# Extraction de lexiques bilingues à partir de corpus comparables spécialisés : étude du contexte lexical

**Amir Hazem — Emmanuel Morin**

*Université de Nantes, LINA UMR CNRS 6241*
*2 rue de la Houssinière, BP 92208*
*F-44322 Nantes cedex 3*
*{amir.hazem,emmanuel.morin}@univ-nantes.fr*

RÉSUMÉ. *Ce travail s'intéresse à la notion de contexte lexical qui est au cœur de l'approche fondatrice en extraction de lexiques bilingues à partir de corpus comparables spécialisés. D'une part, nous revenons sur les deux principales stratégies, dédiées à la caractérisation du contexte lexical, qui reposent sur l'exploitation de représentations graphique ou syntaxique. Nous montrons que l'exploitation conjointe de ces deux représentations a un intérêt particulier pour la tâche de construction de lexiques bilingues. D'autre part, nous abordons la difficulté de disposer d'observations significatives du contexte des mots en corpus comparables spécialisés. Pour répondre à cette difficulté, nous proposons de mettre en œuvre des stratégies de réestimation des observations de cooccurrences de mots par méthode de lissage ou par prédiction. Les différentes contributions associées à ce travail engendrent une amélioration significative de la qualité des lexiques extraits.*

ABSTRACT. *This work focuses on the concept of lexical context that is central to the historical approach of bilingual lexicon extraction from specialized comparable corpora. First, we revisit the two main strategies dedicated to lexical context characterization, that rely on the use of window-based and syntactic-based representations. We show that the combination of these two representations has a particular interest in the task of building bilingual lexicons. Second, we address the problem of the reliability of context words observations in specialized comparable corpora. To answer this difficulty, we propose to implement strategies to re-estimate words cooccurrence observations by smoothing or prediction techniques. Our results show a significant improvement in the quality of extracted lexicons.*

MOTS-CLÉS : *corpus comparable, extraction de lexiques bilingues, contexte lexical.*

KEYWORDS: *Comparable corpus, bilingual lexicon extraction, lexical context.*

TAL. Volume 55 – n° 1/2014, pages 13 à 44

## 1. Introduction

L'extraction de lexiques bilingues à partir de corpus a initialement été réalisée en s'appuyant sur des textes en correspondance de traduction (c'est-à-dire des corpus parallèles) (Véronis, 2002). Cependant, et en dépit des bons résultats obtenus, ces corpus demeurent des ressources rares, notamment pour les domaines spécialisés et pour des couples de langues ne faisant pas intervenir l'anglais. Dans ce contexte, les recherches en extraction de lexiques bilingues se sont penchées sur d'autres corpus composés de textes partageant différentes caractéristiques telles que le domaine, le genre, la période... sans être en correspondance de traduction (c'est-à-dire des corpus comparables) (Bowker et Pearson, 2002).

Si les corpus comparables sont des ressources bien plus abondantes que les corpus parallèles, les lexiques bilingues extraits de corpus comparables sont d'une qualité bien inférieure à ceux obtenus à partir de corpus parallèles. Cette différence s'explique principalement par l'absence d'élément d'ancrage dans les corpus comparables (l'alignement préalable des paragraphes, des phrases... n'y est pas possible). L'approche historique pour extraire des termes en correspondance de traduction à partir de corpus comparables, connue sous le nom d'*approche directe* (Fung et McKeown, 1997 ; Rapp, 1999), repose sur l'observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes contextes lexicaux. Un mot est alors décrit par les mots de son voisinage. La traduction d'un mot de la langue source est, quant à elle, obtenue en comparant la traduction de ce voisinage avec les voisinages de la langue cible.

Dans ce travail, nous nous intéressons à la notion de contexte lexical qui est au cœur de l'approche directe en extraction de lexiques bilingues à partir de corpus comparables. Plus précisément, nous nous intéressons à l'extraction de lexiques bilingues français-anglais en domaines spécialisés. D'une part, nous présentons les deux principales stratégies dédiées à la caractérisation du contexte lexical dans l'approche directe qui reposent sur l'exploitation de représentations graphique ou syntaxique (section 2). Nous verrons que l'exploitation conjointe de ces deux représentations permet une amélioration significative de la qualité des lexiques bilingues extraits de corpus comparables spécialisés (section 3). D'autre part, nous abordons la difficulté de disposer d'observations significatives du contexte des mots en corpus comparables spécialisés par comparaison avec un corpus comparable de langue générale. Pour répondre à cette difficulté, nous proposons de mettre en œuvre des stratégies de réestimation des observations de cooccurrences de mots par méthode de lissage ou par prédiction (section 4). Ces différentes contributions sont évaluées en section 5 d'une manière individuelle mais aussi de manière conjointe. Enfin, la section 6 vient conclure ce travail.

## 2. Méthode et ressources

Dans cette section, nous commençons par présenter l'approche fondatrice en extraction de lexiques bilingues à partir de corpus comparables. Nous décrivons ensuite les différentes ressources mobilisées pour ce travail.

### 2.1. *Approche directe*

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. C'est l'idée du linguiste J. R. Firth (1957, p. 11) selon laquelle « *On reconnaît un mot à ses fréquentations*[1] ». La mise en œuvre de cette observation s'appuie sur la caractérisation du contexte des mots pour faire émerger un ensemble de traits singuliers reflétant l'usage des mots en contexte. Cette caractérisation, qui s'inscrit dans l'hypothèse distributionnelle de Harris (1971), revient à identifier des *affinités du premier ordre* : « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné*[2] » (Grefenstette, 1994a, p. 279).

Les contextes lexicaux sont observés d'un point de vue monolingue (c'est-à-dire les documents des langues source et cible du corpus comparable) à travers le prisme d'une fenêtre plus ou moins étroite qui peut aller de quelques mots (Rapp, 1999 ; Gamallo, 2007) à quelques phrases (Déjean *et al.*, 2002 ; Daille et Morin, 2005). Quelle que soit la représentation mise en œuvre (graphique comme syntaxique), les traits associés à un mot à caractériser sont d'autres mots. Dans le cadre d'une représentation graphique, un mot est caractérisé par les mots avec lesquels il apparaît. À ce niveau, les mots grammaticaux (qui ne sont pas considérés comme porteur de sens) ne sont généralement pas pris en compte. En outre, les mots sont le plus souvent lemmatisés afin de renforcer la caractérisation des contextes lexicaux. Dans le cadre d'une représentation syntaxique, un mot sera, quant à lui, caractérisé par les relations de dépendances syntaxiques qu'il entretient avec ses voisins.

Un simple dénombrement permet de quantifier la relation qu'entretient un mot à caractériser avec les autres mots du corpus. Une mesure de récurrence contextuelle est généralement préférée à un simple dénombrement pour limiter l'influence des mots fréquents. Enfin, la traduction d'un mot de la langue source est obtenue en comparant son contexte préalablement transféré en langue cible à l'aide d'un dictionnaire bilingue à l'ensemble des contextes de la langue cible à l'aide d'une mesure de similarité.

1. « *You shall know a word by the company it keeps.* »
2. « *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word.* »

L'implémentation que nous faisons de l'approche directe se décompose de la manière suivante (Rapp, 1995 ; Fung et McKeown, 1997 ; Chiao et Zweigenbaum, 2003 ; Morin *et al.*, 2007 ; Déjean *et al.*, 2002 ; Laroche et Langlais, 2010).

**Identification des contextes lexicaux**

Pour chaque partie du corpus comparable, le contexte de chaque mot $w$ est extrait en repérant les mots qui apparaissent autour de lui selon une caractérisation graphique[3] ou syntaxique[4]. Pour chaque mot $w$, un vecteur de contexte $\mathbf{w}$ est ainsi obtenu. Afin d'identifier les mots caractéristiques des contextes lexicaux et de supprimer l'effet induit par la fréquence des mots, nous normalisons l'association entre les mots sur la base d'une mesure de récurrence contextuelle comme l'*information mutuelle* – IM (Fano, 1961), le *taux de vraisemblance* – TV (Dunning, 1993) ou le *rapport des cotes actualisées* – RCA (Evert, 2005) (cf. les équations 1 à 3 et le tableau 1). Après normalisation, à chaque élément $\mathbf{w_k}$ du vecteur de contexte $\mathbf{w}$ nous attachons le taux d'association $assoc(\mathbf{w_k})$.

**Transfert d'un mot à traduire**

Le transfert d'un mot $w$ à traduire de la langue source à la langue cible repose sur la traduction de chacune des entrées $\mathbf{w_k}$ de son vecteur de contexte au moyen d'un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour une entrée, nous ajoutons au vecteur de contexte traduit l'ensemble des traductions proposées[5] (lesquelles sont pondérées par le nombre d'occurrences de la traduction en langue cible[6]). Dans le cas où l'entrée n'est pas présente dans le dictionnaire, elle ne sera pas exploitée dans le processus de traduction.

**Identification des vecteurs proches du mot à traduire**

Le vecteur de contexte $\overline{\mathbf{w}}$ du mot $w$ ainsi traduit est ensuite comparé à chacun des vecteurs de contexte $\mathbf{t}$ de la langue cible en s'appuyant sur une mesure de similarité comme le *cosinus* – COS (Salton et Lesk, 1968) ou le *Jaccard pondéré* – JAC (Grefenstette, 1994b) (cf. les équations 4 et 5 où $assoc(\overline{\mathbf{w}}_\mathbf{k})$ représente après transfert la mesure d'association du mot $w$ avec une entrée $\mathbf{w_k}$ du vecteur de contexte $\overline{\mathbf{w}}$).

3. Dans le cas d'une représentation graphique, les mots sont extraits dans une fenêtre contextuelle de $n$ mots autour du mot $w$ à caractériser et les mots grammaticaux ne sont pas considérés.

4. Dans le cas d'une représentation syntaxique, ce sont les mots en relation de dépendances syntaxiques avec le mot $w$ à caractériser qui sont sélectionnés pour faire partie de son contexte lexical.

5. Cela correspond à une approche classique lorsque les traductions ne sont pas ordonnées dans le dictionnaire bilingue. D'autres techniques ont été proposées notamment par Bouamor *et al.* (2013).

6. Par exemple, si la mesure d'association de l'entrée *maison* en langue source est de 20 et dispose des deux traductions *home* et *house* avec un nombre d'occurrences en langue cible réciproquement de 5 et 35, l'entrée sera remplacée en langue cible par les deux entrées *home* et *house* avec les mesures d'association réciproquement de 2,5 ($20 \times 5/(5+35)$) et 17,5 ($20 \times 35/(5+35)$). Cette stratégie permet de préserver la mesure d'association initiale tout en tenant compte de l'importance de chaque traduction.

