Annotation of multiword expressions in French

Agnès Tutin
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
agnes.tutin@u-grenoble3.fr

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes
LIG
F-38040 Grenoble
emmanuelle.esperanca-rodier@imag.fr

Manolo Iborra
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
iborra.manolo@gmail.com

Justine Reverdy
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
Justine.reverdy @e.u-grenoble3.fr

Keywords: Multiword expressions – Annotation – Typology of multiword expressions

Abstract

This paper presents an experiment of annotation of MWEs in French. The corpus used is made of several genres (news, novel, scientific report, film subtitles) and includes a rich annotation scheme including several kinds of MWEs from collocations to routines and full phrasemes. The annotation is performed semi-automatically with finite-state transducers. The inter-annotator agreement score shows that the annotation is quite consistent but the difficulty of the task relies heavily on the textual genre: literary texts are harder to annotate than scientific reports. Besides, two types of categories are difficult to differentiate, collocations and full phrasemes.

1. INTRODUCTION

This paper presents an experiment of multiword expression annotation on the French part of a French-English bilingual corpus. Our aim is to achieve three goals: a) building a corpus-based and robust typology of MWEs; b) providing a basis for linguistic studies on MWEs, especially in relation to diverse textual genres; c) building a corpus of evaluation for Machine Translation (MT) tasks, and especially statistical machine translation (SMT) tasks (e.g. Potet *et al.* 2012).

Every scholar working on MWEs knows that defining clearly different types of MWEs is a complex task. But we think that confronting concrete examples will help to refine typologies of MWEs, and enable to better understand how they work.

This will also help to explore the most frequent MWEs, especially according to the specific genres, in order to answer questions such as the following ones:

- Are collocations really more frequent in general expressions than in idiomatic expressions?
- Are true idiomatic expressions, such as *to break the ice*, more frequent in spoken genres?
- Regarding syntax now, we would like to observe in more detail syntactic properties of MWEs. Are real MWEs highly variable, as suggested by Moon, or not?

Considering now practical goals, we know that there are few annotated corpora with MWEs, especially for French. There are two small corpora with nouns and MWE

adverbs (Laporte *et al.* 2008a; Laporte *et al*. 2008b), but these corpora do not include any typology of expressions. The French Treebank (Abeillé *et al*. 2003) includes several kinds of MWEs including verbs, but only on contiguous MWEs such as *faire part* but no discontinuous expressions, e.g. *prendre ce problème en compte*. In English too, there are not many resources. One of the most interesting ones is undoubtedly Schneider *et al*. (2014) social web corpus with MWE annotations, which distinguishes between strong and weak MWEs, but does not include any fine-grained typology.

We obviously need reference corpora for evaluation purposes in MT and especially statistical MT applications, especially to evaluate which MWEs are more difficult to translate. Our hypothesis (which has been partially confirmed) is that contiguous expressions are easier to translate.

But the annotation of MWEs is not a straightforward task, mainly due to three types of problems:

1. Deciding whether an expression is a MWE expression or not. This is easy for compounds such as *as long as* but complex for collocations and routines.
2. Delimiting the expression is also complex: for example, do determiners belong or not to the expression? In our annotation scheme, we decided to include fixed determiners, e.g. *la* in *il fait la fête* (lit. 'he makes the party') but not in examples such as e*lle donne un cours* ('she gives a class') where we could have any determiner.
3. Deciding the type of the expression is also complex and needs clear-cut criteria.

We will first present the corpora and the annotation scheme, and also the semi-automatic annotation process. We will then present the quantitative and qualitative results of the evaluation of the annotation, including inter-annotator agreement.

2. CORPORA AND ANNOTATION SCHEME

Our project is to build a corpus of about 45,000 tokens, freely available, of the French part of a bilingual corpus. This paper only focuses on the annotation of the French part. The next step will be the bilingual alignment.

Our corpus includes several genres of texts, namely:

- a scientific writing (*BAF Citi 1* from Baf corpus, about 14,500 tokens),
- news (journalese, from several news corpora of the WMT evaluation campaign (from 2006 to 2010), almost 12,300 tokens),
- subtitles of a French film, *Amélie Poulain* (9,900 tokens), and
- an extract of Zola's *Thérèse Raquin* (7,260 tokens).

Several MWEs are included and a semi-automatic annotation using Nooj system has been carried out.

