Acquisition semi-automatique de collocations à partir de corpus monolingues et multilingues comparables

Vincent Archer

GETA-CLIPS-IMAG (UJF & CNRS)
vincent.archer@imag.fr

Résumé

Cet article présente une méthode d'acquisition semi-automatique de collocations. Notre extraction monolingue estime pour chaque co-occurrence sa capacité à être une collocation, d'après une mesure statistique modélisant une caractéristique essentielle (le fait qu'une collocation se produit plus souvent que par hasard), effectue ensuite un filtrage automatique (en utilisant les vecteurs conceptuels) pour ne retenir que des collocations d'un certain type sémantique, puis effectue enfin un nouveau filtrage à partir de données entrées manuellement. Notre extraction bilingue est effectuée à partir de corpus comparables, et a pour but d'extraire des collocations qui ne soient pas forcément traductions mot à mot l'une de l'autre. Notre évaluation démontre l'intérêt de mêler extraction automatique et intervention manuelle pour acquérir des collocations et ainsi permettre de compléter les bases lexicales multilingues.

Mots-clés : collocations, acquisition semi-automatique, corpus comparables.

Abstract

This paper presents a method for the semi-automatic acquisition of collocations. Our monolingual extraction estimates the ability of each co-occurrence to be a collocation, using a statitical measure which represents an essential property (the fact that a collocation occurs more often than would be expected by chance), then makes an automatic filtering (using conceptual vectors) to keep only one semantic type of collocation, and finally makes a new filtering, using manually entered data. Our bilingual extraction uses comparable corpora, and is aiming to extract collocations which are not necessarily word-to-word translations. Our evaluation shows the interest of mixing automatic extraction and manual intervention to obtain collocations and, in this manner, to complete multilingual lexical databases.

Keywords: collocations, semi-automatic acquisition, comparable corpora.

1. Introduction

Il est nécessaire, dans les tâches de traitement automatique des langues, d'avoir des connaissances linguistiques. Celles-ci sont souvent insuffisantes : les systèmes actuels de traduction automatique ont de mauvaises performances, le plus souvent dues à une reconnaissance syntaxique erronée de la phrase à traduire et à une mauvaise gestion des collocations. Les *collocations* sont des expressions particulières, où l'un des termes (la *base*) est utilisé pour son sens habituel, et où l'autre (le *collocatif*) est employé, non pas pour son sens propre, mais pour modifier le sens de la base. Un exemple de collocation est « *pluie battante* », où le collocatif *battante* est employé pour exprimer l'intensification de la base *pluie*.

Les systèmes actuels traduisent généralement mot à mot les collocations : ainsi, *pluie battante*

sera traduit en *beating rain* alors qu'il faudrait obtenir *heavy* ou *driving rain*. Un dictionnaire bilingue ne peut proposer *heavy* ou *driving* parmi les traductions de *battante*, cela n'étant valable que dans un cas particulier. La base lexicale multilingue *Papillon* vise à modéliser les collocations comme des fonctions lexicales telles que *Magn*(*pluie*)={*battante*} (Mel'čuk *et al.*, 1995). Ainsi, un traducteur automatique utilisant une telle base saura qu'en français *gros fumeur* est une intensification de *fumeur*, et qu'en anglais *smoker* s'intensifie par *heavy* (lourd), et sera alors capable de traduire correctement *gros fumeur* en anglais par *heavy smoker*, ce qui n'est actuellement pas le cas.

Comment construire cette base ? On ne peut le faire ni manuellement (il y a trop de collocations à lister) ni automatiquement (il y a besoin de précision). L'idée est donc de permettre à l'homme et à l'ordinateur de combiner leurs compétences pour remplir la base. Les pistes sont diverses (extraction, apprentissage automatique, interaction avec des non-spécialistes, etc.). Nous allons présenter ici les travaux déjà mis en œuvre, qui consistent en deux expérimentations : la première est une extraction monolingue, se basant sur certaines propriétés linguistiques pour mettre en évidence des couples candidats pour être des collocations ; la seconde est bilingue et cherche à exploiter le fait que la base d'une collocation se traduit, mais pas nécessairement le collocatif.