**Obtention des traductions candidates**

En fonction des précédentes valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour le mot à traduire.

| | $j$ | $\neg j$ |
|---|---|---|
| $i$ | $a = cooc(i, j)$ | $b = cooc(i, \neg j)$ |
| $\neg i$ | $c = cooc(\neg i, j)$ | $d = cooc(\neg i, \neg j)$ |

**Tableau 1.** *Table de contingence où $cooc(i, j)$ désigne le nombre d'occurrences des mots $i$ et $j$*

$$\text{IM}(\mathbf{w}_k) = log\frac{a}{(a+b)(a+c)} \quad [1]$$

$$\begin{aligned}\text{TV}(\mathbf{w}_k) = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \quad [2]$$

$$\text{RCA}(\mathbf{w}_k) = \log\frac{(a+\frac{1}{2})\times(d+\frac{1}{2})}{(b+\frac{1}{2})\times(c+\frac{1}{2})} \quad [3]$$

$$\text{COS}(\overline{\mathbf{w}}, \mathbf{t}) = \frac{\sum_k assoc(\overline{\mathbf{w}}_\mathbf{k})\, assoc(\mathbf{t_k})}{\sqrt{\sum_k assoc(\overline{\mathbf{w}}_k)^2}\sqrt{\sum_k assoc(\mathbf{t_k})^2}} \quad [4]$$

$$\text{JAC}(\overline{\mathbf{w}}, \mathbf{t}) = \frac{\sum_k min(assoc(\overline{\mathbf{w}}_\mathbf{k}), assoc(\mathbf{t_k}))}{\sum_k max(assoc(\overline{\mathbf{w}}_\mathbf{k}), assoc(\mathbf{t_k}))} \quad [5]$$

L'approche directe mobilise différents paramètres comme la représentation contextuelle utilisée et les mesures d'association et de similarité exploitées qui ont une influence directe sur la qualité des résultats. L'étude la plus complète sur l'influence de ces paramètres a été réalisée par Laroche et Langlais (2010).

### 2.2. *Ressources et outils*

Dans nos expérimentations, nous avons besoin de trois ressources linguistiques, à savoir : i) un corpus comparable, ii) un dictionnaire bilingue et iii) une liste d'évaluation. Nous commençons par présenter ces trois ressources, puis nous décrivons l'outil exploité pour identifier les relations de dépendances syntaxiques.

#### 2.2.1. *Corpus comparables*

Nous utilisons trois corpus comparables spécialisés français-anglais pour évaluer nos différentes propositions.

**Corpus du cancer du sein**

Ce corpus a été construit à partir de documents extraits du portail Elsevier[7]. L'ensemble des documents collectés relève du domaine médical restreint à la thématique du « cancer du sein » et comportent, dans le titre ou dans les mots-clés, le terme *cancer du sein* en français et *breast cancer* en anglais pour la période de 2001 à 2008. Ce corpus est composé de 130 documents pour le français (7 376 mots distincts) et de 103 documents pour l'anglais (8 457 mots distincts) pour une taille d'environ un million de mots.

**Corpus des énergies renouvelables**

Ce corpus a été construit à partir du Web à l'aide du crawler *Babouk* (Groc, 2011). Pour la recherche et l'extraction des différents documents du corpus, le crawler s'est appuyé sur des mots-clés du domaine tels que *wind, energy, rotor...* en anglais et *vent, énergies, éolien, renouvelables...* en français. Nous disposons ici d'un corpus d'environ 600 000 mots avec 5 606 mots distincts pour le français et 6 081 mots distincts pour l'anglais.

**Corpus de vulcanologie**

Ce corpus a été construit manuellement par Amélie Josselin-Leray du Laboratoire CLLE-ERSS. Il contient des documents extraits du Web, des manuels universitaires, des ouvrages de vulgarisation scientifique, des quotidiens généralistes, des magazines plus ou moins vulgarisés ainsi que des magazines de voyages/découvertes et des glossaires. La taille du corpus est d'environ 800 000 mots avec 9 142 mots distincts pour le français et 8 623 mots distincts pour l'anglais.

En plus de ces trois corpus comparables spécialisés, nous utilisons un corpus comparable de langue générale pour construire nos modèles d'apprentissage.

**Corpus journalistique**

Ce corpus a été construit à partir des publications électroniques du journal français *Le Monde* et du journal anglais *Los Angeles Times* de l'année 1994. Nous avons extrait deux versions de ce corpus, une première version d'une taille d'environ 500 000 mots et une seconde version d'une taille d'environ 10 millions de mots.

Pour chaque corpus, les documents sont nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique (pour les deux langues nous utilisons l'étiqueteur de Brill (1994)) et lem-

7. `www.elsevier.com`

matisation (pour le français nous utilisons *Flemm* de Namer (2000) et pour l'anglais un outil interne fondé sur la base lexical CELEX[8]).

2.2.2. *Dictionnaire bilingue*

Nous avons utilisé le dictionnaire français-anglais ELRA-M0033 de 200 000 entrées. Le tableau 2 indique le nombre de couples de traduction en fonction de leurs catégories grammaticales : adjectifs, noms et verbes (première colonne) dans le dictionnaire (colonne ELRA), puis après projection du dictionnaire sur chacun des trois corpus comparables spécialisés (trois dernières colonnes).

| Ctg. FR-EN | ELRA | Cancer du sein | Énergies renouvelables | Vulcanologie |
|---|---|---|---|---|
| N-N | 127 236 | 5 881 | 6 216 | 10 770 |
| N-V | 3 | 0 | 0 | 0 |
| N-A | 29 | 1 | 2 | 5 |
| V-V | 52 697 | 3 940 | 4 200 | 7 675 |
| V-N | 2 | 0 | 0 | 0 |
| V-A | 0 | 0 | 0 | 0 |
| A-A | 47 827 | 2 358 | 2 149 | 4 229 |
| A-N | 429 | 35 | 45 | 64 |
| A-V | 10 | 0 | 0 | 0 |

**Tableau 2.** *Comparaison des couples de traduction du dictionnaire après projection sur les corpus comparables*

Ce tableau montre que la différence en termes de nombre de couples de traduction est nettement moins prononcée après projection du dictionnaire sur les trois corpus comparables. On y remarquera la prédominance des couples N-N. Concernant les trois corpus comparables, le corpus de vulcanologie contient le plus de couples de traduction distincts. Les corpus du cancer du sein et des énergies renouvelables sont plus ou moins équivalents de ce point de vue.

2.2.3. *Listes d'évaluation*

Pour construire les listes de couples de traduction nécessaires à l'évaluation de nos approches, nous nous appuyons sur des nomenclatures attestées des termes du domaine. Cette méthode de création d'une liste de référence est différente de celle proposée par Déjean *et al.* (2002) qui construisent leur liste à partir d'un sous-ensemble du dictionnaire bilingue. Nous pensons que cette approche, plus fiable d'un point de vue statistique, ne correspond pas aux véritables difficultés rencontrées avec des corpus spécialisés. En domaines spécialisés, les termes qui représentent une difficulté de traduction n'appartiennent que rarement au dictionnaire de langue générale. Nous sélectionnons les couples de traduction pour lesquels le mot français apparaît au moins

8. `catalog.ldc.upenn.edu/LDC96L14`

cinq fois dans la partie française et sa traduction au moins cinq fois dans la partie anglaise du corpus comparable. Ce choix est motivé par la nécessité d'avoir un minimum de contexte pour déployer l'approche directe, ce qui est clairement difficile, voire impossible, pour des termes ayant une fréquence très faible.

Pour le corpus du cancer du sein, nous obtenons 321 couples de termes simples français-anglais à partir du métathésaurus UMLS[9] et du *Grand dictionnaire terminologique*[10]. Nous obtenons 150 couples pour le corpus des énergies renouvelables et 158 pour celui de vulcanologie à l'aide de lexiques et glossaires de chaque domaine.

Il n'est pas aisé de juger de la difficulté de la tâche d'extraction de termes bilingues à partir de corpus comparables. Le fait d'utiliser plusieurs ressources linguistiques signifie que la difficulté peut émaner de chacune d'elles voire de toutes les ressources. Il n'existe pas, à notre connaissance, une manière de mesurer la difficulté de notre tâche. Néanmoins, nous pouvons dire que si le corpus ou le dictionnaire bilingue est de mauvaise qualité alors la tâche est difficile. Nous pouvons aussi dire que si la liste d'évaluation contient beaucoup de termes peu fréquents ou des couples de traduction ayant une forte spécificité (Ahmad *et al.*, 1994), alors la tâche est plus difficile. Il est en effet souvent plus facile de traduire des termes très fréquents. Or, en domaines spécialisés, nous sommes principalement confrontés à des termes moyennement voire peu fréquents. Pour avoir une idée de la fréquence des termes à traduire et de leur bonne traduction, nous présentons une représentation des trois listes d'évaluation selon plusieurs plages d'occurrences, dans le tableau 3.

| **Plages d'occurrences** | **Cancer du sein** | | | **Énergies renouvelables** | | | **Vulcanologie** | | |
|---|---|---|---|---|---|---|---|---|---|
| | #EN | #FR | #EN∩FR | #EN | #FR | #EN∩FR | #EN | #FR | #EN∩FR |
| [5, 10] | 57 | 67 | 25 | 10 | 15 | 3 | 5 | 9 | 0 |
| ]10, 50] | 132 | 137 | 71 | 38 | 48 | 20 | 60 | 60 | 38 |
| ]50, 100] | 50 | 40 | 9 | 41 | 29 | 15 | 24 | 26 | 9 |
| ]100, 500] | 73 | 64 | 32 | 47 | 47 | 22 | 54 | 50 | 30 |
| ]500, 1 000] | 4 | 7 | 1 | 8 | 6 | 1 | 8 | 9 | 4 |
| ]1 000, 5 000] | 5 | 6 | 5 | 6 | 5 | 3 | 7 | 4 | 4 |

**Tableau 3.** *Comparaison des listes d'évaluation par plage d'occurrences*

Ce tableau indique la distribution du nombre d'occurrences des termes source et cible des listes d'évaluation ainsi que des couples de traduction selon différentes plages de valeurs. Nous pouvons remarquer qu'une bonne partie des couples de traduction ne partage pas les mêmes plages de valeurs. Si nous prenons par exemple la plage [5, 10] de la liste d'évaluation du cancer du sein, seulement 25 couples de traduction appartiennent à cette plage. Si l'appartenance des couples de traduction aux mêmes plages de valeurs ne peut en aucun cas suffire à mesurer le degré de difficulté d'une liste d'évaluation, celle-ci peut néanmoins indiquer une tendance dans des cas

9. `www.nlm.nih.gov/research/umls`
10. `www.granddictionnaire.com`

spécifiques. Par exemple, si un terme source appartient à la plage ]10, 50] et sa traduction à la plage ]1 000, 5 000], alors la taille du vecteur de contexte du terme source serait au maximum de 300 mots et celle de sa traduction de 6 000 mots au minimum, et ceci pour une taille de fenêtre égale à 7 (3 mots avant et 3 mots après le terme à caractériser). La différence de taille des vecteurs de contexte engendrée par une trop grande différence d'occurrences pourrait expliquer la difficulté à traiter des couples de traduction se trouvant dans cette configuration.

#### 2.2.4. *Identification de relations de dépendances syntaxiques*

Afin de mieux représenter le contexte d'un mot, plusieurs travaux se sont tournés vers les relations de dépendances syntaxiques (Lin, 1998 ; Gamallo, 2008a ; Garera *et al.*, 2009). Dans ce cas, il s'agit de décrire un mot par les relations de dépendances qu'il entretient avec les mots avoisinants. Une relation de dépendance est décrite comme une relation binaire asymétrique entre un premier mot appelé tête ou parent et un second mot appelé modificateur ou dépendant. Les relations de dépendances forment ainsi un arbre qui interconnecte l'ensemble des mots d'une phrase. Un mot dans une phrase pourra avoir plusieurs modificateurs mais chaque mot ne pourra modifier au plus qu'un seul mot (Lin, 1998). Le nœud de l'arbre de dépendance qui ne modifie aucun mot de la phrase est naturellement appelé racine[11].

