Extraction de collocations et leurs équivalents de traduction à partir de corpus parallèles

Violeta Seretan

Laboratoire d'analyse et de technologie du langage, Université de Genève
2, rue de Candolle, CH-1205 Genève, Suisse

Violeta.Seretan@unige.ch

RÉSUMÉ. *Identifier les collocations dans le texte source (par exemple,* break record*) et les traduire correctement (*battre record *contre* *casser record*) constituent un réel défi pour la traduction automatique, d'autant plus que ces expressions sont très nombreuses et très flexibles du point de vue syntaxique. Cet article présente une méthode permettant de repérer des équivalents de traduction pour les collocations à partir de corpus parallèles, qui sera utilisée pour augmenter la base de données lexicales d'un système de traduction. La méthode est fondée sur une approche syntaxique « profonde », dans laquelle les collocations et leurs équivalents potentiels sont extraits à partir de phrases alignées à l'aide d'un analyseur multilingue. L'article présente également les outils qui sont utilisés par cette méthode. Il se concentre en particulier sur les efforts déployés afin de rendre compte des divergences structurelles entre les langues et d'optimiser la performance de la méthode, notamment en ce qui concerne la couverture.*

ABSTRACT. *Identifying collocations in a text (e.g.,* break record*) and correctly translating them (*battre record *vs.* *casser record*) represent key issues in machine translation, notably because of their prevalence in language and their syntactic flexibility. This article describes a method for discovering translation equivalents for collocations from parallel corpora, aimed at increasing the lexical coverage of a machine translation system. The method is based on a "deep" syntactic approach, in which collocations and candidate translations are identified from sentence-aligned text with the help of a multilingual parser. The article also introduces the tools on which this method relies. It focuses in particular on the efforts made to account for structural divergences between languages and to improve the method's performance in terms of coverage.*

MOTS-CLÉS : *collocations, équivalents de traduction, analyse syntaxique, alignement de texte.*

KEYWORDS: *collocations, translation equivalents, syntactic parsing, text alignment.*

TAL. Volume 50 – n° 1/2009, pages 305 à 332

1. Introduction

Les collocations (ou associations habituelles de mots, telles que *jouer un rôle*, *battre un record*, *jeter les bases*, *combler une lacune*, *remplir une condition*) constituent un sous-type d'expressions à mots multiples qui présentent un intérêt particulier dans des domaines comme l'étude des langues étrangères et la lexicographie, et qui ont attiré aussi, durant les dernières années, une attention toujours plus grande dans le domaine du traitement automatique des langues. Deux raisons majeures contribuent à cet intérêt croissant : d'une part, la présence massive des collocations dans le langage – plusieurs auteurs affirment que les collocations sont les plus nombreuses parmi les expressions à mots multiples, par exemple, (Mel'čuk, 1998), tandis qu'une étude de corpus (Howarth et Nesi, 1996) avait montré que, effectivement, toute phrase est susceptible de contenir au moins une collocation –, d'autre part, l'idiosyncrasie des collocations. Ainsi, même si les collocations sont assez semblables aux constructions régulières du langage, elles partagent certains traits des expressions figées, notamment le caractère idiomatique de l'encodage.

Par exemple, l'expression *battre un record* est facile à décoder, son sens étant relativement transparent et compréhensible ; néanmoins, tout comme les expressions idiomatiques, elle est difficile à encoder. Le collocatif du nom *record*, le verbe *battre*, est imprédictible pour les locuteurs non natifs du français. En accord avec (Fillmore *et al.*, 1988), nous considérons que dans les collocations, l'encodage (à la différence du décodage) est idiomatique. La conséquence de cette idiomaticité est que les collocations n'admettent généralement pas de traduction littérale vers une autre langue. Ainsi, l'équivalent anglais de *battre un record* est *to break a record*, plutôt que **to beat a record* ; inversement, la collocation anglaise *to break a record* ne peut pas être traduite littéralement en français, comme **casser un record*. L'expression allemande *[eine] falsche Entscheidung treffen* ('prendre une mauvaise décision', lit. *[une] fausse décision rencontrer*) est un autre exemple qui est particulièrement éloquent. Cette collocation à trois termes est composée de deux collocations binaires imbriquées, une de type adjectif-nom (*falsche Entscheidung*), l'autre de type verbe-objet (*Entscheidung treffen*), aucune des deux n'admettant de traduction littérale vers le français[1].

Au cours des dernières décennies, des nombreux travaux de recherche se sont concentrés sur l'extraction des collocations à partir de corpus textuels, et en particulier sur l'évaluation de la performance des mesures d'association lexicale employées ; voir, entre autres, (Lafon, 1984 ; Church et Hanks, 1990 ; Smadja, 1993 ; Daille, 1994 ; Kilgarriff *et al.*, 2004 ; Evert, 2004 ; Tutin, 2004 ; Charest *et al.*, 2007), ou encore (Seretan, 2008) pour un compte-rendu détaillé.

1. Le terme *collocation* étant ambigu, nous tenons à préciser que nous adoptons ici son acception syntaxique (comme association de mots liés syntaxiquement), plus restrictive que son acception statistique qui est plus répandue (comme association de mots apparaissant à courte distance dans le texte). Parmi les traits principaux des collocations, nous mentionnons leur caractère récurrent, préfabriqué, arbitraire et imprédictible. Pour une discussion plus détaillée sur la définition des collocations, nous renvoyons le lecteur intéressé à (Seretan, 2008).

Beaucoup moins de travaux ont en revanche été réalisés sur le traitement ultérieur des résultats d'extraction, afin de rendre possible leur intégration dans des applications clés, telles que la traduction automatique, la génération de textes, la désambiguïsation lexicale, ou l'analyse syntaxique. Parmi les exceptions, on peut citer des essais de classification sémantique des collocations (Wanner *et al.*, 2006), de détection de synonymes pour les collocations (Wu et Zhou, 2003), de détection de traductions pour les collocations (Smadja *et al.*, 1996 ; Lü et Zhou, 2004), ou d'utilisation des collocations pour l'analyse syntaxique (Hindle et Rooth, 1993), pour la traduction (Smadja *et al.*, 1996 ; Lü et Zhou, 2004), ainsi que pour la génération (Heid et Raab, 1989) et la classification de textes (Williams, 2002). Il s'agit, en général, de travaux qui restent préliminaires et isolés, malgré la reconnaissance du rôle crucial que ces expressions jouent dans le traitement du langage (Sag *et al.*, 2002), et malgré le développement soutenu des techniques d'extraction.

Le travail décrit dans cet article vise à détecter des équivalents de traduction pour les collocations à partir de corpus de textes parallèles, afin de peupler le lexique d'un système de traduction multilingue. La méthode développée repose – tout comme ce système – sur une approche essentiellement syntaxique, rendue possible par le progrès réalisé dans le domaine de l'analyse syntaxique en général, et en particulier par le niveau de développement atteint par l'analyseur multilingue Fips, créé dans notre laboratoire (Wehrli, 2007). Les avantages de cette méthode par rapport aux travaux précédents sont, principalement, qu'elle est capable de prendre en charge des collocations plus flexibles du point de vue syntaxique (en plus des constructions plus rigides), et qu'elle est opérationnelle même pour des données de taille réduite, ou en l'absence de dictionnaires bilingues.

L'article est organisé de la manière suivante. La section 2 passe en revue les travaux existants sur l'extraction d'équivalents de traduction pour les collocations, y compris les travaux connexes concernant d'autres types d'expressions. La section 3 décrit brièvement Fips, l'analyseur syntaxique qui est à la base de notre méthode d'extraction. Ensuite, dans la section 4, nous présentons cette méthode, ainsi que les différents modules de traitement qu'elle utilise – notamment, le module d'extraction des collocations à partir de corpus et le module d'alignement de phrases. La section 5 présente des résultats expérimentaux accompagnés d'une évaluation de la performance de notre méthode. La section 6 se penche sur l'analyse des erreurs en discutant, également, les possibles améliorations de la méthode. La dernière section conclut l'article en comparant notre approche aux approches existantes.

2. Travaux précédents

Le travail de (Kupiec, 1993) peut être considéré comme l'un des premiers travaux sur l'extraction, à partir de corpus, d'équivalents de traduction pour les collocations. Ce travail était centré sur des groupes nominaux comme *late spring* (*la fin du printemps*). Des correspondances bilingues sont identifiées à partir du corpus parallèle français-anglais Hansard aligné au niveau de la phrase. Les deux corpus, source et

cible, sont étiquetés à l'aide d'un étiqueteur morphosyntaxique fondé sur un modèle markovien (HMM), et des groupes nominaux sont identifiés sur la base de la catégorie lexicale des mots en utilisant un reconnaisseur à états finis. Ensuite, des correspondances bilingues sont formées à l'aide d'un algorithme itératif de réestimation, Expectation Maximization (EM). La précision[2] de cette méthode, rapportée pour les 100 premières correspondances obtenues, est de 90 %.

Une méthode similaire d'identification de groupes nominaux a été employée par (van der Eijk, 1993) pour la paire de langues néerlandais-anglais. L'appariement est réalisé en employant deux heuristiques principales : le groupe nominal cible est choisi en fonction i) de sa fréquence dans le sous-ensemble de phrases cibles correspondant au groupe nominal source, et ii) de la corrélation entre sa position (dans la phrase cible) et la position du groupe nominal source (dans la phrase source). Évaluée sur 1 100 correspondances sélectionnées de manière aléatoire, la méthode a atteint une précision de 68 % et une couverture[3] de 64 %.

Dans le cadre du système Termight (Dagan et Church, 1994), on identifie également des équivalents bilingues pour les groupes nominaux à partir de corpus parallèles, mais en faisant appel à l'alignement des mots plutôt qu'à l'alignement des phrases. Une fois les correspondances des mots trouvées, ce système considère tout simplement la plage de mots entre les équivalents du premier et du dernier mot d'un syntagme nominal comme sa traduction potentielle. Les syntagmes sont déterminés auparavant en considérant des séquences de noms dans un corpus anglais étiqueté. Les différentes traductions potentielles obtenues pour un syntagme sont triées dans l'ordre des fréquences des têtes syntaxiques et sont affichées dans un concordancier bilingue. La précision obtenue pour les 192 correspondances anglaises-allemandes testées est de 40 %, mais les auteurs soulignent qu'en descendant dans la liste des résultats, une traduction correcte a toujours été trouvée.

