* mérite le statut de collocation.

## 3. Définition de la pertinence d'une collocation

La notion de pertinence est définie :

- par rapport à la tâche considérée,
- par rapport au thème du texte (par exemple : « élection primaire » et « campagne électorale » sont des collocations thématiquement pertinentes pour des textes politiques se rapportant aux élections dans les démocraties),
- par rapport aux concepts saillants (Hillary Clinton, « première dame » et « l'épouse du Président » désignent le même concept saillant, qui est la « candidate Hillary Clinton aux élections sénatoriales »).

La pertinence par rapport au thème, aussi bien que par rapport aux concepts saillants, est définie au sein d'un type de tâche particulier : la *classification conceptuelle* ou *thématique*. C'est pourquoi nous nous contenterons ici de définir la pertinence par rapport à la tâche afin de proposer un cadre général d'évaluation de cette pertinence.

### *3.1. Pertinence par rapport à la tâche*

Une collocation peut être pertinente pour une tâche donnée mais inappropriée pour une autre. Ce constat n'est pas nouveau, la linguistique de corpus a eu le mérite de montrer la dépendance de la nature des traitements par rapport aux objectifs. Cependant, si jusqu'à présent la collocation a été bien exploitée par les différents auteurs cités en introduction et par d'autres également, elle l'a été dans un contexte majoritairement mono-tâche. En outre, les mesures statistiques utilisées dans la plupart des articles de linguistique de corpus relèvent

plus de cadres théoriques (la sémantique distributionnaliste, l'approche « sac de mots », la démarche LSA, etc) que des spécificités applicatives (et pas seulement du domaine). L'originalité de cet article consistera surtout à étudier de manière transversale et multitâches l'adaptation de la mesure statistique au problème soulevé.

C'est pourquoi nous montrons ici deux types de tâches dont les objectifs et les moyens sont fondamentalement différents :

- Pré-traitement de textes pour la mise en conformité, encore appelée **normalisation** (voir section 4).
- Classification conceptuelle (voir section 5).

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, par exemple :

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

Nous considérons en section 5 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.

### *3.2. Evaluation de la pertinence*

Nous chercherons à évaluer expérimentalement l'hypothèse : *une suite de mots présente dans les textes, qui respecte une structure grammaticale déterminée obéissant aux règles de formation des collocations de Mel'čuk est une* ***collocation pertinente*** *si elle est utile pour une tâche ou une sous-tâche à réaliser (par un expert ou un système d'assistance à la tâche). L'utilité en question sera mesurable par des méthodes statistiques*.

Dans les sections suivantes, l'extraction des collocations fondée sur différentes mesures statistiques sera traitée pour deux tâches : la *normalisation des textes* et la *classification conceptuelle*. Nous montrons de quelle manière ces mesures peuvent aider à évaluer les candidats et à privilégier les collocations propres à une tâche principale.

## 4. Acquisition des collocations pour la normalisation des textes

Une chaîne globale de fouille de textes a été développée dans (Kodratoff et al, 2003 ; Mathiak et Eckstein, 2004), en vue de réaliser divers processus d'extraction de connaissances. Après l'acquisition du corpus, sa « normalisation » est la première tâche importante à effectuer. Elle consiste 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 » « 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. Les noms propres sous forme (« prénom », « nom ») sont utiles pour réaliser une pré-terminologie. Cette dernière consiste à extraire des collocations particulières considérées comme des mots à part entière. Dès qu'elles sont repérées, un trait d'union (ou un « blanc souligné » plus spécifiquement utilisé pour ce type d'entités nommées) peut être placé entre chacun des mots composant ces collocations. Ainsi, « Hillary Clinton » apparaîtra sous la forme « Hillary_Clinton » vue comme une désignation atomique. On remarquera que les couples (« prénom », « nom ») obéïssent à la définition de Mel'čuk pour les collocations avec le nom jouant le rôle de base.

Cette lexicalisation permet de traiter ces couples comme des entrées de dictionnaire banalisées lors des étapes ultérieures (étiquetage grammatical, analyse syntaxique, etc). Bien entendu, ce type de collocation est souvent spécifique aux domaines étudiés.