2.1. Typology of MWEs

In our annotation scheme, we include a typology of MWEs inspired by Granger & Paquot (2008), Ulrich Heid (2008), Mel'čuk (2013) and a previous work on this topic (Tutin 2010).

Following Mel'čuk, we distinguish « Full phrasemes » which are non-compositional expressions of nouns, adjectives, verbs and adverbs (including compounds) e.g. *dead end* or *to take into account* from collocations, which are

compositional but are difficult to predict e.g. a *heavy smoker* in English while we have a *« big » smoker, gros fumeur* in French (e.g. Tutin & Grossmann, 2002).

Then, we have a specific category for functional words such as prepositions, conjunctions, determiners, pronouns and discourse and negation adverbs, e.g. *on the one hand, in front of, insofar as, a large number of*.

We also use the category of pragmatemes, which is very frequent in dialogues, for expression with a specific pragmatic function e.g. *You're welcome, See you later*.

Furthermore, we include complex terms, specific to a field, named entities e.g. *Université Stendhal, Laboratoire d'Informatique de Grenoble*, and routine formulae, e.g. *As previously said, force est de constater …* which are prefabricated verbal expressions frequent in a specific genre.

2.2. Annotation scheme

Our annotation scheme is a very simple surface annotation. We are aware that a stand-off annotation would be more suited for our purpose, but such an annotation scheme requires complex annotation tools and the annotation is not very convenient for linguists or people who are not familiar with NLP. Nevertheless, a stand-off annotation could be used at the end of the process.

The annotation scheme includes several features which are illustrated fig. 1 with the help of an example:

- An **identifier**, e.g. id="23" on each element of the MWE.
- The **type of MWE** (e.g. full phraseme, collocation, functional word…). In the example, *pris en compte* is a full phraseme.
- **Syntactic categories of the expression** (here the MWE is a verb) and of **the parts of the expression** (Verb + Prep + Noun).

```
Nous avons <epl id="23" type="fphraseme" catepl="verb" catw="verb" lemma="prendre_en_compte">pris</epl> ce problème <epl id="23" catw="prep">en</epl> <epl id="23" catw="noun">compte</epl>
```

Fig. 1 : Annotation of *Nous avons pris ce problème en compte*

Besides, overlapping expressions can be annotated. As we can see in the examples below, partial overlapping such as *pay close attention (pay attention + close attention)* as well as inclusions such as *au minimum* which is included in the collocation *réduire au minimum*, can be annotated with several identifiers, types and syntactic categories.

1) Unlike many theorists, he **[paid$_1$] [close$_2$] [attention$_{1+2}$]** to a broad range of experimental evidence…
2) Afin de **[réduire$_1$] [au$_{1+2}$] [minimum$_{1+2}$]** cet effort …

3. ANNOTATION PROCESS

The annotation process is a semi-automatic process. It uses a finite-state tool called Nooj, developed by M. Silberztein (2013), known to be very suited for the treatment of MWEs. Nooj uses a core dictionary of MWEs extracted from several resources:

- The most frequent MWEs from the French Treebank (Abeillé *et al.*, 2003), where MWEs are decomposed;
- Other dictionaries of MWEs such as the *Dictionnaire Electronique des Mots* (Dubois & Dubois Charlier, 2010), Wiktionary or the DELAC (Courtois *et al.* 1997).

Here is an example on Fig. 2 of the annotation process with the finite-state tool.

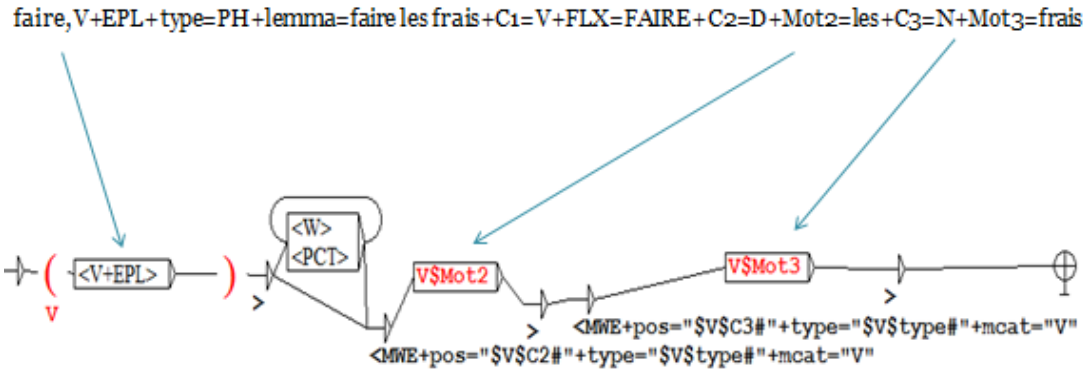

Fig. 2: Annotation of MWEs with the NooJ processing system

In the dictionary, the MWE *faire les frais* is decomposed as follow:

