Pour une modélisation dynamique des collocations dans les textes

Agnès Tutin
Agnès Tutin, LIDILEM, BP 25,
38040 Grenoble Cédex, France
tutin@u-grenoble3.fr

Abstract
Lexical collocations in texts obviously do not appear as described in dictionaries. They are prone to several kinds of syntactic and lexical variations. We propose a simple finite-state transducer based model using Intex (Silberztein 1993) which easily enables the automatic detection and annotation of collocations in texts. The linguistic model takes advantages of the syntactic and semantic properties of lexical functions (Mel'čuk, Clas & Polguère 1995) and operates on syntactic classes of collocations instead of individual lexical entries. The output is an annotated corpus including syntactic and semantic information on collocations, which could be used for pedagogical purposes.

Introduction

Les collocations, expressions semi-figées à mi-chemin entre unités lexicales figées et expressions libres, constituent dorénavant un champ d'étude bien circonscrit en lexicologie. Pour le français, des ressources lexicographiques conséquentes sont désormais disponibles (Base de données de Fontenelle 1997 ; DAFLES du GRELEP Verlinde *et al.* 2003 ; LAF de Polguère 2000). La modélisation linguistique de ces phénomènes n'est pas en reste : elle fait l'objet de nombreuses discussions, principalement autour des Fonctions Lexicales du *Dictionnaire Explicatif et Combinatoire* d'Igor Mel'čuk et ses collègues (Mel'čuk *et al.* 1995, Wanner 1996, Kahane & Polguère 2001, Grossmann & Tutin 2003).

Mais si les collocations sont étudiées et décrites d'un point de vue statique (en tant que données lexicales), leur comportement dans les textes n'est guère observé. On sait pourtant que, n'étant pas foncièrement figées du point de vue syntaxique et sémantique, elles présentent une grande variabilité (Segond & Breidt 1995). Il nous paraît essentiel d'aborder l'étude des collocations dans les textes, tant dans une perspective didactique (apprentissage inductif à partir d'exemples sur corpus) que pour des applications comme le TAL, le repérage des collocations dans les textes supposant qu'en soit faite une description " dynamique ". L'utilisation de ressources textuelles permet par ailleurs d'affiner les données lexicographiques qui, à leur tour, peuvent être projetées sur les textes.

Le projet que nous présentons ici vise une modélisation fine des collocations, permettant de rendre compte de leur variabilité, sans proposer une description *ad hoc*. Ce traitement débouchera sur la constitution d'un corpus à visée didactique qui sera utilisé dans le cadre d'un projet d'aide à la rédaction[1]. Dans un premier temps, nous abordons les liens étroits qui unissent les collocations aux corpus. Puis, nous présentons les différentes variations sous lesquelles se présentent les collocations dans les textes et proposons un modèle de traitement inspiré du DiCo de Polguère. Nous décrivons enfin l'implémentation des données lexicales et

le processus d'annotation à l'aide de grammaires locales du système Intex de Silberztein (1993) et présentons un schéma d'annotation des collocations dans les textes intégrant des informations syntaxiques et sémantiques.

1 Corpus et collocations

Dans la tradition contextualiste anglaise (Williams 2003), les collocations, entendues comme des mots qui tendent à souvent apparaître ensemble, sont indissociables de la notion de corpus puisque c'est la récurrence même des associations lexicales dans les textes qui leur donne le statut de collocation (Sinclair 1991).

En TAL, les ressources textuelles ont été largement exploitées pour l'extraction des collocations. De nombreux outils statistiques reposant sur une définition très souple de la notion de collocation ont ainsi été mis au point pour repérer les couples de mots qui tendent apparaître ensemble (par exemple, Church & Hanks 1990, Smadja 1993). Des techniques symboliques, utilisant des patrons syntaxiques (par exemple, Heid 1996, Ludewig 2001) ont permis d'extraire de façon plus ciblée des collocations d'un type linguistique donné. Enfin, des techniques mixtes, utilisant règles syntaxiques et modèles statistiques (Kilgarriff & Tugwell 2001) ont servi à l'élaboration d'outils sophistiqués.

Cependant, si les corpus servent à l'extraction et à l'élaboration de données lexicographiques, ces dernières sont, à notre connaissance, fort peu exploitées pour des traitements sur corpus. Une exception notable toutefois : le travail pionnier de Thierry Selva (2002) qui projette les collocations codées dans le DAFLES dans des corpus de presse pour générer automatiquement des exercices pour le FLE.

