

Annotating Lexical Functions in Corpora: Showing Collocations in Context

Agnès Tutin

Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM),

Université Grenoble 3

BP 25, 38040 Grenoble Cédex 09

e-mail : agnes.tutin@u-grenoble3.fr; Tel : 04 76 82 43 00 Fax: 04 76 82 43 95

Summary

In this paper, we show how the DiCo database of French Collocations (Polguère 2000) can be adapted in order to annotate collocational information in corpora for pedagogical purposes. We first present syntactic information that should be added to the database for this application and show how simple finite-state transducers associated with metagraphs (lexicon-grammars) can be successfully used to annotate collocations.

Introduction

Collocations are part of the class of frozen expressions and present remarkable linguistic properties for they do not seem to be entirely frozen nor fully compositional. They eliminate the clear-cut distinction between lexicon and syntax, since a collocation is usually associated with a syntactic construction (Lewis 2000, Gross 1975, Mel'čuk *et al.* 1995). However, despite the existing detailed syntactic descriptions (Hausmann 1989; Lexical Functions of Mel'čuk, Clas & Polguère 1995), these generally fail to show the syntactic and morphological variations of collocations. For example, in the LF model, a prototypical LF such as Oper_i is described from the syntactic viewpoint as a light verb having the noun as a second actant,. In both the French ECD (Mel'čuk *et al.* 1995) and in the DiCo (Polguère 2000), the value of the LF is usually described merely with its standard construction. Nevertheless, within the framework of second language acquisition/teaching, it is crucial to place the stress on the verbal collocations' most representative syntactic patterns so as to facilitate the acquisition of the frozen expressions together with their respective syntactic use. For instance, a collocation such as *fonder une analyse sur* ('to base analysis on') can be observed in the active voice but also, and more often, in the passive voice (*l'analyse est fondée sur*) or in the passive pronominal construction (*l'analyse se fonde sur*). With the use of large corpora together with NLP robust tools may also indeed facilitate the acquisition of such syntactic properties and thus could also be integrated in a learner tool. In this paper, we experiment automatic collocation annotation by adapting Polguère's DiCo (Polguère 2000) combined with patterns described with finite state transducers (Silberstein 1999). We first present what kind of syntactic information should be added to the Lexical Function database for the NLP application. We implement this database in a "lexicon-grammar" model and finally we evaluate the results of the automatic annotation in the semantic field of emotions.

1. Using the DiCo database to encode the LF

In the present study, our aim is to build a corpus of collocations in the semantic field of emotions for learners of French as a first or second language... These annotated examples could be related to a lexical database and enable to introduce information about the relative frequency of collocations, the most used syntactic patterns, morphosyntactic properties, etc. At this first stage, we focused only on Verb-Noun patterns such as *faire peur* ('to frighten'), *nourrir de la haine* ('to feel hatred'), *surmonter ses craintes* (lit 'to overcome fears') to assess the feasibility of the method. The Verb-Noun syntactic pattern is the most frequent and the most complex, as there are numerous alternations which can be observed:

- passive voice: *la haine est surmontée* (lit. 'the hatred is overcome').
- pronominal passive constructions: *la peur se lit sur son visage* ('fear can be read on his face').
- pseudo-passive constructions: *une peur difficile à surmonter* ('a fear hard to overcome').

- etc.

Our hypothesis is that the present approach can be easily extended to other verbal collocations. Adjectival collocations are predictably easier to annotate since they are less prone to variations.

We used the DiCo¹ database developed by Alain Polguère and his colleagues (Polguère 2000; Steinlin *et al.* 2004) as a basis for the annotation of our corpus. We then converted the latter into an Excel-like table, as presented in table 1 (we have presented only the fields relevant to our study). We selected nouns belonging to the semantic field of emotions (semantic labels such as ‘sentiment’ (*feeling*), ‘sentiment négatif’ (*negative feeling*), ‘émotion’ (*emotion*), ‘sentiment positif’ (*positive feeling*)) with LFs corresponding to Verb-Noun patterns such as: *provoquer le regret, susciter le regret, faire l’admiration...* and obtained a subset of collocations².