- The elements of the expressions: for example, verb, determiner and noun for the expressions *faire les frais.*
- The type of MWE: here, it is a « full phraseme » since the meaning is not compositional.

Then the parts of the expression are associated to elements of a finite state graph, in order to enable the annotation in texts.

In order to achieve the annotation process, a core lexicon of about 5,000 MWEs is used. The semi-automatic annotation is thus performed and checked on concordances. It is then completed with manual annotation by two linguists with an XML editor (Oxygen). About 35% to 50% of MWEs, depending on the kind of text, are semi-automatically annotated.

The automatic annotation is quite good with functional words and frequent phrasemes. However, it has a weak coverage of collocations, pragmatemes and routines. In order to get better results, we could use a complex term extractor for technical terms and a better named entity recognition system.

In Fig. 3; we can see an example of the annotated expressions with a style sheet of subtitles of the film *Amélie Poulain*.

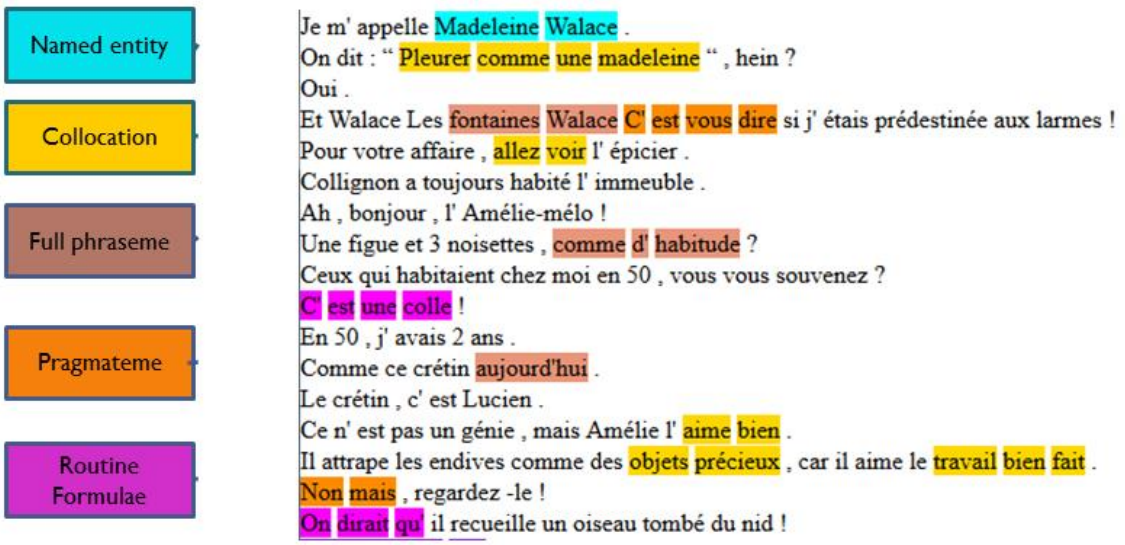

Fig 3.: An example of annotation of MWEs: *Amélie Poulain*'s subtitles

Named entities with proper nouns are in light blue e.g. *Madeleine Walace*, full phrasemes e.g. *fontaines Walace* or *comme d'habitude* are in brown. In orange, we can find pragmatemes such as *non mais* or *C'est vous dire* which are typical of dialogues, in yellow, collocations e.g. *pleurer comme une Madeleine* or *aime bien* and finally in purple routines e.g. *C'est une colle !*.

Once the manual annotation performed by the two linguists according to the typology already described, we have evaluated the annotated expressions in texts.