Gamallo (2008b) aborde trois notions élémentaires de dénotation : i) les mots lexicaux, ii) les dépendances syntaxiques (sujet, objet direct, relation prépositionnelle entre deux noms, relation prépositionnelle entre un verbe et un nom...) et iii) le modèle lexico-syntaxique qui consiste à combiner les mots et leurs catégories syntaxiques en termes de dépendances (nom + sujet + verbe). Les mots lexicaux représentent des ensembles de propriétés (noms, verbes, adjectifs, adverbes...) alors que les dépendances et les modèles lexico-syntaxiques sont définis comme des opérations sur ces ensembles. Une dépendance est une relation binaire qui prend comme entrée deux ensembles de propriétés et donne en sortie un ensemble plus restreint qui est l'intersection des ensembles d'entrées. On retrouve sept types de relations de dépendances résumées dans le tableau 4 (Gamallo, 2008a).

Pour le syntagme anglais *local recurrence* par exemple, il existe une relation Lmod entre l'adjectif *local* et le nom *recurrence*. Donc, dans le processus de construction du contexte de *recurrence*, nous comptabilisons le nombre de fois où l'adjectif *local* apparaît à gauche de *recurrence* dans le corpus. Nous faisons de même pour les autres relations de dépendances syntaxiques. Concernant l'extraction des relations de dépendances syntaxiques, nous avons utilisé l'outil proposé par Gamallo (2008b)[12].

11. Pour plus de détails sur les dépendances syntaxiques, et plus particulièrement pour les tâches de désambiguïsation de mots et de résolution de dépendances (attachement et résolution), nous renvoyons le lecteur à Gamallo (2008b).

12. `gramatica.usc.es/pln/tools/deppattern.html`

| Relation | Type | Exemple |
|---|---|---|
| Lmod | Modificateur gauche si relation A-N | *local - recurrence* |
| Rmod | Modificateur droit si relation N-A | *number - insufficient* |
| modN | Modificateur de nom si relation N-N | *breast - cancer* |
| Lobj | Objet à gauche si relation N-V | *study - demonstrate* |
| Robj | Objet à droite si relation V-N | *have - effect* |
| PRP | Si relation prépositionnelle N-Prep-N | *malignancy - in - woman* |
| iobj | Si relation objet indirect V-Prep-N | *occur - in - portion* |

**Tableau 4.** *Relations de dépendances syntaxiques*

## 3. Combinaison de contextes

Une première manière pour combiner les deux représentations contextuelles est une combinaison *a posteriori*, c'est-à-dire la combinaison des scores retournés pour chaque représentation de l'approche directe. Une seconde manière consiste à mettre en œuvre une combinaison *a priori* qui intègre les deux informations contextuelles dans le même vecteur pour ensuite appliquer l'approche directe sur l'ensemble du corpus.

### 3.1. *Combinaison* a posteriori *des contextes*

Dans le domaine de la recherche d'information, la combinaison de plusieurs listes renvoyées par différents moteurs de recherche est souvent utilisée pour améliorer les performances d'un système de questions/réponses (Aslam et Montague, 2001). Dans notre cas, nous nous trouvons à devoir combiner deux approches distinctes. La première correspond à l'approche directe fondée sur une représentation graphique et la seconde correspond à l'approche directe fondée sur une représentation syntaxique. Une manière classique de fusionner ces deux approches est de prendre comme entrée, la sortie de chaque approche individuelle. Nous prenons donc comme entrée pour un mot à traduire la liste de ses traductions candidates (où à chaque traduction est associé un score de similarité) retournée par chacune des deux approches, puis nous fusionnons les deux listes par une simple combinaison arithmétique des scores des mêmes traductions candidates. Ceci nous donne une nouvelle liste de traduction candidates ordonnées (les scores fusionnés sont compatibles dans la mesure où nous utilisons la même mesure de similarité pour les deux approches). Nous calculons ainsi le score de similarité d'une traduction candidate comme étant la somme des scores renvoyés par chacune des deux approches comme suit :

$$\mathrm{S}_{comb}(w) = \mathrm{S}_{gra}(w) + \mathrm{S}_{syn}(w) \quad [6]$$

où $\mathrm{S}_{comb}(w)$ est le score final de la traduction candidate $w$, $\mathrm{S}_{gra}(w)$ est le score retourné par l'approche directe fondée sur une représentation graphique et $\mathrm{S}_{syn}(w)$ est le score retourné par l'approche directe fondée sur une représentation syntaxique.

Cette équation peut aussi s'écrire comme suit :

$$\mathrm{S}_{comb}(w) = \lambda \times \mathrm{S}_{gra}(w) + (1 - \lambda) \times \mathrm{S}_{syn}(w) \quad [7]$$

avec $\lambda$ comme indice de confiance donné à chaque méthode ($\lambda \in [0,\ 1]$). Dans notre cas, $\lambda = 0,5$, notre but n'étant pas de trouver la valeur optimale de $\lambda$ pour obtenir les meilleurs résultats. Les différentes expériences réalisées indiquent que les meilleurs résultats sont globalement obtenus pour $\lambda \in [0,55,\ 0,65]$. D'autres méthodes de combinaison de scores ont été testées comme la combinaison harmonique des rangs et des scores (Zweigenbaum et Habert, 2006 ; Morin, 2009), mais la méthode retenue par combinaison arithmétique des scores est celle qui donne les meilleures performances dans nos expériences.

### 3.2. *Combinaison* a priori *des contextes*

Le vecteur de contexte a pour but d'enregistrer un ensemble d'informations sur le contexte d'un mot $w$ donné. Dans le cas de la représentation graphique, ces informations sont les mots qui apparaissent avec le mot $w$. Dans le cas d'une représentation syntaxique, ce sont les mots en relation de dépendances syntaxiques avec $w$ qui sont sélectionnés pour faire partie de son vecteur de contexte. Dans un cadre plus générique, nous pourrions imaginer plusieurs autres sources d'informations à exploiter. Cependant, si chaque nouvelle information engendre un nouveau vecteur de contexte, nous pourrions vite être dépassés par le nombre de sources à fusionner. Pour remédier à cela, une autre manière serait de représenter dans un seul vecteur de contexte toutes les informations concernant le mot $w$. C'est la position adoptée avec la combinaison *a priori* des contextes.

| **Contexte graphique** | **Contexte syntaxique** | **Combinaison** |
|---|---|---|
| $regional_{13}$ | $regional_{Lmod_2}$ | $regional_{13}$<br>$regional_{Lmod_2}$ |
| $local_5$ | $local_{Lmod_1}$ | $local_5$<br>$local_{Lmod_1}$ |
| $oestrogen_1$ | | $oestrogen_1$ |
| $rate_{32}$ | $rate_{modN_{29}}$<br>$rate_{PRPV_3}$ | $rate_{32}$<br>$rate_{modN_{29}}$<br>$rate_{PRPV_3}$ |

**Tableau 5.** *Exemple de représentation du contexte du mot* recurrence *et du nombre de ses cooccurrences, en fonction des représentations graphique et syntaxique ainsi que de leur combinaison*

Dans cette technique de combinaison, nous considérons le vecteur de contexte d'un mot comme un descripteur qui contient plusieurs informations pour chaque entrée du

vecteur. Dans notre cas, nous avons deux types d'informations : (i) une information de cooccurrences globale fournie par la représentation graphique et (ii) une information plus spécifique fournie par la représentation syntaxique. Si nous prenons par exemple le mot *regional* (représenté dans le tableau 5), nous pouvons voir qu'il apparaît treize fois avec le mot *recurrence* selon la représentation graphique et deux fois comme modificateur gauche (Lmod) selon la représentation syntaxique. La combinaison prend en compte les deux informations comme deux entrées distinctes dans le vecteur de contexte résultant[13].

Une information importante à souligner est que l'approche directe s'appuyant sur les relations de dépendances syntaxiques considère $rate_{modN_{29}}$ et $rate_{PRPV_3}$ par exemple, comme étant deux mots distincts. Le premier avantage de la combinaison *a priori* est que si l'une des méthodes manque une information (un mot), comme nous pouvons le constater avec le mot *oestrogen*, par exemple, la fusion permet de pallier ce manque (grâce ici à la représentation graphique). Le deuxième avantage réside dans sa capacité à rendre les vecteurs de contexte plus discriminants et ceci, en considérant une information globale sur la relation de cooccurrence qui existe entre deux mots, et une information syntaxique plus précise sur la nature de ces cooccurrences. Ainsi, nous considérons les deux informations (graphique et syntaxique) comme complémentaires. Cette complémentarité peut être vue de deux manières. La première est liée aux erreurs qui peuvent être engendrées par l'une des représentations (erreur de l'analyseur syntaxique par exemple) et la seconde consiste en la prise en compte dans un même vecteur de l'information globale (information graphique) renforcée par une information plus fine (information syntaxique).

Dans les tableaux 6, 7 et 8, nous illustrons les premières entrées du vecteur de contexte du mot *recurrence* extrait du corpus du cancer du sein, en fonction de trois mesures d'association (TV, RCA et IM). La notation $(+/-)$ indique l'apport positif ou négatif de la combinaison *a priori*. L'indice $+$ indique qu'un mot classé dans les premières entrées du vecteur de contexte de la représentation graphique ou syntaxique, conserve son classement dans les premières entrées après combinaison. Le signe $-$ en revanche, indique l'apparition d'un mot non classé dans les premières entrées du vecteur de contexte. Nous notons par $graphique_k$ la représentation graphique avec une fenêtre de taille $k$ et par $syntaxique$ la représentation syntaxique.