2 Modéliser le comportement dynamique des collocations dans les textes

Dans les textes, les collocations se présentent rarement sous la forme figée décrite dans les dictionnaires. Plusieurs types de variations peuvent être observés. Nous reprenons ici partiellement la typologie proposée dans Segond & Breidt (1995) :

- Variations lexicales de mots pleins ou de mots grammaticaux. Parmi les variantes de mots pleins, les variantes synonymiques de collocatifs, avec ou sans changement de registre, sont extrêmement productives : ***périr*** ou ***mourir*** *d'ennui*, ***mourir*** ou ***crever*** *de peur*, etc. Dans les formations des collocations, on observe également des variantes dans les prépositions introduisant les arguments (*être paralysé* ***de/par*** *la peur*, ***de/par*** *dépit*) ou dans la présence/absence de déterminants (*avoir* ∅ *peur/avoir* ***une*** *peur bleue/ avoir* ***la*** *peur de sa vie*), ce dernier phénomène étant partiellement régi par des contraintes syntaxiques[2].
- Variations morphologiques. Les verbes des collocations verbales subissent bien évidemment des variations de personne/nombre. Plus intéressantes sont les contraintes liées au nombre des noms, souvent déterminées par des caractéristiques sémantiques. Par exemple, dans le champ sémantique des sentiments, la plupart des noms, indénombrables, ne se mettent pas facilement au pluriel (*éprouver une joie immense*, ? *éprouver des joies immenses*), mais ce n'est cependant pas toujours le cas : *avoir une/des appréhension(s)*.

- Insertion de modifieurs. La plupart des collocations peuvent être étendues à l'aide de modifieurs adjectivaux ou adverbiaux : *avoir peur* vs *avoir une peur bleue* ; *le froid tétanisait Lulu* vs *un froid sibérien tétanisait Lulu* ; *il n'éprouvait qu'une légère appréhension.* On distinguera deux types de modifieurs : (a) les modifieurs qui font partie d'une collocation comme *avoir une peur bleue.* Ces expressions peuvent être considérées comme des superpositions de collocations (*avoir une peur bleue* = *avoir peur* + *une peur bleue*)[3] (b) les modifieurs " externes " comme les négations, les adverbes, etc.
- Variations distributionnelles. Si les dictionnaires présentent généralement les éléments de la collocation dans un ordre donné, la plupart des collocatifs adjectivaux et adverbiaux apparaissent dans plusieurs positions. Par exemple, certains adjectifs d'intensité sont d'un emploi assez souple (*une immense anxiété, une anxiété immense, son anxiété était immense*). D'autres combinaisons sont plus restreintes (**une bleue peur, une peur bleue, *sa peur était bleue*).
- Variations liées aux alternances syntaxiques. Dans les ressources lexicales, les collocations verbales sont représentées sous une configuration syntaxique privilégiée, généralement la construction standard à la voix active. Bien entendu, les collocations verbales apparaissent généralement dans d'autres constructions comme le passif (*elle était paralysée de/par la peur*), les constructions pronominales (*se ronger d'angoisse, l'inquiétude s'accroît* ...), les phrases relatives (*la peur qu'il ressentait*), ... Par ailleurs, le verbe peut lui-même apparaître dans une construction complexe à auxiliaire ou pseudo-auxiliaire comme dans *la peur devait commencer à le paralyser* ...

Ces variations diverses doivent être modélisées dans les entrées lexicales, en particulier dans la perspective d'un traitement automatique. Les techniques d'automates d'états finis s'avèrent bien adaptées à cette tâche et permettent facilement d'associer des grammaires locales aux entrées. Le formalisme IDAREX (Breidt & Segond 1995) permet ainsi de coder les variantes sous la forme d'expressions régulières. Par exemple, la formule :

```
perdre Verbe: <la: tête: | la: boule: | les: pédales:>
```

indique que *perdre* comme verbe peut être suivi de trois suites d'expressions possibles. Des variables renvoyant à d'autres expressions régulières peuvent être introduites pour des insertions régulières comme les auxiliaires ou les modifieurs productifs.

