

Constitution d'un corpus annoté autour du lexique des émotions: collocations et fonctions lexicales

Magdalena Augustyn

Laboratoire de Linguistique et Didactique
des Langues Etrangères et Maternelles
(LIDILEM), Université Grenoble 3
BP 25, 38040 Grenoble Cedex 09
Magdalena.Augustyn@u-
grenoble3.fr

Agnès Tutin

Laboratoire de Linguistique et Didactique
des Langues Etrangères et Maternelles
(LIDILEM), Université Grenoble 3
BP 25, 38040 Grenoble Cedex 09
Agnes.Tutin@u-grenoble3.fr

Abstract

In this paper, we report an experiment of annotation of collocations in texts for pedagogical purposes using the Lexical Function model. We first show why showing collocations in context is according to us a fruitful method, and we present our annotation scheme and the corpora used. We then present some problems raised by the annotation process: delimitation of collocations, consistency of Lexical Functions, treatment of metaphors. All in all, the LF model appears to be operational, since over than 90% of collocations could be annotated with standard LFs in our corpus. We think that the model would probably benefit from being simplified (and homogenized) for a wider use for pedagogical purposes.

1 Introduction

Cette étude sur l'annotation des collocations dans le champ lexical des émotions s'inscrit dans le prolongement d'un projet plus global autour de l'étude des marqueurs linguistiques de la subjectivité, dans une perspective didactique¹. Dans ce cadre, deux grands types d'études linguistiques ont été effectuées : d'une part, l'étude des marques du discours rapporté et des passages entre guillemets, autour des phénomènes de polyphonie (Rinck & Tutin, 2007) ; d'autre part, une étude autour du lexique des émotions, qu'il s'agisse du repérage et traitement sémantique des formes simples (Augustyn et al., 2008), sujet qui a également été traité à plusieurs reprises dans notre équipe (Goossens, 2005 ; Tutin et al., 2006), ou comme dans la présente communication, des collocations.

Notre objectif est ici d'utiliser et d'adapter le modèle des fonctions lexicales, afin d'annoter les collocations dans des corpus textuels. Les corpus annotés pourraient être utilisés à des fins didactiques, afin d'illustrer le phénomène collocatif dans son environnement « naturel », le texte. Nous avons souhaité exploiter les descriptions lexicographiques de la lexicologie explicative et combinatoire, en particulier la modélisation proposée par les Fonctions Lexicales, pour les projeter sur les occurrences textuelles, tout en souhaitant fournir des descriptions simplifiées pour des publics non spécialistes. Cependant, cette procédure s'est avérée plus complexe qu'envisagé et l'annotation textuelle a permis de mettre en évidence un ensemble de points problématiques que nous souhaitons exposer ici.

Après avoir présenté l'intérêt que représente pour nous l'élaboration de corpus annotés intégrant des informations phraséologiques, nous exposerons la méthodologie et les corpus utilisés. Nous aborderons ensuite les difficultés posées par l'annotation de ces phénomènes à l'aide du modèle des fonctions lexicales, et montrerons que le processus d'annotation permet d'enrichir la réflexion sur la modélisation des collocations.

¹ PPF piloté par le LIDILEM (2003-2007) (F. Grossmann et G. Antoniadis) : « Développement et exploitation de ressources linguistiques pour la didactique du français à l'aide d'outils de TAL. Etude des marqueurs linguistiques de la subjectivité et de la polyphonie. »

2 L'annotation des collocations dans les textes

Les associations lexicales décrites sous le terme de collocations posent de nombreuses difficultés, maintenant bien connues, aux apprenants en langue maternelle et étrangère (Cf. par exemple, Granger, 1998; Nesselhauf, 2005). Sans en examiner toutes les facettes, nous définirons la notion de collocation comme une association mémorisée de deux éléments linguistiques sémantiquement pleins qui entretiennent une relation sémantique directe, et généralement une relation syntaxique directe. Suivant l'analyse désormais classique de Hausmann (1978), nous distinguons dans ces associations binaires deux éléments au statut distinct : la base est l'élément stable de la collocation, et le collocatif est choisi contextuellement en relation avec la base. Dans le cadre de cette étude, nous avons choisi et adapté le modèle des fonctions lexicales de la lexicologie explicative et combinatoire (Mel'čuk et al., 1995), le modèle qui nous paraît le plus abouti dans la modélisation des collocations, afin d'annoter ces éléments lexicaux dans les textes.