4. QUANTITATIVE AND QUALITATIVE RESULTS

We now come to the evaluation of the annotation process. The objectives of this evaluation are to improve the typology by splitting or merging complex categories, or by adding criteria for the annotation process.

For that, an experiment was carried out on two different extracts of our corpus:

1) the literary text (*Thérèse Raquin*, a Zola's novel) around 2,000 words

2) the scientific report (CITI 1) around 2,000 words

Two annotators, two linguists familiar with the issue of MWEs, were involved in that task. We mainly focused on the type of MWE and not on grammatical issues (e.g. the grammatical category of the MWE).

4.1. Quantitative Results

The first parameter to examine is the proportion of MWEs in texts. On Fig. 4, we can see the percentage of words which belong to a MWE according to each annotator and common to both annotators.

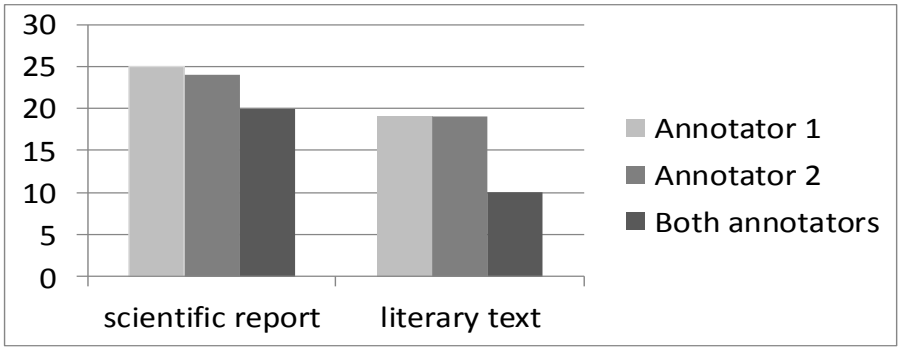

Fig. 4: % of words belonging to a MWE

For the scientific report, we have more than 20% of the words from the text which are considered as belonging to a MWE and the agreement (in %) is good for this text. But for the literary text, we have less than 20% of the words and it falls to 10% for both annotators. Therefore, we observe more MWEs in the scientific text, probably due to the formulaic style of this genre, than in the literary extract, which was expected. We also notice that the agreement (in %) on what is a MWE is less good for the extract of the novel.

We observe the same tendencies on the agreement concerning the type of MWE. We computed the Fleiss Kappa score to compare the annotation of the type of MWE (only for MWEs selected by both annotators) (see. Fig 5).

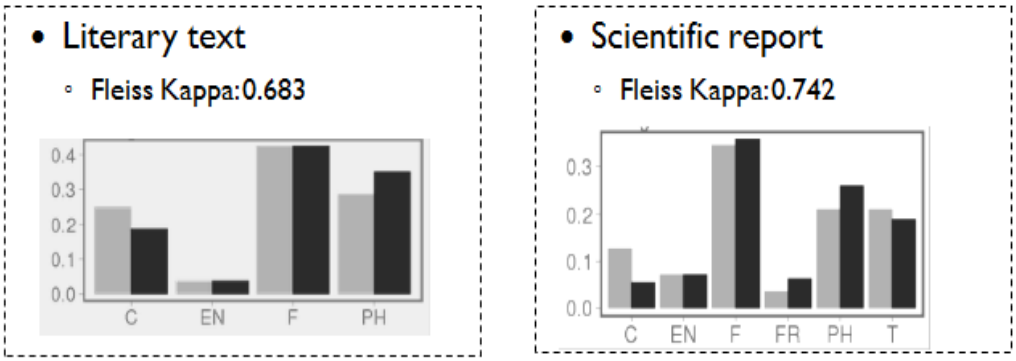

Fig. 5: Inter-annotator agreement for types of MWEs

We can observe that the results are not bad: we have a kappa of 0.683 for the literary text and a kappa of 0.741 for the scientific report. There is a good agreement for named entities (EN) and functional words (F) but this is more complicated for collocations (F) and phrasemes (PH), which is not very surprising.

4.2. Qualitative Results

If we now have a look at the qualitative results, we can observe that, as expected, the different kinds of MWEs depend on the genres. This is quite obvious and expected. For example, we do not have any terms (T) or routine formulae (FR) in the literary text.

The disagreements between the two annotators often concern collocations or full phrasemes for several reasons.