Ce codage efficace permet de rendre compte de toutes les variations observées, mais apparaît malgré tout assez lourd, en particulier pour les collocations verbales où toutes les alternances doivent être énumérées au cas par cas. Par ailleurs, dans ce type de codage, la relation sémantique et syntaxique liant les éléments de la collocation n'est pas traitée. L'expression est codée comme une unité lexicale, ce qui se justifie pour le cas des expressions figées comme *perdre la boule* mais apparaît discutable dans le cas des collocations où une certaine compositionnalité sémantique est perceptible. Le recours à un modèle linguistique plus fin comme celui des Fonctions Lexicales de Mel'čuk *et al.* (1995, Polguère 2000) permet à la fois de systématiser les associations syntaxique et sémantique et de proposer un traitement informatique moins *ad hoc*, par classe de phénomènes.

3. Codage des collocations

Le modèle que nous utilisons s'inspire du modèle combinatoire du DEC, en particulier de l'adaptation DiCo proposé par Alain Polguère (2000). Notre première implémentation porte sur les collocations verbales portant d'un sous-ensemble de noms de sentiments, ce champ sémantique étant développé dans plusieurs projets de notre équipe.

Le codage que nous proposons exploite le modèle des Fonctions Lexicales (désormais FL). Il tire parti des propriétés syntaxiques et sémantiques associées à ces fonctions, propriétés qui peuvent être traduites par des schémas syntaxiques spécifiques nous permettant de décrire les données dans les textes. Les propriétés syntaxiques sont codées dans des tables organisées dans une base de données relationnelles. Les tables, après avoir été fusionnées, sont ensuite exploitées de façon dynamique à l'aide du logiciel Intex dans des grammaires locales (transducteurs d'états finis) qui permettent de repérer et d'annoter les collocations dans les textes (Cf. Figure 1).

3.1 Fonctions lexicales syntagmatiques et patrons syntaxiques

Les FL syntagmatiques sont utilisées pour encoder les collocations, expressions semi-figées (du type *froid sibérien* ou *prêter attention*) où l'un des éléments, la base, conserve son sens habituel (*froid* ou *attention* nos exemples), l'autre, le collocatif (*sibérien* et *prêter* dans nos exemples), étant sélectionné en fonction de la base.

L'originalité du modèle des FL est de systématiser cette association par l'utilisation d'une étiquette, la Fonction Lexicale, associée à la base pour produire un collocatif. L'étiquette indique les propriétés syntaxiques et sémantiques qui lient le collocatif à la base. Ainsi, la FL **Magn** encode à la fois une propriété sémantique (expression de l'intensité) et une propriété syntaxique (c'est un modifieur). Il est ainsi possible, bien que la grammaire des FL ne soit pas parfaitement définie de façon formelle (Cf. Alonso Ramos & Tutin 1996 ; Kahane & Polguère 2000), de prédire les patrons syntaxiques associés à ces FL (Cf. aussi Heid 1996). Par exemple, la FL Magn associée à un nom produira un adjectif ; la FL $Oper_1$[4] (verbe support) produira un verbe dont le nom sera le premier complément dont le premier actant du nom est le sujet ; la FL $IncepFunc_1$ (le N commence à se manifester chez qqun) produira un verbe dont le nom argument sera sujet. Le tableau ci-dessous schématise ces associations.

Nom de la fonction	**Catégorie syntaxique de la base B**	**Catégorie syntaxique du collocatif C**	**Patron syntaxique prototypique de la collocation**	**Exemple**
Magn (intensif)	Nom	Adjectif	B –ATTR→ C (le collocatif est épithète de la base)	*Froid* –ATTR→ *sibérien* (*un froid sibérien*)
IncepFunc$_1$ ('commencer à atteindre')	Nom	Verbe	B ← I – C (le nom est le sujet du verbe)	*Doute* ← I – *gagner* –II → X (*Le doute gagne* X)
Oper$_1$ (verbe support)	Nom	Verbe	C—II→ B (le nom est le premier complément du verbe)	X ←I–*éprouver* – II → *chagrin* (*éprouver du chagrin*)

Tableau 1 : Patrons syntaxiques associés aux FL

Cependant, les patrons syntaxiques, tels qu'ils sont décrits ou suggérés dans le DEC apparaissent rarement tels quels dans les textes, de nombreuses alternances syntaxiques et modifications étant possibles, comme on l'a vu précédemment (passivation, relativisation, position attribut des adjectifs, etc.). Le modèle est cependant séduisant dans la mesure où il systématise les relations sémantiques et permet de regrouper sous une même étiquette les variations possibles, ce qui évite les traitements au cas par cas et garantit une cohérence dans le traitement. Par ailleurs, les définitions syntaxiques assez précises des FL nous permettent de regrouper les collocations ayant un patron syntaxique comparable, comme nous le verrons dans la section suivante.