Nous pensons en effet que, sur le plan didactique, l'utilisation de corpus annotés qui présentent la phraséologie dans un environnement textuel naturel est pertinente pour plusieurs raisons :

1. L'observation des collocations dans les textes permet une meilleure réutilisation des expressions. La plupart des didacticiens du lexique insistent sur la contextualisation du lexique pour l'apprentissage (par exemple, Tréville & Duquette, 1996 ; Cavalla & Labre, à paraître).
2. Les collocations étant généralement assez transparentes sur le plan sémantique, une mise en contexte pertinente est parfois plus éclairante pour les apprenants qu'une explication ou un métalangage complexes. Dans cette perspective, une réalisation remarquable est celle du dictionnaire des cooccurrences du logiciel Antidote (Charest et al., 2007), qui relie de façon systématique des collocations (appelées cooccurrences) à des exemples sélectionnés sur corpus. Un traitement sémantique systématique permet cependant aux apprenants de comparer les collocations et leur donne un accès onomasiologique, qui peut aussi être profitable sur le plan didactique.
3. L'observation sur corpus permet de mémoriser les propriétés syntaxiques des collocations (types de déterminants, tournures actives ou passives) qui doivent être apprises en même temps que les éléments du lexique, et que les dictionnaires mentionnent rarement (Tutin, 2004). Pour le champ qui nous concerne, on relève par exemple que des collocations comme *la peur paralyse* ou *le remords ronge* sont bien plus fréquentes au passif réduit (*paralysé par la peur, rongé de remords*) qu'à la voix active, information qui sera directement observable pour l'apprenant sur le corpus.

Malgré l'intérêt des corpus annotés comportant une information phraséologique, ce type de ressources n'a pas été énormément développé pour les collocations, à notre connaissance, hormis les expérimentations de Ludewig (2001), Fellbaum & Geyken (2005) et Tutin (2005)². Cela est probablement dû à la difficulté de cette tâche qui exige un traitement extrêmement précis des objets lexicaux traités et ne peut en aucun cas être entièrement automatisé, comme nous le verrons.

3 Méthodologie

3.1 Corpus

Le corpus utilisé pour notre étude comporte un ensemble de textes variés, principalement des écrits libres de droits puisque les corpus élaborés dans notre projet devaient être utilisables librement pour les usagers³. Cela restreint malheureusement le corpus à des œuvres plutôt anciennes. Notre objectif étant didactique, nous avons sélectionné pour les œuvres littéraires – qui constituent l'essentiel du corpus – des romans largement utilisés dans le cadre scolaire. Le corpus annoté, dont la composition détaillée apparaît dans le tableau 1, comporte ainsi des ouvrages classiques au programme des collèges comme *Les lettres de mon moulin* d'Alphonse Daudet, ou *La petite Fadette* de George Sand.

² Dans cette expérimentation, nous avons adapté un sous-ensemble des collocations du Dicouèbe de Polguère pour une annotation dans des textes littéraires. Ici, le corpus et le lexique traités sont nettement plus vastes.

³ Rappelons que dans le droit français, les œuvres tombent dans le domaine public au bout de 70 ans.

Texte	Type	Nombre de mots
<i>La petite Fadette</i> (George Sand)	Littéraire	74456
<i>Les Lettres de mon moulin</i> (Alphonse Daudet)	Littéraire	46950
<i>Colomba</i> (Prosper Mérimée)	Littéraire	52619
<i>Le petit chose</i> (Alphonse Daudet)	Littéraire	83561
<i>Le mystère de la chambre jaune</i> (Gaston Leroux)	Littéraire	86 271
<i>Contes</i> (Perrault)	Littéraire	21684
<i>Les contes du lundi</i> (Alphonse Daudet)	Littéraire	68088
<i>L'Ile mystérieuse</i> (Jules verne)	Littéraire	199426
<i>Le droit à la paresse</i> (Jules Lafargue)	Essai polémique	12269
2 articles de la revue LIDIL	Ecrits scientifiques	10074
1 rapport scientifique <i>La place de la LSF dans l'intégration scolaire des enfants sourds</i> (Agnès Millet)	Ecrits scientifiques	13642
TOTAL		669040

Tableau 1. Composition du corpus annoté.

3.2 Ressources lexicales

Pour annoter les collocations au niveau du corpus, nous avons utilisé une procédure semi-automatique utilisant la base de données de collocations développée par Th. Fontenelle (1997)⁴, que nous avons choisie du fait de sa grande couverture lexicale. Cette base de données, qui codifie les associations lexicales à l'aide du modèle des fonctions lexicales, a été constituée semi-automatiquement par Th. Fontenelle à partir du dictionnaire anglais-français Collins-Robert (pour la méthodologie utilisée, voir Fontenelle, 1997). Le tableau 2 ci-dessous donne un aperçu de la base utilisée.