2. Le projet Papillon

Le projet *Papillon* a pour but de créer une base lexicale multilingue (Mangeot-Lerebours *et al.*, 2003), consultable à l'adresse *http ://www.papillon-dictionary.org*. La base Papillon fonctionne comme un dictionnaire de plusieurs volumes (un pour chaque langue, et un pour la structure de pivot interlingue). L'architecture du dictionnaire s'organise selon deux niveaux :

– la macro-structure est le lien entre les entrées des différents volumes. La base a un pivot abstrait fait d'acceptions interlingues, ou *axies* (Sérasset, 1994) : chaque acception monolingue est reliée à une axie (les différences de raffinement sémantique entre les langues sont modélisées, ainsi l'axie liée à *river* se raffine en deux axies distinctes liées à *rivière* et *fleuve*).
– la micro-structure (celles des unités du lexique) reprend la structure du dictionnaire DiCo, que nous décrirons dans la suite de l'article, en y rajoutant quelques informations.

Le principe essentiel du projet Papillon est de construire la base lexicale à partir de dictionnaires existants (pour amorcer la base) et de contributions d'utilisateurs (ajout, suppression, modification) qui, une fois validées par des spécialistes, pourront être intégrées à la base. Il n'est pas nécessaire que les contributeurs soient des spécialistes de la langue : chaque locuteur francophone a une connaissance suffisante de sa langue maternelle pour savoir qu'une *peur bleue* est une peur intense. Les travaux présentés ici ont pour but de faciliter le remplissage de la base.

3. Modélisation

3.1. Les collocations

Avant l'extraction, il convient de définir précisément ce qu'est une collocation pour nous. En effet, tous les auteurs n'entendent pas la même chose lorsqu'ils emploient ce terme. Ainsi, (Sinclair *et al.*, 1970) définissent la collocation comme « la co-occurrence de deux unités dans un contexte, à l'intérieur d'un environnement spécifié » et parlent de collocation significative pour désigner « une collocation habituelle entre deux unités, telle qu'elles se trouvent ensemble plus souvent que leurs fréquences respectives », sans parler de la prédominance d'un terme sur l'autre. (Hausmann, 1989) introduira plus tard les termes de *base* et de *collocatif*. Nous adoptons

ici la définition de (Tutin et Grossmann, 2002) selon laquelle une collocation est « l'association d'une lexie L et d'un constituant C (généralement une lexie, mais parfois un syntagme) entretenant une relation syntaxique telle que C (le collocatif) est sélectionné en production pour exprimer un sens donné en cooccurrence avec L (la base), et que le sens de L est habituel ».

3.2. Modélisation. Les fonctions lexicales

Il existe de nombreuses modélisations du lexique qui ont, pour certaines, été mises en œuvre pour créer une base lexicale informatisée. Certaines d'entre elles ont des structures correspondant à des co-occurrences, pouvant parfois renfermer des collocations (le dictionnaire de co-occurrences d'*EDR*, la structure qualia du *Lexique Génératif* ou les troponymes de *Wordnet*) mais rien ne permet de les distinguer des autres co-occurences.

La seule modélisation qui cherche réellement à modéliser les collocations en tant que telles est celle des *Fonctions Lexicales*, qui fait partie de la théorie Sens-Texte d'Igor Mel'čuk. Une fonction lexicale donnée associe à une lexie *L* un ensemble de lexies qui a un lien particulier avec *L*. Les fonctions lexicales peuvent ainsi modéliser des relations paradigmatiques (liens sémantiques) comme la synonymie (par exemple Syn(*voiture*)=*automobile*), l'antonymie, la dérivation verbale, etc. Elles permettent également de modéliser les relations syntagmatiques (liens combinatoires, donc collocations) comme l'intensification (par exemple Magn(*vente*)=*grande*, *grosse*, *importante*), l'évaluation positive ou négative, les verbes supports, etc. Cette théorie a été mise en œuvre par la réalisation de *Dictionnaires Explicatifs et Combinatoires* sur papier afin d'affiner la théorie (Mel'čuk *et al.*, 1995), puis dans une version informatisée (et simplifiée) : le *DiCo* (Dictionnaire de Combinatoire), dont il existe une interface en ligne, le Dicouèbe, à l'adresse *http ://olst.ling.umontreal.ca/dicouebe*, et dont la structure des entrées a été reprise en grande partie dans la micro-structure du dictionnaire Papillon.