3.2 Le codage lexical des collocations

Le codage lexical des collocations est effectué dans une base de données relationnelles. Les propriétés lexicales sont codées à partir d'observations sur corpus (prose de *Frantext* depuis 1950) et sur dictionnaires (*Le petit Robert, Le TLFI)*. A chaque entrée lexicale, par exemple *appréhension* ou *angoisse*, on associe la classe sémantique (dans nos exemples, " peur "), le type de fonction lexicale et les propriétés syntaxiques qui lui sont associées. Par exemple, pour *angoisse*, les valeurs qui correspondront à la FL IncepFunc$_1$ ('le sentiment commence à se manifester chez qqun') sont <*saisir*> ou <*prendre*> (sous forme lemmatisée). On traite par ailleurs un ensemble de propriétés syntaxiques sous forme de valeurs binaires ou ouvertes. Les informations syntaxiques à renseigner sont communes à un paradigme de FL répondant au même schéma syntaxique. Par exemple, les valeurs correspondant aux FL Func$_1$ (*L'angoisse se manifeste chez lui*), IncepFunc$_1$ (*l'angoisse le saisit*), Magn+Fact$_1$ (*L'angoisse le paralyse*), seront traitées d'une façon identique car elles sont construites sur le même modèle syntaxique : la base de la collocation est le sujet, le collocatif est le verbe

conjugué qui est muni d'un complément (schéma MC V SN). Ce regroupement permettra un traitement unitaire dans les textes de ces classes de FL à l'aide d'un transducteur d'états finis manipulant les enregistrements (Cf. Figure 1). Une table correspondant à un paradigme de FL regroupant les noms, leurs collocatifs et leurs propriétés est ainsi facilement créée comme on le voit dans le tableau 2 ci-dessous. Le modèle est illustré dans la figure 1.

