Le problème des collocations en TAL

Luka Nerima, Violeta Seretan, Eric Wehrli
Laboratoire d'analyse et de technologie du langage
Département de linguistique, Université de Genève
{Luka.Nerima, Violeta.Seretan, Eric.Wehrli}@lettres.unige.ch

Résumé

Cet article présente le modèle de traitement des expressions à mots multiples tel qu'il est implémenté dans les travaux en TAL du LATL. Il discute le repérage automatique, le stockage dans le lexique, ainsi que la prise en charge de ces expressions dans le parser Fips, le traducteur Its-2 et dans le système d'assistance terminologique TWiC. En se focalisant sur les collocations, les plus flexibles et les plus fréquentes de ces expressions, il met en évidence la nécessité d'effectuer une analyse syntaxique détaillée du texte afin d'assurer le traitement approprié de ces expressions et de garantir une meilleure performance à l'analyse et à la traduction.

Mots-clés *: traitement automatique des langues naturelles, extraction de collocations, analyse syntaxique, traduction automatique, traduction de mots en contexte, aide terminologique*

1. Introduction

Un des problèmes reconnus en traitement automatique de la langue (TAL) est celui des expressions à mots multiples, c'est-à-dire des unités lexicales constituées par plusieurs mots orthographiques, non nécessairement contigus. Dans Sag et al. (2002, 2), ces expressions sont définies comme « des interprétations idiosyncrasiques qui dépassent la limite du mot » et on estime que leur proportion dans la langue est — d'après Jackendoff (1997, 156) — comparable à celle des mots individuels.

Bien que ces expressions constituent une difficulté pour de nombreuses applications en TAL, c'est à coup sûr dans le domaine de la traduction automatique et de l'aide terminologique que l'absence d'un traitement adéquat de ces expressions se fait le plus cruellement sentir. La traduction littérale d'une expression — suite à sa non-identification par le système — de même que la traduction idiomatique d'un groupe de mots — suite à la reconnaissance erronée d'une expression par le système — constituent une cause fréquente de traduction incompréhensible. Par exemple, si l'expression *traffic light* (en anglais) n'est pas identifiée par le système, ce dernier proposera la traduction littérale *lumière du trafic* à la place d'un meilleur équivalent

français, tel que *feu rouge*. À l'inverse, si l'expression *lumière du trafic* est traduite comme *traffic light*, on aurait employé une expression idiomatique au lieu de la traduction littérale.

Les systèmes TAL qui prennent en charge les expressions à mots multiples font d'habitude recours à des dictionnaires (souvent spécifiques à un domaine donné), tandis que le traitement effectué est plutôt basique, car on considère ces expressions tout simplement comme des mots avec des espaces. Comme le montrent Sag et al. (2002, 2), ce traitement manque de la flexibilité nécessaire dans le cas des expressions moins figées, parmi lesquelles les collocations.

Dans cet article, nous présentons le modèle de traitement des expressions à mots multiples tel qu'il est implémenté dans nos travaux en TAL et en particulier dans les systèmes d'analyse syntaxique Fips, dans le traducteur Its-2 et dans le système d'assistance terminologique TWiC. Nous présentons d'abord rapidement une classification des expressions à mots multiples (section 2), ensuite nous abordons tour à tour les trois questions majeures liées au traitement de ces expressions en TAL : l'extraction à partir de corpus (section 3), le stockage dans le lexique (section 4) et l'utilisation dans les système de traitement (section 5).

2. Classification des expressions à mots multiples

Dans cette section nous proposons un découpage des expressions à mots multiples en trois sous-classes, sur la base de leurs propriétés catégorielles, ainsi que de leur degré de figement syntaxique et sémantique.

•Mots composés

Les mots composés sont des unités de catégorie lexicale (nom, verbe, adverbe, etc.), dont tous les constituants sont contigus, et dont la signification n'est pas nécessairement compositionnelle.

Exemples : *pomme de terre, sèche-linge, c'est-à-dire, d'ores et déjà.*

•Expressions idiomatiques

Les expressions idiomatiques sont des unités de catégorie syntaxique (VP, TP, NP), susceptibles d'une certaine flexibilité syntaxique (modification, extraposition, etc.) et dont la signification est habituellement non compositionnelle.

Exemples : *poser un lapin à quelqu'un, enterrer la hache de guerre, casser sa pipe, tomber dans les pommes*.

•Collocations

Les collocations sont définies comme des associations conventionnelles de mots, arbitraires et récurrentes, dont les éléments ne sont pas

nécessairement contigus et dont la signification est largement transparente.

Exemples : *gros fumeur*, *caresser l'espoir*, *encourir un risque*, *exercer une profession*.

Typiquement, la substitution d'un terme d'une collocation par un synonyme est possible mais souvent ressentie comme peu (ou moins) appropriée (*exercer* vs. *pratiquer une profession*).

Les collocations peuvent subir un large éventail de transformations grammaticales, comme illustré dans les exemples suivants pour l'expression *décerner un prix :*

- modification adjectivale : *décerner un important prix*
- passivisation : *le prix Nobel de la Paix 2005 a été décerné hier*
- relativisation : *le prix qui lui a été décerné l'année passé*
- clivage : *c'est le prix le plus important qui sera décerné demain soir*
- interrogation : *quels prix ont été décernés lors de ce festival ?*

Les collocations sont, selon plusieurs chercheurs, les plus nombreuses des expressions à mots multiples : « Les collocations se taillent la part du lion dans l'inventaire des phrasèmes » (Mel'čuk 1998, 24). Une étude de Howarth & Nesi (1996) a montré, également, que la plupart des phrases contiennent au moins une collocation.

À cause de leur omniprésence dans la langue, ainsi que de leur haut degré de variabilité nécessitant un traitement complexe, les collocations occupent une place importante dans nos systèmes de traitement du langage. Les trois prochaines sections décrivent, respectivement, l'extraction de collocations à partir de corpus, leur stockage dans une base de données lexicale et l'utilisation de collocations dans quelques unes de nos applications TAL.