Some collocations can have some specific syntactic properties, for example a lack of determiner, which is close from full phrasemes; e.g. in *tirer profit* (’take advantage’) even though they are compositional.

A second problem occurs with « complex terms » which are very close to full phrasemes: the decision is determined by the genre of the text: is *emplois salariés* a term or a phraseme? Complex terms generally seem more compositional than phrasemes.

Binomials such as *jour et nuit* (‘night and day’), *pur et simple* (‘pure and simple’) also raise specific problems: should they be considered as a type of collocation or a type of phraseme? This is not a very frequent type of MWEs and creating a specific category would seem artificial.

Finally, nominal hyperonymic collocations (non predicative collocations) such as *cuillère à soupe* (‘tablespoon’), *boîte à outils* (‘toolbox’) also raise interesting problems. They are partially compositional but refer to a single referent and have an intermediate status.

These problems should be examined thoroughly in order to refine, with the help of clear-cut examples, the annotation guidelines.

5. CONCLUSION AND PERSPECTIVES

Annotating MWEs is a stimulating task (it is a deep reflection about MWEs) but this is also a feasible task, even though this is a complex one. The complexity of the task depends on the genre of the text and on the types of MWEs. This is especially difficult for some categories for which we need to develop more detailed criteria. For collocations and phrasemes, the issue of compositionality should be considered more in detail with analyze of complex examples described in the guidelines.

We also saw that this kind of annotation is complex and it seems necessary to carry out a double annotation, and with more annotators in case of disagreement. Therefore the annotation task requires the use of dedicated tools that allow collaborative annotation in order to maintain a good inter-annotator agreement. To our knowledge, such tools are not available yet, and one of our future works will be to characterize specifications for implementing an annotation tool. In addition, the automatic annotation needs to be developed, and this can be done incrementally with annotated corpora.

Observation on MWEs in real texts, especially on challenging cases such as routines or atypical collocations, is necessary to make progress in the field of phraseology and to propose well-suited encoding schemes.

References

ABEILLÉ, A., CLÉMENT, L. AND L. TOUSSENEL, 2003. Building a treebank for French. In: Treebanks. Springer Netherlands. pp. 165-187.

COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, M., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, SILBERZTEIN, M. AND VIVÈS, R., 1997. Dictionnaire électronique DELAC : les mots composés binaires. Technical Report 56, University Paris 7, LADL.

DUBOIS, J. AND DUBOIS-CHARLIER, F. 2010. La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, 179(3, pp 31-56.

GRANGER, S. AND PAQUOT, M. 2008. Disentangling the phraseological web. In Granger, S. & Meunier, F. *Phraseology: An interdisciplinary perspective.* Amsterdam/Philadelphia: John Benjamins.

HEID, U. 2008. Computational phraseology. An overview. In: S. Granger and F. Meunier, *Phraseology. An interdisciplinary perspective*. Amsterdam: Benjamins. pp 337-360.

LAPORTE, E., NAKAMURA, T., AND VOYATZI, S. 2008a. A French corpus annotated for multiword nouns. *In Language Resources and Evaluation Conference*. Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30.

LAPORTE, E., NAKAMURA, T., AND VOYATZI,. 2008b. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC)*. Linguistic Annotation Workshop. pp. 48-51.

MEL'ČUK, I. 2013. Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie*, 102, pp. 129-149.

MOON, R. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford:Oxford University Press.

POTET, M., ESPERANÇA-RODIER, E., BESACIER, L., BLANCHON, H. 2012. Collection of a Large Database of French-English SMT Output Corrections, *(LREC 2012)*, Istambul, 2012, 21-27 mai.

SCHNEIDER, N., ONUFFER, S., KAZOUR, N., DANCHIK, E., MORDOWANEC, M. T., CONRAD, H., AND SMITH, N. A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC*. Reykjavík, Iceland.

SILBERZTEIN, M., KHURSHUDIAN, V., AND DONABÉDIAN, A., 2013. *Formalizing Natural Languages with Noo*. Cambridge: Cambridge Scholar Press.

TUTIN, A. 2010. *Sens et combinatoire lexicale: de la langue au discours*. Unpublished Dossier en vue de l'habilitation à diriger des recherches). Grenoble: Université Stendhal.

TUTIN, A., AND GROSSMANN, F. 2002. Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, *7*(1), pp. 7-25.