The Difficult Identification of Multiworld Expressions: From Decision Criteria to Annotated Corpora

Agnès Tutin[1(✉)] and Emmanuelle Esperança-Rodier[2(✉)]

[1] Université Grenoble Alpes, LIDILEM, 38000 Grenoble, France
agnes.tutin@univ-grenoble-alpes.fr
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP (Institute of Engineering Univ. Grenoble Alpes), LIG, 38000 Grenoble, France
emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr

Abstract. Multiword expressions (MWEs) are known to be widespread in most languages, but how can we actually identify them in texts? How can we account for the diverse nature of these multiword expressions with a consistent typology? This paper aims to address this topic by carrying out two tasks: (1) conducting a small survey with expert linguists on several types of MWEs; (2) building an annotated corpus which includes several genres incorporating a large panel of multiword expressions. The results show that experts can reach a consensus and that the annotation task can be performed satisfactorily, as long as the typology is not overly complex and clear guidelines are provided. Finally, a useful application of the annotated corpus for statistical machine translation is presented, showing significant differences among multiword expressions considered in the study.

Keywords: Typology of multiword expressions · Annotated corpora · Statistical machine-translation

1 Introduction

It is now common to consider that language is largely prefabricated and that phraseology plays a major role. “Tout est idiomatique dans la langue” [all is idiomatic in language] claims Hausmann (1997). Sinclair’s (1991) idiom principle and interest in formulaic language (Wray 2009) seem to attract a growing number of followers. Many consider that a large part of texts are prefabricated, if not the majority (for example, Altenberg 1998). But what exactly is the proportion of MWEs in corpora? Where do multiword expressions (now MWEs) begin and end? Can they be easily delineated and can we apply consistent criteria easily to develop a typology?

This paper aims to address these issues. Beyond the convenient prototypes presented in the articles (*heavy smoker* and *spill the beans*), it seemed essential to us to better understand the phraseological phenomenon in all its diversity. In this perspective, two tasks were carried out in order to clarify the issue. The first was a small survey with expert linguists to measure the degree of agreement among them to classify a set of expressions using the Explanatory and Combinatorial Lexicology model (Mel’čuk

G. Corpas Pastor and R. Mitkov (Eds.): Europhras 2019, LNAI 11755, pp. 404–416, 2019.
https://doi.org/10.1007/978-3-030-30135-4_29

et al. 1995; 2012). The second one was an annotation task, based on a typology of MWEs that we consider "realistic".

This work had several objectives: (1) determining whether or not there were prototypes on which everyone could agree; (2) refining the delimitation criteria among MWEs and proposing "realistic" criteria for linguists and NLP scholars; (3) providing an annotated corpus with MWEs in order to illustrate the phenomena for linguistics, NLP and lexicography.

2 A Small Expert Survey: To What Extent Do Linguists Agree on a MWE Typology?

Multiword expressions, which we define here as lexical bundles perceived as preconstructed, are a very diverse class of elements (Cowie 1998; Granger and Paquot 2008; Heid 2008; Mel'čuk 2012). However, for linguistic and NLP applications, we think it is essential to propose a typology which takes into account the diversity of this heterogeneous group of elements.

Nevertheless, while this variety is essential (pragmatic expressions such as *see you later* have little in common with grammatical multiword expressions such as *insofar as*), how far shall we go in the typology of MWEs? It is advisable that such a typology should be both detailed enough but not too complex, on the one hand, and consistent among linguists, on the other hand.

Through a small survey with expert linguists, we wanted to evaluate to what extent a consensus could be reached among specialists. Another aim of this survey was to highlight prototypes among MWE classes.

2.1 The Typology, Experts, and Questionnaire

For this survey, the typology used the Explanatory and Combinatorial Lexicology (now ECL) (Mel'čuk *et al*. 1995, 2012). This model was chosen because it is, in our opinion, one of the most mature and detailed approaches to MWEs. It also has the merit of having been applied to a large subset of expressions treated in the volumes of the Explanatory and Combinatorial Dictionary (Mel'čuk et al. 1992, 1994, 1996, 1998) and more recently, as part of the electronic dictionary *Réseau Lexical du Français* (Polguère 2014).