3. Extraction de collocations

3.1. Méthodes d'extraction

Les collocations étant souvent définies comme des associations typiques de mots — « items lexicaux qui apparaissent souvent ensemble » (Cruse 1986), « combinaisons arbitraires et récurrentes » (Benson 1990), ou « expressions qui correspondent à une manière conventionnelle de dire » (Manning & Schütze 1999) — une méthode fréquemment utilisée pour leur identification dans la langue est l'analyse statistique des textes, permettant de mettre en évidence les associations typiques de mots.

L'essor des technologies de l'information, de même que la prolifération de textes en format numérique a favorisé le développement de

la linguistique de corpus et, en particulier, des outils d'extraction terminologique basés sur l'analyse statistique de grandes collections de textes. Ces outils sont employés déjà depuis une ou deux décennies pour la création des dictionnaires phraséologiques. C'est le cas notamment des dictionnaire COBUILD (Sinclair 1995), entièrement construit à partir des exemples extraits d'un corpus en anglais, The Bank of English, et DEC — Dictionary of English Collocations (Kjellmer 1994) —, basé sur une analyse de la fréquence des mots. Les outils d'extraction terminologique jouent actuellement un rôle primordial en lexicographie.

Du point de vue statistique, le problème de l'extraction des collocations revient à l'application des tests de signification aux mots provenant d'un corpus, tests qui visent à déterminer le degré de dépendance entre chaque paire de mots apparaissant relativement proches l'un de l'autre dans le texte. Une paire de mots est statistiquement significative si la co-occurrence des deux mots n'est pas due au hasard[1]. D'après une caractérisation fournie par Smadja (1993), les mots d'une collocation « apparaissent ensemble plus fréquemment que par pur hasard », c'est pourquoi ces tests statistiques sont considérés comme appropriés pour l'extraction. Parmi les tests les plus fréquemment utilisés on peut citer le test du rapport de vraisemblance (*likelihood ratio*), le *Z-test*, ou le *khi-carré* (une description détaillée de ces tests peut être trouvée par exemple dans Manning & Schütze 1999).

Une autre méthode d'identification des collocations est celle basée sur l'*information mutuelle* (une mesure probabiliste dérivée de la théorie de l'information). Cette mesure quantifie l'information partagée par deux variables aléatoires. Appliquée aux mots, elle indique combien d'information un mot contient sur un autre mot ; or, c'est la même idée qui est exprimée, en effet, par le célèbre slogan contextualiste se référant aux collocations : « You shall know a word by the company it keeps ! » (Firth 1968, 179).

La qualité des résultats obtenus par les méthodes citées peut varier en fonction de plusieurs paramètres, tels que la taille du corpus considéré, la catégorie grammaticale des mots testés, ou aussi la langue source. L'étude comparative de la performance de ces méthodes a fait l'objet de nombreuses recherches dans les dernières années (Kilgarriff 1996, Pearce 2002, Evert 2004).

À côté des méthodes déjà mentionnées et qui ont été largement employées dans différents travaux d'extraction (Church & Hanks

[1] Un événement est appelé *statistiquement significatif* quand il ne se produit pas par hasard.

1990, Smadja 1993, Fontenelle et al. 1994), d'autres techniques ont été aussi développées pour l'extraction des collocations :

- des techniques basées sur le *data mining* (Lafon 1984, Feldman et al. 1998, Rajman & Besançon 1998) ;

- des techniques d'apprentissage automatique (Yang 2003, Zinsmeister & Heid 2003) ;

- des méthodes basées sur la sémantique lexicale, exploitant le critère de non-substitution par des synonymes (Pearce 2001, Wermter & Hahn 2004).

3.2. Architecture générale des systèmes d'extraction

Les différents systèmes d'extraction cités dans la section précédente suivent, dans les grandes lignes, la même procédure d'extraction :

1. dans une première étape, ils identifient les paires de mots qui seront testés ;

2. dans une deuxième étape, ils appliquent leur propre procédure pour associer à chaque paire de mots un score indiquant sa probabilité de constituer une collocation.

Le rôle de la deuxième étape est d'ordonner selon un critère bien défini les paires choisies dans l'étape précédente ; cet ordre représentera, en effet, le résultat de l'extraction. La première étape est elle aussi très importante pour la performance des systèmes d'extraction, car c'est là qu'on effectue le choix initial des paires, ce qui a une influence cruciale sur les résultats obtenus lors de la deuxième étape.

Pour des raisons pratiques, comme le risque d'explosion combinatoire, les systèmes d'extraction ne retiendront pas toutes les combinaisons possibles de mots comme candidats (dans l'étape 1). L'espace des combinaisons est ainsi limité à une fenêtre de mots de dimension prédéterminée (d'habitude 5 mots). Optionnellement, ces systèmes appliquent aussi un filtre de nature linguistique sur les paires considérées, afin d'exclure certaines combinaisons jugées indésirables, comme celles incluant des catégories fermées : articles, prépositions, conjonctions, auxiliaires, etc.

3.3. Sélection des candidats — aspects linguistiques

Une question importante qui se pose par rapport au choix initial de paires de mots est de savoir s'il existe des indications précises de nature linguistique pour guider le processus de sélection, afin de distinguer les collocations des combinaisons triviales ou des autres types d'expressions (mots composés, expressions idiomatiques).

Mais les recherches en phraséologie ne fournissent actuellement pas les critères nécessaires pour une distinction nette entre les différentes

sous-classes d'expressions à mots multiples. Généralement, on considère que les expressions idiomatiques, les collocations et les combinaisons triviales de mots forment un continuum et que les frontières séparant ces trois classes d'expressions ne sont pas clairement établies (Wehrli 2000, McKeown & Radev 2000).

Les définitions même de la collocation n'offrent généralement pas de descriptions linguistiques, quoique, dans certains cas, on spécifie les structures syntaxiques impliquées — par exemple, dans la définition de F.J. Hausmann :

> « On appellera collocation la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète) ; b) substantif + verbe ; c) verbe + substantif (objet) ; d) verbe + adverbe ; e) adjectif + adverbe ; f) substantif + (prép.) + substantif » (Hausmann 1989, 1010).