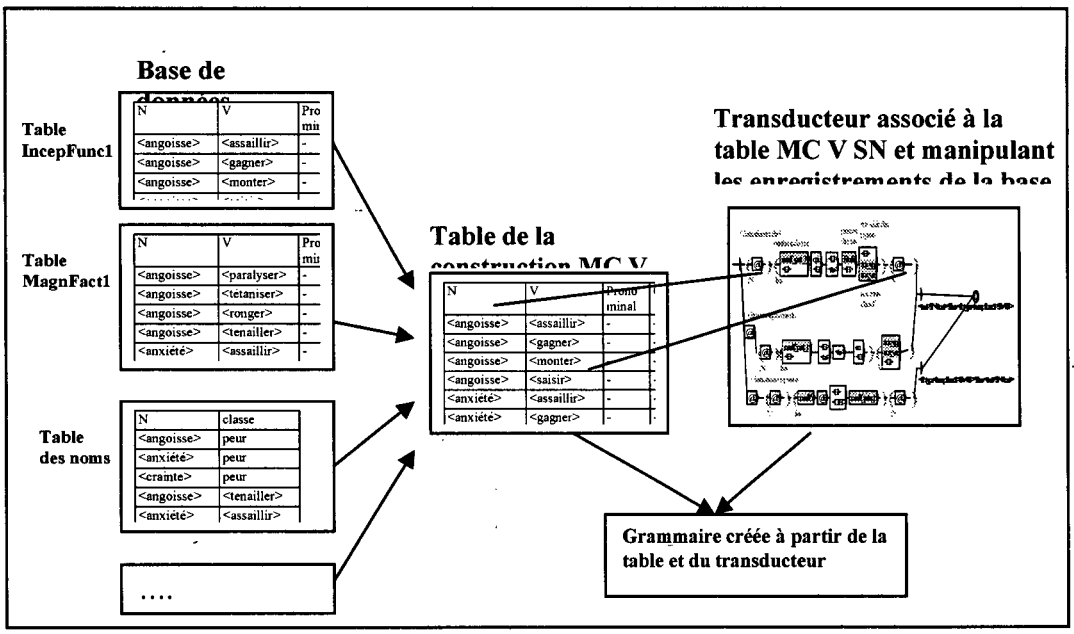

Figure 1 : Modèle du traitement des collocations

N	V	FL	Classe	Pronominal	Cind	Prep ind	passif	Prep passif
<angoisse>	<assaillir>	IncepFunc1	peur	-	-	-	+	:de_par
<angoisse>	<gagner>	IncepFunc1	peur	-	-	-	+	par
<angoisse>	<monter>	IncepFunc1	peur	-	-	:loc	-	-
<angoisse>	<saisir>	IncepFunc1	peur	-	-	-	+	:de_par
<angoisse>	<habiter>	Func1	peur	-	-	-	+	:de_par
<angoisse>	<manifester>	Func1	peur	+	+	chez	+	:de_par
<angoisse>	<consumer>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<dévorer>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<étreindre>	Magn_Fact1	peur	-	-	-	+	par
<angoisse>	<hanter>	Magn_Fact1	peur	-	-	-	+	par
<angoisse>	<paralyser>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<ronger>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<saisir>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<submerger>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<tenailler>	Magn_Fact1	peur	-	-	-	+	:de_par
<angoisse>	<torturer>	Magn_Fact1	peur	-	-	-	+	:de_par
<anxiété>	<assaillir>	IncepFunc1	peur	-	-	-	+	:de_par
<anxiété>	<emparer>	IncepFunc1	peur	-	+	<de>	-	-
<anxiété>	<gagner>	IncepFunc1	peur	-	-	-	+	:de_par
<anxiété>	<monter>	IncepFunc1	peur	-	-	en	-	-
<anxiété>	<saisir>	IncepFunc1	peur	-	-	-	+	par
<anxiété>	<étreindre>	Magn_Fact1	peur	-	-	-	+	par
...	...	...	...	...	...	...	...	...

Tableau 2 : Table des collocations correspondant au schéma " mot-clé verbe SN "

Les noms et les verbes apparaissent sous forme lemmatisée (entre chevrons), ce qui permettra de manipuler les formes fléchies dans les transducteurs. On indique pour chaque collocation le type de FL et la sous-classe sémantique du nom (dans notre exemple, le sous-champ de la peur). Les propriétés syntaxiques et les variantes sont indiquées à l'aide de valeurs binaires (" + ", " - ") ou à l'aide de valeurs spécifiques. On précise ainsi si le verbe apparaît dans une construction pronominale (*l'angoisse se manifeste ...*), si la construction est directe ou indirecte (et quel type de préposition est associé dans ce dernier cas), si la construction est passivable (et à l'aide de quelle préposition) et si le nom peut apparaître au pluriel (ce qui n'est pas toujours le cas pour les noms de sentiment, plutôt de type massif). Pour ce paradigme de FL, seules les propriétés strictement liées aux verbes seront indiquées.

Les variations systématiques comme l'insertion de modifieurs ou la relativisation ne sont pas traitées puisqu'elles peuvent être facilement modélisées dans les transducteurs.

4 Implémentation des données lexicales dans les grammaires locales pour l'annotation dans les textes

4.1 Les grammaires locales des collocations sous forme de transducteurs

Les données codées dans les tables sont exploitées dans des grammaires locales sous forme de transducteurs. Chaque paradigme de FL (sous forme de table) fait l'objet d'un transducteur qui indique les patrons textuels correspondant aux propriétés énoncées dans la table. Une grammaire est générée par le système Intex (Silberztein 1993) à partir de ce transducteur associé à une table[5]. Ce traitement permet de rendre compte du comportement souple des collocations dans les textes, tout en vérifiant les propriétés syntaxiques des collocations.

La figure 2 ci-dessous présente le transducteur associé au schéma " MC V SN " présenté dans la table correspondante (Cf. tableau 2) (les sorties n'apparaissent pas ici pour faciliter la lisibilité. L'annotation est présentée dans la section suivante). Une grammaire " MC V SN " est automatiquement générée à partir du transducteur associé à une table. Cette grammaire produit un chemin pour chaque série de valeurs d'un enregistrement vérifiant les conditions indiquées dans le chemin. Par exemple, dans le transducteur de la figure 2, le chemin du haut, repèrera les suites où apparaît un élément du premier champ de la table (qui correspond à " @A ". Pour la table concernée, ce sera un nom, par exemple *angoisse*) suivi d'un verbe apparaissant dans le même enregistrement pour le deuxième champ de la table (" @B ", un verbe, par exemple *assaillir*). Plusieurs éléments peuvent séparer le nom du verbe : le nom peut être suivi de modifieurs (décrits dans une grammaire locale appelé modif_post_N, apparaissant ici en grisé), d'un pronom relatif sujet, d'une négation, de pronoms clitiques (décrits dans une grammaire locale Proclit) et d'auxiliaires (grammaires aux_vpp et aux_ver). Tous ces éléments peuvent être facultatifs. Ce chemin permettra d'analyser des collocations comme *l'angoisse le saisit* ou *l'angoisse qui l'a violemment paralysé*. Le deuxième chemin permettra d'analyser les constructions pronominales (la propriété apparaissant dans le cinquième champ de la table, le champ " @E " sera vérifiée), par exemple *l'angoisse se manifeste dans son comportement*. Le troisième chemin permet d'analyser les constructions passives (en vérifiant que le verbe est passivable (champ " @H ") et la préposition utilisée dans la passivation (champ " @I ")) comme dans *Jeanne sanglotait, dévorée d'angoisse et de douleur*.