Catégorie syntaxique	Collocatif	Fonction Lexicale	Glose sémantique
vt	Abîmer	Causdegrad	Dégrader
adj	Abject	Antiver	Mauvais
adj	Abominable	magn+antibon	intense et négatif
vt	Absorber	Causpredminus	faire diminuer
n	Absorption	s0causpredminus	Diminution
vt	Accabler	Nocer ⁵	affecter vivement
vpron/vt	Accélérer	inceppredplus/causpredplus	faire augmenter/augmenter
vt	Accentuer	Causpredplus	faire augmenter
n	Accès	Culm	Maximum
n	Accroissement	s0inceppredplus	Augmentation
vpron/vt	Accroître	causpredplus/inceppredplus	faire augmenter/augmenter
n	Accumulation	s0inceppredplus	Augmentation
vt	Accumuler	Causpredplus	faire augmenter
vt	Achever	culmreal1/real1	Réaliser

⁴ Un très grand merci à Thierry Fontenelle de nous avoir autorisées à utiliser cette base lexicale, qui s'est révélée extrêmement riche. Nous regrettons que cette base n'ait pas davantage été exploitée dans le cadre des travaux sur la lexicologie explicative et combinatoire.

⁵ Cette FL n'est pratiquement jamais utilisée dans le DEC ou le Dicouèbe.

Tableau 2. Liste de collocatifs associés aux noms d'affect extraite de la base de Fontenelle 1997 (et gloses correspondantes).

Nous avons extrait de cette base tous les collocatifs employés en cooccurrence avec les noms d'affect, avec l'indication de FL associée, ainsi qu'une glose de la fonction lexicale (cf. plus loin la discussion sur les gloses). Cette base, sous forme de table, a ensuite été appliquée à notre corpus en utilisant le système Intex développé par Max Silberstein (1998) et corrigé semi-automatiquement, en parcourant l'environnement lexical des noms d'affect. Les données du DEC (Mel'čuk et al., 1984, 1988, 1992, 1999) ainsi que celles du Dicouèbe ont également été consultées, mais leur utilisation n'a pas toujours été aisée pour les raisons que nous verrons plus bas.

3.3 Principes de base de l'annotation

Comme signalé plus haut, le modèle des Fonctions Lexicales nous paraît être un modèle riche pour le codage syntaxique et sémantique des collocations. Cependant, la modélisation apparaît évidemment complexe pour les non spécialistes que nous visons et nous avons souhaité la simplifier, tout en conservant la philosophie, un peu à la façon du *Lexique Actif du Français* (Mel'čuk & Polguère, 2008), mais en poussant encore plus loin la simplification.

Les collocations sont annotées dans le corpus à l'aide du langage de balisage XML. La collocation est annotée sur le collocatif (élément <COLLOC>), la même base pouvant fréquemment être associée à plusieurs collocatifs, comme dans *avoir une peur terrible* (*avoir peur* + *une peur terrible*), comme on peut l'observer dans l'exemple (1).

Sur le collocatif, les éléments suivants sont annotés :

- Le **lemme de la base** (attribut BASE), qui permettra une recherche aisée dans le corpus, par exemple, pour obtenir tous les contextes où l'on a des collocations avec *amour*. Dans notre exemple, la base des collocations *avoir peur* et *une peur terrible* est *peur*.
- La **catégorie syntaxique (et sous-catégorie)** du collocatif (attribut CAT). On pourra ainsi rechercher toutes les collocations qui comportent un verbe transitif.
- La **fonction lexicale** (attribut FL) est également codée.
- Enfin, une **glose sémantique** (attribut TYPE_SEM), devant permettre un accès onomasiologique à la collocation, est également proposée. Dans notre exemple, la glose pour *avoir (peur)* est /éprouver/ alors que la glose pour *(peur) terrible* est /intense et mauvais/.

(1) Pour le coup le petit Chose <COLLOC BASE="peur" CAT="vt" FL="Oper1" TYPESEM="éprouver">eut</COLLOC> une <LEXIQUE TYPE="affect" CAT="N" DOMAINE="peur" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="négatif" >peur</LEXIQUE> <COLLOC BASE="peur" CAT="adj" FL="Magn+AntiBon" TYPESEM="intense et mauvais">terrible</COLLOC>; il se voyait déjà dans la rue, sans ressources... (*Le Petit Chose*, A. Daudet)

Sur la base (<LEXIQUE>), apparaissent de nombreuses informations sémantiques et syntaxiques concernant les mots d'affect : la catégorie, le champ sémantique, le niveau de langue, l'intensité, la polarité (pour une description détaillée, voir Augustyn et al., 2008). Précisons toutefois que le lexique de l'affect est traité dans notre corpus indépendamment des collocations.