The 7 consulted experts[1] are all linguists experienced in phraseological issues and familiar or experts in the ECL model. Three are non-native French speakers, but all have an excellent level of French or are perfectly bilingual.

In order to simplify the task, the survey did not cover all types of MWEs, but only idioms (semi or weak idioms), collocations or free expressions in French. The definitions of these elements, taken from Mel'čuk (2013[2]), were provided to the experts.

[1] Special thanks to (in the alphabetic order): Margarita Alonso Ramos, Cristelle Cavalla, Francis Grossmann, Véronika Lux, Salah Mejri, Igor Mel'čuk, Alain Polguère.

[2] See Mel'čuk (2012) for definitions in English.

They were asked to determine the types of 23 MWEs in French, all taken from the Zola's novel *Thérèse Raquin* and presented in a large context, among the following options:

1. **Full idiom** (e.g., *'by heart'*, *'black sheep'*). The meaning of the idiom does not include any meaning of its components.
2. **Semi-idiom**, for example *'sea anemone'*. The idiom includes the meaning of one element but not the other.
   or
   weak idiom, for example *'barbed wire'*. It includes the meaning of the components but also an additional meaning. The two types were merged in our survey because of their semantic proximity.
3. **Collocations**, such as *seriously injured,* or *heavy smoker*, which are lexical phrases with a combinatorial restriction.
4. **Other cases**, with possible comments from the authors.

These different types of MWEs can easily be classified according to their "degree of frozenness" (Fraser 1970; Gross 1996; Moon 1998; Granger and Paquot 2008), from the most compositional (free phrases) to the least compositional or more fixed (full idioms), as shown in the diagram below (Fig. 1).

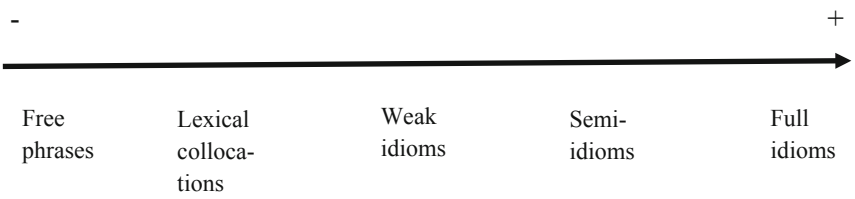

Fig. 1. Continuum scale among MWEs

The survey was conducted by email sent to the 7 experts, with a list of 23 expressions[3] for which, the experts had to decide the type. Table 1 presents an excerpt of this survey with some examples:

2.2 Analysis of the Results

For lack of space, we will not go into detail here about the analysis of the results, but we observed that the distinction between weak/semi-idioms and strong idioms was particularly hard to make and this fine-grained difference did not reach a very satisfactory agreement. When merging all types of idioms (full, weak and semi idioms), the

[3] The complete list of expressions is: *becs de gaz, vert bouteille, à angle droit, beaux jours, en face de, le long de, tablier de travail, coup d'œil, lampes funéraires, clair-obscur, de haut en bas, de l'autre côté, sans doute, en forme de, d'ordinaire, taches de rousseur, un peu de, vieille dame, en même temps, pendant ce temps, au fond de.*

Table 1. Excerpts of the questionnaire sent to the expert linguists

MWEs	Context	Full idiom	Weak idiom or semi-idiom	Collocation	Other (comment)
bec de gaz	Le soir**, trois becs de gaz ['gaslight'],** enfermés dans des lanternes lourdes et carrées, éclairent le passage				
Vert bouteille	Il y a quelques années, en face de cette marchande, se trouvait une boutique dont les boiseries d'un **vert bouteille ['bottle green']** suaient l'humidité par toutes leurs fentes				

results appear to be more satisfactory. Given the small number of expressions analyzed, we did not use inter-annotator measures commonly used in NLP. For analyzing the results, we distinguished three main types of values to evaluate the agreement between experts: a very good agreement (above 70% agreement), a fairly good agreement (above 50%) and a bad agreement (below 50%). In Fig. 2, we set out the values distributed according to this classification, with the predominant type.