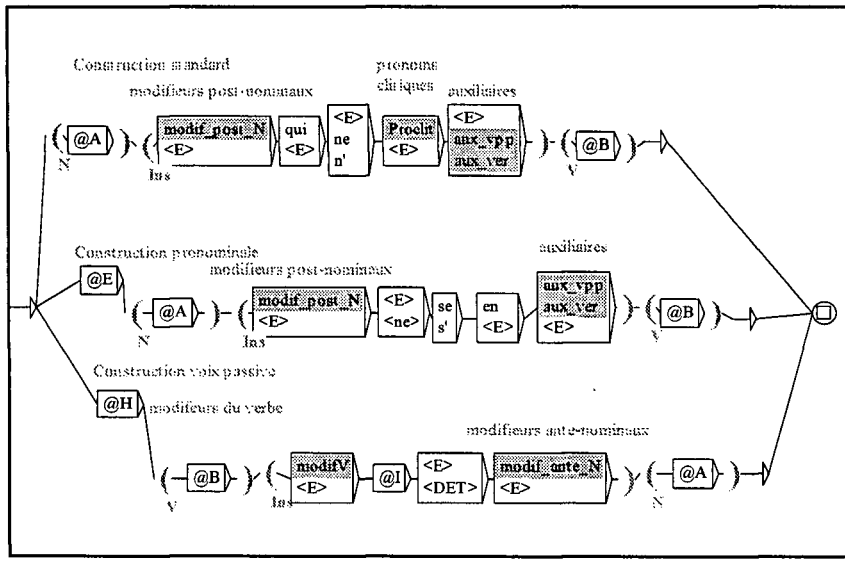

Fig. 2 : Transducteur pour le schéma de collocation " MC V SN "

4.2 Le schéma d'annotation des collocations en sortie

Les transducteurs comportent une sortie qui permet l'annotation automatique des collocations (ce sont, entre autres, les variables N, Ins, V qui sont prises en compte dans les sorties)[6]. Le schéma d'annotation, en XML, isole la base de la collocation et le collocatif. Des attributs sont associés à ces éléments :

- Le type de classe sémantique du nom.
- Le type de FL (le champ sémantique des émotions se prête bien à ce type de traitement).
- Le type de construction de la collocation. Ce paramètre est intéressant pour un travail fin sur la syntaxe des collocations.
- Le lemme du collocatif.
- Le lemme de la base.

Ces informations sont tirées du contexte linguistique (par exemple, les collocations analysées à l'aide du premier chemin dans la Figure 2 ont une construction directe) ou des champs apparaissant dans la table (par exemple, les lemmes, la classe du nom ou le type de FL). Pour rendre possible le traitement des superpositions de collocations, par exemple dans *une effroyable angoisse saisit son âme* (où la base est impliquée dans deux collocations : *l'angoisse saisit* et *effroyable angoisse*), nous avons décidé de coder les collocations sur les collocatifs qui ne sont en principe reliés qu'à une base[7]. La figure 3 présente quelques exemples de collocations de schéma " MC V SN " annotées automatiquement par Intex.

une <base classe="peur">peur</base>la<colloc type="IncepFunc1" const="stand" colloc="prendre" base="peur">prenait</colloc>, elle appelait Djali, …

Peu à peu, ces <base classe="peur">craintes</base>de Rodolphe la<colloc type="IncepFunc1" const="stand" colloc="gagner" base="crainte">gagnèrent</colloc>.