Mais à la différence de Hausmann, la plupart des chercheurs considèrent que la collocation peut se manifester dans n'importe quelle configuration syntaxique :

> « Le terme collocation se réfère à la combinaison syntagmatique des items lexicaux et elle est indépendante de la catégorie des mots ou de la structure syntaxique » (Fontenelle 1992, 222).

En effet, certaines recherches offrent des arguments contre l'exclusion a priori des catégories fermées (van der Wouden 2001), tandis que le BBI, le dictionnaire de collocations le plus complet à l'heure actuelle (Benson et al. 1986), montre aussi une très grande variété de configurations syntaxiques pour les collocations répertoriées.

En dehors de la configuration syntaxique, on pourrait envisager qu'une caractérisation syntaxique des collocations serait possible à l'aide des tests syntaxiques vérifiant si certaines transformations grammaticales seraient applicables à une collocation (pluriel, modification adjectivale, pronominalisation, etc.). Pourtant, à cause de la flexibilité morphosyntaxique marquée des collocations, il est impossible d'aboutir à une description des collocations au moyen de ces tests. Comme Heid (1994, 234) le remarque, « les propriétés syntaxiques semblent ne pas avoir de pouvoir discriminatoire ».

De manière générale, on peut considérer que l'absence de description linguistique des collocations est due à l'incapacité des théories linguistiques actuelles de modéliser les informations statistiques provenant de la linguistique de corpus. Les collocations se révèlent, par des observations de nature empirique dérivées grâce à l'analyse des corpus de texte, en tant qu'associations *conventionnelles* de mots établies par l'usage dans la langue (ou plutôt, en termes saussuriens, dans la parole). Elles sont autrement imprédictibles par rapport à un

système régulateur, c'est-à-dire, par la simple application des règles d'une grammaire à ses unités lexicales. L'élément essentiel dans la définition de la collocation est en effet de nature empirique et il prime sur les prescriptions d'ordre linguistique.

Une définition de la collocation qui englobe des notions empiriques (tels que l'usage, ou le choix de mots par les locuteurs) et qui décrit très bien le phénomène collocationnel est celle proposée par Mel'čuk (1998, 2003). La collocation est entendue comme une expression bipartite dans laquelle l'un de deux constituants est choisi librement pour exprimer le sens global de l'expression, tandis que l'autre est choisi de manière contrainte, en fonction du premier constituant et du sens à exprimer[2].

3.4. Extraction de collocations et analyse syntaxique

La première étape dans le processus d'extraction (la sélection des paires candidates, cf. section 3.2) fait appel à un filtre linguistique qui retient pour l'étape suivante seulement les paires définies par certaines combinaisons catégorielles, telles que les combinaisons énumérées dans la définition de Hausmann dans la section 3.3. De plus, on recourt souvent à la lemmatisation, afin de reconnaître toutes les formes fléchies d'un mot comme des instances de la même forme de base (comparer par exemple *dresser des bilans* et *dressait le bilan* pour la paire *dresser - bilan*).

Ce traitement ne correspond pourtant qu'à un niveau superficiel d'analyse du texte, qui peut conduire à des résultats erronés. À partir de la phrase :

(1) Ce gâteau se mange avec les doigts.

on extrairait, par exemple, les paires *gâteau - mange* comme combinaison substantif + verbe et *mange - doigt* comme verbe + substantif ; on proposerait alors erronément comme collocations *le gâteau mange* (sujet - verbe) et *manger le doigt* (verbe - objet). Cet exemple montre pourquoi l'extraction des collocations doit impérativement être précédée par l'analyse syntaxique du texte afin d'assurer la qualité des résultats.

Notre travail s'inscrit dans la sphère des approches qui considère l'analyse syntaxique profonde comme une pré-condition essentielle pour l'extraction. Cette pré-condition a été déjà reconnue dans le passé (Smadja 1993, 151) :

« Idéalement, pour identifier les relations lexicales dans un corpus on de-

[2] Traditionnellement, le premier constituant est appelé *base* et l'autre *colloqué* (conformément à la terminologie introduite par Hausmann).

> vrait d'abord analyser le texte pour vérifier que les mots sont utilisé dans le même syntagme »,

mais elle a été ignorée à cause de l'absence, à l'époque, d'outils d'analyse suffisamment performants. Cette situation s'est ensuite perpétuée malgré le récent progrès de la technologie en matière d'analyse syntaxique, avec seulement peu d'exceptions (Lin 1998, Daille 1994) en dehors de la nôtre.

3.4.1. FipsCo

Dans cette section nous présentons de manière succincte FipsCo, un extracteur de collocations (Goldman et al. 2001, Nerima et al. 2003, Seretan et al. 2004a) basé sur le système d'analyse syntaxique Fips développé au LATL (Laenzlinger & Wehrli 1991, Wehrli 1997). Fips est un analyseur symbolique qui suit le formalisme GB et qui prend en charge le français, l'anglais, l'italien et l'allemand, ainsi que de manière partielle d'autres langues comme l'espagnol, le grec (d'autres encore étant en développement).

Du point de vue architectural, FipsCo est intégré à Fips, l'extraction de collocations étant considérée comme une option de sortie de Fips, après l'analyse syntaxique des textes source. Au fur et à mesure que les fichiers d'un corpus sont analysés, on sélectionne, pour chaque phrase, les paires de mots qui représentent des candidats possibles comme collocations (ce qui correspond à la première étape d'extraction, la sélection des candidats). Une fois l'analyse du corpus achevée, tous les candidats sélectionnés passent (dans la deuxième étape d'extraction) un test statistique, le *likelihood ratios* (cf. section 3.3). Ce test associe à chaque paire de mots un score qui indique sa probabilité de constituer une collocation. Le résultat de l'extraction sera donc représenté par une démarcation graduelle des candidats selon ce score, plutôt que par une séparation nette entre collocations et non-collocations.

La sélection initiale des paires candidates est faite en tenant compte de la structure syntaxique de la phrase, telle que fournie par l'analyseur Fips. Comme montré au début de la section 3.4, il est essentiel que les deux mots d'une paire soient reliés par une relation syntaxique.