Ces méthodes sont conçues pour un type d'expressions plutôt rigide, les groupes nominaux simples. En revanche, Champollion (Smadja *et al.*, 1996), le premier système de traduction des collocations proprement dit, est aussi capable de trouver un équivalent pour les collocations flexibles, qui comportent des verbes[4]. Les colloca-

2. La *précision* mesure le pourcentage de réponses correctes parmi les résultats d'un système.

3. Cette méthode est la seule pour laquelle on dispose de détails sur la *couverture* (c'est-à-dire, le pourcentage de données pour lesquelles une traduction a été proposée). L'auteur précise que celle-ci est affectée par le fait que l'équivalent d'un groupe nominal n'est pas toujours un groupe nominal. À noter que la couverture est différente du *rappel*, autre mesure standard utilisée conjointement à la précision dans la recherche documentaire, exprimant le pourcentage de réponses correctes fournies par un système parmi toutes les réponses correctes attendues. Le calcul du rappel de la traduction est plus difficile, car il présuppose l'existence d'un jeu de traductions de référence, ce qui n'est pas toujours le cas.

4. Ces collocations sont caractérisées par une grande permissivité syntaxique. Les mots qui les composent peuvent être inversés et séparés par plusieurs autres mots, comme par exemple *break* et *record* dans ***records*** *are made to be* ***broken*** (*les records sont faits pour être battus*) ; voir aussi les exemples fournis dans la section 3.

tions sont d'abord identifiées dans la version anglaise du Hansard à l'aide du système Xtract (Smadja, 1993). Pour trouver l'équivalent français, Champollion sélectionne les mots des phrases cibles qui sont le plus corrélés avec la collocation source. La corrélation est mesurée avec le coefficient statistique de corrélation Dice, qui prend en compte le nombre d'occurrences de deux termes simultanément dans une paire de phrases alignées, ainsi que le nombre d'occurrences total de chaque terme dans le corpus. Cette méthode nécessite une étape ultérieure qui consiste à parcourir les phrases cibles afin de déterminer l'ordre des mots de la traduction proposée, car le système ne dispose pas d'informations d'ordre syntaxique. L'évaluation effectuée sur deux jeux de test différents, contenant chacun 300 collocations, a montré une précision de 77 % dans le premier cas et de 61 % dans le second. Cette baisse de performance a été expliquée par les auteurs par la plus faible fréquence des collocations du second jeu de test.

Finalement, le travail de (Lü et Zhou, 2004) sur la paire de langues anglais-chinois prend aussi en charge les collocations flexibles, car ces dernières sont identifiées dans le corpus source et cible à l'aide d'un analyseur syntaxique. Trois configurations syntaxiques ont été considérées : verbe-objet, adjectif-nom, et adverbe-verbe. À la différence des méthodes précédentes, cette méthode peut s'appliquer aussi sur des corpus non parallèles. Elle utilise un modèle de traduction statistique dans lequel les probabilités de traduction des mots sont estimées à l'aide de l'algorithme EM. Les traductions initiales pour chaque mot sont assignées à partir de dictionnaires bilingues. La précision de cette méthode, mesurée sur un échantillon de 1 000 collocations, varie entre 51 % et 68 %, selon la configuration syntaxique.

La méthode que nous allons présenter est une méthode générique qui peut traiter un large éventail de configurations syntaxiques et plusieurs paires de langues. Relativement simple, elle est efficace même pour des collocations peu fréquentes, et même si des dictionnaires bilingues ne sont pas disponibles ; d'un autre côté, elle nécessite des outils d'analyse syntaxique pour les langues traitées, tels que l'analyseur multilingue Fips présenté dans la section suivante.

3. L'analyseur Fips

Fips (Laenzlinger et Wehrli, 1991 ; Wehrli, 1997 ; Wehrli, 2004 ; Wehrli, 2007) est un analyseur symbolique « profond » développé au LATL, le Laboratoire d'analyse et de technologie du langage de l'Université de Genève. Initialement conçu pour le français – l'acronyme Fips dérive de l'anglais *French Interactive Parsing System* –, il a été étendu successivement à l'anglais, l'italien, l'allemand, et plus récemment à l'espagnol et au grec[5].

Fips s'appuie sur une adaptation des concepts de grammaire générative inspirés par des théories chomskyennes (Chomsky, 1995 ; Haegeman, 1994). Le constituant

5. D'autres langues, parmi lesquelles le romanche, le roumain et le japonais, seront ajoutées par la suite dans le cadre d'un projet d'extension multilingue ultérieur.

syntaxique est représenté sous la forme d'une structure $\overline{X}$ simplifiée, [$_{XP}$ L X R], limitée à deux niveaux : XP – la projection maximale, et X – la tête de la projection. L et R dénotent des listes (éventuellement vides) de sous-constituants gauches et droits de X. X représente les catégories lexicales usuelles, N (nom), A (adjectif), V (verbe), Adv (adverbe), D (déterminant), P (préposition), Conj (conjonction), etc. Les entrées lexicales disponibles dans les lexiques créés manuellement contiennent des informations morphosyntaxiques détaillées, telles que des propriétés sélectionnelles, des informations de sous-catégorisation, et des traits syntactico-sémantiques susceptibles d'influencer l'algorithme d'analyse.

L'analyseur est implémenté dans le langage orienté objet Component Pascal[6]. Du point de vue de l'architecture, Fips est constitué d'un noyau générique qui définit des structures de données abstraites et des opérations s'appliquant à toutes les langues traitées, auquel s'ajoutent des modules de traitement spécifiques à chaque langue. Les opérations principales de l'analyseur sont :

Project, l'opération de projection, qui est associée à un élément lexical (tel qu'il apparaît dans le lexique) et qui crée un constituant syntaxique avec la tête lexicale correspondante (par exemple, pour un nom, on crée une projection NP) ;

Merge, l'opération de combinaison des constituants, qui permet l'ajout d'un nouveau constituant à une structure existante par attachement à gauche ou à droite dans un nœud actif (le site d'attachement) ; les attachements sont contraints par des conditions définies dans les règles de grammaire spécifiques à chaque langue ;

Move, l'opération de déplacement, qui établit un lien entre un élément extraposé et un constituant abstrait dans une position canonique[7].

La stratégie d'analyse est fondée sur un algorithme essentiellement ascendant, de type gauche à droite, dirigé par les données. À chaque pas, l'analyseur applique une des trois opérations mentionnées ci-dessus. Les alternatives sont traitées en parallèle, et un filtre descendant est utilisé conjointement avec des heuristiques d'élagage afin de limiter l'espace de recherche.

La sortie fournie par l'analyseur pour une phrase d'entrée consiste en une structure riche, englobant, outre la structure des constituants, i) l'interprétation des constituants comme arguments (représentée sous la forme d'une table d'arguments similaire à la f-structure de la LFG[8]), ii) l'interprétation des clitiques, pronoms interrogatifs et re-

6. L'environnement de développement utilisé est BlackBox Component Builder (http://www.oberon.ch/blackbox.html).

7. Selon la théorie chomskyenne, les éléments extraposés sont des éléments déplacés par une transformation de mouvement à partir d'une position dite *canonique*, gouvernée par un prédicat. L'identification de ce mouvement permet de déterminer le rôle thématique de l'élément extraposé.

8. Dans la théorie LFG, *Lexical functional grammar* - « Grammaires lexicales fonctionnelles » (Bresnan, 2001), la f-structure sert à la représentation des fonctions grammaticales.

latifs, et iii) les chaînes de co-indexation mettant en relation les éléments extraposés avec leur position canonique. Par exemple, pour la phrase donnée en (1a) Fips propose l'analyse en (1b), où l'index *i* montre la chaîne connectant le nom *records* au constituant vide e qui suit la forme verbale *broken*. Dans la table d'arguments, on trouvera ce constituant vide dans la position de complément d'objet direct du prédicat *to break*. À travers l'index i, il est ensuite possible d'identifier la relation verbe-objet entre *broken* et *records*.

(1) a. Records are made to be broken.
b. $[_{TP}[_{NP}Records]_i are[_{VP}made[_{NP}e]_i[_{TP}[_{DP}e]_i to[_{VP}be[_{VP}broken[_{DP}e]_i]]]]]$

Très robuste, l'analyseur Fips peut traiter des grands corpus de texte sans restrictions, en un laps de temps acceptable (environ 150-200 symboles sont traités par seconde). Sa large couverture grammaticale permet de traiter des phénomènes aussi complexes que ceux exemplifiés ci-dessous[9] :

(2) a. Interrogation : *Quel **objectif** spécifique le projet doit-il **atteindre** comme contribution aux objectifs globaux ?*
b. Passivation : *Premièrement, parce que l'**objectif** du système des écopoints a pratiquement été **atteint**.*
c. Relativisation : *Face aux nombreux **objectifs** que ce programme doit **atteindre**,* [...]
d. Topicalisation : *Ce sont là des **objectifs** tout à fait essentiels qui ne peuvent être **atteints** qu'au terme d'efforts communs.*

4. Détection d'équivalents

L'application de détection d'équivalents bilingues pour les collocations est construite comme l'extension d'un système plus complexe d'aide à la traduction que nous avons développé pendant les dernières années (Seretan *et al.*, 2004 ; Seretan, 2008), et qui intègre les modules principaux suivants :

1) un extracteur hybride de collocations, alliant aux calculs statistiques d'association lexicale des informations de nature syntaxique fournies par l'analyseur Fips ;

2) un concordancier monolingue/bilingue, qui visualise les résultats d'extraction en présentant la phrase d'origine et, simultanément, la phrase alignée si des corpus parallèles sont disponibles ;

3) un module (sous-jacent) d'alignement, qui met en correspondance la phrase d'origine avec la phrase qui représente sa traduction dans le document parallèle (la phrase alignée) ;

4) un module de validation des résultats, qui permet la création, par l'utilisateur, d'une base de données monolingue/bilingue de collocations, pouvant servir comme référence pour les traductions futures.