Though the DiCo is probably the principal source and the most detailed database which exists for French collocations, there remain certain imprecisions concerning formal properties of collocates for NLP applications, as we show below. To provide a long term solution dealing with collocations in an NLP context, one should focus in establishing a relation between the value (e.g. *susciter*, CausFunc₁ value for REGRET II.1) and the lexical entry of the LF value. Hence, syntactic and semantic properties would be registered and handled with inheritance mechanisms (Tutin 1997). We shall not develop this idea any further. We shall list only the linguistic properties which must be added to the collocates (the LF values) for our annotation application.

Lexeme	N°	grammatical information	semantic label	gloss	Lexical Function	value	grammatical properties of the LF value
REGRET 'regret'	II.1	nom, masc	sentiment	[Qqch.] causer un R. chez X	CausFunc1	provoquer	__ [ART ~ Loc-in N=X]
REGRET 'regret'	II.1	nom, masc	sentiment	[Qqch.] causer un R. chez X	CausFunc1	susciter	__ [ART ~ Loc-in N=X]
HAINE 'hatred'		nom, fém	sentiment négatif	[Qqch.] causer 'Y] être visé par la H. de X'	CausFunc2	attirer	__ [ART ~ sur N=Y]
HAINE 'hatred'		nom, fém	sentiment négatif	[Qqch.] causer 'Y] être visé par la H. de X'	CausFunc2	canaliser	__ [ART ~ sur N=Y]
ÉMOI 'agitation'		nom, masc, surtout sing	émotion	[Qqch.] rendre l'É. de X apparent	CausManif	trahir	__ [ART ~]
PEINE 'sorrow'	II.1	nom, fém	sentiment négatif	[Qqch.] causer non 'P. commence à exister en X'	CausNonIncepFunc1	épargner	__ [ART ~ à N=X]
ADMIRATION('admiration'	I	nom, fém, pas de pl	sentiment positif	[Y] être l'objet de l'A. de X	Oper2	faire	__ [l'~ de N=X]

Table 1: Records extracted from the DiCo database

- **Lemmas of the LF value**

In the DiCo database, some values are not lemmas but inflected forms. It is the case for adjectival values in particular. For example, Magn(*anxiété*) ['anxiety'] = *folle, absolue* ...['mad, absolute']

This inflectional presentation is obviously useful for pedagogical purposes and everyday use and thus must be preserved. However, the lemma must be added systematically in order to analyse other inflected forms which occur in the texts. As for verbal values, the problem does not concern the latter as they are always presented as lemmas.

¹ The Dicouèbe database developed by Polguère and his colleagues from the DiCo database has a more regular structure and could have been used for this experiment, but it was not yet available when this study began.

² 755 collocations.

- **Morpho-syntactic information of the LF value**

In the DEC and the DICO, information regarding the part of speech of the LF value remains implicit: a given LF applied to a specific part of speech produces a specific part of speech. For example, when **Magn** placed before a verb it produces an adverb, yet when it is applied to a noun it produces an adjective. It gets more complex when the value is not a single word but a phraseme or even an expression which cannot be considered as a lexical unit.

Eng. Magn(*rain*): *cats and dogs*.

Fr. Magn(*pleuvoir*): (colloquial) *comme vache qui pisse* [lit. 'like a pissing cow']

Moreover, phrasal and single units do not behave the same way from the syntactic viewpoint³. Phrasal units have to be detailed according to their syntactic behaviour as is the case for the pronominal status of verbs such as *s'accroître* ('to increase').

To deal with this, two fields should be introduced:

- **a general part of speech information** including the phrasal status of the value, e.g. adverb (*badly*), adverbial "phraseme" (Fr. *par monts et par vaux* ('up hill and down dale')), adverbial expression (e.g. not a lexical unit) (*comme vache qui pisse*).
- **a detailed subtype information of the phrasal unit**. We have not tackled this problem in the present paper and more research has to be carried out on this specific topic.