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 candidats à la collocation (voir tableau 1). Pour les extraire, une étape préalable consiste à étiqueter grammaticalement les mots des textes (nous avons utilisé ici l'étiqueteur de Brill (Brill, 1994)). Nous avons obtenu aisément des candidats à la collocation respectant des patrons précis suivants : *[Nom]-[Nom]*, *[Nom]-[Prep]-[Nom]*, *[Nom]-[Adjectif]*, *[Adjectif]-[Nom]*, etc. Ces candidats à la collocation qui sont lemmatisés seront ensuite évalués sur deux points :

- leur compatibilité avec les règles de bonne formation des collocations de Mel'čuk,
- et dans les candidats bien formés, considérés donc comme des collocations à part entière, ceux qui sont les plus pertinents par rapport à la tâche.

L'approche d'extraction de la terminologie que nous utilisons suit un processus itératif, détaillé dans (Roche, 2004), capable de construire des termes complexes. Par exemple, si à la première itération, le terme « fouille de texte » de type *[Nom]-[Prep]-[Nom]* est extrait, à la deuxième itération, nous pouvons extraire le terme « logiciel de fouille de texte ».

| **Collocation (candidats)** | **Nombre total** | **Collocation (candidats)** | **Nombre total** |
|---|---|---|---|
| [Nom] − [Prep] − [Nom] | 5 340 | [Nom] − [Adjectif] | 2 904 |
| [Nom] − [Nom] | 9 394 | [Adjectif] − [Nom] | 878 |

*Tableau 1. Nombre de candidats à la collocation extraits dans le corpus de CVs non normalisé.*

Afin de classer les candidats extraits, nous nous appuyons sur deux mesures classiques du domaine : l'Information Mutuelle (Church et Hanks, 1990) et le Rapport de Vraisemblance (Dunning, 1993). Ces mesures qui calculent une certaine forme de dépendance de chacun des mots composant les candidats à la collocation sont détaillées dans (Roche, 2004). Comme cela est montré dans (Daille, 1994), le Rapport de Vraisemblance privilégie les candidats les plus fréquents, contrairement à l'Information Mutuelle qui place en tête les candidats les plus rares.

Dans cet article nous nous sommes concentrés sur deux mesures ayant un comportement radicalement différent et qui sont très couramment utilisées dans les travaux liés à la terminologie. Le lecteur intéressé pourra se référer aux travaux de (Daille, 1994 ; Roche, 2004) qui décrivent de manière précise d'autres mesures statistiques adaptées à la terminologie qui ont été évaluées (Dice, Information Mutuelle au Cube, etc).

Dans le tableau 2, nous présentons les résultats de l'analyse manuelle des 100 premiers candidats de patron *[Nom]-[Nom]* en utilisant ces deux mesures statistiques. Les expérimentations y sont réalisées avec l'ensemble des candidats sans élagage (sans suppression des collocations ayant un nombre d'occurrences faible). 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 premiers candidats extraits 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 normalisation. Soulignons 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. Cependant, nous discuterons en section 6 le fait que, dans certains contextes, cette mesure peut avoir des limites pour identifier les entités nommées.

Dans cette section, les expérimentations présentées sont relatives au corpus de CVs. Sur d'autres corpus, l'Information Mutuelle extrait également des collocations pertinentes pour la constitution de lexiques spécifiques utiles pour l'étape de normalisation. Cependant, ces lexiques peuvent être de nature différente. Par exemple, un lexique constitué de « termes littéraires » (expressions linguistiques) peut être construit à partir du corpus des Ressources Humaines. L'auteur de ce corpus utilise un vocabulaire caractéristique composé de nombreux termes littéraires extrêment bien classés par l'Information Mutuelle : par exemple, « *statu quo* », « *voeux pieux* », « *carte blanche* », « *bâton rompus* », « *coudées franches* », etc. De tels candidats sont des expressions figées ou quasi-figées d'un point de vue linguistique, mais de la même manière que « moyenne pondérée » peut l'être dans un corpus de statistiques ou d'économétrie, alors que cette dernière aura un statut de collocation. La collusion entre littérature et linguistique efface ici le fait que le statut de collocation est terriblement dépendant du domaine et du genre textuel. Le corpus des Ressources Humaines semble dénoter un caractère littéraire et pourra donc se référer au dictionnaire des expressions figées comme complément de son lexique terminologique.