4. Extraction de collocations

4.1. Découverte de collocations

Nous avons cherché ici à acquérir des collocations d'intensification dont la base est un verbe et le collocatif un adverbe. (Sébillot, 2002) a montré l'intérêt de l'apprentissage sur corpus de relations lexicales sémantiques. Cependant, nous ne disposons actuellement pas d'une base d'apprentissage. De plus, nous voulons une méthode dont la mise en oeuvre soit facile, notamment dans d'autres langues. C'est pourquoi nous proposons une acquisition de collocations par extraction, qui puisse permettre, entre autres, d'obtenir une base d'apprentissage. (Léon et Millot, 2005), qui cherchaient à acquérir des relations lexicales bilingues, font passer leur précision finale de 7,5 % à 83,3 %, grâce à une simple validation manuelle des relations lexicales anglaises obtenues après un filtrage automatique, ce qui montre bien l'intérêt de l'intervention humaine pour compléter l'extraction automatique. Notre démarche repose sur le même principe : nous souhaitons appliquer un filtrage manuel simple (et ainsi utiliser les connaissances humaines du langage) à notre acquisition automatique (à partir de leur emploi dans des corpus).

4.2. Extraction monolingue

Le système *Xtract* (Smadja, 1993), même s'il ne cherchait pas à extraire la même chose que nous (les collocations y étaient définies comme « arbitraires et récurrentes », sans prédominance

d'un terme sur l'autre), montre l'intérêt d'utiliser une méthode hydride combinant une analyse linguistique (syntaxique) et un filtre statistique. C'est pourquoi notre démarche aussi est hybride. Nous avons utilisé pour notre expérimentation le corpus des articles du *Monde* de l'année 1995 (47 646 documents, 1 016 876 phrases, 24 730 579 mots, et 200 093 lemmes distincts).

Contextes d'extraction La partie linguistique de notre démarche repose sur une analyse syntaxique. En effet, il ne suffit pas que deux termes soient dans la même phrase pour qu'il s'agisse de collocations : il faut qu'entre la base et son collocatif existe une relation particulière (modification, dans notre cas). (Lin, 1998) a montré qu'on pouvait obtenir des résultats satisfaisants en utilisant les informations de dépendance. Nous avons considéré deux types de contexte : un contexte « simple »(couples où le verbe est immédiatement suivi par l'adverbe) et un contexte de dépendances (couples où le verbe est, d'après l'analyse de dépendance, modifié par l'adverbe).

4.2.1. Mesure de corrélation - Information mutuelle pondérée

Une des principales propriétés des collocations est de se produire plus souvent que par hasard. Nous choisissons donc d'employer une mesure mettant en évidence cela. La mesure classique d'*information mutuelle* ($IM(x,y) = \log_2[P(x,y)/(P(x)P(y))]$)[1] permet de quantifier cette corrélation entre les deux éléments (les dépendances se situant au niveau de la phrase, les probabilités utilisées ici et par la suite correspondent à l'apparition du phénomène dans une phrase). Cependant, cette mesure présente un inconvénient pour le traitement des langues : elle a tendance à favoriser les couples dont chaque terme a une fréquence très rare. Pour faire face à ce problème, (Fung et McKeown, 1997) définissent une mesure d'*information mutuelle pondérée* : $W(w_1,w_2) = P(w_1,w_2).\log\frac{P(w_1,w_2)}{P(w_1)P(w_2)}$. Nous utilisons ici l'adaptation de cette dernière mesure à des couples en relation, proposée par (Wu et Zhou, 2003) [2] :

$$WMI(w_1,r,w_2) = P(w_1,r,w_2).\log\frac{P(w_1,r,w_2)}{P(w_1|r)P(w_2|r)P(r)}$$

4.2.2. Filtrage des résultats. Les Vecteurs Conceptuels

avoir	pas	être	là	entrer	en vigueur
mettre	fin	coûter	cher	être	pas
avoir	jamais	arriver	en tête	vouloir	pas
avoir	déjà	mettre	en avant	avoir	toujours
valoir	mieux	aller	loin	lire	ci-dessous