Les principes de base de l'annotation étant posés, nous passerons maintenant aux difficultés rencontrées dans la mise en œuvre de l'annotation.

4 Problèmes rencontrés dans l'annotation et solutions apportées

4.1 Les gloses des Fonctions Lexicales syntagmatiques

Le *Lexique Actif du Français* (Mel'čuk & Polguère, 2007) reprend les principes du *Dictionnaire Explicatif et Combinatoire* en les simplifiant et en les didactisant. Les associations lexicales, qu'elles soient syntagmatiques ou paradigmatiques, sont introduites à l'aide de « formules de description »

permettant un accès onomasiologique. Par exemple, dans l'article de EFFROI, les collocations *l'effroi prend, gagne, saisit* X est glosée de la façon suivante dans le LAF :

E. commence à être éprouvé par X envahir, gagner, prendre, saisir [N_x], s'emparer [de N_x]/[de l'âme de N_x]

Pour l'annotation, nous utilisons des gloses plus courtes, que nous supposons plus faciles à comprendre. Par exemple, pour la collocation précédente, nous utilisons la glose sémantique /envahir/, comme dans l'exemple suivant, tiré de *L'Ile Mystérieuse* :

- (2) En effet, les singes, <COLLOC BASE="effroi" CAT="vt" FL="IncepFunc1" TYPESEM="envahir">**pris**</COLLOC> **d'un** <LEXIQUE TYPE="affect" CAT="N" DOMAINE="peur" NV_LANGUE="littéraire" INTENSITE="haut" POLARITE="négatif">**effroi**</LEXIQUE> subit, provoqué par quelque cause inconnue, cherchaient à s'enfuir. (*Ile mystérieuse*, J. Verne)

Les gloses sémantiques sont au nombre d'une cinquantaine, et apparaissent pour les collocations les plus productives. Elles sont moins précises sur le plan sémantique que les « formules de description » du LAF, mais la plupart des collocations étant assez transparentes sur le plan sémantique, nous pensons que le contexte permet de restituer facilement le sens. Dans notre démarche, nous privilégions ainsi la facilité d'accès au sens plutôt que la précision de la description, mais cette facilité d'usage doit être testée de façon concrète auprès d'utilisateurs.

Pour obtenir un encodage plus homogène et respecter une certaine facilité de lecture, nous avons réduit les expressions à un système basé sur des étiquettes simples comme *commencer*, *intense*, *positif*, *négatif*, opérateur causatif *faire*, par exemple. Nous indiquons ci-dessous quelques-unes des correspondances rencontrées entre fonctions lexicales et gloses :

Glose sémantique	Fonction lexicale
/commencer/	FL="IncepFunc0"
/commencer_éprouver/	FL="IncepOper1" ; FL="IncepPred"
/augmenter/	FL="IncepPredPlus"
/faire augmenter/	FL="CausPredPlus"
/affecter/	FL="Func1"
/affecter vivement/	FL="Magn+Fact1" ; FL="Magn+Func1"
/état/	FL="A1" ; FL="Adv1/2"
/intense/	FL="Magn"
/peu intense/	FL="AntiMagn"

Tableau 3. Exemple de correspondances glose/FL

Le problème des étiquettes sémantiques s'est aussi posé lors de l'annotation de fonctions non standard et de fonctions standard difficilement paraphrasables. Par exemple, nous avons recensé des collocations dans les textes où la fonction lexicale apparaît non standard, parce que la relation qui relie les éléments de la collocation n'est pas productive, par exemple : *liens d'amitié, crainte enfantine / puérile, amour maternel, amitié fraternelle*. Par ailleurs, même si la fonction lexicale est annotée, il nous est apparu difficile dans certains cas de trouver une glose simple, comme dans les cas suivants :

Oper1+Bon : *épanoui de N affect*
NonOper1 : *soyez sans (inquiétude, crainte)*
A1Real2 : *stupéfait d'admiration*

Pour ces deux cas de figure, il aurait été possible d'attribuer les fonctions lexicales non-standard à la manière du Dicouèbe, mais dans notre système simplifié de codage, nous avons préféré signaler la

collocation, sans lui associer de traitement spécifique, comme dans l'exemple suivant pour *amitié fraternelle* :

- (3) Peu à peu, quand il les vit honnêtes, énergiques, liés les uns aux autres par une <LEXIQUE TYPE="affect+relation" CAT="N" DOMAINE="affection" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif" JUGEMENT=""/" ATTITUDE="positif">**amitié**</LEXIQUE> <COLLOC BASE="amitié" CAT="adj" FL="" TYPESEM="">**fraternelle**</COLLOC> , il s'intéressa à leurs efforts. (*Ile mystérieuse*, J. Verne)

On peut cependant signaler que statistiquement, les fonctions lexicales non standard sont relativement peu nombreuses dans notre corpus (elles ne représentent que 6,1% des collocations annotées). De la même façon, peu de collocations annotées ne reçoivent pas d'étiquette sémantique (seulement 4,4% de l'ensemble des fonctions standard), ce qui montre que notre système d'étiquettes sémantique est dans l'ensemble très couvrant.