Le tableau 6 montre que la combinaison *a priori* a un apport positif car elle engendre un vecteur de contexte qui respecte le classement des méthodes $graphique_5$ et $syntaxique$ en utilisant la mesure TV. Le tableau 7 indique aussi que la combinaison *a priori* a un apport positif en utilisant la mesure RCA. Nous remarquons néanmoins que la combinaison a avantagé la méthode $syntaxique$, car il n'y a que ses entrées qui sont présentes dans les premières entrées du vecteur de contexte de la méthode de combinaison *a priori*. Enfin, le tableau 8 montre que la combinaison *a priori* a

13. La fusion des deux vecteurs engendre un nouveau vecteur dont les entrées sont les mots issus des représentations graphique et syntaxique. Les scores d'association du nouveau vecteur sont différents car une fois la fusion réalisée nous normalisons à nouveau le vecteur.

| $\mathbf{graphique_5}$ | | **syntaxique** | | $\mathbf{a\_priori_5}$ | | +/− |
|---|---|---|---|---|---|---|
| local | 818,98 | $local_{Lmod}$ | 618,17 | $local_{Lmod}$ | 936,05 | + |
| rate | 119,71 | $risk_{PRPN}$ | 96,02 | local | 791,15 | + |
| distant | 72,62 | $rate_{modN}$ | 68,34 | $risk_{PRPN}$ | 153,14 | + |
| risk | 61,00 | $tumor_{modN}$ | 62,82 | rate | 113,96 | + |
| salvage | 39,15 | $rate_{PRPN}$ | 40,18 | $rate_{modN}$ | 110,28 | + |
| year | 39,08 | $time_{PRPN}$ | 32,85 | $tumor_{modN}$ | 104,71 | + |
| time | 31,84 | $disease_{modN}$ | 28,76 | distant | 70,23 | + |
| tumor | 31,04 | $isolated_{Lmod}$ | 24,29 | $rate_{PRPN}$ | 64,69 | + |
| isolate | 30,15 | $distant_{Lmod}$ | 24,28 | risk | 54,89 | + |
| inoperable | 28,16 | $patient_{PRPN}$ | 23,64 | $time_{PRPN}$ | 53,13 | + |

**Tableau 6.** *Illustration des premières entrées du vecteur de contexte du mot* recurrence *en fonction du taux de vraisemblance (TV) pour les représentations $graphique_5$ et syntaxique ainsi que pour la combinaison $a\_priori_5$*

| $\mathbf{graphique_5}$ | | **syntaxique** | | $\mathbf{a\_priori_5}$ | | +/− |
|---|---|---|---|---|---|---|
| isolated | 5,10 | $freedom_{PRPN}$ | 7,83 | $freedom_{PRPN}$ | 8,12 | + |
| geographic | 4,62 | $heat_{Robj}$ | 6,72 | $fat_{PRPN}$ | 7,02 | + |
| adjudication | 4,44 | $operable_{Rmod}$ | 6,72 | $threat_{PRPN}$ | 7,02 | + |
| conspicuous | 4,44 | $fat_{PRPN}$ | 6,72 | $operable_{Rmod}$ | 7,02 | + |
| liberate | 4,44 | $threat_{PRPN}$ | 6,72 | $heat_{Robj}$ | 7,02 | + |
| evade | 4,44 | $local_{Lmod}$ | 5,89 | $local_{Lmod}$ | 6,02 | + |
| inoperable | 4,38 | $fear_{PRPN}$ | 5,63 | $fear_{PRPN}$ | 5,93 | + |
| quarter | 4,29 | $suspicion_{PRPN}$ | 5,63 | $suspicion_{PRPN}$ | 5,93 | + |
| local | 4,28 | $inoperable_{Lmod}$ | 5,63 | $inoperable_{Lmod}$ | 5,93 | + |

**Tableau 7.** *Illustration des premières entrées du vecteur de contexte du mot* recurrence *en fonction du rapport des cotes actualisées (RCA) pour les représentations $graphique_5$ et syntaxique ainsi que pour la combinaison $a\_priori_5$*

un apport négatif pour au moins 5 mots. Ces mots qui n'étaient pas classés dans les premières entrées des méthodes $graphique_5$ et $syntaxique$ le sont maintenant avec la combinaison *a priori* utilisant l'information mutuelle. Il semble donc que la mesure IM ne soit pas appropriée dans ce cas, car elle ne préserve pas le classement des entrées de $graphique_5$ et $syntaxique$. Elle affecte des scores élevés à des mots qui avaient des scores faibles comme pour $rate$ ou $cancer$ par exemple, qui passent respectivement de $5,59$ à $14,39$ et de $2,28$ à $14,04$[14].

14. Nous partons du principe que chaque entrée du vecteur de contexte avec un score élevé constitue une information importante, que ce soit pour la représentation graphique ou syntaxique. Ainsi, dans le processus de combinaison le but est de ne pas reléguer ces entrées à des rangs inférieurs et de voir apparaître d'autres mots qui étaient moins bien classés aupa-

| $\mathbf{graphique_5}$ | | **syntaxique** | | $\mathbf{a\_priori_5}$ | | +/− |
|---|---|---|---|---|---|---|
| isolated | 8,73 | $local_{Lmod}$ | 14,77 | local | 16,17 | + |
| geographic | 8,15 | $tumor_{modN}$ | 13,84 | $local_{Lmod}$ | 15,83 | + |
| inoperable | 8,00 | $risk_{PRPN}$ | 12,84 | breast | 14,64 | − |
| local | 7,82 | $time_{PRPN}$ | 12,44 | rate | 14,39 | − |
| adjudication | 7,73 | $distant_{Lmod}$ | 12,09 | tumor | 14,15 | − |
| conspicuous | 7,73 | $rate_{modN}$ | 11,91 | cancer | 14,04 | − |
| reconcile | 7,73 | $year_{modN}$ | 11,80 | $risk_{PRPN}$ | 13,90 | + |
| liberate | 7,73 | $rate_{PRPN}$ | 11,63 | patient | 13,75 | − |
| quarter | 7,73 | $tumour_{modN}$ | 11,63 | $cancer_{modN}$ | 13,15 | − |
| | ⋮ | | ⋮ | | ⋮ | ⋮ |
| rate | 5,59 | $cancer_{modN}$ | 10,51 | | | |
| survival | 4,12 | | ⋮ | | | |
| tumor | 3,69 | | | | | |
| patient | 3,21 | | | | | |
| breast | 2,92 | | | | | |
| cancer | 2,28 | | | | | |

**Tableau 8.** *Illustration des premières entrées du vecteur de contexte du mot* recurrence *en fonction de l'information mutuelle (IM) pour les représentations* $graphique_5$ *et syntaxique ainsi que pour la combinaison* $a\_priori_5$

## 4. Réestimation de cooccurrences

Partant de l'hypothèse que les cooccurrences des mots ne sont pas fiables, surtout pour des corpus de petite taille (Zipf, 1949 ; Evert et Baroni, 2007) et sachant que les méthodes de réestimation visent à pallier ce problème, nous proposons une extension de l'approche directe en introduisant une étape intermédiaire qui consiste à réestimer les cooccurrences des mots observés, soit par des techniques de lissage soit par des techniques de prédiction des cooccurrences. Chaque valeur de cooccurrence (notée $cooc(i, j)$) est réestimée en fonction d'une des méthodes présentées dans les sections 4.1 et 4.2. La nouvelle estimation (notée $cooc^*(i, j)$) est alors utilisée pour calculer l'association entre les mots $i$ et $j$. Nous nous sommes limités ici à l'estimation des mots observés dans le corpus. Concernant les mots inconnus, Pekar *et al.* (2006) ont montré l'efficacité des méthodes de lissage pour l'alignement de termes peu fréquents.

Dans ce travail, la réestimation correspond à deux aspects différents. Un premier aspect porte sur des techniques de lissage qui permettent une redistribution des poids des mots selon certains critères. Ainsi plusieurs techniques de lissage de l'état de l'art sont abordées. Le deuxième aspect est motivé par le manque de corpus de grande

ravant (avec des scores faibles), ce qu'a tendance à faire la mesure de l'information mutuelle (favoriser les cooccurrences de faible fréquence).

taille en domaines spécialisés. L'idée consiste à se projeter dans un corpus de grande taille sans en avoir un à disposition, et ceci partant de l'hypothèse que si deux mots apparaissent $n$ fois dans un petit corpus, et que ces mots sont fortement liés alors ils apparaîtront ensemble $n \times k$ fois dans un corpus de plus grande taille. Ceci permet d'asseoir plus précisément leur association. Dans ce cas, nous cherchons une fonction qui permet de prédire les cooccurrences des mots dans un grand corpus partant des observations faites sur un corpus de taille plus modeste.

### 4.1. *Réestimation par méthode de lissage*

Les méthodes de lissage ont montré leur efficacité dans plusieurs domaines et notamment dans la traduction automatique. Nous présentons les différentes méthodes exploitées, à savoir la méthode de Laplace (Add-One), la méthode Good-Turing, l'estimation par interpolation linéaire de Jelinek-Mercer, la méthode de repli de Katz et celle de Kneser-Ney.

#### 4.1.1. *Méthode de Laplace (ou estimation Add-One)*

La méthode de Laplace (Lidstone, 1920 ; Johnson, 1932 ; Jeffreys, 1948) estime la probabilité $P$ en supposant que chaque type de mot absent ou non vu apparaît une fois. Dans ce cas, si nous avons $N$ événements et $V$ mots possibles alors, la probabilité du mot $w$ est :

$$P(w) = \frac{occ(w)}{N} \quad [8]$$

où $occ(w)$ est le nombre d'occurrences de $w$.

L'estimation de $P$ devient alors :

$$P_{Laplace}(w) = \frac{occ(w) + 1}{N + V} \quad [9]$$

L'utilisation de l'estimation de Laplace pour les cooccurrences des mots suppose que si deux mots apparaissent $n$ fois, alors ils peuvent apparaître $n + 1$ fois. Selon l'estimation du maximum de vraisemblance :

$$P(w_{i+1}|w_i) = \frac{cooc(w_i, w_{i+1})}{occ(w_i)} \quad [10]$$

En utilisant l'estimation de Laplace nous obtenons :

$$cooc^*_{Laplace}(w_i, w_j) = \frac{cooc(w_i, w_j) + 1}{occ(w_i) + V} \quad [11]$$

Nous pouvons constater que l'estimation de Laplace, qui est une technique très simple, coïncide avec notre objectif d'augmenter la valeur de cooccurrence des mots observés, et ceci, dans le but de renforcer leurs relations. Ceci étant dit, l'estimateur de Laplace comporte plusieurs désavantages :

– la probabilité des $n$-grammes fréquents est sous-estimée ;

– la probabilité des $n$-grammes rares ou absents est surestimée ;

– tous les $n$-grammes absents sont lissés de la même manière ;

– une trop grande masse de probabilité est réservée aux $n$-grammes absents (Gale et Church, 1994).

Une amélioration éventuelle serait d'ajouter une valeur inférieure à 1, selon l'équation suivante :

$$cooc^*_{Lidstone}(w_i, w_j) = \frac{cooc(w_i, w_j) + \delta}{occ(w_i) + \delta \times V} \quad [12]$$

avec $\delta \in [0, 1]$.

#### 4.1.2. *Estimateur de Good-Turing*

L'estimateur de Good-Turing (Good, 1953) fournit une autre manière de lisser les probabilités. Il stipule que pour chaque $n$-gramme apparaissant $r$ fois, nous pouvons prétendre qu'il apparaît $r^*$ fois. Cette hypothèse converge avec notre idée de prédire les cooccurrences des mots dans un corpus de grande taille en partant des observations faites à partir d'un corpus de petite taille. L'estimateur de Good-Turing utilise les comptes de ce que l'on observe une fois pour estimer les comptes de ce que l'on n'a jamais observé. Pour estimer les fréquences de cooccurrences des mots, nous aurons besoin de calculer $N_c$ qui est le nombre d'événements observés $c$ fois (ceci suppose que tous les événements suivent une loi binomiale). Soit $N_r$ le nombre de $n$-grammes qui apparaissent $r$ fois. $N_r$ peut être utilisé pour fournir une meilleure estimation de $r$. Étant donnée une distribution binomiale, la fréquence estimée devient alors :

$$r^* = (r + 1)\frac{N_{r+1}}{N_r} \quad [13]$$

Notre adaptation de l'estimateur de Good-Turing s'écrit comme suit :

$$cooc^*_{GT}(w_i, w_j) = (cooc(w_i, w_j) + 1)\frac{N_{cooc(w_i,w_j)+1}}{N_{cooc(w_i,w_j)}} \quad [14]$$

#### 4.1.3. *Estimation par interpolation linéaire*

Comme alternative aux $n$-grammes absents, Mercer (1980) proposent d'utiliser l'interpolation linéaire avec le modèle de Good-Turing. Notre adaptation aux cooccurrences des mots est définie selon la formule qui suit :

$$cooc^*_{int}(w_i, w_j) = \lambda \times cooc^*_{GT}(w_i, w_j) + (1 - \lambda) \times occ^*_{GT}(w_i) \quad [15]$$

avec $\lambda$ qui correspond au facteur de confiance du $n$-gramme. Il est souvent utile d'interpoler les $n$-grammes d'ordre supérieur avec les $n$-grammes d'ordre inférieur, car quand il y a un manque de $n$-grammes d'ordre supérieur, les $n$-grammes d'ordre inférieur peuvent compléter et apporter une information non négligeable.