Tableau 1. 15 premiers couples de l'expérimentation sur le contexte de dépendances, sans filtrage

Une première extraction naïve est réalisée sans aucun filtrage, en classant simplement tous les couples possibles avec la mesure WMI. Le tableau 4.2.2 montre deux facteurs de bruit : les verbes d'état ou de changement d'état (*être, arriver, aller*, etc.), dont nous supposons qu'ils ne peuvent jamais être modifiés par un adverbe, et de nombreux adverbes (*pas, jamais, déjà,*

[1] $P(w)$: probabilité d'occurrence de w ; $P(x,y)$; probabilité de co-occurrence de x et y en contexte

[2] $P(r)$ probabilité d'appartion de la relation r ; $P(w_1,r,w_2)$: probabilité d'apparition du couple (w_1,w_2) en relation r ; $P(w|r)$ probabilité d'apparition de w parmi tous les couples en relation r

mieux, etc.) qui ne peuvent jamais correspondre à une intensification. Pour résoudre le premier problème, nous rajoutons un filtre qui supprime tous les couples dont la base est un tel verbe.

Les Vecteurs Conceptuels On peut, pour obtenir des résultats de qualité, se baser sur des données sémantiques existantes : par exemple, (Pearce, 2001) utilise les synsets de Wordnet pour savoir qu'on emploie avec un terme donné, un terme plutôt que son synonyme, et extrait ainsi des collocations. Ici, nous souhaitons avoir un filtre sémantique, pour nous concentrer sur le concept d'intensification. Nous utilisons pour cela les *Vecteurs Conceptuels*, qui représentent les termes du langage sous forme de vecteurs dont les dimensions sont des concepts de base de la langue (Schwab *et al.*, 2002) ; la distance entre deux termes est la distance angulaire entre les vecteurs correspondants. Une expérimentation des vecteurs conceptuels est menée sur le français au LIRMM, à partir des 873 concepts décrits dans (Larousse, 1992). Ils sont construits à partir de différentes sources (listes de synonymes, dictionnaires, indexation manuelle) qui permettent de calculer de nouveaux vecteurs. Le système est amorcé à la main avec des vecteurs pré-calculés.

En effet, nous pensons que l'ensemble des adverbes les plus proches (d'après la distance angulaire) de *c4.intensité*[3] peut constituer une « classe »d'adverbes exprimant l'intensification, permettant ainsi un filtrage sémantique. Nous récupérons les 500 termes qui sont les plus proches de *c4.intensité* à l'adresse *http ://www.lirmm.fr/∼lafourcade*, et filtrons les résultats précédemment obtenus, en ne conservant que les couples dont l'adverbe fait partie de la liste. Le tableau 4.2.2 montre qu'on retrouve alors bien plus facilement des collocations d'intensification (*régner sans partage*, *changer radicalement*), mais il reste toujours des couples dont l'adverbe n'exprime jamais l'intensification (*en partie*, *peu*, etc.). La classe d'adverbes obtenue en utilisant les Vecteurs Conceptuels, bien que globalement satisfaisante, ne correspond donc pas exactement à la notion d'intensification : ils doivent être encore affinés.

juger	trop	ancrer	à gauche	devoir	beaucoup
expliquer	en partie	changer	radicalement	coûter	très
régner	sans partage	réduire	considérablement	ignorer	superbement
montrer	très	réduire	d'autant	assumer	pleinement
savoir	trop	financer	en partie	voter	à gauche