9. Les phrases discutées dans cet article sont toutes attestées, et proviennent des corpus utilisés.

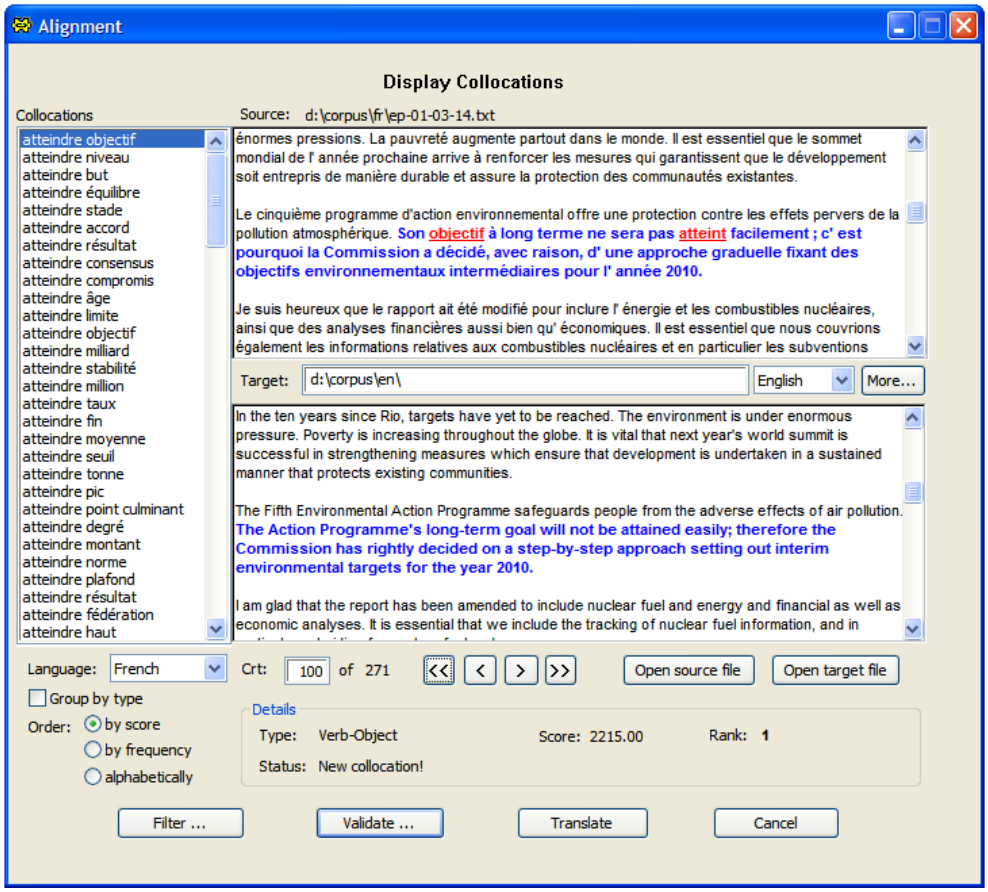

Figure 1. *L'interface du concordancier bilingue pour les collocations*

Les figures 1 et 2 montrent l'interface du concordancier bilingue et celle du module de validation de notre système. Dans la première, nous avons affiché (dans la liste à gauche) des constructions de type verbe-objet avec le verbe *atteindre*, que nous avons triées par leur score statistique d'association. La zone de texte en haut à droite présente, pour la première collocation obtenue, *atteindre - objectif*, l'occurrence numéro 100 dans le corpus source (sur 271 identifiées au total). Le système fait défiler automatiquement le texte dans le document d'origine jusqu'à l'occurrence recherchée et sa phrase source, les deux étant mises en évidence par des couleurs contrastées. La zone de texte en bas présente de manière similaire la phrase correspondante dans le document parallèle en anglais. La deuxième interface (figure 2) montre que cette collocation a été choisie pour la validation, et que l'utilisateur a introduit une traduction vers l'anglais, *attain - goal*, dans le champ correspondant.

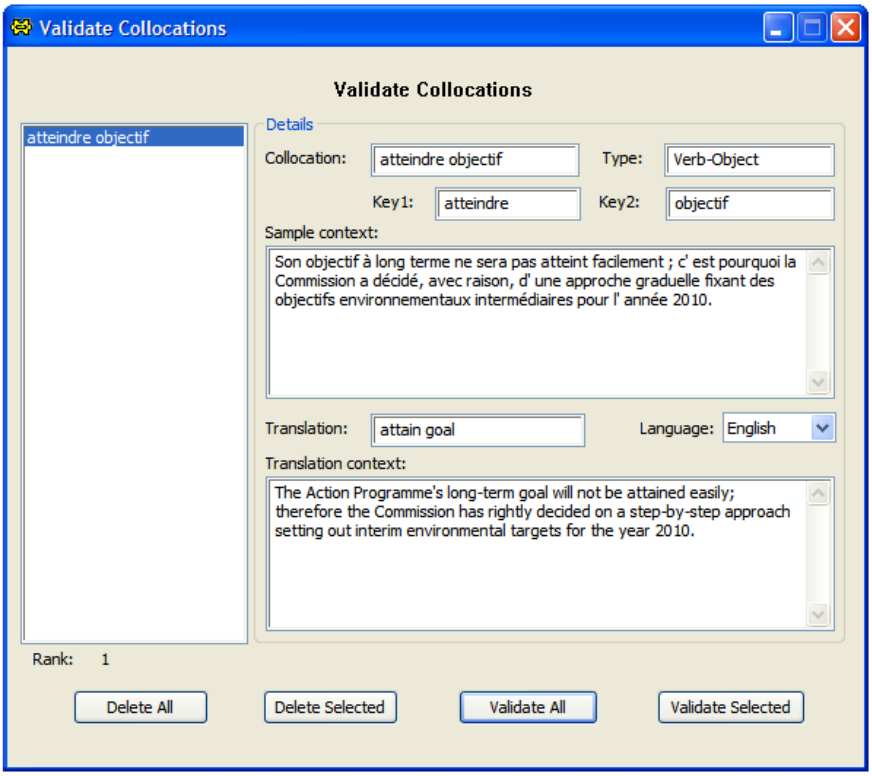

Figure 2. *L'interface du module de validation de collocations*

Comme notre exemple le montre, à l'aide de notre système l'utilisateur peut repérer la traduction d'une collocation en consultant les phrases alignées correspondant aux différentes occurrences de cette collocation, et ensuite stocker cette traduction dans la base de données bilingue, avec des contextes qui constituent des exemples d'utilisation. Le but de notre méthode est d'automatiser ce travail, et d'identifier automatiquement la traduction d'une collocation en utilisant la technologie dont on dispose.

Ainsi, la méthode conçue consiste en la constitution, pour chaque collocation source, d'un mini-corpus de phrases alignées, et en son analyse afin d'y détecter, à l'aide de certaines heuristiques, une traduction potentielle pour cette collocation. La procédure de base a été décrite dans (Seretan et Wehrli, 2007). Dans le présent article, nous décrivons une version améliorée et étendue de cette procédure de base, et nous présentons des nouveaux résultats d'évaluation. Aussi, nous fournissons une description plus détaillée des outils sur lesquels cette méthode repose.

4.1. *L'extracteur de collocations*

À la différence de la plupart des extracteurs de collocations existants (certains d'entre eux ont été mentionnés dans la section 1), notre système d'extraction fait appel à une analyse syntaxique complète du texte source, comme étape préliminaire à la procédure de calculs statistiques qui, en mesurant le degré d'association entre les mots, identifie des collocations potentielles. D'autres systèmes pouvant aussi être qualifiés d'hybrides s'appuient sur une analyse moins détaillée, souvent fondée sur l'étiquetage des catégories lexicales et la reconnaissance de certaines relations syntaxiques en utilisant des expressions régulières sur les étiquettes assignées aux mots.

Dans notre système, l'analyse syntaxique fournie par Fips (*cf.* section 3) permet un traitement plus uniforme et fiable de ces expressions. Premièrement, les paires de mots qui constituent des collocations candidates sont identifiées à partir de la structure normalisée associée à une phrase par l'analyseur : indifféremment de la manière dont une relation syntaxique est réalisée dans le texte, le système considère toujours l'ordre canonique des mots et la forme de base à la place de la forme fléchie, ceci permettant un traitement plus général et plus homogène des données. Deuxièmement, grâce à une analyse plus complète qui permet l'interprétation dans un contexte plus large et la désambiguïsation au niveau des catégories lexicales (ainsi qu'au niveau des lectures pour une même catégorie), l'analyse syntaxique profonde contribue à améliorer la précision des résultats. Aussi, le regroupement des variantes morphosyntaxiques d'une collocation sous le même type permet l'application plus fiable des mesures d'association lexicale ; il est connu que ces mesures ont un comportement insatisfaisant pour les candidates dont le nombre d'occurrences dans le corpus est très petit (moins de 5, selon (Evert, 2004)), et que la plupart des candidates dans un corpus n'apparaissent que très peu de fois. En regroupant, en revanche, les variantes dispersées d'un même candidat, on réduit la fragmentation des données (surtout pour les langues ayant une morphologie riche), ce qui permet d'obtenir des résultats statistiques plus fiables.

La procédure d'extraction peut être décrite, dans les grandes lignes, comme suit : au fur et à mesure que le corpus source est analysé, le module d'extraction parcourt récursivement les structures obtenues[10] afin d'y repérer des collocations candidates. La sélection d'une paire obéit au critère principal suivant : l'un des éléments de la paire, celui qui représente la tête lexicale X de la structure courante [$_{XP}$ L X R], peut former une paire avec la tête lexicale d'un des sous-constituants droit ou gauche (de L ou de R). Ce critère assure la présence d'une relation syntaxique entre les deux éléments de la paire, c'est-à-dire leur proximité syntaxique (en opposition à la proximité linéaire utilisée par les approches traditionnelles). Pour les verbes, les relations de type prédicat-argument sont récupérées directement à partir de la table d'arguments. Ainsi, les paires *break - record* et *atteindre - objectif* sont facilement identifiées à partir de

10. Lorsqu'une phrase n'est pas complètement analysée (c'est-à-dire que l'analyseur n'aboutit pas à une analyse globale de la phrase, mais seulement à des analyses partielles de ses sous-parties), plusieurs structures sont fournies, correspondant aux morceaux analysés.

phrases comme celles présentées en (1) et (2), puisque tout le calcul nécessaire pour identifier le lien verbe-objet est déjà effectué par l'analyseur.