Elle fut <colloc type="IncepFunc1" const="passif" colloc="saisir" lemme="appréhension">saisie</colloc>d'une<base classe="peur">appréhension</base>, et,…

maintenant à genoux, Jeanne sanglotait, <colloc type="Magn_Fact1" const="passif" colloc="dévorer" lemme="angoisse">dévorée</colloc>d'<base classe="peur">angoisse</base> et de douleur.

Fig. 3 : Exemples de collocations annotées

4.3 Evaluation de l'annotation automatique

Pour évaluer la couverture de nos grammaires, nous avons testé un ensemble de collocations verbales (correspondant aux FL IncepFunc$_1$, Magn+Fact$_1$, Oper$_1$, Oper$_2$, Sympt$_1$, Real$_1$, c'est-à-dire trois schémas syntaxiques de collocations) sur un ensemble de noms liés à la peur (*angoisse, anxiété, crainte, effroi, épouvante, frayeur, inquiétude peur*). Le corpus de test est un ensemble de cinq romans français classiques du 19ème siècle[8], ce qui représente 567 000 mots.

L'évaluation (Cf. Tableau 3) donne des résultats encourageants.

Nbre de collocations dans le texte (A)	164
Nbre de collocations correctement annotées (B)	145
Nbre de collocations annotées (C)	150
Rappel (B/A)	88,4
Précision (B/C)	96,6

Tableau 3 : Evaluation des résultats de l'annotation automatique

Les collocations non repérées apparaissent dans plusieurs cas de figure :

- Des constructions syntaxiques non prévues par la grammaire, comme les cas de coordination, des constructions complexes à inversion du sujet ou des cas de " gapping ".
- Des données lexicales absentes comme l'expression *avoir grand'peur* que l'on trouve à plusieurs reprises dans le corpus.

Les collocations mal repérées (la collocation est repérée mais l'annotation n'est pas adéquate) sont principalement liées à des contraintes trop lâches, souvent au niveau des temps et modes verbaux qui ne sont pas précisés.
Les résultats nous permettent néanmoins facilement d'envisager une annotation automatique qui sera complétée manuellement.

Conclusion

Nous avons montré qu'un codage lexical simple des collocations, et des outils de TAL robustes, des transducteurs d'états finis (manipulés par le système INTEX), permettent un repérage satisfaisant de ces phénomènes, rendant bien compte des propriétés syntaxiques. Le codage tabulaire de l'information permet de traiter les informations lexicales par groupes de phénomènes, les FL, évitant ainsi les traitements fastidieux au cas par cas et permet de garantir la cohérence de l'information traitée. Les transducteurs, dont l'élaboration est assez rapide, permettent de manipuler facilement des classes syntaxiques de collocations. Une annotation automatique des collocations à des fins pédagogiques apparaît ainsi facilement envisageable.

Le modèle rencontre néanmoins quelques limites. Tout d'abord, la description lexicale des collocations est longue et fastidieuse, mais il est peut-être possible de l'alléger. Deux directions pourraient être explorées à cette fin. D'une part, il est probablement possible de relâcher les contraintes syntaxiques sur les collocations (comme le type de préposition utilisé pour l'introduction des compléments), mais il faudrait vérifier que cela n'introduise pas trop de bruit dans le repérage. D'autre part, il serait peut-être possible d'associer les collocatifs à des classes de noms plutôt qu'à des entrées lexicales. Par exemple, le collocatif *paralyser* serait associé à la classe des noms d'émotion, et non à des entrées particulières comme *angoisse* ou *épouvante*. Par ailleurs, le modèle rencontre les limites liées à l'utilisation des outils d'états finis, qui sont peu efficaces pour le repérage des dépendances à distance que l'on rencontre fréquemment dans les collocations verbales. Nous envisageons d'évaluer prochainement dans quelle mesure des outils syntaxiques robustes, utilisant des grammaires de dépendance, comme XIP (Aït-Mokhtar *et al.* 2002), permettent un traitement plus élégant et plus efficace pour l'analyse et l'annotation des collocations.

Remerciements

Un grand merci à mes collègues Cristelle Cavalla et Francis Grossmann qui ont pris le temps de lire ce document et m'ont fait des remarques judicieuses.

Notes

1. Le corpus développé sera exploité dans le cadre de deux projets :
- Projet "Ecole et Sciences Cognitives" piloté par le laboratoire LIDILEM (Université Stendhal-Grenoble 3) (Favoriser le développement des compétences lexicales et métalexicales en vue d'une aide à la production de textes au cycle 3 et au collège).
- Projet "plan pluri-annuel de formation" piloté par le LIDILEM (Développement de ressources pour la didactique du français à l'aide d'outils de TAL : étude des marqueurs linguistiques de la subjectivité).