- **Syntactic information of the LF value**

In the DiCo database, some grammatical information is available for the LF value. For example, for the lexeme REGRET II.1, the CausFunc₁ value has the following properties: __ [ART ~ Loc-in N=X]. It means that the verbal value is followed by a determiner and a complement introduced by a locative preposition (this is the first actant of the key-word). Yet, this grammatical information does not suffice, for it should be broken down further into several fields for it is too concise.

First of all, the **deep syntactic relation** between the value and the key-word should be indicated. Yet, this information remains implicit in the ECD model, for the syntactic relation is taken into account in the LF definition, and can easily be acquired automatically⁴. In addition, the **surface syntactic relation** should also be clearly indicated (with values such as Object, Prepositional, etc), with the specific preposition(s) used. Then, **constraints on determiners** should appear in a specific field, with a consistent description (definite or indefinite determiners, lack of determiners, ..). The **number** of the value has also to be filled in a specific field. Constraints on other **actants** (depending on the verb or the key word) should also be described.

For our application, we just replaced the DiCo's grammatical information with several fields (highlighted in the table of the following section) that we detail in the following section⁵.

2. A lexicon-grammar based method to extract and annotate collocations

In order to annotate the collocations in our corpora, we use a shallow analysis based on finite state transducers, more particularly on a methodology inspired by Maurice Gross's lexicon-grammar (Gross 1975). In this method, data registered in an Excel-like table are related with metagraphs (metatransducers) which are transducers (Roche 1993, Silberztein 1999, Laporte 2005). The tool used for this task is INTEx⁶

³ For example, use of auxiliaries, use of modifiers, etc.

⁴ Here is the list of syntagmatic LFs where the key-word is the second actant of the verbal value: AntiReali, Caus(i)Manif, Caus(i)NonManif, CausContFunci, CausFacti, CausFunci, CausOperi, CausPredMinus, CausPredPlus, Conv21Manif, FinOperi, LiquFunci, Magn--temp.OperI2, NonPermIFact0, NonPermIManif, Operi, PredAblei, Reali.

⁵ We did not introduce constraints on the actant of the value, because this field was not essential in our application.

⁶ Another tool called UNITEx (very similar to Intex) implements the same functionality (Laporte 2005). This functionality was first implemented in Intex.

(Silberztein 1993, Silbertzein 1999). Each cell in the table is related to a transition in the metatransducer and a large finite state transducer is generated from these data. The table below shows examples of entries of nouns of emotion. Due to the limited amount of space, we have only presented a simplified version of the table.

A = LEXEME	B= N°	C=lemma	C= semantic tag	D= LF	E=LF value	F= cat. value	G = Deep Synt. Rel.	H = Surf. Synt. Rel.	I = O Det?	J= type of DET	L=si ng	M= plur
ADMIRATION 'admiration'		admira- tion	sentiment positif	IncepOper23	emporter	V	II	OBJ	-	:le	+	-
HONTE 'shame'	1.1	honte	sentiment positif	NonPerm1 Manif	cache	V	II	OBJ	-	:det	+	+
HONTE 'shame'	1.1	honte	sentiment négatif	Oper1	avoir	V	II	OBJ	+	:det	+	-

Table 2: Examples of records (derived from the DiCo database) used by the metagraph

We shall focus on the first entry in our table, namely the collocation *emporter l'admiration* ('to be the admiration of'). The value which can be analysed as the value of an IncepOper₂₃ LF, is a verb (F column) and is related with the key-word *admiration* with the help of a deep syntactic second actant relation (G column) realised as an object surface relation (H column). The noun cannot be introduced without a determiner (value "-" in the I column) but with the help of a definite article ("le") (J column). The metagraph exploits the recorded data in the cells of the lexical entry and consequently produces a transducer which includes all the data recorded in the tables. For reasons of readability, the following is a simplified version of a metagraph showing how these values are used.