| ***Tâches*** | ***Sous-tâches*** | ***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 société,* | | | | |
| | *noms d'établissements,* | *25%* | | *20%* | |
| | *noms d'organisations* | | | | |
| | *association prénoms/noms* | *30%* | | *3%* | |

*Tableau 2. Pourcentage des 100 premiers candidats à la collocation de type [Nom]-[Nom] (première itération) vérifiant des sous-tâches de la phase de normalisation. Ils sont classés avec deux mesures : l'Information Mutuelle (IM) et le Rapport de Vraisemblance (RV). Expérimentations à partir du corpus de CVs non normalisé et sans appliquer d'élagage par rapport à la relation [Nom]-[Nom].*

Soulignons enfin que sur des corpus relevant de domaines de spécialités scientifiques (biologie, médecine, etc) de telles tâches de normalisation (par exemple, la normalisation de variantes de termes médicaux) se révèlent essentielles pour un certain nombre de traitements : indexation, questions-réponses, traduction, etc. (Grabar et Zweigenbaum, 2005 ; Claveau et Zweigenbaum, 2005). Ainsi, dans ces domaines, la constitution de dictionnaires est utile pour normaliser le vocabulaire des textes. Cela est particulièrement courant en biologie moléculaire (par exemple, *carboxyl-terminal* peut avoir de très nombreuse formes lexicales : *carboxy termini*, *carboxy terminal*, *COOH-terminal*, *COOH-termini*, *C02H-terminal*, etc). Cette même situation peut se retrouver avec des corpus en Ancien Français car cette langue ne possède pas de norme orthographique fixe (Cazal et al*.,* 2007). Notons que le processus de normalisation qui s'appuie sur l'utilisation de dictionnaires spécifiques peut parfois être lié aux tâches de construction de classifications conceptuelles qui seront décrites dans la section ci-dessous.

## 5. Acquisition des collocations pour la classification conceptuelle

Pour construire une classification conceptuelle, les collocations évoquant des concepts du domaine, définis par des experts, sont extraites puis regroupées. Le tableau 3 présente des exemples de collocations associées à des concepts à partir des deux corpus étudiés. Afin de valider les candidats extraits, plusieurs catégories de pertinence (ou de non pertinence) ont été identifiées :

- **catégorie 1 :** Le candidat est une collocation pertinente pour la classification conceptuelle (exemple du corpus de CVs : « *baccalauréat littéraire* »)
- **catégorie 2 :** Le candidat est une collocation très spécifique et pas nécessairement pertinente pour le domaine (exemple du corpus de CVs : « *écosystème marin* »)
- **catégorie 3 :** Le candidat est une collocation très générale et pas nécessairement pertinente pour la classification (exemple du corpus de CVs : « *situation actuelle* »)
- **catégorie 4 :** Le candidat n'est pas une collocation (exemple du corpus de CVs : « *jour quotidienne* »)
- **catégorie 5 :** Le candidat peut être une collocation, mais l'expert ne peut pas juger de sa pertinence pour le domaine (exemple du corpus de CVs : « *master franchisé* »).

| CURRICULUM VITAE | | 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* |

*Tableau 3. Extrait des classifications conceptuelles.*

Les collocations représentant des traces linguistiques de concepts doivent être pertinentes par rapport à un objectif, donc une sous-tâche à réaliser. Ainsi, les collocations qui sont des instances de concepts peuvent être utilisées pour découvrir des règles d'association entre concepts présents dans les textes. Cela permet alors de déterminer la force des associations entre les concepts. Ces derniers peuvent également être utilisés pour construire des patrons d'extraction.

### *5.1. Découverte des règles d'association*

Dans une première sous-tâche propre à la classification conceptuelle, les concepts utilisés doivent être précis afin de déterminer des associations éventuelles. Ce travail a des similarités avec les approches de (Srikant et Agrawal, 1997) qui consistent à utiliser une taxonomie pour généraliser des règles d'association extraites. Dans (Kodratoff et al., 2003) et (Azé, 2003), les règles d'associations découvertes sont de la forme $concept_1 \ldots concept_{n-1} \rightarrow concept_n$ où $n$ est le nombre de concepts impliqués dans les règles d'association extraites. Le détail de l'algorithme est présenté dans (Azé, 2003). A titre d'exemple, nous donnons une règle d'association extraite à partir du corpus des Ressources Humaines : « Stress » → « Environnement ». Elle 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 pour la découverte des règles d'association entre concepts sont celles de la catégorie 1. Les candidats issus des catégories 2, 3 et a fortiori 4 sont jugés comme non pertinents. Enfin, les collocations de la catégorie 5 ne sont pas prises en considération car non validées par l'expert.