Tableau 2. 15 premiers couples de l'expérimentation sur le contexte simple, avec filtrage par Vecteurs Conceptuels

4.2.3. Filtrage manuel

Il faut donc effectuer un nouveau filtrage pour affiner nos résultats : nous établissons la liste des adverbes proches du concept d'intensité d'après les Vecteurs Conceptuels, mais qui n'intensifient en réalité jamais les verbes qu'ils modifient : *trop*, *en partie*, *très*, *par trop*, *à gauche*, *peu*, *de plus*, *d'autant*, *tant*, *partiellement*, *tellement*. Nous éliminons alors de nos précédents résultats tous les couples dont l'adverbe fait partie de cette liste. Le tableau 4.2.3 montre que les résultats désormais obtenus semblent satisfaisants (le bruit semble relativement faible), ils seront évalués dans la suite de l'article.

[3] *c4* signifie concept de base de *4*ème niveau dans la hiérarchie de concepts

régner	sans partage	assumer	pleinement	varier	considérablement
changer	radicalement	parler	beaucoup	consacrer	entièrement
réduire	considérablement	aimer	beaucoup	augmenter	considérablement
devoir	beaucoup	exercer	pleinement	jouer	pleinement
ignorer	superbement	montrer	particulièrement	faire	beaucoup

Tableau 3. 15 premiers couples de l'expérimentation sur le contexte simple, avec filtrage par Vecteurs Conceptuels, puis affinage manuel

4.3. Extraction bilingue contrastive

La traduction d'un terme complexe n'est pas toujours obtenue en traduisant mot à mot ses composants (Morin *et al.*, 2004). Nous essayons donc dans cette démarche d'extraire des couples *collocation française - collocation anglaise* où l'une est traduction de l'autre de manière *contrastive*, c'est-à-dire où l'un des deux termes (celui qu'on suppose être la base) se traduit, mais où l'autre (celui qu'on suppose être le collocatif) ne se traduit pas nécessairement directement.

4.3.1. Comparabilité des documents

Nous menons nos travaux sur des corpus de documents comparables (qui parlent du même thème), qui ont pour avantage le fait qu'une collocation et sa traduction peuvent se retrouver dans deux documents qui parlent du même thème sans que ces documents soient forcément traductions l'un de l'autre, comme c'est le cas avec les corpus de documents alignés. Nous utilisons pour notre expérimentation le corpus des articles du *Monde* de l'année 1995 présenté dans l'expérimentation monolingue, et le corpus des articles du Glasgow Herald de la même année (56 472 documents, 1 321 323 phrases, 28 122 780 mots, et 175 207 lemmes distincts), qui décrit donc en partie les mêmes événements.

Nous ne savons pas si la convergence thématique des documents est nécessaire pour l'extraction de collocations, une évaluation serait nécessaire pour le savoir. Mais, si nous gardions tous les couples bilingues possibles de documents dans notre expérimentation, il faudrait alors en gérer environ 2,7 milliards : afin de réduire la taille des calculs, nous prenons comme hypothèse qu'il est plus probable de trouver des collocations traductions l'une de l'autre quand on considère des documents comparables. Pour savoir quels documents sont comparables, nous nous fions aux indices suivants : deux articles de journaux de langue différente parlent du même événement si leurs dates de publication sont proches, si les entités nommées se retrouvent dans les deux, et si les syntagmes nominaux (les termes qui expriment le thème du document) sont traduits littéralement. Pour évaluer la similarité entre deux ensembles de mots, nous utilisons la mesure overlap ($overlap(x, y) = |X \cap Y| / \min(|X|, |Y|)$) qui permet à un document court et un document long parlant d'un même thème d'avoir une forte similarité. De plus, nous considérons que deux documents ne peuvent être similaires que s'ils ont été publiés à deux jours ou moins d'intervalle. Nous calculons la similarité d'un document français et d'un document anglais ainsi :

$$sim(D_{fr}, D_{en}) = \frac{overlap(D_{fr}, D_{en}) + overlap(Trad_{en}(D_{fr}), D_{en})}{2}$$