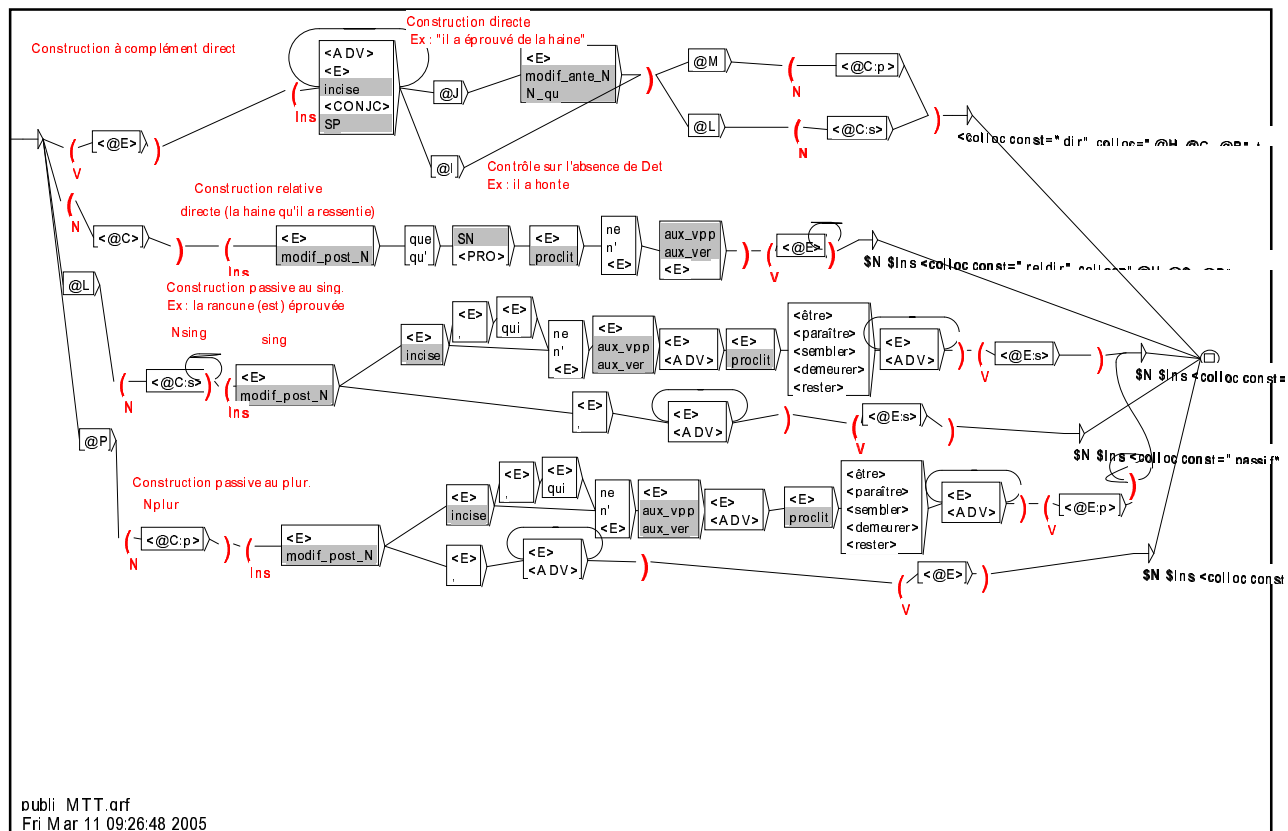


Fig. 1: The V-N collocation metagraph

The first path, on the top of the graph, shows how a collocation such as *emporter l'admiration* can be analysed in the active voice and how it has been annotated. The first transition (the variable @E) must be a

verb in the table (column E) (the “< >” indicate a lemmatized form), possibly followed by a modifier such as an adverb or a prepositional phrase. The graph then assesses whether a determiner can be introduced and consequently chooses which one (column J), and then examines whether the noun chosen can be a plural or not (columns L and M). Thus, a collocation like *emporter l’admiration* can be analysed, but *emporter les admirations* will be rejected, since this collocation does not allow for a plural (column M).