4.1.4. *Repli de Katz*

Katz (1987) a étendu l'intuition de l'estimateur de Good-Turing en combinant les modèles d'ordre supérieur avec ceux d'ordre inférieur. Notre adaptation aux cooccurrences des mots est donnée par l'équation suivante :

$$cooc^*_{katz}(w_i, w_j) = \begin{cases} r^* & \text{si } r > 0 \\ \alpha(w_i) occ^*_{GT}(w_i) & \text{si } r = 0 \end{cases} \quad [16]$$

et :

$$\alpha(w_i) = \frac{1 - \sum_{w_i : cooc(w_i, w_j) > 0} cooc_{katz}(w_i, w_j)}{1 - \sum_{w_i : cooc(w_i, w_j) > 0} occ(w_i)} \quad [17]$$

Selon Katz (1987), l'estimation $r^*$ de Good-Turing n'est pas utilisée pour toutes les valeurs de $r$. En effet, les très grandes valeurs de $r$ sont supposées fiables à partir d'un certain seuil $k$. Katz (1987) suggère un $k = 5$. Ainsi, $r^* = r$ pour $r > k$ et :

$$r^* = \frac{(r+1)\frac{N_{r+1}}{N_r} - r\frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}} \quad [18]$$

pour $r \leq k$.

4.1.5. *Kneser-Ney*

Kneser et Ney (1995) ont proposé une extension de l'estimation par *absolute discounting*. La distribution d'ordre supérieur est estimée en soustrayant une valeur fixe $D$ de chaque valeur de la distribution. La différence avec la méthode *absolute discounting* standard réside dans la manière d'estimer la distribution d'ordre inférieur, comme le montre l'équation suivante :

$$cooc^*_{kney}(w_i, w_j) = \begin{cases} \frac{Max(cooc(w_i, w_j) - D, 0)}{\sum_{w_i} cooc(w_i, w_j)} & \text{si } cooc(w_i, w_j) > 0 \\ \\ \alpha(w_i, w_j) occ^*_{GT}(w_i) & \text{si } cooc(w_i, w_j) = 0 \end{cases} \quad [19]$$

où $\alpha(w_i, w_j)$ est choisi de sorte que la somme de la distribution soit égale à 1 (Chen et Goodman, 1999).

### 4.2. *Réestimation par prédiction*

Nous partons de l'hypothèse que chaque couple de mots dont la cooccurrence n'est pas le fruit du hasard dans un corpus de petite taille, devrait avoir le même comportement dans un corpus de plus grande taille avec une plus grande valeur de cooccurrence. Notre but est d'estimer cette nouvelle valeur. Gardons en tête la question qui motive cette démarche, à savoir qu'étant donné l'ensemble des valeurs de cooccurrences observées, $E^p = \{o^p_1, o^p_2, ..., o^p_N\}$, pour les couples de mots d'un corpus de petite taille,

quelles seront les valeurs attendues $E^g = \{o_1^g, o_2^g, ..., o_N^g\}$ de cet ensemble dans un corpus de grande taille ?

Nous proposons, dans ce qui suit, plusieurs méthodes pour estimer l'ensemble $E^g$ :

– une première technique fondée sur l'augmentation moyenne des cooccurrences des mots pour chaque valeur de cooccurrence ;

– une deuxième technique fondée sur un modèle de régression linéaire simple ;

– deux techniques fondées sur le maximum et la moyenne des valeurs observées sur un corpus d'apprentissage de grande taille.

#### 4.2.1. *Prédiction par l'augmentation moyenne des cooccurrences (AMC)*

Nous considérons l'utilisation de la moyenne comme une solution intuitive ayant pour principal objectif de fournir une information sur la tendance moyenne de l'augmentation des cooccurrences. Pour estimer les valeurs de l'ensemble $E^g$, nous utilisons un corpus d'apprentissage divisé en deux échantillons. Le premier échantillon correspond au corpus journalistique de petite taille (500 000 mots), et le second échantillon correspond au corpus journalistique de grande taille (10 millions de mots). Ainsi, nous calculons l'augmentation moyenne observée dans le corpus de grande taille pour tous les couples de mots ayant une valeur de cooccurrence $k$ dans le corpus de petite taille. Nous répétons cette procédure pour chaque valeur de cooccurrence, jusqu'à obtenir un vecteur d'augmentation pour toutes les valeurs de cooccurrences.

Soit $E_k^p = \{(w_i, w_j) \,|\, cooc_k^p(w_i, w_j) = k \; et \; i \in [1, N] \; et \; j \in [1, M]\}$ l'ensemble des couples de mots qui apparaissent $k$ fois dans le corpus de petite taille (chaque valeur de chaque couple est représentée par $cooc_k^p(w_i, w_j)$). Soit $E_k^g = \{(w_i, w_j) \,|\, cooc_k^g(w_i, w_j) = o_{ij} \; et \; i \in [1, N] \; et \; j \in [1, M]\}$ l'ensemble des couples de mots qui apparaissent avec leurs valeurs observées $o_{ij}$ dans le corpus de grande taille (chaque valeur de chaque couple est représentée par $cooc_k^g(w_i, w_j)$). La valeur moyenne d'augmentation $\mu_k$ est calculée comme suit :

$$\mu_k = \frac{1}{|E_k^p|} \sum_{i=1}^{N} \sum_{j=1}^{M} (cooc_k^g(w_i, w_j) - cooc_k^p(w_i, w_j)) \qquad [20]$$

Une fois les valeurs $\mu_1, \mu_2, ...\mu_l$ calculées, et pour un corpus de test donné, nous augmentons chaque valeur de cooccurrence observée ($cooc(w_i, w_j)$) par la valeur moyenne correspondante comme suit :

$$cooc_{AMC}^*(w_i, w_j) = cooc(w_i, w_j) + \mu_k \qquad [21]$$

avec AMC qui désigne l'augmentation moyenne des cooccurrences et $\mu_k$ la valeur moyenne correspondante si $cooc(w_i, w_j) = k$.

#### 4.2.2. *Prédiction par régression linéaire (REG)*

La régression linéaire est souvent utilisée pour étudier l'influence d'une variable quantitative $X$ sur une autre variable quantitative $Y$. La première est souvent appelée

variable explicative et la seconde est appelée variable expliquée. Dans notre problématique de prédiction des cooccurrences de mots dans un corpus de grande taille, en partant des observations faites sur un corpus de petite taille, nous sommes confrontés à un type de problème qui peut être résolu en utilisant une régression linéaire. Si nous désignons par la variable $X$ l'ensemble des observations des couples de cooccurrences dans un corpus de petite taille et par la variable $Y$ l'ensemble des observations des mêmes couples de cooccurrences dans un corpus de grande taille, l'objectif est alors de rechercher une liaison linéaire entre les variables $X$ et $Y$.

Dans notre cas, effectuer une régression linéaire signifie que l'on émet l'hypothèse que la fréquence de cooccurrences des mots doit croître proportionnellement à la taille du corpus. La droite de régression linéaire $y = ax + b$ a pour but de décrire au mieux la tendance du nuage observé et constitue donc un modèle de prédiction. Ainsi, la réestimation de la cooccurrence des mots $w_i$ et $w_j$, par exemple, est représentée par la formule :

$$cooc^*_{REG}(w_i, w_j) = a \times cooc(w_i, w_j) + b \quad [22]$$

avec $a$ et $b$ qui représentent les paramètres de régression.

#### 4.2.3. *Prédiction par la moyenne (MOY) et le maximum (MAX)*

À la différence de la méthode AMC, les modèles de prédiction MOY et MAX que nous proposons maintenant se fondent exclusivement sur les valeurs observées dans le corpus de grande taille pour estimer les nouvelles valeurs. L'estimation par la moyenne ($MOY$) est représentée par l'équation suivante :

$$MOY_k = \frac{1}{N} \sum_{i=1}^{N} compte(k, C_i) \quad [23]$$

où $k$ représente la valeur observée sur le corpus d'apprentissage de petite taille et $compte(k, C_i)$ représente la valeur observée pour un couple de mots $C_i$ sur le corpus de grande taille ayant une valeur de cooccurrence égale à $k$ dans le corpus d'apprentissage de petite taille. $N$ correspond au nombre de ces couples. La nouvelle estimation de la cooccurrence des mots $w_i$ et $w_j$, par exemple, est représentée par la formule :

$$cooc^*_{MOY}(w_i, w_j) = MOY_k \quad [24]$$

si $cooc(w_i, w_j) = k$.

De la même manière, en utilisant un processus d'estimation par le maximum ($MAX$) chaque valeur de cooccurrence est estimée selon l'équation suivante :

$$MAX_k = \max_{i=1}^{N} compte(k, C_i) \quad [25]$$

La nouvelle estimation de la cooccurrence des mots $w_i$ et $w_j$ est alors :

$$cooc^*_{MAX}(w_i, w_j) = MAX_k \quad [26]$$

si $cooc(w_i, w_j) = k$.

## 5. Évaluation

Dans cette section, nous présentons les résultats des expériences réalisées sur les trois corpus spécialisés. Nous commençons par évaluer individuellement l'apport de la combinaison de contextes et de la réestimation des cooccurrences avant d'évaluer leur complémentarité.

Les résultats sont présentés en fonction de la mesure de la précision moyenne MAP (*Mean Average Precision*) (Manning *et al.*, 2008) qui donne une bonne estimation de la qualité des listes bilingues extraites :

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{Rang_i} \quad [27]$$

avec $|Q|$ qui représente le nombre de termes à traduire, $Rang_i$ le rang de la traduction correcte renvoyée par le système. Les bornes de variation de MAP sont entre 0 et 1 mais par souci de lisibilité les résultats sont donnés sur une échelle de 100 (à noter que plus les scores de MAP sont hauts, meilleurs sont les résultats).