3.4.2. Proximité structurelle vs. proximité textuelle

À la différence de la plupart des autres travaux, le critère utilisé dans notre système pour la sélection des paires candidates n'est pas la proximité linéaire des mots dans le texte, mais la présence d'une relation structurelle entre les deux mots, soit la proximité structurelle.

Cette approche présente, premièrement, l'avantage évident

d'améliorer les résultats d'extraction, en réduisant le nombre des faux positifs (c'est-à-dire, des paires considérées par le système comme des collocations potentielles, malgré leur agrammaticalité — dans l'exemple 1, *manger - doigt* et *gâteau - manger*).

Deuxièmement, en renonçant au critère de la proximité en faveur du critère syntaxique, il est possible de détecter aussi les collocations dont les constituants se retrouvent, suite aux transformations grammaticales, à une longue distance dans le texte[3].

D'autres avantages apportés par l'analyse complète de la phrase sont :

a) la lemmatisation,

b) la normalisation des structures syntaxiques, par laquelle on peut traiter les cas d'extraposition telles que celles montrées dans les exemples de la section 2,

c) la désambiguïsation lexicale : dans le cas où plusieurs lectures sont disponibles pour un mot, l'analyseur choisit la variante qui est compatible avec l'analyse en cours pour la phrase.

Utiliser un critère syntaxique pour la sélection des collocations potentielles à la place du critère de la proximité textuelle correspond à la dichotomie entre une acceptation fondamentalement linguistique et une autre, purement stochastique, qui est observable dans la manière dont la collocation a été définie par différents auteurs. Une bonne partie des définitions de collocation met en évidence l'idiosyncrasie statistique de la collocation, sans préciser qu'il s'agit d'une expression bien-formée de la langue : « combinaisons arbitraires et récurrentes » (Benson 1990), « l'occurrence de deux ou plusieurs mots à proximité l'un de l'autre dans le texte » (Sinclair 1991, 170). Mais d'autres définitions insistent sur le statut linguistique de la collocation, entendue comme unité syntaxique et du sens : « éléments dans un patron syntaxique » (Cowie 1978), « unité syntaxique et sémantique » (Choueka 1988), « manière conventionnelle de dire » (Manning & Schütze 1999), cf. aussi la définition de Hausmann (1989) citée dans la section 3.3.

On remarque ainsi un passage de l'acceptation initiale, contextualiste, qui est purement stochastique, vers une approche linguistique, passage qui, d'après nous, doit être également suivi par la pratique. Étant donné les progrès réalisés dans le domaine de l'analyse syntaxique, il n'y a plus de raisons pour ignorer la structure du texte et imposer la contrainte de la proximité textuelle à la place du critère syntaxi-

[3] C'est le cas notamment dans les langues romanes : dans un corpus français, FipsCo a détecté une collocation dont les mots étaient séparés par pas moins de 39 autres mots (Goldman et al. 2001).

que.

3.4.3. Importance de l'extraction

FipsCo a été utilisé pour extraire des collocations à partir de corpus français et anglais et a permis ainsi d'enrichir le lexique de l'analyseur Fips. Les collocations repérées sont ensuite utilisées par Fips durant l'analyse, d'un côté pour désambiguïser les attachements et, d'un autre, pour extraire des collocations de taille plus grande, où un des termes peut à son tour être une collocation ; étant donné le caractère récursif des collocations (Heid 1994, 232), les collocations peuvent s'étendre, en effet, sur plus que deux mots : *armes de destruction massive, se constituer partie civile, accorder une attention particulière, jouer un rôle crucial,* etc.

De manière générale, l'extraction des collocations est largement utilisée dans la lexicographie et dans les applications majeures en TAL, telles que la traduction automatique et la génération du texte (voir McKeown & Radev 2000 pour un survol des systèmes qui prennent en charge les collocations). Selon Orliac et Dillinger (2003, 292), « les collocations sont la clé pour produire des résultats plus acceptables dans les systèmes commerciaux ».

L'extraction apporte aussi des bénéfices considérables à l'analyse syntaxique (Hindle & Rooth 1993, Alshawi & Carter 1994, Collins 1997, Wehrli 2000) et d'autres applications TAL, telles que la désambiguïsation lexicale (Brown et al. 1991), l'OCR (Church & Hanks 1990), ou la recherche d'informations (Hull & Grefenstette 1998).

4. Stockage des collocations dans le lexique Fips

Il est bien connu que les ressources lexicales constituent une composante clé des projets de TAL. De plus, les lexiques électroniques réutilisables sont une ressource rare et, même lorsqu'ils existent, leur structure s'avère généralement inadéquate pour un projet de TAL spécifique et l'information qui se trouve dans le lexique souvent lacunaire. C'est ce qui nous a poussé dans le cadre du projet Fips à constituer nos propres lexiques. Sans entrer dans les détails, la structure de notre base de données lexicale s'articule comme suit : pour chaque langue nous avons (i) un lexique des mots, contenant toutes les formes fléchies des mots de cette langue, (ii) un lexique des lexèmes, contenant les informations syntaxiques de chaque unité lexicale (une unité lexicale correspond plus ou moins à une entrée de dictionnaire classique) et (iii) un lexique des collocations, que nous décrivons en détail plus loin dans cette section. Pour être complet, il faut encore signaler que la base de données contient un lexique de correspondances bilingue pour chaque paire de langues.

4.1. Le stockage des collocations dans les lexiques

Pour lever d'emblée toute ambiguïté, précisons que les collocations ont été entrées manuellement dans nos lexiques. Même si un outil d'assistance de repérage de collocations tel que FipsCo décrit dans la section précédente a été largement utilisé, nous sommes convaincus que l'insertion de nouvelles collocations dans le lexique de manière 100% automatique n'est pas judicieuse du point de vue qualitatif et qu'il appartient en fin de compte au lexicographe de juger si un groupe de mots constitue ou non une collocation.