2. La présence du modifieur rend en effet obligatoire l'insertion du déterminant.
3. Les superpositions sont productives, mais sont parfois bloquées : *il éprouvait une peur bleue, cela lui a fait une peur bleue* mais **il a pris une peur bleue* vs *il a pris peur.*
4. L'indice renvoie à la place de l'argument. Dans le cas d'$Oper_1$, le 1 indique que c'est le premier argument du nom qui occupe la place de sujet.
5. Cette fonctionnalité particulièrement intéressante a été au départ développée pour traiter les données du lexique-grammaire. Elle est parfaitement adaptée à notre problématique des collocations.
6. Par exemple, la sortie associée au premier chemin est la suivante : <base classe="@D">$N</base>$Ins<colloc type="@C" const="stand" colloc="@B" base="@A">$V</colloc> : les éléments précédés de " @ " correspondent aux enregistrements de la base, les éléments précédés de " $ " sont les variables analysées.
7. Cela n'est pas tout à fait juste, un collocatif pouvant être relié à plusieurs bases dans le cas de coordinations. Ex : *Jeanne sanglotait, dévorée d'angoisse et de douleur* (*Une vie*, Maupassant). Il faudra prévoir une extension de notre schéma d'annotation pour traiter ces cas.
8. Il s'agit de *La petite Fadette* (George Sand), *La femme de trente ans* (Balzac), *Colomba* (Mérimée), *Une vie* (Maupassant), *Madame Bovary* (Flaubert).

Bibliographie

Aït-Mokhtar S., Chanod J-P., Roux C. (2002), Robustness beyond shallowness: incremental dependency parsing. *Special issue of the NLE Journal*

Alonso Ramos, M., Tutin, A. (1996), A Classification and Description of Lexical Functions for the Analysis of their Combinations. in Wanner L. (éd.). *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphie, Benjamins, 147-167.

Church, K., Hanks, P. (1990), Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, vol 16/1, 22-29.

Fontenelle Th. (1997) : *Turning a bilingual dictionary into a lexical-semantic database.* Tübingen, Max Niemeyer Verlag.

Grossmann F., Tutin A. (2003), *Les collocations lexicales : analyse et traitement.* Amsterdam, De Werelt.

Kilgarriff A., Tugwell D. (2001), Word sketch : Extraction and Display of Significant Collocations for Lexicography, *Proc. Collocations Workshop, ACL 2001*. Toulouse, France, 32-38.

Kahane, S., Polguère, A. (2001), Formal foundation of lexical functions. *Proc. Collocations Workshop, ACL 2001*. Toulouse, France, 8-15.

Ludewig P., (2001), LogoTax - un outil exploratoire pour l'étude des collocations en corpus. *Linguistique de corpus*, *TAL*, 2001, sous la direction de B. Daille et L. Romary, vol 42 n°2, 623-642.

Mel'čuk I., Clas, A., Polguere A. (1995), *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve, AUPELF/UREF, Duculot.

Polguère, A. (2000), Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. *Proceedings of EURALEX'2000.* Stuttgart, 517-527.

Selva Th. (2002), Génération automatique d'exercices contextuels de vocabulaire. *Actes de TALN*. Nancy 2002, 185-194.

Segond F., Breidt E. (1995), (rapport technique MLTT22), *IDAREX : Formal Description of German and French Multi-Word Expressions With Finite State Technology.*

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes* : *le système INTEX*. Masson, Paris, 1993.
Sinclair J. (1991) : *Corpus, Collocation, Collocation.* Oxford, Oxford University Press.
Smadja, F.. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, Vol. 19, No. 1, March 1993, 143-178.
Verlinde S., Selva Th., Binon J. (2003), Les collocations dans les dictionnaires d'apprentissage : repérage, présentation et accès. *Les collocations lexicales : analyse et traitement*, F. Grossmann & A. Tutin, De Werelt, Amsterdam.
Wanner, L. (éd.) (1996) : *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam /Philadelphia, John Benjamins.
Williams G. (2003), Les collocations et l'école contextualiste britannique. *Les collocations lexicales : analyse et traitement*. F. Grossmann & A. Tutin, De Werelt, Amsterdam.