À ce critère principal s'ajoutent des contraintes plus spécifiques sur la paire candidate, qui s'appliquent soit à la configuration syntaxique de la paire, soit à chacun de ses éléments individuellement. Plus précisément, une paire est retenue seulement si elle est dans une configuration prédéfinie, telles que celles montrées dans le tableau 1, et si elle ne contient pas comme élément un verbe modal ou un nom propre. À la différence de beaucoup d'autres systèmes existants, notre système n'applique pas de filtre fondé sur les *stop-words*, ni sur la fréquence des paires candidates ; un tel filtre peut être appliqué ultérieurement par l'utilisateur en fonction de ses propres exigences.

Configuration	**Abréviation**	**Exemple**
Adjectif-nom	A-N	haute technologie
Nom-adjectif	N-A	voie ferrée
Nom-[prédicat-]adjectif	N-Pred-A	livre [semble] intéressant
Nom(tête)-nom	N(tête)-N	bouc émissaire
Nom-préposition-nom	N-P-N	danger de mort
Nom-préposition-verbe	N-P-V	machine à laver
Nom-préposition	N-P	précaution quant
Adjectif-préposition-nom	A-P-N	fou de rage
Adjectif-préposition	A-P	tributaire de
Sujet-verbe	S-V	incendie se déclarer
Verbe-objet	V-O	présenter risque
Verbe-préposition-argument	V-P-N	répondre à besoin
Verbe-préposition	V-P	centrer sur
Verbe-adverbe	V-Adv	refuser catégoriquement
Verbe-adjectif	V-A	sonner creux
Verbe-verbe	V-V	faire suivre
Adverbe-adjectif	Adv-A	grièvement blessé
Adverbe-adverbe	Adv-Adv	très bien
Déterminant-nom	D-N	ce matin
Préposition-nom	P-N	sur mesure
Préposition-nom-adjectif	P-N-A	à titre indicatif
Adjectif-*et*-adjectif	A&A	bête et méchant
Nom-*et*-nom	N&N	frères et sœurs

Tableau 1. *Configurations syntaxiques permises pour les paires candidates*

La seconde étape du processus d'extraction consiste en l'application des mesures d'association lexicale afin d'estimer le degré d'affinité entre les éléments des paires candidates. Le résultat d'extraction est représenté par la liste de paires candidates triées par ordre décroissant du score d'association obtenu. Les paires situées en haut de la liste sont les plus susceptibles, théoriquement, de constituer de vraies collocations et de présenter un intérêt lexicographique. Seulement un nombre limité de paires peuvent être examinées en pratique par les utilisateurs, d'où les efforts pour perfectionner les

mesures d'association afin de placer les paires intéressantes au sommet de la liste et celles moins intéressantes vers la fin.

Notre système implémente une douzaine de mesures décrites dans la littérature récente sur l'extraction de collocations. Parmi celles-ci, le système propose par défaut la mesure *log-lihelihood ratio* (LLR), ou le rapport de vraisemblance (Dunning, 1993). Son choix est motivé par le fait qu'elle a reçu des évaluations positives de la part de beaucoup d'autres auteurs, et qu'elle est jugée comme particulièrement appropriée aux données les moins fréquentes.

Avant d'être soumises au calcul du score, les paires candidates sont divisées en des ensembles syntaxiquement homogènes sur la base de leur configuration syntaxique (*cf.* tableau 1) ; cette stratégie est censée avoir des effets bénéfiques sur la performance des mesures d'association (Heid, 1994). Un autre trait distinctif de notre système d'extraction est qu'il est capable de reconnaître, grâce aux informations fournies par l'analyseur syntaxique, si l'un des termes d'une paire candidate fait partie d'une autre collocation, et ensuite de traiter ce terme complexe comme une seule entité. En conséquence, le système peut retourner à la place des collocations binaires des collocations plus longues, comme on peut le remarquer dans la figure 1 dans le cas de la construction *atteindre point culminant*.

Dans notre approche, la première étape (la sélection des paires candidates) reste l'étape clé du processus d'extraction, car c'est en premier lieu de la qualité des paires proposées que la qualité des résultats dépend. La supériorité des approches d'extraction hybride fondées sur l'analyse syntaxique complète, longtemps postulée par la théorie[11] mais souvent questionnée à cause des échecs et des erreurs inhérentes d'analyse, a été démontrée dans le cas de notre extracteur par des études d'évaluation que nous avons menées pour plusieurs langues. Nos expériences sur des corpus en français, anglais, italien et espagnol (Seretan, 2008 ; Seretan et Wehrli, 2009) ont démontré que, en effet, l'analyse syntaxique conduit à une réduction substantielle des faux positifs agrammaticaux parmi les résultats d'extraction, par rapport à une méthode standard d'extraction fondée sur l'étiquetage morphosyntaxique et sur le critère de proximité linéaire : 99 % des 500 meilleurs résultats de la méthode syntaxique sont valides, contre 76,4 % de la méthode standard ; si l'on considère des résultats situés à différents niveaux[12] dans la liste complète des résultats, on obtient un pourcentage moyen de 88,8 % pour la première méthode, et de seulement 33,2 % pour la seconde. En outre, nos expériences ont montré que le calcul statistique est lui aussi allégé, puisque le filtre syntaxique réduit considérablement la taille des données candidates.

11. Beaucoup de chercheurs avaient indiqué que, idéalement, l'identification des collocations devrait tenir compte de l'information syntaxique fournie par les analyseurs, surtout vu le progrès réalisé dans le domaine de l'analyse syntaxique (Smadja, 1993 ; Krenn, 2000 ; Pearce, 2002 ; Evert, 2004).

12. Un nombre de 50 résultats adjacents on été testés au début (0 %) et ensuite à 1 %, 3 %, 5 % et 10 % des listes retournées par les deux méthodes d'extraction.

4.2. *Alignement de phrases*

Afin de détecter la traduction d'une phrase source dans le document cible, on utilise notre propre méthode d'alignement intégrée au système d'aide à la traduction, qui a été décrite dans (Nerima *et al.*, 2003). La spécificité de cette méthode consiste dans le calcul *à la volée* d'un appariement partiel, seulement pour la phrase couramment visualisée par l'utilisateur dans le concordancier. Très rapide, cette méthode n'a pas besoin d'un traitement préalable des documents parallèles disponibles ; notamment, elle ne présuppose pas un alignement macro-structurel (par exemple, au niveau des sections ou paragraphes), comme le font d'autres méthodes d'alignement de phrases.

Étant donné le document source et la position d'occurrence d'une collocation dans ce document, la méthode identifie les positions dans le document cible qui délimitent la traduction de la phrase d'origine. Le calcul d'appariement est fait principalement selon le critère de conservation des longueurs relatives des paragraphes dans les documents source et cible, exprimées en nombre de caractères. Plus les proportions des longueurs dans le voisinage d'une phrase candidate ressemblent aux proportions des longueurs dans le voisinage de la phrase source, plus cette phrase est considérée comme une candidate valide. La méthode fait également appel à l'analyse du contenu des documents source et cible, mais uniquement pour vérifier la compatibilité de la numérotation, s'il y en a une[13]. La précision de cette méthode est d'environ 90 % pour des textes relativement difficiles[14] à aligner, qui peuvent contenir non seulement du texte mais aussi des tableaux (en format HTML) et qui comportent des paragraphes structurés de manière différente à travers les langues, ainsi que des paragraphes manquants[15].

4.3. *La méthode de base de détection d'équivalents*

Les prérequis de notre méthode de traduction de collocations sont, d'un côté, la disponibilité d'un corpus parallèle, et de l'autre, la disponibilité d'un analyseur syntaxique pour la langue cible (bien que dans notre cas les collocations sources soient aussi extraites à l'aide de l'analyse syntaxique, du point de vue de la méthode il n'est pas nécessaire de disposer d'un analyseur pour la langue source).

Étant donnée une collocation source CS et ses occurrences dans le corpus source, notre méthode essaie de trouver une traduction adéquate pour CS dans le corpus cible en appliquant la stratégie suivante :

13. D'autres méthodes s'appuient plus sur la comparaison de contenu, par exemple, la méthode de (Simard *et al.*, 1992) fondée sur l'identification de mots semblables ou *cognate words*.

14. La précision obtenue par les meilleurs systèmes d'alignement est proche de 100 % pour les textes « normaux » qui sont identiques au niveau structurel, mais elle peut descendre jusqu'à environ 65 % pour les textes difficiles (Véronis et Langlais, 2000).

15. Dans le futur, notre système sera étendu afin de permettre l'utilisation d'autres méthodes d'alignement existantes, et de prendre en charge des corpus parallèles préalignés au niveau de la phrase.

1) dans une première étape, on constitue un mini-corpus de contextes cibles associés à CS en utilisant la méthode d'alignement de phrases décrite dans la section 4.2 ; pour chaque occurrence de CS, on met ainsi en correspondance la phrase source avec la phrase équivalente dans le corpus cible ;

2) dans la seconde étape, on utilise ce mini-corpus comme l'espace de recherche de CC, la collocation cible.