4.2 Le problème de la délimitation des collocations dans les textes

Dans les textes bien entendu, les collocations ne se présentent pas sous la forme canonique qu'elles ont dans les dictionnaires. Les délimiter par des éléments XML n'est pas une tâche triviale et cette procédure doit suivre un ensemble de principes cohérents.

Le premier problème concerne les phénomènes de « mise en facteur » d'éléments de la collocation. Tout d'abord, une base peut être utilisée par plusieurs collocatifs, comme dans l'exemple (1), *avoir une peur terrible* (fusion de *avoir peur* + *peur terrible*). Ce cas de figure est traité assez simplement puisque la collocation est mentionnée sur le collocatif. Inversement – et moins fréquemment – il arrive qu'un collocatif porte sur plusieurs bases, comme dans l'exemple suivant, *il pleurait de rage et de désespoir*, où les collocations *pleurer de rage* et *pleurer de désespoir* sont fusionnées. Dans ce cas, la disjonction de la base sera indiquée dans un attribut de l'élément XML <COLLOC> (BASE="rage/désespoir").

Une autre question à traiter est la délimitation du collocatif même, lorsqu'il apparaît dans des formes composées, qu'il s'agisse de formes pronominales, de temps composés ou de formes passives. Nous avons décidé de n'intégrer dans l'élément du collocatif que la forme lexicale pleine, à l'exclusion des mots purement grammaticaux. Ainsi, pour les formes pronominales, les pronoms réfléchis ont été intégrés dans le collocatif pour les verbes intrinsèquement pronominaux comme *se pâmer* (*de joie*), alors que pour des constructions pronominales, seul le verbe a été isolé comme collocatif.

- (4) Ses aides de camp l'entourent, empressés, respectueux, <COLLOC BASE="admiration" CAT="vpronti" FL="Sympt1" TYPESEM="manifester_physiquement">**se pâmant**</COLLOC> d'<LEXIQUE TYPE="affect+manif" CAT="N" DOMAINE="admiration" NV_LANGUE="courant" INTENSITE="haut" POLARITE="positif" JUGEMENT="positif" ATTITUDE="positif">**admiration**</LEXIQUE> à chacun de ses coups. (*Contes du lundi*, A. Daudet)

Outre les formes verbales composées, le collocatif peut inclure plusieurs lexies. Lorsqu'elles apparaissent obligatoires, elles ont été intégrées dans l'élément <COLLOC>. Ainsi, pour la collocation *ne pas se tenir de joie*, *ne pas se tenir* a été annoté comme collocatif, comme on peut l'observer dans l'exemple suivant.

- (5) Pencroff <COLLOC BASE="joie" CAT="locvti" FL="Magn+Oper1" TYPESEM="éprouver vivement">**ne se tenait pas**</COLLOC> **de** <LEXIQUE TYPE="affect+manif" CAT="N" DOMAINE="gaieté" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif">**joie**</LEXIQUE>, et chaque matin et chaque soir il ... (*Ile mystérieuse*, J. Verne)

4.3 Adaptation des FL syntagmatiques à l'encodage des collocations

4.3.1 Les FL : un inventaire à augmenter ?

Le système des fonctions lexicales disponibles permet de rendre compte d'un nombre important de collocations dans notre tâche, et remplit ainsi bien son objectif. Cependant, quelques types de collocations récurrentes rencontrées dans les textes à annoter ne sont pas décrits par une FL standard. Pour traiter ces

cas (relativement rares), nous avons proposé, si certaines conditions étaient remplies, de dégager de nouvelles fonctions standard lors de l'encodage de relations lexicales.