où D_{fr} et D_{en} sont les ensembles de syntagmes nominaux de chaque document, et $Trad_{en}(D_{fr})$ l'ensemble des traductions des éléments de D_{fr}. Nous calculons la similarité pour tous les couples possibles, et considérons que D_{fr} et D_{en} sont similaires si $sim(D_{fr}, D_{en})$ dépasse 0,2. Le choix de la mesure est certainement discutable, et celle-ci pourrait sans aucun doute être

améliorée, mais nos travaux concernent surtout l'extraction bilingue à partir de corpus de documents comparables. Le choix du seuil se justifie par la volonté d'avoir un nombre raisonnable d'associations de comparabilité entre documents : nous obtenons ainsi 63621 associations (environ 1,34 association par document français et 1,13 association par document anglais).

4.3.2. Extraction à partir de documents comparables

Une fois établie la comparabilité des documents, nous pouvons réaliser l'extraction proprement dite. Nous considérons désormais des couples de co-occurrences (une co-occurrence française et une co-occurrence anglaise) candidates pour être des collocations contrastives. Nous exprimons le fait que si une collocation apparaît dans un document, il est fort probable que sa traduction apparaisse dans un document comparable. Nous adaptons pour cela la mesure classique de similarité cosinus ($cos(x, y) = |X \cap Y|/\sqrt{|X| \times |Y|}$) à des ensembles distincts comparables :

$$cos_{bilingue}(c_{fr}, c_{en}) = \frac{|CORRES(c_{fr}, c_{en})|}{\sqrt{|CFR| \times |CEN|}}$$

où CFR et CEN sont les ensembles de documents où apparaissent respectivement c_{fr} et c_{en}, et $corres(c_{fr}, c_{en})$ l'ensemble des couples de documents (d_{fr}, d_{en}) tels que d_{fr} et d_{en} soient comparables et que c_{fr} et c_{en} apparaissent respectivement dans d_{fr} et d_{en}. Il faut également que les deux couples de termes que l'on cherche à associer soient tous les deux des collocations, nous utilisons donc une pondération visant à mettre cela en évidence (la mesure WMI modélisant la principale caractéristique des collocations, qui est de se produire plus souvent que par hasard) :

$$poids_{bilingue}(c_{fr}, c_{en}) = (WMI(c_{fr}) + WMI(c_{en})) \times cos_{bilingue}(c_{fr}, c_{en})$$

Ce poids est calculé pour tous les quadruplets possibles (verbe français, adverbe français, verbe anglais, adverbe anglais) où les deux verbes sont traductions. Les premiers résultats, une fois que tous les filtrages ont été effectués, sont présentés dans le tableau 4.3.2.

mettre	beaucoup	take	seriously	mettre	beaucoup	take	long
jouer	pleinement	work	hard	jouer	pleinement	work	regularly
vouloir	particulièrement	want	really	exercer	pleinement	train	specially
accomplir	particulièrement	perform	strongly	jouer	pleinement	act	unlawfully
regretter	énormément	regret	deeply	jouer	pleinement	work	responsibly

Tableau 4. 10 premiers quadruplets de l'expérimentation bilingue

5. Évaluation

Nous évaluons ici la qualité des résultats produits. Pour l'extraction monolingue, le candidat est-il bien une collocation ? Pour l'extraction bilingue, le quadruplet est-il bien un couple de deux collocations traductions l'une de l'autre ? Nous calculons pour chaque expérimentation une mesure de précision (ne pouvons pas calculer de rappel, ne disposant pas de base de référence). Il n'y a pas de mesure d'évaluation standardisée pour l'acquisition de collocations (contrairement à la traduction automatique, par exemple). De plus, nous ne cherchons pas à acquérir exactement la même chose que d'autres : dans l'extraction monolingue, nous n'avons pas la même définition de *collocation* que (Smadja, 1993) ; dans l'extraction bilingue, nous cherchons

à extraire des associations bilingues où les collocations ne sont pas nécessairement traductions l'une de l'autre, là où (Wu et Zhou, 2003) extrayaient des associations où les collocations le sont. Nous ne pouvons donc pas comparer objectivement ces différents travaux.