La méthode de base d'identification de CC présentée dans (Seretan et Wehrli, 2007) était fondée sur les trois principes suivants :

P1) CC conserve la configuration syntaxique de CS : en conformité avec des travaux précédents – par exemple, (Lü et Zhou, 2004) –, on assumait que, avec peu d'exceptions, une collocation peut être traduite par une expression du même type (ainsi, une paire verbe-objet et traduite par une paire verbe-objet, etc.) ;

P2) la fréquence des paires ayant la même configuration syntaxique que CS dans le mini-corpus cible est un bon indicateur de la traduction de CS : vu la manière dont le mini-corpus a été construit, la paire la plus fréquente est très probablement la traduction de CS ;

P3) les collocations sont partiellement compositionnelles et leur base (l'élément dont le sens est conservé dans le sens de la collocation) peut être traduite littéralement, tandis que le choix du collocatif reste arbitraire[16]. Par exemple, la base de la collocation *battre un record* (le nom *record*) est traduite littéralement vers l'anglais, à la différence du collocatif verbal *battre*, dont l'équivalent est *to break* (lit. *casser*) ; le collocatif *mauvaise* de la collocation *mauvaise décision* est traduit en allemand comme *falsche* (lit. *fausse*), etc.

Ces principes étaient utilisés pour limiter successivement l'espace de recherche de CC, afin d'aboutir à une seule expression candidate à la traduction. Le mini-corpus cible est d'abord analysé avec Fips, et l'extracteur de collocations (décrit dans la section 4.1) identifie, dans la sortie de l'analyseur, des cooccurrences syntaxiques dans toutes les configurations appropriées aux collocations (*cf.* tableau 1). Ces cooccurrences constituent l'espace initial de recherche, qui est ensuite réduit en retenant seulement les cooccurrences du même type que le type de CS. Si des dictionnaires bilingues sont disponibles pour la paire de langues en cause, on restreint cet espace aux cooccurrences dont la base (par exemple, le nom dans une combinaison verbe-objet) est une des traductions trouvées dans le dictionnaire pour la base de CS. Finalement, la décision finale qui permet de choisir CC est de considérer la paire la plus fréquente parmi les paires qui restent. La réponse de la méthode est unique ; en cas de fréquences identiques, le système ne fournit aucune traduction (des résultats expérimentaux avaient indiqué que le système perdait en précision lorsqu'il proposait

16. Ce principe est compatible avec les stipulations théoriques considérant les collocations comme des combinaisons polaires formées d'une base *autosémantique* et d'un collocatif *synsémantique* choisi en fonction de la base (Hausmann, 1989 ; Mel'čuk, 1998 ; Polguère, 2000). La base d'une collocation peut être déterminée de manière univoque en fonction de sa configuration syntaxique.

des traductions multiples, même s'il réussissait ainsi à proposer plusieurs traductions synonymiques valides)[17].

Cette méthode de détection d'équivalents pour les collocations a été appliquée à des paires de type verbe-objet extraites du corpus parallèle Europarl (Koehn, 2005). L'extraction a été effectuée sur un sous-corpus d'environ 4 millions de mots en moyenne, pour 4 langues : le français, l'anglais, l'italien et l'espagnol. La méthode de traduction a été appliquée pour la totalité des 12 paires de langues possibles. Son évaluation effectuée sur 4 000 collocations (500 collocations[18] pour 8 paires de langues) a montré des résultats prometteurs : en moyenne pour les paires de langues évaluées, la précision obtenue était de 89,8 % et la couverture de 70,9 % (Seretan et Wehrli, 2007). Il a été aussi montré que la précision n'est que légèrement affectée par la diminution de la fréquence des collocations (elle varie entre 91,8 % pour les collocations avec au moins 30 occurrences et 84,6 % pour celles ayant une fréquence entre 1 et 15), alors que la couverture descend plus vite avec la fréquence (de 79,3 % à 53 %).

4.4. *Extensions*

Dans cette section, nous décrivons les améliorations que nous avons apportées à la méthode de base, afin d'étendre son champ d'application à d'autres configurations syntaxiques (autres que celle déjà traitée, verbe-objet) ; de permettre des divergences syntaxiques entre les collocations sources et cibles ; et d'augmenter sa performance, notamment en ce qui concerne la couverture (qui était plus déficitaire).

Par souci de simplicité, on appelle *méthode A* la méthode de base d'obtention d'équivalents, décrite dans la section précédente (section 4.3). Puisque le principe P1 qu'elle utilise est trop contraignant, afin d'augmenter les chances qu'une collocation source puisse être traduite, la première extension prévue consiste à définir des correspondances entre des configurations syntaxiques sources et cibles, en fonction des langues traitées. Ainsi, pour la paire de langues français-anglais on spécifie, par exemple, qu'une collocation verbe-objet peut être traduite non seulement comme verbe-objet, mais aussi comme verbe-préposition-argument, et *vice versa* : *relever défi - respond to challenge*, *profiter de occasion - take opportunity*. De la même manière, on spécifie la liste de configurations cibles acceptées pour la configuration source adjectif-nom quand on traduit de l'anglais vers le français : adjectif-nom (*serious problem - grave problème*), nom-adjectif (*humanitarian aid - aide humanitaire*), etc.

Les correspondances syntaxiques spécifiées sont considérées dans le système selon une échelle de préférences préétablie. Des poids[19] différents sont associés à chaque

17. Cependant, le choix du numéro de réponses désirées pourrait se faire en fonction du type d'application considérée : souvent (par exemple, dans un contexte lexicographique), on décide de favoriser le rappel au détriment de la précision.

18. Il s'agit, plus précisément, des 500 meilleures paires extraites, selon le score LLR.

19. Actuellement, ces poids privilégient la ressemblance avec la configuration syntaxique source, mais il est aussi envisageable que leur détermination soit faite expérimentalement.

configuration syntaxique cible, prenant des valeurs numériques entre 0 et 1. Une fois que le système a appliqué le filtre syntaxique correspondant, ces poids sont pris en compte conjointement à la fréquence des collocations candidates restantes pour sélectionner la collocation cible (CC) : l'espace de recherche est restreint aux paires candidates c pour lesquelles l'expression $\mathit{fréquence(c)} \times \mathit{poids(configuration(c))}$ atteint sa valeur maximale.

La deuxième extension de la méthode consiste en l'introduction d'un critère supplémentaire qui s'applique pour départager les traductions candidates dans le cas où la valeur maximale est atteinte par plusieurs d'entre elles. La solution qui a été adoptée est de calculer le score LLR d'association lexicale pour toutes les paires cibles identifiées, et de départager les candidates à égalité selon ce score. Cette variante étendue de la méthode A est appelée *méthode B*. Comme on le verra dans la section suivante, les deux extensions considérées ont conduit à une augmentation considérable de la proportion des collocations pour lesquelles une traduction a pu être proposée.

Une extension ultérieure a ensuite été réalisée, en partant de l'observation que le filtre lexical – conformément au principe P3 – peut avoir une influence négative sur la performance de notre méthode, si la couverture du dictionnaire bilingue est insatisfaisante pour les entrées consultées. En effet, l'analyse des résultats de la méthode A avait mis le doigt sur des situations où le manque d'alternatives supplémentaires de traduction pour la base d'une collocation conduisait à l'échec de la traduction de celle-ci. Par exemple, les traductions trouvées dans le dictionnaire français-anglais pour le mot *remarque* (*remark*, *comment*, et *note*) étaient insuffisantes pour pouvoir traduire la collocation *faire une remarque*, car le corpus proposait souvent *make a point* comme équivalent, et le mot *point* n'apparaissait pas dans le dictionnaire comme traduction de *remarque*. Quoique généralement valable, le principe P3 devrait être relâché si on admet que certaines collocations sont relativement moins transparentes et que leur base peut ne pas se traduire littéralement.

Une nouvelle version de la méthode B a été ainsi développée (appelée *méthode C*), qui prévoit que si la traduction échoue à cause du filtre fondé sur le dictionnaire, le système devrait alors renoncer à ce filtre et appliquer seulement les autres critères de sélection. En suivant cette stratégie, la couverture du système est encore augmentée jusqu'à 94,2 %, ce qui représente, par rapport à la méthode de base, une hausse de presque 25 %. L'impact de ces extensions sur la qualité des résultats obtenus fait l'objet de l'étude d'évaluation présentée dans la section suivante.

5. Résultats et évaluation

Afin de faciliter la comparaison de la performance obtenue par les nouvelles versions de notre méthode (méthodes B et C décrites dans la section 4.4), nous avons utilisé le même cadre expérimental que celui de la méthode de base (méthode A, *cf.* section 4.3). La méthode de traduction a été appliquée sur les 500 premières collocations extraites automatiquement (sans aucune validation manuelle préalable, le pro-

cessus étant entièrement automatique). Le tableau 2 présente quelques équivalences de traduction obtenues pour des collocations de type verbe-objet, adjectif-nom, et verbe-préposition-argument pour la paire de langues français-anglais.

Adjectif-Nom	**Verbe-Objet**	**Verbe-Préposition-Argument**
courte majorité/narrow majority	accuser retard/experience delay	aboutir à conclusion/come to conclusion
étroite collaboration/close cooperation	apporter aide/give aid	acquitter de mission/fulfil duty
faible densité/low density	attirer attention/draw attention	arriver à conclusion/reach conclusion
ferme conviction/strong belief	avoir sens/make sense	assister à réunion/attend meeting
forte augmentation/substantial increase	commettre erreur/make mistake	attaquer à problème/address problem
forte concentration/high concentration	déployer effort/make effort	entrer dans détail/go into detail
forte pression/heavy pressure	effectuer visite/pay visit	entrer en contact/come into contact
grande attention/great attention	entamer dialogue/start dialogue	figurer à ordre du jour/be on agenda
grande diversité/wide range	exercer influence/have influence	insister sur importance/stress importance
grande vitesse/high speed	ménager effort/spare effort	parvenir à compromis/reach compromise
grave erreur/bad error	ouvrir voie/pave way	parvenir à résultat/achieve result
grossière erreur/great mistake	prononcer discours/make speech	répondre à exigence/meet requirement
jeune âge/early age	tirer leçon/learn lesson	traduire en justice/bring to justice

Tableau 2. *Exemples d'équivalences extraites (français-anglais)*

Ces équivalences ont été choisies pour illustrer l'idiomaticité d'encodage dans la langue source et en particulier le choix du collocatif en fonction du mot avec lequel il se combine. L'adjectif *forte*, par exemple, est traduit de trois manières différentes (*substantial*, *high*, *heavy*), selon le nom qu'il modifie. De plus, aucune des trois alternatives ne correspond à sa traduction littérale, *strong*. De manière générale, les traductions littérales des collocations sont soit moins utilisées que leurs équivalents corrects (par exemple, *strong increase* contre *substantial increase*), soit complètement inadéquates et dans ce cas elles pourraient être qualifiées d'anticollocations[20] (par exemple, **short majority* contre *narrow majority*).