### *5.2. Construction des patrons d'extraction*

Cette sous-tâche consiste à déterminer des concepts dans le but de construire des patrons d'extraction. Ceux qui sont relatifs aux noms de personnes sont utiles en extraction d'information. 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 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. Pour cette sous-tâche, les collocations pertinentes sont donc les candidats des catégories 1, 2 et 3. De manière générale, les candidats de la catégorie 4 ne peuvent être considérés ni comme de véritables traces de concepts, ni comme des collocations au sens de Mel'čuk.

Dans la suite, nous allons examiner les 100 premiers candidats de type *[Nom]-[Nom]* du corpus de CVs classés avec l'Information Mutuelle et le Rapport de Vraisemblance. Ces données, analysées pour la tâche principale de classification conceptuelle, correspondent aux mêmes jeux de données que l'étude sur la normalisation (section 4). 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. Cela corrobore les résultats de la section précédente qui montraient que l'Information Mutuelle était une mesure particulièrement bien adaptée pour extraire des collocations afin de constituer ou d'enrichir des lexiques spécifiques.

| Catégories | Nb de candidats à la collocation | Sous-tâche 1 : découvrir des règles d'associaition | Sous-tâche 2 : Construire des patrons d'extraction |
|---|---|---|---|
| **INFORMATION MUTUELLE** | | | |
| **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** | | | |
| **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 |

*Tableau 4. Evaluation des candidats à la collocation pour deux sous-tâches. Les 100 premiers candidats de type [Nom]-[Nom] sont classés et évalués avec l'Information Mutuelle et le Rapport de Vraisemblance sans élagage.*

Deuxièmement, on remarque que selon la sous-tâche à réaliser (construire une classification conceptuelle pour découvrir des règles d'association entre concepts ou pour construire des patrons d'extraction) la qualité des candidats extraits 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 3).

## 6. Conclusion et Perspectives

L'extraction de syntagmes nominaux (groupes nominaux prépositionnels, paires adjectivales, *[Nom]-[Adjectif]* ou *[Adjectif]-[Nom]*, paires nominales de type « prénom » « nom » ou désignation de rubriques) offre un ensemble de candidats au statut de collocation. Syntaxiquement, ces structures sont a priori compatibles avec la définition en *base* et *collocatif* de Mel'čuk si un des éléments joue un rôle de gouverneur. Sémantiquement et pragmatiquement, le statut de collocation doit être validé par une forme de *pertinence*. Or cette dernière est totalement dépendante du problème posé (tâche à réaliser), de la situation en vigueur, du genre textuel et du domaine considéré. De plus, pour une tâche donnée, la

pertinence dépend de l'évaluateur, de l'expert humain ou du système d'assistance à l'activité terminologique. Telle est l'hypothèse que nous avons cherché à évaluer sur des corpus de nature et de style différents : un corpus de CVs, un corpus de commentaires psychologiques dans le milieu des ressources humaines. De nombreux systèmes d'extraction de la terminologie s'appuient seulement sur les textes en les traitant avec des outils statistiques et/ou linguistiques pour obtenir des résultats, en n'introduisant pas les objectifs méthodologiques et applicatifs dans la boucle d'évaluation.

Nos récents travaux relatifs à l'étude des entités nommées spécifiques que sont les sigles et leurs définitions nous amènent à des conclusions propres à l'évaluation de ce type de collocation (Roche et Prince, 2007). Nos travaux s'appuient sur la mise oeuvre d'une mesure de qualité pour une tâche consistant à déterminer la définition pertinente à partir de définitions potentielles pour un sigle (par exemple, notre approche consiste à déterminer si la définition à associer au sigle *JO* présent dans un texte donné est *Jeux Olympiques* ou *Journal Officiel*). Pour une telle tâche, nous avons proposé une mesure statistique fondée sur le contexte (mots « significatifs » présents dans la page dans laquelle le sigle se situe) mais également sur l'Information Mutuelle propre aux mots définissant les sigles. Les fréquences utilisées dans ces mesures s'appuient sur des statistiques issues de moteurs de recherche du web (nombre de pages retrouvées avec les mots composant les définitions des sigles et le contexte). Les sigles étant des entités nommées, l'Information Mutuelle doit potentiellement se révéler adaptée, ce que les travaux présentés dans (Roche et Prince, 2007) ont infirmé. Cette situation peut s'expliquer pour deux raisons :