5.1. Expérimentation monolingue

Nous procédons, pour chaque expérimentation, à l'évaluation des 1000 premiers couples produits avec notre méthode, d'après la mesure WMI. Les résultats sont résumés dans le tableau 5.1. Seulement 17 % des candidats produits sans aucun filtrage sont réellement des collocations.

Expérimentation		Précision
Sans filtrage	Contexte de dépendances	17 %
Filtrage par Vecteurs Conceptuels	Contexte de dépendances	41 %
Filtrage par Vecteurs Conceptuels	Contexte simple	44 %
Filtrage par Vecteurs Conceptuels + affinage	Contexte simple	83 %

Tableau 5. Évaluation des expérimentations monolingues

Pour le même type de contexte, la précision est multipliée par 2,5 (41 %) grâce au filtrage automatique. Le filtrage nous permet donc bien de gagner en précision, même s'il fait baisser le rappel (on perd alors des vraies collocations, comme *refuser obstinément*, *défendre bec et ongles*, *reprendre de plus belle*, etc.). Autre enseignement de cette évaluation : la connaissance des dépendances ne permet pas d'améliorer les résultats ; en effet, le contexte simple (verbe suivi immédiatement d'un adverbe) donne une meilleure précision que le contexte de dépendances (verbe modifié par l'adverbe). Cela peut s'expliquer par le fait qu'en utilisant les analyses de dépendances, on récupère plus d'adverbes éloignés du verbe (on augmente ainsi le rappel), mais on est également plus sensible aux erreurs de l'analyseur (on perd donc en précision). Enfin, le simple fait d'ajouter un nouveau filtrage pour éliminer les adverbes qui ne peuvent jamais être utilisés pour une intensification (*trop*, *en partie*, *partiellement*, etc.) permet d'éliminer environ 47 % des couples candidats, et de faire grimper la précision des résultats de 44 % à 83 %.

5.2. Expérimentation bilingue

Nous obtenons 80 298 quadruplets. Après filtrage en français (par vecteurs conceptuels) et en anglais (avec des stoplists pour éliminer les adverbes de lieu, de temps, de doute, etc.), on ne conserve que 501 candidats. Enfin, un filtrage manuel d'après les premiers retours de notre évaluateur (*by* était reconnu comme un adverbe au lieu d'une préposition, *too* reconnu seul alors qu'il fait partie de syntagmes comme *too many* ou *too much*, et les "*phrasal verbs*" comme *get up* ou *keep apart* où l'adverbe fait partie du syntagme et ne peut donc pas être un intensificateur) nous ramène à 201 candidats seulement : cette démarche est simple à mettre en œuvre (il est plus facile de trouver des corpus comparables), mais il y a plus de bruit, on doit donc filtrer beaucoup plus pour avoir des résultats intéressants. Le tableau 5.2 présente le résultat de l'évaluation des 201 associations candidates (dont le poids calculé est positif).

La proportion de candidats corrects décroît au fil du classement, ce qui est satisfaisant. De même, la proportion d'associations correctes par collocation correcte est divisée environ par 2 entre la première et la seconde moitié. Le fait que les couples français et anglais soient tous deux des collocations n'implique pas qu'il s'agisse de traductions, ce qui explique que la précision

Intervalle	Collocations anglaises correctes	Associations où les 2 couples sont traductions	Proportion d'associations correctes par collocations anglaises correctes
1-50	20	8	40 %
51-100	20	9	45 %
101-150	14	3	21 %
151-201	9	2	22 %

Tableau 6. Évaluation des expérimentations bilingues

est plus faible ici. La première explication vient des verbes polysémiques, où la traduction en anglais du verbe correspond à une acception qui n'est pas celle dans laquelle il est employé en français (*exercer pleinement* vs. *train specially*). L'autre cause d'erreur est le fait que la modification induite par le collocatif peut ne pas porter sur le même argument du prédicat (*accomplir particulièrement* intensifie l'action accomplie, *perform strongly* la manière dont elle est accomplie).