Le tableau 3 montre quelques équivalences qui ne conservent pas la structure syntaxique d'une langue à l'autre, et dont l'extraction a nécessité la prise en charge de correspondances structurelles non isomorphes entre les langues.

Anglais (V-O) → Français (V-P-N)		**Anglais (V-P-N) → Français (V-O)**	
answer question	répondre à question	adhere to deadline	respecter délai
attend meeting	assister à réunion	bring to attention	attirer attention
have access	bénéficier de accès	bring to end	mettre terme
lose sight	perdre de vue	come to decision	prendre décision
meet demand	satisfaire à exigence	comply with legislation	respecter législation
meet need	répondre à besoin	lead to improvement	entraîner amélioration
reach compromise	parvenir à compromis	meet with resistance	rencontrer résistance
reach conclusion	aboutir à conclusion	provide with opportunity	fournir occasion
reach consensus	aboutir à consensus	respond to challenge	relever défi
stress need	insister sur nécessité	touch on point	aborder point

Tableau 3. *Exemples d'équivalences ne conservant pas la configuration syntaxique*

20. Ce terme, introduit par (Pearce, 2001), désigne les variantes paradigmatiques des collocations perçues comme inadéquates par les locuteurs natifs d'une langue.

L'évaluation des nouvelles méthodes proposées a été effectuée pour la paire de langues anglais-français, sur les mêmes données de test qui ont servi pour l'évaluation de la méthode A, c'est-à-dire sur les 500 premières collocations sources de type verbe-objet extraites. Les équivalences proposées par les méthodes B et C ont été classées en deux catégories, chaque équivalence étant considérée soit comme correcte, soit comme incorrecte. L'annotation des équivalences a été faite à l'aide du concordancier bilingue (décrit dans la section 4), ce qui a permis leur interprétation en contexte et rendu plus facile le processus d'évaluation[21].

Un large pourcentage des résultats fournis par les deux nouvelles méthodes sont identiques aux résultats de la méthode de base (environ 60 %). La comparaison des résultats distincts pour les trois méthodes a permis l'identification d'équivalents de traduction synonymiques, à partir de situations où plusieurs traductions ont été jugées correctes pour une même collocation source. Quelques exemples sont présentés dans le tableau 4.

Collocation source	**Équivalent 1**	**Équivalent 2**
achieve consensus	parvenir à consensus	établir consensus
bridge gap	combler fossé	combler lacune
do job	accomplir travail	faire travail
draw distinction	établir distinction	faire distinction
make effort	faire effort	déployer effort
meet need	répondre à besoin	satisfaire besoin
miss opportunity	rater occasion	perdre occasion
provide aid	fournir assistance	apporter soutien
raise issue	aborder problème	soulever question
reach compromise	parvenir à compromis	trouver compromis
reap benefit	retirer avantage	récolter fruit
submit proposal	présenter proposition	soumettre proposition

Tableau 4. *Quelques équivalents synonymiques trouvés en comparant les méthodes*

La performance des méthodes est mesurée en tenant compte de leur capacité à fournir une traduction correcte pour les collocations sources testées. Si *n* est le nombre de collocations sources (dans notre cas, $n = 500$), p le nombre d'équivalences proposées, et c le nombre d'équivalences correctes proposées, les notions standard de *précision* et *couverture* sont définies comme suit :

$$P = \frac{c}{p} \; ; \; C = \frac{p}{n} \qquad [1]$$

Si l'on considère que la tâche à réaliser consiste à trouver *une* traduction possible pour chaque collocation source, on peut alors définir le *rappel* comme le pourcentage d'équivalences correctes retournées parmi toutes les équivalences correctes at-

21. Compte tenu de la relative objectivité de la tâche et de la possibilité de vérifier les traductions dans le corpus aligné, nous avons effectué nous-mêmes l'annotation des équivalences proposées.

tendues (n), et on peut employer la F-mesure[22] comme mesure globale de la performance :

$$R = \frac{c}{n} \; ; \; F = \frac{2PR}{P+R} \qquad [2]$$

	Méthode A	**Méthode B**	**Méthode C**
c (vrais positifs)	337	367	396
p (résultats proposés)	357	402	471
C (couverture)	71,4 %	80,4 %	94,2 %
P (précision)	94,4 %	91,3 %	84,1 %
R (rappel)	67,4 %	73,4 %	79,2 %
F (F-mesure)	78,6 %	81,4 %	81,6 %

Tableau 5. *Performance des trois méthodes sur des données de type verbe-objet*

Les résultats comparatifs de l'évaluation sont présentés dans le tableau 5. Ils montrent que l'objectif principal d'amélioration de la méthode de base (méthode A) a été atteint, la couverture du système augmentant graduellement de 71,4 % à 80,4 % jusqu'à 94,2 %. Cette importante augmentation de la couverture (de 22,8 % par rapport à la valeur initiale) engendre une diminution de la précision du système, d'abord d'environ 3 %, ensuite d'environ 10 % ; toutefois, en balance, la performance globale est maintenue (la F-mesure croît même légèrement, de 78,6 % à 81,6 %), car le rappel connaît une amélioration substantielle, de 11,8 % (de 67,4 % à 79,2 %).

La différence entre la nouvelle méthode (B) et sa variante plus évoluée (la méthode C) consiste en l'utilisation de techniques pour contourner les éventuelles lacunes des dictionnaires bilingues. De manière plus générale, nous nous sommes intéressés à l'impact de l'utilisation d'un dictionnaire sur la performance de notre méthode. Une évaluation secondaire a ainsi été menée, afin de comparer la performance obtenue par la méthode B en la présence et en l'absence d'un dictionnaire bilingue. Les résultats sont présentés dans le tableau 6.

Comme on peut l'observer, la présence du dictionnaire contribue à l'augmentation de la précision mais, en même temps, à la diminution de la couverture ; globalement, on obtient une légère hausse du rappel et de la F-mesure. Notre précédente comparaison (Seretan et Wehrli, 2007) entre la performance obtenue pour les paires de langues pour lesquelles on disposait de dictionnaires bilingues et celles pour lesquelles on n'en disposait pas avait trouvé des écarts assez semblables de précision (92,9 % contre 84,5 %), et des différences mineures de couverture (71,7 % contre 69,5 %). Sur la base de ces résultats, on peut conclure que l'absence de dictionnaires (monolexèmes) bilingues, même si indésirable, n'entraîne qu'une perte limitée en performance.

22. Dans la recherche documentaire, la F-mesure englobe dans un seul score les valeurs de précision et rappel, en considérant leur moyenne harmonique.

	Méthode B$_{dict-}$	**Méthode B$_{dict+}$**
c (vrais positifs)	343	367
p (résultats proposés)	447	402
C (couverture)	89,4 %	80,4 %
P (précision)	76,7 %	91,3 %
R (rappel)	68,6 %	73,4 %
F (F-mesure)	72,4 %	81,4 %

Tableau 6. *Impact de la présence du dictionnaire sur la performance*

La performance de la version la plus récente de notre méthode (la méthode C) a été également mesurée sur des données dans d'autres configurations syntaxiques, qui sont actuellement prises en charge par le système. Les premières colonnes du tableau 7 affichent les résultats d'évaluation pour les 500 premières collocations de type adjectif-nom et nom-préposition-nom, pour la même paire de langues (anglais-français). Ces valeurs sont comparables aux valeurs obtenues pour les données de type verbe-objet, reprises dans l'avant-dernière colonne. La dernière colonne synthétise les résultats d'évaluation sur l'ensemble des données testées.

	Adjectif-Nom	**Nom-Prép-Nom**	**Verbe-Objet**	**Total**
c (vrais positifs)	378	337	396	1 111
p (résultats proposés)	466	443	471	1 380
C (couverture)	93,2 %	88,6 %	94,2 %	92,0 %
P (précision)	81,1 %	76,1 %	84,1 %	80,5 %
R (rappel)	75,6 %	67,4 %	79,2 %	74,1 %
F (F-mesure)	78,3 %	71,5 %	81,6 %	77,2 %

Tableau 7. *Performance de la méthode C par type de données*

6. Analyse des erreurs

Le processus d'extraction d'équivalents est entièrement automatique et ne nécessite pas d'intervention humaine pour la validation des résultats intermédiaires dans le flux de traitements. En conséquence, les erreurs qui subsistent parmi les résultats finaux peuvent être dues aux différents modules de traitement utilisés.

Nous avons repéré, ainsi, une première catégorie d'erreurs qui sont causées par des erreurs d'analyse syntaxique de la langue source. Les équivalences contenant des paires sources grammaticalement incorrectes ont été elles aussi marquées comme incorrectes : par exemple, l'équivalence *develop country* (EN) - *frapper pays* (FR), où la paire *develop country* est incorrectement analysée comme verbe-objet dans des contextes tels que *the developing countries* (« les pays en voie de développement »). Dans la sortie de la méthode C pour des données de type verbe-objet, que nous avons

examinées en détail, ces erreurs représentent 32,7 % du nombre total de 104 erreurs ou échecs de traduction (les autres 396 paires étant des vrais positifs, *cf.* tableau 5)[23].

Une deuxième catégorie d'erreurs est représentée par les erreurs dues à une mauvaise analyse syntaxique de la langue cible. Assez rarement, il peut arriver que le système propose des équivalents agrammaticaux, comme par exemple *libre de personne* (FR) pour l'expression source *movement of person* (EN) ; cette erreur est due à la mauvaise analyse du syntagme d'origine *libre mouvement des personnes*. Pour les données de type verbe-objet examinées, la taux de ces erreurs est de 9,6 %.