1. la fréquence des mots issue du web est souvent trop importante et pas assez adaptée aux domaines spécialisés étudiés (par exemple, un terme très spécifique pour un corpus peut se révéler très fréquent à l'échelle du web),
2. les mots composant les sigles sont souvent généraux et donc fréquents ce qui dégrade l'Information Mutuelle (par exemple, les mots propres au sigle « JADT » sont très généraux : *Journées*, *internationales*, *Analyse*, *statistique*, *Données*, *Textuelles*).

Les expérimentations menées confirment toute la complexité de la notion de pertinence des collocations qui s'appuie sur l'utilisation de mesures statistiques qui sont adaptées à un corpus et à un domaine pour une tâche.

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, ce qui permettra :

- de sélectionner un type de mesure adapté à l'évaluation de la pertinence,
- s'il ne se dégage pas une mesure particulière, de proposer un couple (tâche, mesure) optimisé pour cette pertinence.

## Remerciements

Les auteurs remercient Yves Kodratoff (LRI) pour le travail mené sur la chaîne de fouille de textes présentée dans cet article et Serge Baquedano (société PerformanSe) pour la constitution d'un des corpus étudiés dans cet article.

## Références

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

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

Cazal E., Serp C., Roche M. and Laurent L. (2007). Extraction de terminologie pour l'ancien français : la quête du graal. In *Proceedings of atelier FDC'07 (Fouille de Données Complexes dans un processus d'extraction des connaissances) à la conférence EGC'2007*, p.11-20.

Church K. W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, Vol. 16, p.22-29.

Clas A. (1994). Collocations et langues de spécialité. *Meta*, 39(4) : 576-580.

Claveau V. and L'Homme M. C. (2006) Discovering and Organizing Noun-Verb Collocations in Specialized Corpora Using Inductive Logic Programming. *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, Vol. 11, No 2, p.209-243.

Claveau V. and Zweigenbaum P. (2005) Translating biomedical terms by inferring transducers. In *Actes 10th Conference on Artificial Intelligence in Medicine Europe*, Aberdeen.

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

Grabar N. and Zweigenbaum P. (2005). Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale. In *Proceedings of Traitement Automatique de Langues Naturelles (TALN)*.

Grossmann F. and Tutin A. (2003). Les collocations : analyse et traitement. Editions De Werelt. *Collection : Travaux et recherches en linguistique appliquée. Série E : lexicologie et lexicographie*, Amsterdam.

Heid U. and Freibott G. (1991). Collocations dans une base de données terminologiques et lexicales. *Meta*, 36-1, p.77-91.

Janetzko D., Cherfi H., Kennke R., Napoli A. and 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, 247-252.

Kodratoff Y., Azé J., Roche M. and 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.

Orliac B. (2006) Colex : un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales. In *Proceedings of Terms in Specialized Dictionaries,* L'Homme (ed.), p.261-280.

Mathiak B. and 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.

Mel'čuk I. A., Arbatchewsky-Jumarie N., Elnitsky L. and Lessard A. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain.* Presses de l'Université de Montréal, Montréal, Canada. Vol. 1, 2, 3, 4.

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

Roche M. and Prince V. (1997). *AcroDef :* A Quality Measure for Discriminating Expansions of Ambiguous Acronyms. In *Proceedings of CONTEXT, Springer-Verlag, LNCS*, p.411-424.

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics,* n°19, 1, p.143-177.

Soboroff I. and Harman D. (2003). Overview of the TREC 2003 novelty track. *NIST Special Publication TREC.*

Soboroff I. and Harman D. (2005). Novelty Detection: The TREC Experience. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* p.105-112.

Srikant R. and Agrawal R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13(2-3): 161-180.

Thanopoulos A, Fakotakis N. and Kokkianakis G. (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Vol. 2, p.620-625.

Yang Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval 1*(1-2), p.69-90.