### 5.1. *Combinaison de contextes*

Nous évaluons ici i) l'approche directe fondée sur une représentation graphique notée $graphique_k$ ($k$ correspond à la taille de la fenêtre contextuelle et prend les valeurs 5, 7, 9, 11 et 15), ii) l'approche directe fondée sur une représentation syntaxique notée $syntaxique$, iii) la combinaison *a posteriori* notée $a\_posteriori_k$ qui combine les scores des approches $graphique_k$ et $syntaxique$ et iv) la combinaison *a priori* notée $a\_priori_k$ qui exploite les contextes fournis par une fenêtre contextuelle $graphique_k$ et les relations de dépendances $syntaxique$ dans un même vecteur pour ensuite appliquer l'approche directe. L'évaluation est réalisée sur les trois couples de mesures les plus utilisées dans l'état de l'art : TV-JAC (Morin, 2009), RCA-COS (Laroche et Langlais, 2010) et IM-COS (Gamallo, 2008a).

Les tableaux 9, 10 et 11 présentent pour chacun des trois corpus comparables les résultats des expériences pour les approches *a priori* et *a posteriori* pour les trois combinaisons de mesures d'association et de similarité. Pour la configuration TV-JAC, l'approche par combinaison *a priori* des contextes est supérieure à l'approche directe de base et ce pour les trois corpus spécialisés. Les meilleurs résultats sont obtenus en combinant $syntaxique$ avec des fenêtres de taille 5, 9 et 11. Concernant la configuration IM-COS, la combinaison *a priori* n'est pas efficace et dégrade même les résultats dans certains cas. Les résultats obtenus par la combinaison *a priori* associée à la configuration RCA-COS suivent le même comportement que les résultats de la configuration TV-JAC avec une nette amélioration des résultats. En ce qui concerne la combinaison *a posteriori*, nous voyons une nette amélioration des résultats selon différentes tailles de fenêtres.

| | **IM-COS** | **RCA-COS** | **TV-JAC** |
|---|---|---|---|
| $graphique_5$ | 25,3 | 24,5 | 27,7 |
| $graphique_7$ | 22,6 | 24,8 | 27,9 |
| $graphique_9$ | 17,3 | 23,9 | 26,6 |
| $graphique_{11}$ | 15,1 | 22,9 | 26,2 |
| $graphique_{15}$ | 11,1 | 20,4 | 23,7 |
| $syntaxique$ | 18,9 | 14,6 | 19,2 |
| $a\_priori_5$ | 20,0 | 31,5† | 33,3† |
| $a\_priori_7$ | 17,9 | 32,8† | 34,2† |
| $a\_priori_9$ | 13,1 | 33,0† | **34,5**† |
| $a\_priori_{11}$ | 11,7 | **34,0**† | 33,5† |
| $a\_priori_{15}$ | 09,8 | 31,8† | 31,3† |
| $a\_posteriori_5$ | **30,3**† | 31,8† | 32,0† |
| $a\_posteriori_7$ | 29,4† | 32,0† | 32,8† |
| $a\_posteriori_9$ | 27,2† | 33,5† | 33,4† |
| $a\_posteriori_{11}$ | 25,9† | 32,6† | 33,9† |
| $a\_posteriori_{15}$ | 23,2† | 32,5† | 32,0† |

**Tableau 9.** *Combinaisons de contextes* a priori *et* a posteriori *en MAP (%) pour le corpus du cancer du sein († : amélioration significative avec un indice de confiance de 0,01 pour le test de Student)*

| | **IM-COS** | **RCA-COS** | **TV-JAC** |
|---|---|---|---|
| $graphique_5$ | 19,9 | 18,6 | 23,9 |
| $graphique_7$ | 15,6 | 20,2 | 24,2 |
| $graphique_9$ | 12,5 | 18,2 | 24,3 |
| $graphique_{11}$ | 11,6 | 17,7 | 21,9 |
| $graphique_{15}$ | 09,1 | 14,9 | 20,0 |
| $syntaxique$ | 16,5 | 13,7 | 23,0 |
| $a\_priori_5$ | 14,9 | 27,8† | **34,1**† |
| $a\_priori_7$ | 13,8 | 29,6† | 32,2† |
| $a\_priori_9$ | 12,1 | 28,9† | 31,4† |
| $a\_priori_{11}$ | 10,5 | 28,5† | 29,8† |
| $a\_priori_{15}$ | 09,9 | 28,4† | 27,4† |
| $a\_posteriori_5$ | **30,0**† | 29,1† | 30,6† |
| $a\_posteriori_7$ | 26,4† | 31,3† | 31,0† |
| $a\_posteriori_9$ | 23,7† | 30,3† | 32,2† |
| $a\_posteriori_{11}$ | 21,2† | 31,5† | 29,5† |
| $a\_posteriori_{15}$ | 19,4† | **31,6**† | 28,7† |

**Tableau 10.** *Combinaisons de contextes* a priori *et* a posteriori *en MAP (%) pour le corpus des énergies renouvelables († : amélioration significative avec un indice de confiance de 0,01 pour le test de Student)*

| | **IM-COS** | **RCA-COS** | **TV-JAC** |
|---|---|---|---|
| $graphique_5$ | 29,3 | 33,3 | 43,5 |
| $graphique_7$ | 21,7 | 30,3 | 46,8 |
| $graphique_9$ | 20,5 | 34,7 | 46,1 |
| $graphique_{11}$ | 17,7 | 33,0 | 45,8 |
| $graphique_{15}$ | 13,3 | 28,3 | 44,2 |
| $syntaxique$ | 23,3 | 18,4 | 30,2 |
| $a\_priori_5$ | 20,6 | 44,9† | 49,8† |
| $a\_priori_7$ | 18,3 | 44,8† | 50,6† |
| $a\_priori_9$ | 15,2 | 48,1† | 50,5† |
| $a\_priori_{11}$ | 14,5 | **50,2**† | **53,2**† |
| $a\_priori_{15}$ | 13,1 | 49,2† | 50,1† |
| $a\_posteriori_5$ | **43,3**† | 45,7† | 49,1† |
| $a\_posteriori_7$ | 39,6† | 45,9† | 51,0† |
| $a\_posteriori_9$ | 37,4† | 48,4† | 51,9† |
| $a\_posteriori_{11}$ | 35,1† | 48,2† | 52,6† |
| $a\_posteriori_{15}$ | 33,1† | 49,0† | 52,5† |

**Tableau 11.** *Combinaisons de contextes a priori et a posteriori en MAP (%) pour le corpus de vulcanologie († : amélioration significative avec un indice de confiance de 0,01 pour le test de Student)*

Les deux nouvelles méthodes de combinaison proposées obtiennent globalement de meilleurs résultats que chaque représentation contextuelle prise individuellement, avec un léger avantage pour la méthode *a priori*. Nous avons aussi pu constater que la méthode *a priori* était plus sensible aux modifications des mesures d'association et de similarité que la méthode *a posteriori*. Cela provient probablement de la manière de réaliser la combinaison, la méthode *a posteriori* agit sur les scores alors que la méthode *a priori* agit directement sur le contenu des vecteurs de contexte.

#### 5.1.1. *Discussion*

Dans ce travail, nous cherchions dans un premier temps à comparer les deux principales représentations contextuelles utilisées avec l'approche directe, puis dans un second temps à proposer deux nouvelles manières de les combiner pour en augmenter les performances. La première remarque concerne l'utilisation de la représentation graphique $graphique_k$. Il est évident que le choix de la taille de la fenêtre joue un rôle important, comme nous avons pu le constater dans les différentes expériences. Dans la plupart des cas, ce sont des fenêtres de taille $5$, $7$ et $9$ qui donnent les meilleurs résultats. Ceci montre que la caractérisation du contexte des mots par ceux qui leur sont très proches semble être un choix adéquat, si l'on se fonde sur une caractérisation par fenêtre contextuelle. Le fait de choisir des fenêtres de taille plus grande n'améliore pas significativement les résultats de nos expériences.

La deuxième remarque concerne la représentation syntaxique. Dans Gamallo (2008a), cette représentation donne les meilleurs résultats. Dans nos expériences, en revanche, la méthode $syntaxique$ reste globalement en deçà de $graphique_k$. Ceci s'explique par trois facteurs. Le premier concerne la taille des corpus. Gamallo (2008a) avait utilisé des corpus de très grande taille (10 millions de mots environ) contrairement à nos corpus spécialisés qui sont de petite taille (600 000, 800 000 et 1 million de mots). Le deuxième facteur, qui est directement lié au premier, concerne la manière de considérer les entrées des vecteurs de contexte de la méthode $syntaxique$. Si dans le vecteur de contexte d'un mot $w_i$, il existe un mot $w_j$ avec une relation Lmod de $w_i$ ayant un score $S_{w_i^{Lmod}}$ et une autre relation Robj avec un score $S_{w_j^{Robj}}$, alors dans ce vecteur de contexte $w_j^{Lmod}$ et $w_j^{Robj}$ sont considérés comme étant deux mots différents, bien que ce soit le même mot avec deux relations de dépendances distinctes, cela rend la méthode $syntaxique$ plus sensible à la taille des corpus que $graphique_k$. Le troisième facteur est lié aux erreurs que peut produire l'analyseur en dépendances syntaxiques. Le fait de considérer un mot avec des étiquettes différentes comme des mots distincts peut expliquer les performances de la méthode de combinaison *a priori* des contextes. En effet, la méthode $a\_priori_k$ comble le manque de la méthode $syntaxique$, car elle considère les deux informations véhiculées par les deux représentations contextuelles. Ainsi, le fait d'exploiter une fenêtre de taille $k$ va permettre d'avoir une information sur le nombre de fois qu'un mot apparaît dans le contexte d'un autre et, comme deuxième information plus fine, la nature des relations qui existent entre les deux mots rendant ainsi les vecteurs de contexte plus discriminants.

## 5.2. *Réestimation de cooccurrences*

### 5.2.1. *Méthodes de lissage*

Nous comparons l'approche directe notée *AD* ($graphique_7$) avec les différentes techniques de lissage, à savoir la technique de Laplace ($Add1$), l'estimateur de Good-Turing ($GT$), la technique par interpolation linéaire de Jelinek-Mercer ($JM$), le repli de Katz ($Katz$) et la technique de Kneser-Ney ($Kney$).

Le tableau 12 montre les résultats des expériences réalisées sur les trois corpus spécialisés. Concernant le corpus du cancer du sein, la première observation concerne l'approche directe. Les meilleurs résultats sont obtenus pour la configuration TV-JAC avec une MAP de 27,9 %. Nous pouvons aussi remarquer que seule l'approche $Add1$ améliore significativement les résultats avec une MAP de 30,6 %. En revanche, les autres techniques de lissage dégradent les résultats. La deuxième observation concerne la configuration RCA-COS où aucune des techniques de lissage n'améliore significativement les résultats. Bien que les techniques $GT$ et $Katz$ montrent une légère amélioration avec une MAP de 25,2 % et 25,3 % respectivement, ces résultats ne sont pas significatifs. Le résultat le plus intéressant concerne la configuration IM-

| | **AD** | **Add1** | **GT** | **JM** | **Katz** | **Kney** |
|---|---|---|---|---|---|---|
| Cancer du sein | | | | | | |
| IM-COS | 22,6 | 24,8 | 25,6 | **29,5** | 25,9 | 09,1 |
| RCA-COS | 24,8 | 24,4 | 25,2 | 23,3 | **25,3** | 14,1 |
| TV-JAC | 27,9 | **30,6** | 21,4 | 21,2 | 21,2 | 22,9 |
| Énergies renouvelables | | | | | | |
| IM-COS | 17,8 | 23,6 | 22,9 | **30,1** | 24,2 | 14,1 |
| RCA-COS | 21,8 | **26,5** | 19,8 | 20,8 | 19,7 | 11,1 |
| TV-JAC | 25,7 | **29,7** | 20,5 | 21,3 | 21,3 | 22,9 |
| Vulcanologie | | | | | | |
| IM-COS | 21,7 | 29,7 | 25,2 | **32,3** | 29,7 | 15,7 |
| RCA-COS | 30,3 | 28,7 | **30,4** | 29,3 | **30,4** | 15,6 |
| TV-JAC | 46,8 | **47,3** | 44,2 | 39,4 | 39,4 | 40,0 |

**Tableau 12.** *Résultats en MAP (%) des expériences sur les trois corpus spécialisés*

COS pour laquelle quatre des cinq techniques de lissage améliorent les performances. La meilleure technique étant Jelinek-Mercer ($JM$) avec une MAP de 29,5 %.