Ainsi, nous avons proposé une nouvelle fonction lexicale Intent⁶ (/en_vue_de/) calquée sur le modèle Propt (/à_cause_de/), qui nous paraît bien adaptée pour décrire les collocations suivantes : *pour le plaisir, pour l'amour*, par exemple :

- (6) (...) quant à Jacques, trop jeune encore pour comprendre nos malheurs - il avait à peine deux ans de plus que moi -, il pleurait par besoin, <COLLOC BASE="plaisir" CAT="prep" FL="Intent" TYPESEM="en_vue_de">**pour**</COLLOC> **le** <LEXIQUE TYPE="affect" CAT="N" DOMAINE="plaisir" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif">**plaisir**</LEXIQUE>. (*Le Petit Chose*, A. Daudet)

Même si le besoin d'instaurer la fonction Intent s'est posé d'une manière empirique et si cette fonction a été appliquée à un champ sémantique précis (lexique des émotions), nous pouvons justifier notre proposition par le fait que ce type de lien est assez systématique et qu'elle pourrait être ainsi généralisée et appliquée à d'autres lexies dans d'autres champs comme : *pour son intérêt, pour son compte, pour son bien*, etc.

4.3.2 Uniformisation du codage des FL⁷

De la même façon que pour les gloses, nous avons opté pour un encodage homogène des fonctions lexicales afin de proposer un métalangage plus simple. Nous avons essayé de réduire la diversité de FL quand cela était possible. Par exemple, à partir de deux fonctions lexicales équivalentes proposées par le Dicouèbe NonPerm1Fact0 et AntiReal1, nous avons gardé une seule valeur, AntiReal1, qui apparaît plus facilement décodable, par exemple dans :

Surmonter l'angoisse: NonPerm1Fact0 - /[X] ne pas se laisser influencer par son A./

Surmonter la crainte : AntiReal1 - /[X] ne pas se laisser influencer par sa C./

Dans certains cas où les FL s'avéraient difficiles à coder, nous avons eu recours au Dicouèbe. Cependant, le traitement proposé était assez complexe et parfois non uniforme, ce qui semble montrer la difficulté d'utilisation de ce métalangage. Ainsi, par exemple pour *épargner la déception / épargner la peine*, le dictionnaire propose deux FL différentes :

Epargner la déception : NonPermOper21 / [Qqch./Qqn.] empêcher que X éprouve une D./

Epargner la peine : CausNonIncepFunc1 / [Qqch.] empêcher que X éprouve de la P./

Pour uniformiser le codage, nous proposons d'encoder dans les deux cas la fonction NonPermFunc1 pour les exemples comme : *épargner, éviter, empêcher* + *N affect*.

En bref, pour certaines fonctions complexes, il nous semble que le modèle gagnerait peut-être à proposer un encodage plus systématique.

4.3.3 Fonction lexicale syntagmatique ou paradigmatisque ?

Une autre difficulté d'adaptation du modèle des FL a été le recours à certaines fonctions paradigmatisques pour exprimer des relations de cooccurrences (Cf. aussi Alonso Ramos & Tutin 1996 sur ce point). Certaines fonctions comme A₁ sont en effet essentiellement décrites comme ayant un fonctionnement paradigmatisque :

A₁ : adjectif typique pour le premier actant du mot clé. Exemple : A₁(bonheur) = heureux.

⁶ Du lat. *intentio*, étiquette proposée afin de suivre les appellations latines du DEC.

⁷ Nous tenons à remercier Alain Polguère pour ses suggestions et commentaires sur le traitement.

Il arrive cependant très fréquemment dans le champ des émotions que des collocations qui incluent le mot clé puissent être aussi décrites à l'aide de cette fonction standard : *en colère*, *en joie*, *en amour* (fr. québécois), *dans le désespoir*. Nous préférierions donc dans ce cas recourir à une fonction qui indique une relation syntagmatique, plutôt qu'utiliser de façon détournée une FL paradigmatique⁸. Une solution serait de proposer une nouvelle fonction syntagmatique qui indique la préposition typique pour l'état du premier actant du mot-clé. On aurait ainsi :

Loc₁ : préposition qui décrit l'état pour l'actant 1 du mot-clé.

Loc₁(désespoir) = *dans le* ~

Loc₁(colère) = *en* ~

Cependant, pour éviter la prolifération des FL et conserver une cohérence avec le traitement du DEC qui reste notre base, nous avons conservé les FL A₁ et Adv₁, tout en proposant des gloses spécifiques. Du point de vue de la cohérence du modèle cependant, cela ne nous paraît pas parfaitement satisfaisant.

4.4.4 Le traitement des métaphores

En travaillant sur le lexique abstrait que sont les émotions, nous avons été souvent confrontées au problème de la valeur figurée de certaines occurrences. En effet, certains collocatifs des noms d'affect pourraient être qualifiés de métaphoriques. C'est en particulier le cas des collocatifs véhiculant une valeur intensive, comme des nominaux : *éclair de joie*, *feu de l'amour*, *transport de fureur* ou les collocations verbales : *plonger dans la tristesse*, *se noyer dans le chagrin* ou dénotant une valeur aspectuelle : *refroidir l'enthousiasme*.