The other paths on the graph are used for other syntactic constructions (relative clause, passive, etc.). Once the collocation is analysed, it is annotated with XML tags (see the last empty transition). Examples of annotated collocations are presented below.

```
"Et je vous jure qu'au fracas du coup de fusil que je n'attendais point,
j'<colloc const="dir" colloc="avoir-angoisse_1" type_verbe="Oper1"
type_nom="sentiment négatif" lemme_nom="angoisse">eus<colloc> une telle
angoisse du coeur, de l'âme et du corps..
['I swear to you that by the crush of a rifle blow that I did not wait, I had much fear
in heart, soul ad body']

C'était peut-être la première fois qu'un désir <colloc const="passif"
colloc="manifester_désir_1" type_verbe="verbalement Caus1Manif"
type_nom="sentiment" lemme_nom="désir">manifesté</colloc> par le
colonel eût obtenu l'approbation ...
['Perhaps it was the first time that the desire shown by the colonel got approval']

Vous avez expliqué l'intérêt que pouvait avoir Tomaso à menacer monsieur
Barricini au nom d'un bandit redoutable, par le désir qu'il <colloc
const="reldir" colloc="avoir-désir_1" type_verbe="Oper1"
type_nom="sentiment" lemme_nom="désir">avait</colloc> de conserver à son
frère Théodore ...
['You have shown the interest Tomaso could have in threatening Mr Barricini, in the name
of a formidable gangster, for the mere pleasure that he had to keep for the sake of his
brother Théodore.']
```

Fig. 2: Examples of annotated collocations

The verbal collocate is annotated with the help of the tag <collocate>. Several attributes give information about the verbal collocation:

- **const** indicates the type of syntactic construction : direct object (“dir”), passive construction (“passive”), relative clause (“rel”), ...
- **colloc** gives the lemmatized form of the collocation.
- **type_verbe** indicates the Lexical Function.
- **type_nom** gives the semantic type of the noun.
- **lemme_nom** gives the lemma of the noun.

These tags can be used for looking for specific examples in the corpus, in order to observe the syntactic and semantic behaviour of collocations.

3. Evaluation of the methodology and perspectives

Our evaluation was brought about through a set of novels and short stories of French classical authors, often studied at the level of “collège” (11-15 years of secondary school), and freely available. We would like to stress on the fact that our corpus should be used for pedagogical purposes, both for college pupils and for FFL students. The corpus includes around 233 000 words⁷.

⁷ The corpus includes the following texts:

- *Les Lettres de mon Moulin*, Alphonse Daudet.
- *Le petit chose*, Alphonse Daudet.
- *La petite fadette*, George Sand.
- *Colomba*, Prosper Mérimée.

Our lexicon-grammar is based on an extract of Polguère's DiCo database of nouns of emotions⁸ and the corresponding transducers have been applied to the present corpus (see previous section). The output has been analysed with both the precision and the recall rates according to the data available. In general, the results are encouraging⁹.

The precision rate is quite satisfactory since we achieve 90% correct analyses. Several errors observed are due to polysemy and even the more sophisticated techniques, such as dependency parsers, could not easily avoid such errors. For example, the polysemous collocation *avoir de la peine* (lit. 'to have sorrow' or 'to encounter a hardship') is encountered in the text and it has been analysed erroneously.

Les policemen [avaient beaucoup de peine](#) à contenir le populaire, et à mesure que s'avançait l'heure à laquelle devait arriver Phileas Fogg ... (*Le tour du monde en 80 jours*, Jules Verne)
[‘The policemen struggled to contain the masses and as the hour approached in which Phileas Fogg was expected to arrive’]

The recall rate is approximately 86,2 %. Most errors are due to complex syntactic patterns such as coordination, interrogative structure or gapping, which have not been accounted for in our graphs, but could be, as shown in the following example:

-- Mais pourquoi [as-tu une si grosse peine](#) ?
[‘but why are you so troubled ?’]