Concernant le corpus des énergies renouvelables, d'une manière générale les résultats suivent le même comportement que celui observé dans la précédente expérience. Les meilleurs résultats de l'approche directe sont obtenus en utilisant la configuration TV-JAC avec une MAP de 25,7 %. Là encore, seule la technique du $Add1$ améliore les résultats avec une MAP de 29,7 %. Concernant la configuration RCA-COS, aucune des méthodes de lissage n'améliore significativement les résultats excepté la technique $Add1$. Finalement, le résultat le plus probant concerne à nouveau la configuration IM-COS où quatre des cinq techniques de lissage améliorent les résultats. La meilleure étant $JM$ avec une MAP de 30,1 %.

Pour le corpus de vulcanologie, les résultats confirment aussi le comportement de l'approche directe vis-à-vis des techniques de lissage. La technique du $Add1$ améliore les résultats avec une MAP de 47,3 % (configuration TV-JAC). Cependant, l'augmentation est moins importante en comparaison avec les autres corpus. Concernant la configuration RCA-COS, aucune des méthodes de lissage n'améliore significativement les résultats. Pour finir, le résultat le plus important concerne la configuration IM-COS pour laquelle encore une fois les mêmes quatre des cinq techniques de lissage améliorent les résultats. La meilleure étant $JM$ avec une MAP de 32,3 %.

### 5.2.2. *Techniques de prédiction*

Nous comparons maintenant l'approche directe avec les différentes techniques de prédiction, à savoir la technique par sélection du maximum et de la moyenne ($MAX$, $MOY$), le modèle de régression linéaire ($REG$) et l'augmentation moyenne des co-occurrences ($AMC$). Nous rajoutons aussi les résultats du modèle de Good-Turing

($GT$) à titre comparatif, sachant que celui-ci peut être considéré comme un modèle de prédiction. Le tableau 13 illustre les résultats des expériences menées sur les trois corpus spécialisés.

| | **AD** | **MAX** | **MOY** | **REG** | **AMC** | **GT** |
|---|---|---|---|---|---|---|
| | | | Cancer du sein | | | |
| IM-COS | 22,6 | **27,2** | 20,3 | 26,7 | 26,4 | 25,6 |
| RCA-COS | 24,8 | 22,9 | 19,8 | **27,6** | 20,9 | 25,2 |
| TV-JAC | **27,9** | 11,6 | 24,6 | 22,6 | 15,6 | 21,4 |
| | | | Énergies renouvelables | | | |
| IM-COS | 17,8 | 23,1 | 19,2 | **28,0** | 25,0 | 22,9 |
| RCA-COS | 21,8 | 15,7 | 17,0 | **23,3** | 18,0 | 19,8 |
| TV-JAC | **25,7** | 14,0 | 25,1 | 22,9 | 23,7 | 20,5 |
| | | | Vulcanologie | | | |
| IM-COS | 21,7 | 27,4 | 14,7 | **27,5** | 25,8 | 25,2 |
| RCA-COS | 30,3 | 26,1 | 19,5 | 29,2 | 28,4 | **30,4** |
| TV-JAC | **46,8** | 23,8 | 24,5 | 34,2 | 35,4 | 44,2 |

**Tableau 13.** *Résultats en MAP (%) des expériences sur les trois corpus spécialisés*

Concernant le corpus du cancer du sein, nous pouvons remarquer que pour la configuration TV-JAC aucun des modèles de prédiction n'améliore les résultats. Bien au contraire, les résultats sont même dégradés. Concernant la configuration RCA-COS seuls les modèles naïfs $MOY$, $MAX$ et $AMC$ sont en dessous de l'approche de référence $AD$. Le meilleur score est obtenu en utilisant le modèle $REG$ avec une MAP de 27,6 %. Le résultat le plus caractéristique est celui de la configuration IM-COS pour laquelle quatre des cinq modèles de prédiction améliorent les résultats. Le meilleur étant le modèle du $MAX$ qui atteint une MAP de 27,2 %.

Concernant le corpus des énergies renouvelables, les résultats suivent globalement le même comportement que l'expérience menée sur le corpus du cancer du sein. Là encore, aucun modèle de prédiction n'est efficace quand il est associé à la configuration TV-JAC. Les modèles $MOY$, $MAX$ et $AMC$ sont les seuls à montrer des résultats moins bons que ceux de l'approche directe en utilisant la configuration RCA-COS. Les meilleures performances sont données par l'approche $REG$ avec une MAP de 23,3 %. La configuration IM-COS reste indéniablement celle qui se combine le mieux avec quatre des cinq modèles de prédiction. Contrairement à l'expérience précédente, c'est le modèle $REG$ qui donne les meilleurs résultats avec une MAP de 28,0 %.

Concernant le corpus de vulcanologie, nous remarquons là encore qu'aucun modèle de prédiction associé à la configuration TV-JAC n'est efficace. Contrairement aux autres corpus, les résultats de la configuration RCA-COS sont moins bons que ceux de l'approche directe. La configuration IM-COS obtient les meilleurs résultats sur quatre des cinq modèles de prédiction. Là aussi, c'est le modèle $REG$ qui donne

les meilleurs résultats avec une MAP de 27,5 %. Nous notons cependant que ces améliorations sont moins importantes que celles obtenues pour les autres corpus.

#### 5.2.3. *Discussion*

Les techniques de lissage sont souvent évaluées d'après leur capacité à prédire les $n$-grammes non observés dans la phase d'apprentissage. Nous nous sommes exclusivement focalisés dans nos expériences sur les cooccurrences observées. Ainsi, le comportement des différentes techniques de lissage n'est pas forcément en adéquation avec l'état de l'art. C'est le cas par exemple de la méthode de Laplace ($Add1$), souvent considérée comme une approche assez simple, qui montre des résultats intéressants avec la configuration TV-JAC. Les bons résultats de l'approche $Add1$ sont sans doute liés à la seule prise en compte des cooccurrences observées, à l'exclusion des autres, principale faiblesse de $Add1$ pour le lissage de modèles de langue $n$-grammes. Concernant la configuration RCA-COS et malgré une légère amélioration en utilisant $Add1$, les techniques de lissage ont montré des résultats décevants. Si nous ne pouvons expliquer cela de manière formelle, nous pensons néanmoins que cela peut venir de la mesure du rapport des cotes actualisées qui se fonde sur la table de contingence. Ainsi, lisser l'information conjointe (les cooccurrences observées) sans lisser les informations disjointes pourrait expliquer ces résultats. Cette remarque est valable aussi pour la mesure du taux de vraisemblance. Là encore, aucune méthode de lissage (excepté $Add1$) n'a montré une amélioration des résultats pour la configuration TV-JAC. Nous n'avons pas trouvé de réponse évidente à cela. Ainsi, les mesures du rapport des cotes actualisées et du taux de vraisemblance ne semblent pas être compatibles avec la plupart des techniques de lissage.

Le résultat le plus remarquable concerne la configuration IM-COS. Excepté le lissage de Kneser-Ney, toutes les autres techniques ont montré de meilleurs résultats que l'approche directe sans lissage. La technique par interpolation linéaire de Jelinek-Mercer a été la plus performante. Nous expliquons ces résultats par le fait que la mesure de l'information mutuelle a tendance à surestimer les cooccurrences de faible fréquence et à sous-estimer les cooccurrences de fréquence élevée. Utiliser des techniques de lissage permet sans doute de corriger ce biais.

Si le lissage des modèles $n$-grammes a été traité dans de nombreux travaux (Chen et Goodman, 1999), l'estimation du Good-Turing est rarement utilisée seule et sert de base à d'autres techniques comme le repli de Katz ou l'interpolation linéaire de Jelinek-Mercer (deux techniques considérées comme performantes de manière générale). Dans nos expériences, les résultats obtenus par la technique de Kneser-Ney sont très décevants alors qu'elle est connue pour être l'une des meilleures techniques de lissage. L'explication peut venir, d'une part, du fait de ne pas considérer les cooccurrences non observées et, d'autre part, soustraire une valeur fixe pour toutes les cooccurrences observées pourrait altérer le modèle de cooccurrences de base et renforcer la surestimation des cooccurrences de faible fréquence pour l'information mutuelle.

En outre, nous avons présenté une autre manière de réestimer les cooccurrences des mots en considérant cette tâche comme un problème de prédiction. Nous avons pu

constater encore une fois les bonnes performances des modèles de prédiction (hormis *MOY*) pour la configuration IM-COS. Là encore, les bons résultats sont sans doute liés au biais apporté par l'information mutuelle sur les cooccurrences de faible fréquence, biais qui semble être corrigé par les modèles de prédiction. Globalement, les méthodes $MAX$ et $REG$ montrent les meilleures performances. Comme pour les techniques de lissage, les méthodes de prédiction n'ont pas apporté d'amélioration en utilisant la configuration TV-JAC. Alors que les modèles de prédiction tentent d'augmenter les cooccurrences des mots, dans le cas de la mesure du taux de vraisemblance, ceci provoque le contraire de l'effet escompté. Nous remarquons le même comportement concernant la mesure du rapport des cotes actualisées qui, hormis une légère amélioration en utilisant le modèle $REG$, offre de mauvais résultats quand elle est associée aux autres méthodes de prédiction. Si les mesures du taux de vraisemblance et du rapport des cotes actualisées sont fondées sur la table de contingence et s'il nous paraît difficile de construire des modèles de prédiction ou de lissage sur les informations disjointes, une autre solution serait de construire ces modèles de prédiction et/ou lissage une fois les mesures d'association calculées.