Dans la liste des fonctions lexicales proposée par le DEC, on relève une fonction Figur qui renvoie à une « métaphore codifiée par la langue dont la combinaison avec le mot clé est un synonyme (plus étroit) du mot clé » (Mel'čuk et al., 1984:7). Cette fonction lexicale est attribuée par exemple à quelques substantifs, comme une fonction lexicale simple ou dans les configurations avec d'autres fonctions paradigmatiques ou syntagmatiques :

Figur (haine) = feu [de la]	/Métaphore/
MultiFigur (vapeur) = nuage [de]	/Une certaine quantité de ~/
AntiMagn.Figur (espoir) = lueur [d']	/Métaphore d'un ~ peu intense/

Nous pouvons observer que cette fonction n'est pas attribuée d'une manière systématique. Par exemple, *débordement (d'enthousiasme)* est codé dans le Dicouèbe par Magn+Figur, mais *déborder (d'enthousiasme)* est codé Magn+Oper₁ sans mention de l'aspect métaphorique. Ainsi, il y a parfois des variabilités dans le codage de ce type de collocations et par la suite aussi de leur paraphrasage.

Nous trouvons cette fonction assez difficile à manipuler dans le corpus. D'après la définition de Figur dans le DEC, il s'agit d'une des fonctions paradigmatiques qui ne modalisent pas les collocations mais les rapports sémantiques entre les éléments, notamment dans le cas de Figur, une relation de synonymie. Certains précisent qu'il s'agit d'un « synonyme plus riche » de la base, ce qui implique qu'il rajoute une valeur sémantique supplémentaire. Cette définition en termes de synonymie apparaît discutable, ainsi que le statut paradigmatique de cette fonction. En effet, le collocatif figuré instaure une relation syntagmatique avec la base et cette relation devrait être encodée systématiquement. La fonction Figur dénote le plus souvent le haut degré d'intensité et c'est peut-être pour cela qu'elle est parfois identifiée aux autres FL sans indiquer la valeur sémantique véhiculée (p.ex. dans : Figur (haine) = feu [de la]).

Nous proposons de garder la fonction Figur, mais en la surajoutant à une description au niveau syntagmatique. Il faudrait l'appliquer à tous les cas des collocations à valeur figurée. Par exemple :

⁸ De la même façon qu'on distingue dans le DEC V₀ et Oper₁.

MagnOper1+Figur (amour) = brûler [d']
 CausPredMinus+Figur (enthousiasme) = refroidir [ART ~]
 S1Magn+Figur (haine) = feu [de la]

Il serait aussi préférable de coder Figur d'une autre manière, par exemple avec le pointeur ou entre parenthèses, afin de mieux souligner qu'il s'agit d'une information d'un autre niveau, superposée à la description de collocation avec les FL.

5 Bilan et conclusion

Au terme de cette expérimentation, un bilan s'impose. Du point de vue quantitatif, 1892 collocations mettant en jeu des mots d'émotions ont été codées. 93,9% d'entre elles ont pu être traitées à l'aide fonctions lexicales standard. Seulement 4,4% de fonctions lexicales annotées n'ont pas reçu de glose sémantique, comme on peut l'observer dans le tableau 4. Cela montre que le système des FL standard propose une réponse satisfaisante pour l'encodage de la majorité des collocations.

Type de FL	Proportion dans le corpus
FL standard	93,9% (=1776/1892)
FL non standard	6,1% (=116/1892)
FL standard avec glose	96,6% (=1716/1776)
FL standard sans glose	4,4% (=60/1776)

Tableau 4. Proportions de FL annotées

Si l'on se tourne maintenant vers les FL les plus utilisées (Cf. tableau 5), on retrouve bien les FL souvent citées dans la littérature sur les collocations : les verbes supports (Oper1) représentent à eux seuls un quart des collocations annotées et la collocation Magn (surtout quand elle est adjectivale) est également très courante. Les *blessé grave*, *faim de loup* et *peur bleue* souvent cités correspondent donc bien à un prototype productif. Les difficultés d'encodage ne concernent donc qu'un petit nombre de collocations, ce qui en minimise la portée.

Fonction Lexicale	Nombre total	Pourcent
Oper1	442	24,8%
Magn (adj)	295	16,6%
CausFunc(0/1)	205	11,5%
Caus(1)Manif	68	3,8%
Magn (adv)	54	3%
S0Sympt1	53	2,9%
Sympt1	51	2,9%
S0Manif1	44	2,4%

Tableau 5. Répartition des principales FL standard

D'une manière générale, notre expérimentation montre que le système des fonctions lexicales permet de coder la plupart des collocations de façon satisfaisante, même si quelques points gagneraient à être traités pour garantir la cohérence du système. Une simplification et une homogénéisation du modèle en permettraient probablement une meilleure utilisation.