La manière dont les collocations sont stockées dans le lexique doit poursuivre deux objectifs essentiels :

– l'entrée par le lexicographe de nouvelles collocations doit être la plus aisée et la plus rapide possible,

– la description des collocations contenues dans le lexique doit être directement utilisable par l'outil d'analyse, Fips dans notre cas, et par conséquent doit s'appuyer sur le modèle de grammaire sous-jacent.

Le premier objectif découle du fait que nous prévoyons qu'un grand nombre de collocations seront insérées dans nos lexiques, probablement plusieurs milliers par langue[4]. Toute assistance à cette tâche mérite un soin particulier. L'interface du système d'entrée des collocations effectue une analyse syntaxique de l'expression entrée par l'utilisateur et détermine quels sont les lexèmes (unités lexicales) qui la composent, ainsi que le type de la collocation et les traits de figement. C'est ensuite au lexicographe de valider ou de modifier ces paramètres.

Nous avons présenté dans la section précédente l'outil FipsCo qui génère des listes de collocations à partir de corpus choisis. Le lexicographe peut parcourir ces listes et valider les collocations qui lui paraissent pertinentes.

Le deuxième objectif nous a conduit à définir une liste de configurations syntaxiques caractérisant les collocations. Lors de l'analyse, si un groupe de mots est identifié par Fips comme une collocation, la configuration syntaxique est insérée dans l'arbre d'analyse. La section 5 détaillera ce processus.

4.2. La représentation des collocations dans le lexique

Dans la section 2, nous avons discuté de la typologie des expressions (mots composés, collocations, expressions idiomatiques). Nous avons

[4] A titre d'exemple, notre lexique des collocations françaises, qui est le plus abouti, compte actuellement 10'500 entrées.

aussi mis en évidence que les expressions forment un continuum aux frontières parfois difficiles à définir. De plus, du point de vue du traitement automatique de la langue par Fips, autant l'identification des expressions à mots multiples relève d'une importance primordiale, autant la distinction entre les classes d'expressions est d'un intérêt moindre. Au niveau de leur représentation dans notre base de données lexicale, nous avons utilisé deux structures de stockage :

- pour les mots composés, nous avons utilisé la même structure de stockage que pour les mots simples de la langue, c'est-à-dire le lexique des mots et des lexèmes ;

- pour les autres expressions à mots multiples (expressions idiomatiques et collocations), nous avons créé une structure uniforme qui contient essentiellement la référence aux mots qui constituent l'expression et le type de l'expression. Par abus de langage, nous avons appelé cette structure lexique des collocations.

Les mots composés se comportent donc comme des mots simples dans notre base de données de lexicale, c'est-à-dire que, d'une part, leur description morphologique est donnée, toutes les formes fléchies étant présentes dans notre lexique des mots, et que, d'autre part, ils sont définis par leur catégorie lexicale ainsi que par des traits à caractère syntaxique stockés dans le lexique des lexèmes.

Exemple de mot composé et de ses informations lexicales :

pomme de terre
- catégorie : N
- forme fléchie n°1 : pomme de terre (forme de base)
  - nombre : sing
  - genre : fém
  - forme phonétique : [pɔmdətɛʁ]
  - consonne de liaison : Ø
- forme fléchie n°2 : pommes de terre
  - nombre : pluriel
  - genre : fém
  - forme phonétique : [pɔmdətɛʁ]
  - consonne de liaison : Ø
- informations syntaxiques/sémantiques (lexème) :
  - identifiant du lexème : 211000360
  - type de nom : commun
  - traits de nom : {objet physique, comptable}
  - sous-catégorisation : Ø

En ce qui concerne les collocations et les expressions idiomatiques, le lexique des collocations contient :

• la configuration syntaxique de l'expression, appelée type de la collocation (nom + adjectif, nom + nom, nom + préposition + nom, sujet + verbe, verbe + objet, etc.) ;

• la référence aux parties composant l'expression, c'est-à-dire aux lexèmes ;

• la préposition s'il y a lieu ;

• les traits de figement (collocation plurielle, complément sans déterminant, complément figé, etc.).

Exemples d'entrées dans le lexique de collocations :

prendre rendez-vous
type : verbe - objet direct
lexème n°1 : lex211006514 (*prendre* v. transitif)
lexème n°2 : lex2110000373 (*rendez-vous* n. commun)
préposition : Ø
traits de figement : {}

miroir aux alouettes
type : nom - préposition - nom
lexème n°1 : lex211005482 (*miroir* n. commun)
lexème n°2 : lex211019093 (*alouette* n. commun)
préposition : *à*
traits de figement : {compl. avec déterminant, compl. pluriel}

4.3. Critères de classification des mots composés et collocations

Comme nous venons de le voir, deux structures de stockage sont prévues pour les mots composés et les collocations. Le lexicographe peut donc éventuellement être confronté au problème de déterminer dans laquelle des deux entrer une expression à mots multiples. Examinons d'abord les critères en faveur des mots composés :

- les parties de l'expression sont contiguës et ne peuvent pas subir de déplacement ou de modification. L'expression apparaîtra telle quelle dans les corpus. Cette propriété est obligatoire pour les mots composés tel que nous les traitons, sa non-validité entraîne le stockage de l'expression dans le lexique des collocations ;

- les formes fléchies de l'expression ne peuvent pas être calculées à partir des formes fléchies des parties et il faut par conséquent les lexicaliser. Par exemple, le nom *passe-montagne* prend la marque *s* au pluriel alors que *coupe-vent* est invariable. Or *montagne* et *vent* ont tous deux un pluriel avec *s* ;

- les expressions nominales complexes exprimant un concept unique sont à considérer comme des mots composés. Mais ces critères sémantiques et conceptuels sont à manier avec précaution : dans notre

lexique français, *Premier ministre* et *pomme de terre* sont tous deux stockés comme des mots composés. Par contre le test qui consiste à se référer au mot composé par une des parties donnera un résultat positif pour *Premier ministre* mais échouera pour *pomme de terre*. Ici l'explication est simple, un *Premier ministre* est un *ministre* (relation de généralisation) alors qu'une *pomme de terre* n'est pas une *pomme* (mais un *tubercule*).