6. Conclusion

Nous avons décrit une méthode semi-automatique d'extraction de collocations, monolingue et bilingue, basée sur des corpus de documents comparables. Nous avons montré que l'intervention d'un spécialiste était nécessaire dans un tel processus pour avoir des résultats de qualité : cela permet une très bonne précision lors de notre expérimentation monolingue. Pour l'expérimentation bilingue, on rajoute un problème (les collocations d'intensité dont les bases sont traductions ne sont pas nécessairement traductions l'une de l'autre), ce qui explique que la précision soit moins bonne. Ces travaux possèdent de nombreuses perspectives, que ce soit au niveau de l'extraction elle-même (celle présentée ici nécessite un pré-traitement linguistique contraignant, on souhaite donc avoir aussi une méthode entièrement statistique) ou au niveau plus général de la découverte de collocations : apprendre automatiquement les caractéristiques des collocations à partir de la base produite ; permettre d'interagir avec des contributeurs non spécialistes en utilisant les données extraites (comme suggestions, ou pour valider sa proposition) ; étendre les résultats grâce à des thesaurus (retrouver *trouille bleue* si on a déjà *peur bleue*). Enfin, puisque nous avons montré l'utilité de l'intervention humaine, nous souhaitons réaliser des programmes facilement utilisables par des linguistes non-informaticiens.

Références

FUNG P. et MCKEOWN K. (1997). « A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups ». In *Machine Translation*, 12 (1-2), 53–87.

HAUSMANN F. J. (1989). « Le dictionnaire de collocations ». In F. Hausmann, O. Reichmann, H. Wiegand et L. Zgusta (éds.), *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*, p. 1010–1019. Berlin : De Gruyter.

LAROUSSE (1992). *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse.

LIN D. (1998). « Extracting Collocations from Text Corpora ». In *First Workshop on Computational Terminology*. Montréal, Canada.

LÉON S. et MILLOT C. (2005). « Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web ». In *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2005)*. Dourdan, p. 595–604.

MANGEOT-LEREBOURS M., SÉRASSET G. et LAFOURCADE M. (2003). « Construction collaborative d'une base lexicale multilingue, Le projet Papillon ». In *Traitement Automatique des Langues*, 44 (2), 151–176.

MEL'ČUK I. A., CLAS A. et POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot.

MORIN E., DUFOUR-KOWALSKI S. et DAILLE B. (2004). « Extraction de terminolgies bilingues à partir de corpus comparables ». In P. Blache (éd.), *Actes de TALN 2004 (Traitement automatique des langues naturelles)* : LPL. ATALA, Fès, Maroc.

PEARCE D. (2001). « Synonymy in Collocation Extraction ». In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*. CMU.

SCHWAB D., LAFOURCADE M. et PRINCE V. (2002). « Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales ». In J.-M. Pierrel (éd.), *Actes de TALN 2002 (Traitement automatique des langues naturelles)* : ATILF. ATALA, Nancy, p. 125-134.

SINCLAIR J., JONES S. et DALEY R. (1970). *English Lexical Studies : Report to OSTI on Project C/LP/08*. Rapport interne, Departement of English, University of Birmingham.

SMADJA F. (1993). « Retrieving collocations from text : Xtract ». In *Computational Linguistics*, 19 (1), 143–177.

SÉBILLOT P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*. Habilitation à diriger des recherches, Université Rennes 1.

SÉRASSET G. (1994). *Sublim : un système universel de bases lexicales multilingues et Nadia : sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse de doctorat, Université Joseph Fourier - Grenoble 1.

TUTIN A. et GROSSMANN F. (2002). « Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif ». In *Revue Française de Linguistique Appliquée*, VII (1), 7–26.

WU H. et ZHOU M. (2003). « Synonymous Collocation Extraction Using Translation Information ». In E. Hinrichs et D. Roth (éds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* : National Institute of Informatics. Sapporo, Japan, p. 120-127.