Aussi, il est vraisemblable qu'un certain nombre d'erreurs soient causées par des erreurs d'alignement de phrases. Les appariements incorrects ont comme conséquence la diminution du nombre des bons équivalents dans le corpus de phrases cibles, ce qui pose des problèmes spécialement pour les données très peu fréquentes, ainsi que pour les collocations sources n'ayant pas de traduction stable dans le corpus. Certaines expressions sont particulièrement difficiles à traduire, et donc leurs équivalents trouvés dans le corpus cible connaissent une grande variation. Un tel exemple se rencontre dans l'expression *serve purpose* (EN), pour laquelle on trouve des traductions montrant des grandes divergences syntaxiques (*cf.* exemple (3)). Notre analyse a trouvé que 17,3 % des erreurs examinées sont dues à cette cause.

(3) a. an additional cost which *serves no purpose*
un coût supplémentaire *inutile*
b. would *serve little purpose*, in my opinion.
ne *sera* selon moi que *peu efficace*
c. Such a ban will *serve no purpose at all.*
Une telle interdiction *ne sert à personne*.

L'amélioration de tous les modules impliqués permettrait la réduction du taux d'erreurs de notre méthode. Une solution supplémentaire serait d'augmenter le nombre de phrases alignées considérées en entrée, si des données plus larges sont disponibles[24]. Il serait aussi très utile d'utiliser des mécanismes pour trouver toutes les variations grammaticales possibles pour les traductions d'une collocation, afin de pouvoir couvrir un nombre plus grand de configurations cibles que celui actuellement pris en charge par notre système. Notre étude a révélé que 4,8 % des traductions ont échoué à cause de l'insuffisance des configurations cibles prévues : par exemple, pour pouvoir détecter l'équivalence *receive request - être saisi d'une demande*, notre système aurait dû prévoir une configuration cible plus complexe que celles prises en charge actuellement.

Pour poursuivre l'objectif mentionné, nos travaux futurs pourraient s'inspirer des travaux de (Daille, 2003 ; Jacquemin, 2001) sur la variation des termes, ou des travaux de (Ozdowska et Claveau, 2006) sur l'apprentissage de configurations équivalentes

23. Ces erreurs ne devraient pourtant pas être comptabilisées comme erreurs de notre méthode ; si on évalue la méthode C seulement sur les paires sources correctes, on obtient une précision de 88,4 % et un rappel de 85 % (F-mesure = 86,7 %).

24. Actuellement, le système exploite au maximum 50 phrases cibles par collocation, mais pour beaucoup de collocations on ne dispose que d'un nombre limité de contextes.

fondé sur l'alignement des mots et la détection de liens syntaxiques avec l'analyseur en dépendances SYNTEX (Bourigault et Fabre, 2000).

Notre analyse a aussi révélé qu'une certaine partie des erreurs concernent des correspondances incomplètes, telles que celles montrées en (4), où le terme manquant est ajouté entre crochets. Pour les données de type verbe-objet examinées, leur taux est de seulement 2,9 %, mais les autres configurations semblent plus affectées.

(4) résoudre problème - find [solution] to problem
faire travail - carry [out] work
nombreux problèmes - whole host [of problems]
learn from experience - tirer leçon [des expériences]
translate into action - joindre [acte/geste] à parole

Dans notre approche, nous avons traité un seul cas de correspondance, 2:2, dans lequel une collocation à deux termes se traduit aussi par une collocation à deux termes. Il s'agit ici d'un cas typique ; néanmoins, une approche plus sophistiquée devrait prendre en compte le cas général des correspondances de type *m:n*. Le problème des fragments est un problème reconnu dans l'extraction de terminologie (Frantzi *et al.*, 2000), mais l'état actuel de la technologie n'y fournit malheureusement pas des solutions suffisamment adéquates. Notre système est quand même capable de faire face à ce genre de situations de manière limitée, car il peut offrir une traduction plus complète s'il reconnaît, grâce à son lexique et à l'analyse syntaxique, qu'un des termes d'une paire est en réalité un terme complexe (multilexème). Un exemple est l'équivalence *put on agenda* (EN) - *placer à l'ordre du jour* (FR), où *ordre du jour* est une expression multilexème qui apparaît dans le lexique français de l'analyseur.

Lorsqu'une collocation se traduit par un seul élément lexical ($n = 1$), comme par exemple *bear witness* - *témoigner*, le système produit des équivalences erronées qui concernent l'ajout d'éléments indésirables : dans l'équivalence trouvée *bear witness* (EN) - *témoigner de vue* (FR), l'argument *vue* est superflu. L'exemple suivant montre un des contextes source et cible impliqués :

(5) This view *bears witness* to an understanding of the political constraints of the European Union in the Middle East.
Cette approche *témoigne* d'une *vue* claire des limites politiques de l'Union européenne au Moyen-Orient.

Le pourcentage de ce type d'erreurs est de 8,7 % pour les données verbe-objet de la méthode C qui ont été examinées.

Une autre catégorie d'erreurs identifiée est celle des erreurs dues aux limitations de la couverture du dictionnaire bilingue (4,8 % des erreurs analysées). Si le nombre de traductions proposées pour la base est insuffisant (par exemple, notre entrée pour le mot anglais *flag* liste seulement *drapeau*), le système élimines les bonnes corres-

pondances (comme *battre pavillon* pour la collocation source *[to] fly flag*), lorsqu'il applique le filtre lexicale fondé sur le dictionnaire[25].

Finalement, un pourcentage de 6,7 % d'erreurs[26] concerne des équivalences qui sont justes par rapport au corpus utilisé (car le traducteur humain a effectivement employé ces mêmes équivalences et elles sont appropriées aux contextes en cause), mais qui n'ont toutefois pas été jugées comme correctes, car la traduction est plutôt éloignée du sens initial et ne peut pas être considérée comme traduction de référence. C'est le cas, par exemple, de l'équivalence *cover area (EN) - concerner domaine (FR)*. L'exemple (6a) montre une des paires de phrases alignées qui a permis d'inférer cette correspondance ; deux autres paires proposant des équivalents plus appropriés sont montrées en (6b) et (6c).

(6) a. There are six vital *areas covered* by this unprecedented reform proposal
Six *domaines* vitaux sont *concernés* par cette proposition de réforme sans précédent

b. The synthesis report by its nature *covers* a wide *area* of the Commission's activity
Le rapport de synthèse, de par sa nature, *recouvre* une grande *partie* des activités de la Commission

c. so that it *covers areas* not covered by Community legislation
de manière à *couvrir* des *domaines* qui ne sont pas visés par la législation communautaire.

Nos travaux futurs pourraient explorer l'utilisation de modules de traitement complémentaires, tels que la désambiguïsation sémantique, afin d'identifier de manière encore plus précise les équivalents en fonction du contexte. Il serait également utile de proposer des équivalents multiples à la place d'une réponse unique, et de combiner, éventuellement, notre travail avec des travaux connexes de classification sémantique des collocations sources et des méthodes de détection de collocations synonymiques.

7. Conclusion

L'acquisition d'équivalents de traduction pour les expressions à mots multiples qui n'apparaissent pas toujours de manière adjacente dans le texte est un réel défi de la traduction automatique[27]. En nous fondant sur un analyseur syntaxique multilingue robuste et sur des outils que nous avons développés précédemment dans le cadre d'un système d'aide à la traduction, nous avons conçu une méthode de détection d'équi-

25. Notre stratégie pour combler les lacunes du dictionnaire s'applique seulement si le système échoue, mais pas s'il propose tout de même une solution.

26. Le reste de 12,5% sont des erreurs qui n'appartiennent à aucune des classes mentionnées.

27. La nécessité de prendre en compte l'information de nature syntaxique afin de rendre compte des transformations qu'une expression peut subir est de plus en plus reconnue, y compris dans les approches les plus récentes de traduction statistique (comme en témoigne la série d'ateliers *Syntax and Structure in Statistical Translation* - « Syntaxe et structure dans la traduction statistique » (Wu et Chiang, 2007 ; Chiang et Wu, 2008)).

valents de traduction pour un sous-type très fréquent et versatile d'expressions, les collocations. Selon (Orliac et Dillinger, 2003), les collocations représentent le facteur clé pour la production de textes acceptables dans les systèmes de traduction automatique.

La méthode présentée dans cet article est une version évoluée de la méthode de base décrite dans (Seretan et Wehrli, 2007), que nous reprenons également ici et présentons avec plus de détails. Les extensions apportées permettent de rendre compte des différences structurelles entre la langue source et la langue cible et d'améliorer considérablement la couverture et le rappel de cette méthode, tout en maintenant un bon niveau de précision.

Par rapport aux travaux similaires (passés en revue dans la section 2), les avantages de notre méthode sont multiples : prise en charge des constructions syntaxiquement flexibles (contrairement aux méthodes de (Kupiec, 1993 ; van der Eijk, 1993 ; Dagan et Church, 1994) qui, en l'absence d'outils d'analyse syntaxique adéquats, se limitent aux patrons simples) ; détection d'équivalents dans la forme canonique, sans besoin de traitement ultérieur pour trouver le bon ordre des mots (contrairement aux travaux de (Smadja *et al.*, 1996)) ; utilisation optionnelle de dictionnaires monolexèmes bilingues (au contraire, la méthode de (Lü et Zhou, 2004) en est complètement dépendante ; notre méthode fournit des résultats satisfaisants même en l'absence de ce type de dictionnaires).

Notre travail se distingue notamment de celui de (Lü et Zhou, 2004) par l'utilisation partielle des traductions des mots, seulement pour la base et pas pour le collocatif. De ce point de vue, notre méthode se démarque aussi des méthodes fondées sur dictionnaire et validation sur Internet des combinaisons générées, qui ont été proposées dans (Grefenstette, 1999 ; Léon et Millon, 2005). En effet, conformément aux prescriptions théoriques (Mel'čuk, 1998 ; Polguère, 2000), nous considérons que dans une collocation, un des mots – le collocatif – ne peut pas toujours être traduit de façon littérale. Toutefois, ces méthodes sont efficaces pour traduire des collocations relativement compositionnelles mais dont les termes sont des mots très ambigus, pourvu que le dictionnaire ait une couverture satisfaisante[28].