Les critères penchant en faveurs d'une collocation (au sens de notre structure lexicale) sont :

- les parties de l'expression peuvent se déplacer. Cette propriété est surtout caractéristique des expressions de type sujet - verbe et verbe - objet ;

- les parties de l'expression peuvent subir des transformations syntaxiques (passivisation, topicalisation, extraposition, etc), telles que nous les avons vues dans la section 2 ;

- le paradigme flexionnel est calculable à partir des paradigmes flexionnels des parties et la cardinalité de l'ensemble des formes fléchies est élevée. On voit ici l'avantage de cette structure qui s'appuie sur la morphologie des parties et permet l'économie du paradigme flexionnel de l'expression.

Par contre nous n'aborderons pas dans cette section quelles expressions à mots multiples « méritent » d'être insérées dans les ressources lexicales et lesquelles ne le méritent pas. Les objectifs en TAL peuvent en effet être multiples et variés (analyse, étiquetage, génération, traduction, etc.) et pour éviter une discussion stérile il faudrait au préalable soigneusement les définir. Disons simplement (i) qu'aujourd'hui la taille des lexiques électroniques n'est plus un obstacle, ce qui permet d'entrer un grand nombre d'expressions et que (ii) les expressions à mots multiples comptent parmi les entités les plus productives et aussi les plus volatiles des langues naturelles.

5. Les collocations dans nos systèmes TAL

5.1. Analyse

Contrairement à une idée répandue en TAL, nous estimons que le traitement adéquat des expressions à mots multiples ne peut pas se faire avant l'analyse syntaxique — par exemple sous la forme d'un pré-traitement lexical —, mais au contraire doit prendre place à l'issue de l'analyse, sur la base de structures syntaxiques normalisées.

En effet, comme les collocations sont susceptibles de subir des processus syntaxiques de modifications ou de déplacement (montée, extraposition, etc.), leur repérage est beaucoup plus facile, et plus sûr, à

effectuer sur la base d'une structure normalisée que sur la base d'une séquence de mots non analysés[5].

Pour illustrer cet argument, considérons la collocation *forcer la main à quelqu'un*, avec les exemples suivants :

(2a) Jean lui force la main.

(2b) La main lui a été forcée.

(2c) La main semble lui avoir été un peu forcée.

(2b) montre que cette collocation admet le passif et (2c) la montée du sujet (*raising*). On voit immédiatement la difficulté qu'il y aurait à identifier cette collocation dans un pré-traitement, sur la seule base de la chaîne orthographique. En revanche, à l'issue de l'analyse syntaxique, dans les trois cas, la structure obtenue contient un constituant VP avec le lexème *forcer* et un objet direct, le constituant *la main* dans (2a), sa trace dans (2b) et (2c).

5.2. Applications d'aide terminologique

5.2.1. Traduction des mots en contexte : TWiC

Une autre application pour laquelle l'identification des collocations (et des autres formes d'expressions à mots multiples) apporte une contribution qualitative importante est la traduction ou l'aide à la traduction.

TWiC (*translation of words in context*) est un logiciel d'assistance terminologique pour lecteurs de documents en ligne en langue étrangère. Il s'adresse tout particulièrement à des lecteurs ayant une connaissance de base de la langue du document, mais qui sont néanmoins susceptibles de faire face à des problèmes terminologiques. Un double clic sur le terme problématique permet d'afficher sa traduction.

On aurait tort de penser qu'il ne s'agit là que d'un dictionnaire bilingue en ligne. En effet, le système effectue une analyse linguistique détaillée de la phrase dans laquelle se situe le terme sélectionné, avec pour objectif d'identifier au plus près l'unité lexicale concernée. Cette analyse permet de limiter les réponses aux traductions compatibles avec le contexte linguistique et donc d'éviter de manière parfois considérable le 'bruit' inhérent à une recherche dans un bon dictionnaire bilingue.

D'autre part, comme l'analyseur linguistique incorpore un analy-

[5] Par structure normalisée, nous entendons une structure syntaxique dans laquelle les constituants apparaissent dans leur position canonique, ou alors sont liés à des catégories vides en position canonique.

seur morphologique, TWiC traduit n'importe quel mot, qu'il s'agisse d'une forme de base ou d'une forme fléchie. Mais le caractère le plus intéressant et le plus original de ce système est sa capacité à traiter les expressions à mots multiples, y compris celles dont les constituants ne sont pas nécessairement adjacents. Ces expressions, qui comprennent notamment les expressions idiomatiques et surtout les collocations, sont souvent difficiles à trouver dans les dictionnaires courants. L'exemple ci-dessous illustre bien ce cas de figure :

(3) A loss of more than 20M has been sustained during the previous fiscal year.
Une perte de plus de 20M a été essuyée/subie pendant l'année fiscale précédente.

Dans cet exemple, le verbe *to sustain* et le substantif *loss* forment une collocation de type verbe - objet direct *to sustain - loss*, qu'on peut transposer en français sous la forme des collocations (*subir - perte* ou *essuyer - perte*). L'objet direct est antéposé en raison de la forme passive du verbe. Voyons, sur la base de cet exemple, comment fonctionne le système TWiC.

Lorsqu'un mot est sélectionné dans une phrase, la phrase entière est soumise à l'analyseur Fips, qui retourne un fichier d'étiquette sous la forme suivante :

mot	étiquette	position	no. de lexème	no. d'expression
a	DET-SIN	0	111050002	
loss	NOU-SIN	2	111023205	-141005054
has	AUX-VER-PRES	7	111000040	
been	AUX-VER-PPA	11	111000076	
sustained	VER-PPA	16	111037785	141005054

On le voit, à chaque mot est associé :

(i) une étiquette morpho-syntaxique, qui spécifie sa catégorie (DET=déterminant, NOU=nom, AUX-VER=auxiliaire, VER=verbe) et certains traits flexionnels, tels que le nombre pour les déterminants et les noms (SIN=singulier) ou le temps pour les éléments verbaux (PRES=présent, PPA=participe passé),

(ii) un numéro correspondant à la position du premier caractère du mot dans la phrase (à partir de 0),

(iii) le numéro d'identification du lexème dans la base de données lexicale et

(iv) le numéro d'identification d'une expression ou collocation. Ce dernier est représenté sous forme négative pour tous les constituants autres que la tête de la collocation.