Our method, despite its simplicity, provides satisfactory results, and could probably be easily extended to other collocational patterns.

In the near future, we plan to explore the issue further following two directions of research:

- Firstly, to compare our pilot method to more ambitious methods using dependency grammars. A first experiment of dependency grammars using Xerox XIP dependency parser (Aït-Mokhtar *et al.* 2002) has been attempted in our team (Haddara 2004). However, it did not show a significant pre-eminence over finite-state techniques, even if the lexical data used were rather contrasting.
- Secondly, to extend this technique to other collocational patterns and provide a fully annotated corpus with the 50 most frequent nouns of emotion, to be exploited for pedagogical purposes.

Acknowledgements

Special thanks to Alain Polguère who offered me the DiCo database for this experiment.

- *Le tour du Monde en 80 jours*, Jules Verne.

Most texts can be downloaded from the following URL : <http://abu.cnam.fr/>.

⁸ Here are the nouns included at the moment in the DiCo database: *admiration, angoisse, appréhension, aversion, coup de foudre, crainte, déception, dégoût, dépit, désir, effroi, émoi, extase, faveur, fierté, gratitude, haine, honte, hostilité, indignation, irritation, mécontentement, peine, rancune, regret, remords, répugnance, répulsion, ressentiment, sensibilité, vexation.*

⁹ the precision rate has been evaluated on 118 results.

References

- Ait-Mokhtar S., Chanod J-P., Roux C. (2002), Robustness beyond shallowness: incremental dependency parsing. *Special issue o the NLE Journal*
- Bolshakov I.A., Gelbukh A.. Heuristics-based replenishment of collocation databases. In: E. Ranchhold, N. J. Mamede (Eds.) *Advances in Natural Language Processing (PorTAL-2002)*. Lecture Notes in Computer Science, N 2389, Springer-Verlag, p. 25–32.
- Evert S., Heid U., Spranger K. (2004) Identifying Morphosyntactic Preferences in Collocations, in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne.
- Gross M. (1975) *Méthode en syntaxe*. Paris: Hermann.
- Haddara M. (2004), *Extraction automatique des collocations verbales du lexique transdisciplinaire des écrits scientifiques par une approche de type dépendance: étude de faisabilité*, Mémoire de DEA, Université Grenoble3-Stendhal.
- Laporte E. (2005) Graphes paramétrés et lexique-grammaire, *Interface lexique-grammaire et lexiques syntaxiques et sémantiques* (journée ATALA du 12 mars 2005).
- Lewis , M. (2000) (ed.) *Teaching Collocations*. Boston: LTP.
- Mel'čuk I (1998). Collocations and Lexical Functions. In A. P. Cowie (ed.), *Phraseology. Theory, Analysis and Applications*. Oxford: Clarendon Press.
- Mel'čuk I., Clas A., Polguère A.(1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain: Duculot.
- Polguère A. (2000), Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French. In *Proceedings of EURALEX 2000*, Stuttgart.
- Roche E. (1993). Une représentation par automate fini des textes et des propriétés transformationnelles des verbes. In *Lingvisticae Investigationes XVII:1* (pp. 189-222). John Benjamins Publishing Company: Amsterdam/Philadelphia.
- Silberztein (1999) "INTEX: a Finite State Transducer toolbox", in *Theoretical Computer Science* #231:1, Elsevier Science.
- Steinlin J., Kahane S., Polguère A., El Ghali A. (2004), De l'article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux, *Actes d' Euralex*Lorient, 6 au 10 juillet 2004.
- Tutin, Agnès (2004), Pour une modélisation dynamique des collocations dans les textes, *Actes d' Euralex* Lorient, 6 au 10 juillet 2004.