Enfin, il nous reste maintenant à tester l'exploitation de ces ressources annotées pour les applications didactiques que nous visons.

Remerciements

Nous remercions tout particulièrement Alain Polguère, qui nous a fourni de précieuses explications sur le codage des Fonctions Lexicales. Un grand merci également à Thierry Fontenelle qui nous a autorisées à utiliser sa BD de

collocations, extraite du dictionnaire Collins-Robert. Hormis les auteures de ces lignes, le corpus a aussi été annoté par Gwendoline Bloquet et Mériam Haddara que nous remercions aussi vivement. Merci aussi à nos collègues Cristelle Cavalla et Francis Grossmann pour leurs remarques avisées.

Références

- Alonso Ramos, Margarita, & Agnès Tutin. 1996. A Classification and description of the Lexical Functions of the Explanatory Combinatorial Dictionary for the treatment of LF Combinations. In Wanner L. (ed.), *Lexical Functions in Natural Language Processing and Lexicography*. John Benjamins, Amsterdam, 146-167.
- Augustyn, Magdalena, Sabrina Ben Hamou, Gwendoline Bloquet, Vannina Goossens, Fanny Rinck, Mathieu Loiseau. 2008. Constitution de ressources pédagogiques numériques : le lexique des affects. *Autour de la langue et du langage : perspective pluridisciplinaire*. Presses Universitaires de Grenoble, Grenoble.
- Cavalla, Cristelle, & Virginie Labre. A paraître. L'enseignement en FLE de la phraséologie du lexique des affects. In Novakova I. & Tutin A. *Lexique des émotions*. Ellug, Grenoble.
- Charest, Simon, Eric Brunelle, Jean Fontaine, Bertrand Pelletier. 2007. Élaboration automatique d'un dictionnaire de cooccurrences grand public. *Actes de Traitement Automatique des Langues Naturelles 2007*, Toulouse, 282-292.
- Goossens, Vannina. 2005. Les noms de sentiment : esquisse de typologie sémantique fondée sur les collocations verbales. *Lidil*, 32:103-121.
- Fellbaum, Christiane, & Alexander Geyken. 2005. Transforming a Corpus into a Lexical Resource: The Berlin Idiom Project. *Revue Française de Linguistique Appliquée*, X (2):45-62.
- Fontenelle, Thierry. 1997. *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. (Lexicographica/Series maior). Niemeyer Verlag, Tübingen.
- Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. Cowie A. (ed.) *Phraseology: theory, analysis and applications*. Oxford University Press, Oxford, 145-160.
- Hausmann, Franz Josef. 1989. Le dictionnaire de collocations. Hausmann, F.J., Reichmann, O., Wiegand, H.E., Zgusta, L. (eds), *Wörterbücher : ein internationales Handbuch zur Lexicographie*. Dictionaries. Dictionnaires. De Gruyter, Berlin/New-York, 1010-1019.
- Ludewig, Petra. 2001. LogoTax : un outillage exploratoire pour l'étude de collocations en corpus. *Traitement automatique du langage*, 42(2):623-642.
- Mel'čuk, Igor A., & Alain Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. De Boeck Duculot, Louvain-la-Neuve.
- Mel'čuk, Igor A., André Clas, Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Paris.
- Mel'čuk, Igor A. et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain*, Vol. 1, 2, 3, 4. Presses de l'Université de Montréal, Montréal.
- Nesselhauf, Nadja. 2005. Collocations in a Learner Corpus. *Studies in Corpus Linguistics*, 14. John Benjamins, Amsterdam.
- Rinck, Fanny, & Agnès Tutin. 2007. Annoter la polyphonie dans les textes : le cas des passages entre guillemets. *Corpus*, 6:79-100.
- Silberstein, Max. 1999. INTEX: a Finite State Transducer toolbox. *Theoretical Computer Science*, 231-1:33-46. Elsevier Science, Saint-Louis.
- Tréville, Marie-Claude, Lise Duquette. 1996. *Enseigner le vocabulaire en classe de langue*. Hachette, Paris.
- Tutin, Agnès. 2004. Pour une modélisation dynamique des collocations dans les textes. *Actes d'Euralex*. Lorient, 6-10 juillet 2004.
- Tutin, Agnès. 2005. Annotating Lexical Functions in Corpora : Showing Collocations in Context. 2nd International Conference on the Meaning-Text Theory. Moscow, June 23-25 2005.