5.2.2. Extraction et visualisation

Le système d'extraction de collocations FipsCo décrit dans la section 3

a été utilisé dans un outil plus sophistiqué que nous avons développé comme support aux terminologues et aux traducteurs (Nerima et al. 2003, Seretan et al. 2004a). Cet outil permet l'extraction multilingue et la visualisation parallèle du texte source en même temps que le texte correspondant dans une autre langue, via une procédure d'alignement de texte au niveau de la phrase. Il assiste les terminologues dans la création des ressources lexicales bilingues, en montrant d'abord comment les collocations d'une langue ont été traduites vers d'autres langues (par la visualisation des documents parallèles provenant des archives de traduction), et ensuite en permettant le stockage d'une entrée bilingue pour la collocation identifiée dans une base de données terminologique.

L'outil utilise aussi une procédure qui combine les collocations extraites pour une langue en des suites plus longues et détecte ainsi des collocations de longueur arbitraire (Seretan et al. 2004b).

5.2.3. Extraction à partir du Web

La création des corpus pour l'extraction étant un processus long et coûteux, une autre extension que nous avons ajoutée à notre système d'extraction est la création instantanée d'un mini-corpus spécifique à un mot donné, à partir des exemples de texte trouvés sur Internet en utilisant les API de recherche Google. Ce mini-corpus sert à identifier des possibles colloqués pour le mot cherché. L'application (décrite dans Seretan et al. 2004c) est particulièrement appropriée aux lexicographes se focalisant sur la description d'un mot donné et présente aussi l'avantage non négligeable d'offrir une image actuelle de la manière dont il est utilisé dans la langue.

6. Conclusion

Dans cet article, nous nous sommes focalisés sur l'importance cruciale que prennent les expressions à mots multiples dans les systèmes de traitement automatique de langage naturel, autant pour l'analyse syntaxique que pour la traduction. En particulier, nous nous sommes concentrés sur les collocations, qui constituent, d'après beaucoup de chercheurs, la classe d'expressions la plus représentée dans le lexique, et qui sont les plus difficiles à traiter à cause du haut degré de variabilité morphosyntaxique.

Nous avons d'abord discuté l'extraction des collocations à partir des corpus, en présentant les méthodes d'extraction existantes et en décrivant l'architecture générale des systèmes d'extraction. Ensuite, nous avons présenté FipsCo, le système hybride d'extraction développé au LATL. À la différence de la plupart des travaux existants, il est basé sur l'analyse syntaxique détaillée du texte préalablement à

l'application des tests statistiques. Nous avons montré pourquoi cette étape est fondamentale pour l'extraction et quels sont les avantages qu'elle apporte. Nous avons précisé que notre approche oppose la proximité structurelle à la proximité textuelle comme critère majeur pour l'extraction, et que cette opposition est aussi reflétée par une dichotomie entre une définition essentiellement statistique de la collocation et une autre, fondamentalement linguistique, qui est observable dans la littérature. Quant aux désavantages associés à l'approche linguistique, ils sont inhérents à tout système TAL qui est basé sur l'analyse syntaxique : toute évaluation doit mettre en balance, d'un côté, les erreurs commises par l'analyseur, et d'un autre, les bénéfices apportées par l'analyse (dans notre cas, la détection des collocations à longue distance, le traitement de cas complexes d'extraposition, la désambiguïsation lexicale, etc.).

Après l'extraction, nous avons présenté la manière dont les collocations et les autres types d'expressions à mots multiples sont stockées dans le lexique de l'analyseur Fips. Nous avons illustré le rôle que des collocations jouent dans l'analyse, de même que dans la traduction automatique, domaine où les collocations prouvent pleinement leur importance. En particulier, nous avons décrit le système en-ligne TWiC qui effectue la traduction des mots en contexte, ainsi que les autres applications d'aide terminologique développés à LATL qui prennent en charge en particulier les collocations : le système d'extraction multilingue et de visualisation parallèle et l'outil d'extraction de collocations à partir du Web.

Remerciements

Les travaux décrits dans cet article sont en partie liés au projet de recherche « Analyse linguistique et extraction de collocations » financé par RUIG-GIAN (Réseau universitaire international de Genève) et ayant comme partenaires LATL et la Division de traduction de l'OMC. Les auteurs souhaitent remercier en particulier M. Olivier Pasteur pour sa précieuse collaboration à ce projet.

Bibliographie

ALSHAVI H. & CARTER D. (1994), « Training and scaling preference functions for disambiguation», *Computational Linguistics* 20(4), 635-648.

BENSON M. (1990), « Collocations and general-purpose dictionaries », *International Journal of Lexicography* 3(1), 23-35.

BENSON M., BENSON E. & ILSON R. (1986), *The BBI Dictionary of English Word Combinations*, Amsterdam, John Benjamins.

BROWN P.F., DELLA PIETRA S.A., DELLA PIETRA V.J. & MERCER R.L. (1991), « Word-sense disambiguation using statistical methods », in *Proceedings of*

the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991), Berkeley, 264-270.

CHOUEKA Y. (1988), « Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases expressions in large textual databases. », in *Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling*, Cambridge (Mass.)., 609-623.

CHURCH K. & HANKS P. (1990), « Word association norms, mutual information, and lexicography », *Computational Linguistics* 16(1), 22-29.

COLLINS M. (1997), « Three generative, lexicalised models for statistical parsing », in *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, 16-23.

COWIE A.P. (1978), « The place of illustrative material and collocations in the design of a learner's dictionary », in STREVENS P. (ed.), *In Honour of A.S. Hornby*, Oxford, Oxford University Press, 127-139.

CRUSE D.A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge.

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

EVERT S. (2004), *The Statistics of Word Cooccurrences : Word Pairs and Collocations*, thèse de doctorat, Université de Stuttgart.