### 5.3. *Combinaison de contextes et réestimation de cooccurrences*

Dans cette dernière expérience, nous évaluons la complémentarité entre la combinaison de contextes et la réestimation de cooccurrences. Pour cela, nous sélectionnons la meilleure technique de réestimation ($Add1$) pour la meilleure configuration de mesures d'association et de similarité (TV-JAC), et nous la combinons avec les méthodes *a priori* et *a posteriori*. Les résultats sont présentés dans le tableau 14. Nous pouvons y constater une amélioration significative des résultats pour les deux méthodes de combinaison $a\_priori_7 + Add1$ et $a\_posteriori_7 + Add1$. Au regard de la différence de résultats entre ces deux méthodes de combinaison pour le corpus cancer du sein, nous pouvons recommander systématiquement l'utilisation de la combinaison $a\_posteriori_7 + Add1$ quel que soit le corpus envisagé.

| | **Cancer du sein** | **Énergies renouvelables** | **Vulcanologie** |
|---|---|---|---|
| $graphique_7$ | 27,9 | 25,7 | 46,8 |
| $Add1$ | 30,6 | 29,7 | 47,3 |
| $a\_priori_7$ | 34,2 | 32,2 | 50,6 |
| $a\_priori_7 + Add1$ | **38,1** | 35,1 | 52,7 |
| $a\_posteriori_7$ | 32,8 | 31,0 | 51,0 |
| $a\_posteriori_7 + Add1$ | 37,5 | **38,3** | **54,2** |

**Tableau 14.** *Résultats en MAP (%) des expériences pour la configuration* TV-JAC

### 5.4. *Synthèse*

Nous avons réalisé de nombreuses expériences visant à mieux comprendre la notion de contexte lexical sur laquelle repose l'approche directe. S'il est difficile de conclure de manière définitive sur la meilleure stratégie à adopter, nous pouvons néanmoins dégager des tendances utiles aux chercheurs confrontés à l'extraction de lexiques bilingues à partir de corpus comparables spécialisés.

Ainsi, dans le cas de corpus comparables spécialisés d'une taille modeste (autour de 1 million de mots) et pour des termes à traduire avec peu d'occurrences en corpus (entre 10 et 100 occurrences en langue source), il est préférable de privilégier une représentation graphique à une représentation syntaxique à travers les mesures TV-JAC. En outre, la combinaison *a priori* de ces deux représentations pour les mêmes mesures est à recommander pour accroître la qualité des lexiques bilingues extraits.

Une autre manière d'améliorer la qualité d'alignement est de combiner la méthode graphique avec une technique de réestimation de cooccurrences. Dans ce cas, il est préférable de privilégier une technique de réestimation par méthode de lissage à une technique de réestimation par prédiction. Là encore les mesures TV-JAC sont à privilégier pour la représentation graphique et la technique de lissage la plus prometteuse semble être le $Add1$.

Enfin, l'ensemble de ces approches peuvent être combinées entre-elles pour un gain optimum. Dans ce cas, nous ne saurions trop recommander d'utiliser la combinaison $a\_posteriori$ toujours pour les mesures TV-JAC en association avec la technique de lissage $Add1$.

## 6. Conclusion

Dans ce travail, nous avons étudié la notion de contexte lexical qui est au cœur de l'approche fondatrice en extraction de lexiques bilingues à partir de corpus comparables spécialisés. Cette étude a permis de mettre en évidence les limites de l'approche directe lorsqu'elle est confrontée à des corpus comparables spécialisés qui sont traditionnellement d'une taille modeste en comparaison avec un corpus comparable de langue générale. Une première limite concerne la représentation du contexte lexical des mots qui peut se faire à travers une représentation graphique ou syntaxique. À travers différentes expériences, nous avons montré l'intérêt de combiner ces deux représentations, qui induit une amélioration significative de la qualité des lexiques extraits, notamment lorsque cette combinaison est réalisée *a priori*. Une seconde limite concerne la difficulté de disposer d'observations significatives. Pour améliorer ce point, nous avons proposé de nous appuyer sur des stratégies de réestimation par méthode de lissage ou par prédiction des observations de cooccurrences de mots. Ces deux stratégies sont des alternatives pertinentes pour améliorer la représentativité des contextes lexicaux avec un avantage pour les méthodes de lissage. Nous avons finalement étudié la complémentarité de la combinaison de contextes et de la réestimation

des cooccurrences pour tirer parti des avantages de chaque approche. Là encore la qualité des lexiques extraits s'en trouve améliorée.

Même si d'autres approches ont été proposées depuis l'approche fondatrice en extraction de lexiques bilingues à partir de corpus comparables (Gaussier *et al.*, 2004 ; Haghighi *et al.*, 2008 ; Rubino et Linares, 2011 ; Hazem et Morin, 2012), nous pensons que cette étude du contexte lexical sur laquelle repose l'approche directe est pleinement profitable à toutes les approches qui exploitent le contexte lexical des mots.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-08-CORD-009.

## 7. Bibliographie

Ahmad K. A., Fulford H., Rogers M., « What's in a Term ? The Semi-automatic Extraction of Terms from Text. », *in* M. Snell-Hornby, F. Pöchhacker, K. Kaindl (eds), *Translation Studies : An Interdiscipline*, vol. xii, John Benjamins, Amsterdam/Philadephia, p. 267-278, 1994.

Aslam J. A., Montague M., « Models for Metasearch », *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, New Orleans, LA, USA, p. 275-284, 2001.

Bouamor D., Semmar N., Zweigenbaum P., « Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora », *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria, p. 759-764, 2013.

Bowker L., Pearson J., *Working with Specialized Language : A Practical Guide to Using Corpora*, Routeledge, New York, USA, 2002.

Brill E., « Some Advances in Transformation-Based Part of Speech Tagging », *Proceedings of the 12th National Conference on Artifical Intelligence (AAAI'94)*, Seattle, WA, USA, p. 722-727, 1994.

Chen S. F., Goodman J., « An empirical study of smoothing techniques for language modeling », *Computer Speech & Language*, vol. 13, n° 4, p. 359-393, 1999.

Chiao Y.-C., Zweigenbaum P., « The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations », *in* R. Baud, M. Fieschi, P. Le Beux, P. Ruch (eds), *The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, vol. 95 of *Studies in Health Technology and Informatics*, IOS Press, Amsterdam, p. 397-402, 2003.

Daille B., Morin E., « French-English Terminology Extraction from Comparable Corpora », *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, Jeju Island, Korea, p. 707-718, 2005.

Déjean H., Gaussier É., Sadat F., « An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction », *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Taipei, Taiwan, p. 1-7, 2002.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Evert S., The statistics of word cooccurrences : word pairs and collocations, PhD thesis, University of Stuttgart, 2005.

Evert S., Baroni M., « zipfR : Word Frequency Modeling in R », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, 2007.

Fano R. M., *Transmission of Information : A Statistical Theory of Communications*, MIT Press, Cambridge, MA, USA, 1961.

Firth J. R., « A synopsis of linguistic theory 1930–1955 », *Studies in Linguistic Analysis (special volume of the Philological Society)*, Blackwell, Oxford, p. 1-32, 1957.

Fung P., McKeown K., « Finding Terminology Translations from Non-parallel Corpora », *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, Hong Kong, p. 192-202, 1997.

Gale W. A., Church K. W., « What is wrong with adding one ? », *in* N. Oostdijk, P. de Haan (eds), *Corpus-based Research into Language*, Rodopi, Amsterdam, p. 189-198, 1994.

Gamallo P., « Learning bilingual lexicons from comparable english and spanish corpora », *Proceedings of the 11th Conference on Machine Translation Summit (MT Summit XI)*, Copenhagen, Denmark, p. 191—198, 2007.

Gamallo P., « Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora », *Proceedings of the 1st Workshop on Building and Using Comparable Corpora (BUCC'08)*, Marrakech, Marroco, p. 19-26, 2008a.

Gamallo P., « The Meaning of Syntactic Dependencies », *Linguistik Online*, 2008b.

Garera N., Callison-Burch C., Yarowsky D., « Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-speech Equivalences », *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, Colorado, USA, p. 129-137, 2009.

Gaussier E., Renders J.-M., Matveeva I., Goutte C., Déjean H., « A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora », *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, p. 526-533, 2004.

Good I. J., « The Population Frequencies of species and the estimation of population parameters », *Biometrika*, vol. 40, p. 16-264, 1953.

Grefenstette G., « Corpus-Derived First, Second and Third-Order Word Affinities », *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, Amsterdam, The Netherlands, p. 279-290, 1994a.

Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publisher, Boston, MA, USA, 1994b.

Groc C. D., « Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction », *Proceedings of The IEEE WICACM International Conferences on Web Intelligence*, Lyon, France, p. 497-498, 2011.

Haghighi A., Liang P., Berg-Kirkpatrick T., Klein D., « Learning Bilingual Lexicons from Monolingual Corpora », *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, Columbus, OH, USA, p. 771-779, 2008.

Harris Z. S., *Structures mathématiques du langage*, Dunod, 1971. Traduit de l'Américain par C. Fuchs.

Hazem A., Morin E., « ICA for bilingual lexicon extraction from comparable corpora », *Proceedings of the 5th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web (BUCC'12)*, Istanbul, Turkey, p. 126-133, 2012.

Jeffreys H., *Theory of Probability*, Clarendon Press, Oxford, 1948. 2nd edn Section 3.23.

Johnson W., « Probability : the deductive and inductive problems », *Mind*, vol. 41, n° 164, p. 409-423, 1932.

Katz S. M., « Estimation of probabilities from sparse data for the language model component of a speech recognizer », *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, n° 3, p. 400-401, March, 1987.

Kneser R., Ney H., « Improved backing-off for M-gram language modeling », *Proceedings of the 20th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, Detroit, MI, USA, p. 181-184, 1995.

Laroche A., Langlais P., « Revisiting context-based projection methods for term-translation spotting in comparable corpora », *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, Beijing, China, p. 617-625, 2010.

Lidstone G. J., « Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities », *Transactions of the Faculty of Actuaries*, vol. 8, p. 182-192, 1920.

Lin D., « Dependency-based Evaluation of MINIPAR », *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, 1998.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Mercer L. ; Jelinek F., « Interpolated Estimation of Markov Source Parameters from sparse data », *Workshop on pattern recognition in Practice*, Amsterdam, The Netherlands, 1980.

Morin E., « Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues », *Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France, 2009.

Morin E., Daille B., Takeuchi K., Kageura K., « Bilingual Terminology Mining – Using Brain, not brawn comparable corpora », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, p. 664-671, 2007.

Namer F., « FLEMM : Un analyseur flexionnel du français à base de règles », *Traitement Automatique des Langues (TAL)*, vol. 41, n° 2, p. 523-547, 2000.

Pekar V., Mitkov R., Blagoev D., Mulloni A., « Finding translations for low-frequency words in comparable corpora », *Machine Translation*, vol. 20, n° 4, p. 247-266, 2006.

Rapp R., « Identify Word Translations in Non-Parallel Texts », *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, Boston, MA, USA, p. 320-322, 1995.

Rapp R., « Automatic Identification of Word Translations from Unrelated English and German Corpora », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD, USA, p. 519-526, 1999.

Rubino R., Linares G., « A Multi-view Approach for Term Translation Spotting », *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, Tokyo, Japan, p. 29-40, 2011.

Salton G., Lesk M. E., « Computer evaluation of indexing and text processing », *Journal of the Association for Computational Machinery*, vol. 15, n° 1, p. 8-36, 1968.

Véronis J. (ed.), *Parallel Text Processing : Alignment and use of translation corpora*, Kluwer Academic Publishers, Dordrecht, 2002.

Zipf G. K., *Human Behaviour and the Principle of Least Effort : an Introduction to Human Ecology*, Addison-Wesley, 1949.

Zweigenbaum P., Habert B., « Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue », *GLOTTOPOL*, vol. 8, p. 22-44, 2006.