Par rapport aux travaux mentionnés, notre méthode est fonctionnelle pour un nombre plus grand de configurations syntaxiques et de paires de langues. À la différence des méthodes statistiques récentes, elle ne nécessite que très peu de paires de phrases alignées (nos expériences actuelles ont été faites en considérant au maximum 50 phrases alignées par collocation), et aucun effort d'entraînement au changement du domaine textuel. En contrepartie, elle est tributaire de la disponibilité d'outils d'analyse syntaxique pour les langues cibles, et il est aussi souhaitable (même si ce n'est pas nécessaire) qu'elle dispose de dictionnaires bilingues. Nous pensons, tout de même,

28. Par exemple, pour obtenir l'équivalence *groupe de travail - work group*, (Grefenstette, 1999) combine les cinq traductions disponibles dans un dictionnaire bilingue français-anglais pour le mot *groupe* (*cluster, group, grouping, concern, collective*), correspondant chacune à une lecture différente, avec les trois traductions disponibles pour *travail* (*work, labor, labour*).

que l'essor des analyseurs syntaxiques rendra les approches telles que la nôtre de plus en plus courantes, et que notre travail pourra servir, à cet égard, comme référence pour les travaux similaires futurs[29].

L'inconvénient principal de notre approche est la disponibilité relativement réduite des corpus parallèles, ainsi que leur limitation à quelques domaines textuels spécifiques (Léon, 2008). Les développements futurs de notre méthode seront dédiés (en plus des points déjà mentionnés dans la section 6) à son adaptation pour pouvoir être appliquée sur des corpus comparables, à la place des corpus parallèles. Dans cette perspective, nos travaux pourraient s'inspirer des méthodes telles que celle de (Sharoff *et al.*, 2009), fondée sur la détection de contextes similaires dans les corpus source et cible.

Remerciements

Ce travail a pu être réalisé grâce au soutien financier du Fonds national suisse de la recherche scientifique (subside n° 100012-117944). L'auteur souhaite remercier Eric Wehrli pour les discussions fructueuses et ses encouragements, Alexis Kauffmann et Béatrice Pelletier pour la révision répétée de ce manuscrit, ainsi que les trois relecteurs anonymes pour leurs commentaires détaillés et suggestions pertinentes dont la présente version a bénéficié.

8. Bibliographie

Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaire*, vol. 25, p. 131-151, 2000.

Bresnan J., *Lexical Functional Syntax*, Blackwell, Oxford, 2001.

Charest S., Brunelle E., Fontaine J., Pelletier B., « Élaboration automatique d'un dictionnaire de cooccurrences grand public », *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, Toulouse, France, p. 283-292, 2007.

Chiang D., Wu D. (eds), *Proceedings of the ACL-08 : HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Association for Computational Linguistics, Columbus, Ohio, 2008.

Chomsky N., *The Minimalist Program*, MIT Press, Cambridge, Mass., 1995.

Church K., Hanks P., « Word association norms, Mutual Information, and lexicography », *Computational Linguistics*, vol. 16, n° 1, p. 22-29, 1990.

Dagan I., Church K., « *Termight* : Identifying and translating technical terminology », *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*, Stuttgart, Germany, p. 34-40, 1994.

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

29. Les données annotées seront disponibles en ligne.

Daille B., « Terminology Mining », *in* M. T. Pazienza (ed.), *Information Extraction in the Web Era*, Lecture Notes in Artificial Intelligence, Springer, p. 29-44, 2003.

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, 2004.

Fillmore C., Kay P., O'Connor C., « Regularity and idiomaticity in grammatical constructions : The case of *let alone* », *Language*, vol. 64, n° 3, p. 501-538, 1988.

Frantzi K. T., Ananiadou S., Mima H., « Automatic recognition of multi-word terms : the C-value/NC-value method », *International Journal on Digital Libraries*, vol. 2, n° 3, p. 115-130, 2000.

Grefenstette G., « The World Wide Web as a resource for Example-Based Machine Translation tasks », *Proceedings of ASLIB Conference Translating and the Computer 21*, London, 1999.

Haegeman L., *Introduction to Government and Binding Theory*, Blackwell, Oxford, 1994.

Hausmann F. J., « Le dictionnaire de collocations », *in* F. J. Hausmann et al. (ed.), *Wörterbücher : Ein internationales Handbuch zur Lexicographie. Dictionaries, Dictionnaires*, de Gruyter, Berlin, p. 1010-1019, 1989.

Heid U., « On ways words work together – Research topics in lexical combinatorics », *Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX '94)*, Amsterdam, The Netherlands, p. 226-257, 1994.

Heid U., Raab S., « Collocations in multilingual generation », *Proceeding of the Fourth Conference of the European Chapter of the Association for Computational Linguistics EACL'89)*, Manchester, England, p. 130-136, 1989.

Hindle D., Rooth M., « Structural ambiguity and lexical relations », *Computational Linguistics*, vol. 19, n° 1, p. 103-120, 1993.

Howarth P., Nesi H., The teaching of collocations in EAP, Technical report, University of Leeds, 1996.

Jacquemin C., *Spotting and Discovering Terms through Natural Language Processing*, MIT Press, Cambridge MA, 2001.

Kilgarriff A., Rychly P., Smrz P., Tugwell D., « The Sketch Engine », *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, p. 105-116, 2004.

Koehn P., « Europarl : A parallel corpus for Statistical Machine Translation », *Proceedings of The Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, p. 79-86, 2005.

Krenn B., *The Usual Suspects : Data-oriented models for identification and representation of lexical collocations*, vol. 7, German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology, Saarbrücken, Germany, 2000.

Kupiec J., « An algorithm for finding noun phrase correspondences in bilingual corpora », *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, U.S.A., p. 17-22, 1993.

Laenzlinger C., Wehrli E., « Fips, un analyseur interactif pour le français », *TA informations*, vol. 32, n° 2, p. 35-49, 1991.

Lafon P., *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, Genève/Paris, 1984.

Léon S., Acquisition automatique de traductions d'unités lexicales complexes à partir du Web, PhD thesis, Université de Provence, 2008.

Léon S., Millon C., « Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web », *Actes de TALN et RECITAL 2005*, Dourdan, France, p. 595-604, 2005.

Lü Y., Zhou M., « Collocation translation acquisition using monolingual corpora », *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, p. 167-174, 2004.

Mel'čuk I., « Collocations and lexical functions », *in* A. P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, Claredon Press, Oxford, p. 23-53, 1998.

Nerima L., Seretan V., Wehrli E., « Creating a multilingual collocation dictionary from large text corpora », *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, p. 131-134, 2003.

Orliac B., Dillinger M., « Collocation extraction for Machine Translation », *Proceedings of Machine Translation Summit IX*, New Orleans, Lousiana, U.S.A., p. 292-298, 2003.

Ozdowska S., Claveau V., « Inférence de règles de propagation syntaxique pour l'alignement de mots », *TAL*, vol. 47, n° 1, p. 167-186, 2006.

Pearce D., « Synonymy in collocation extraction », *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, Pittsburgh, U.S.A., p. 41-46, 2001.

Pearce D., « A comparative evaluation of collocation extraction techniques », *Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, p. 1530-1536, 2002.

Polguère A., « Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French », *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, Stuttgart, Germany, p. 517-527, 2000.

Sag I. A., Baldwin T., Bond F., Copestake A., Flickinger D., « Multiword Expressions : A pain in the neck for NLP », *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, p. 1-15, 2002.

Seretan V., Collocation extraction based on syntactic parsing, PhD thesis, University of Geneva, 2008.

Seretan V., Nerima L., Wehrli E., « A tool for multi-word collocation extraction and visualization in multilingual corpora », *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, Lorient, France, p. 755-766, 2004.

Seretan V., Wehrli E., « Collocation translation based on sentence alignment and parsing », *Proceedings of TALN 2007*, Toulouse, France, 2007.

Seretan V., Wehrli E., « Multilingual collocation extraction with a syntactic parser », *Language Resources and Evaluation*, vol. 43, n° 1, p. 71-85, 2009.

Sharoff S., Babych B., Hartley A., « 'Irrefragable answers' using comparable corpora to retrieve translation equivalents », *Language Resources and Evaluation*, vol. 43, n° 1, p. 15-25, 2009.

Simard M., Foster G. F., Isabelle P., « Using cognates to align sentences », *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada, p. 67-81, 1992.

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

Smadja F., McKeown K., Hatzivassiloglou V., « Translating collocations for bilingual lexicons : A statistical approach », *Computational Linguistics*, vol. 22, n° 1, p. 1-38, 1996.

Tutin A., « Pour une modélisation dynamique des collocations dans les textes », *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, p. 207-219, 2004.

van der Eijk P., « Automating the acquisition of bilingual terminology », *Proceedings of the Sixth Conference on European chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands, p. 113-119, 1993.

Véronis J., Langlais P., « Evaluation of parallel text alignment systems : The ARCADE project », *in* J. Véronis (ed.), *Parallel text processing : Alignment and use of translation corpora*, Text, Speech and Language Technology Series, Kluwer Academic Publishers, Dordrecht, p. 369-388, 2000.

Wanner L., Bohnet B., Giereth M., « Making sense of collocations », *Computer Speech & Language*, vol. 20, n° 4, p. 609-624, 2006.

Wehrli E., *L'analyse syntaxique des langues naturelles : Problèmes et méthodes*, Masson, Paris, 1997.

Wehrli E., « Un modèle multilingue d'analyse syntaxique », *in* A. Auchlin et al. (ed.), *Structures et discours - Mélanges offerts à Eddy Roulet*, Éditions Nota bene, Québec, p. 311-329, 2004.

Wehrli E., « Fips, a "deep" linguistic multilingual parser », *ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, p. 120-127, 2007.

Williams G., « In search of representativity in specialised corpora : Categorisation through collocation », *International Journal of Corpus Linguistics*, vol. 7, n° 1, p. 43-64, 2002.

Wu D., Chiang D. (eds), *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, Association for Computational Linguistics, Rochester, New York, USA, 2007.

Wu H., Zhou M., « Synonymous collocation extraction using translation information », *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, p. 120-127, 2003.