FELDMAN R., FRESKO M., KINAR Y., LINDELL Y., LIPHSTAT O., RAJMAN M., SCHLER Y. & ZAMIR O. (1998), « Text mining at the term level », in *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, Nantes, 65-73.

FIRTH J.R. (1968), « A synopsis of linguistic theory, 1930–55 », in PALMER F.R. (ed.), *Selected papers of J.R. Firth, 1952-1959*, Bloomington, Indiana University Press.

FONTENELLE T. (1992), « Collocation acquisition from a corpus or from a dictionary : a comparison », in *Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere*, Tampere, 221-228.

FONTENELLE T., BRÜLS W., THOMAS L., VANALLEMEERSCH T. & JANSEN J. (1994), *DECIDE, MLAP-Project 93-19, deliverable D-1 : a survey of collocation extraction tools*, rapport technique, Université de Liège.

GOLDMAN J-P., NERIMA L. & WEHRLI E. (2001), « Collocation extraction using a syntactic parser », in *Proceedings of the ACL '01 Workshop on Collocation*, Toulouse, 61-66.

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

HEID U. (1994), « On ways words work together - research topics in lexical combinatorics », in MARTIN W. et al. (eds), *Proceedings of the VIth Euralex In-*

ternational Congress (EURALEX '94), Amsterdam, 226-257.

HINDLE D. & ROOTH M. (1993), « Structural ambiguity and lexical relations », *Computational Linguistics* 19(1), 103-120.

HOWARTH P. & NESI H. (1996), *The Teaching of Collocations in EAP*, rapport technique, Université de Leeds.

HULL D.A. & GREFENSTETTE G. (1998), « Querying across languages : A dictionary-based approach to multilingual information retrieval », in SPARK JONES K. & WILLETT P. (eds), *Readings in Information Retrieval,* San Francisco, Morgan Kaufmann, 484-492.

JACKENDOFF R. (1997), *The Architecture of the Language Faculty,* Cambridge (Mass.), MIT Press.

KILGARRIFF A. (1996), « Which words are particularly characteristic of a text ? A survey of statistical approaches », in *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex, 33-40.

KJELLMER G. (1994), *A Dictionary of English Collocations*, Oxford, Claredon Press.

LAENZLINGER C. & WEHRLI E. (1991), « Fips, un analyseur interactif pour le français », *TA informations* 32(2), 35-49.

LAFON P. (1984), *Dépouillement et statistique en léxicometrie*, Paris, Slatkine-Champion.

LIN D. (1998), « Extracting collocations from text corpora », in *First Workshop on Computational Terminology*, Montréal, 57-63.

MANNING C. & SCHÜTZE H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge (Mass.), MIT Press.

MCKEOWN K. & RADEV D. (2000), « Collocations », in DALE R. et al. (eds), *A Handbook of Natural Language Processing,* New York, Marcel Dekker, 507-523.

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

MEL'ČUK I. (2003), « Collocations : définition, rôle et utilité », in GROSSMANN F. & TUTIN A. (éds), *Les collocations : analyse et traitement,* Amsterdam, Éditions "De Werelt", 23-32.

NERIMA L., SERETAN V. & WEHRLI E. (2003), « Creating a multilingual collocation dictionary from large text corpora », in *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, 131-134.

ORLIAC B. & DILLINGER M. (2003), « Collocation extraction for machine translation », in *Proceedings of Machine Translation Summit IX*, New Orleans, Lousiana, 292-298.

PEARCE D. (2001), « Synonymy in collocation extraction », in *WordNet and Other Lexical Resources : Applications, Extensions and Customizations (NAACL 2001 Workshop),* Pittsburgh, Carnegie Mellon University, 41-46.

PEARCE D. (2002), « A Comparative Evaluation of Collocation Extraction Techniques », in *Third International Conference on Language Resources and Evaluation*, Las Palmas, 1530-1536.

RAJMAN M. & BESANÇON R. (1998), « Text mining – knowledge extraction from unstructured textual data », in *Proceedings of 6th Conference of International Federation of Classification Societies (IFCS-98),* Rome, 473-480.

SAG I., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2002), « Multiword expressions : A pain in the neck for NLP », in *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002),* Mexico City, 1-15.

SERETAN V., NERIMA L. & WEHRLI E. (2004a), « A Tool for multi-word collocation extraction and visualization in multilingual corpora », in *Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004),* Lorient, 755-766.

SERETAN V., NERIMA L. & WEHRLI E. (2004b), « Multi-word collocation extraction by syntactic composition of collocation bigrams », in NICOLOV N. et al. (eds), *Recent Advances in Natural Language Processing III : Selected Papers from RANLP 2003*, Amsterdam & Philadelphia, John Benjamins, 91-100.

SERETAN V., NERIMA L. & WEHRLI E. (2004c), « Using the web as a corpus for the syntactic-based collocation identification », in *Proceedings of International Conference on Language Resources and Evaluation (LREC 2004),* Lisbonne, 1871-1874.

SINCLAIR J. (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

SINCLAIR J. (1995), *Collins Cobuild English Dictionary,* London, Harper Collins.

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

VAN DER WOUDEN T. (2001), « Collocational behaviour in non content word », in *Proceedings of the ACL Workshop on Collocations,* Toulouse, 16-23.

WEHRLI E. (1997), *L'analyse syntaxique des langues naturelles*, Paris, Masson.

WEHRLI E. (2000), « Parsing and collocations », in CHRISTODOULAKIS D. (ed.) *Natural Language Processing - NLP 2000*, Berlin, Springer-Verlag, 272-282.

WERMTER J. & HAHN U. (2004), « Collocation extraction based on modifiability statistics », in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004),* Genève, Suisse, 980-986.

YANG S. (2003), « Machine Learning for collocation identification », in *International Conference on Natural Language Processing and Knowledge Engineering Proceedings (NPL-KE),* Beijing.

ZINSMEISTER H. & HEID U. (2003), « Significant triples : Adjective+Noun+Verb combinations », in *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003)*, Budapest, Hongrie.