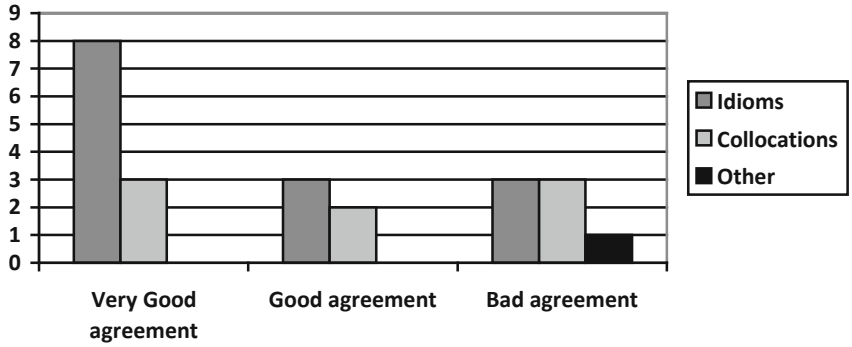

Fig. 2. Expert judgments on the type of MWEs

In line with our expectations, the less consensual types have a higher proportion of adverbs (*pendant ce temps* 'meanwhile' lit. 'during that time'; *sans doute* 'probably' lit. 'without any doubt') and prepositions (*au fond de* 'at the bottom of'). This type of grammatical expression seems to be far more difficult to analyze semantically (grammatical MWEs are really poor cousins of phraseology and have not been much studied). One expert even wishes to treat grammatical MWEs in a specific way, different from other classes of words:

> *"For many prepositional phrases, I would like to consider them as full idioms, even if they include the meaning of one or several of their components."*

These hesitations were confirmed by an evaluation performed on the annotation of collocation (see Sect. 3), which led us to use a specific class for grammatical words. There was also a weak agreement between expressions such as *pays développé* ['developed country'], which can be considered according to the point of view as an idiom (a "developed country" is a type of country) or as a collocation. Fortunately, the agreement was very satisfactory for several MWEs, such as *coup d'œil* ['quick look', lit. 'stroke of eye'], which are clearly non-compositional, or for some expressions which have a very unusual syntactic structure such as *çà et là* ['here and there'] and *clair-obscur* ['half-light', lit. 'light-dark'].

This first survey of experts seems to validate the notion of a continuum along the degree of frozenness. It also shows that a typology with too fine-grained semantic distinctions (full vs. weak idioms) appears difficult to exploit, even with experts. By tightening the types, we obtain a fairly satisfactory degree of agreement, which can be improved by providing more discriminating criteria for grammatical words and adverbs. This survey also confirmed that an annotation of MWEs was a feasible task as long as the typology uses clear and not too complex criteria.

3 Annotation of Multiword Expressions

In conjunction with the expert survey, we conducted a study on the annotation of MWEs in several types of texts. However, although corpora including MWEs are useful, even crucial for several kinds of applications, they still are a scarce resource in NLP and linguistics. This situation is easily understandable since this kind of corpus is challenging to develop as we will see.

To our knowledge, concerning French there are still very few corpora of this type: a small corpus annotated with multiword nouns (Laporte *et al.* 2008a) and another one with multiword adverbs (Laporte *et al*. 2008b), which only annotated one kind of parts of speech. The French Treebank (Abeillé *et al*. 2003) includes several kinds of MWEs but only contiguous ones[4]. More interesting is the resource being developed in the framework of the Parseme-fr ANR project[5] (Candito *et al.* 2017) which aims to provide a large corpus with verbal MWEs, but it only includes at this time[6], verbal MWEs and excludes collocations (except light verb constructions) and pragmatic MWEs, contrary to our annotation scheme, which are very interesting phenomena. In English, more corpora are available. Among them, it is worth mentioning Schneider *et al.* (2014) a

[4] It means that discontinuous verbal expressions, for example in *j'ai pris cela en compte* ['I took that into account'] are not included.

[5] http://parsemefr.lif.univ-mrs.fr/.

[6] A larger set of MWEs is being annotated in the framework of this project (https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Guide-annotation-PARSEME_FR-chapeau). It will include all kinds of frozen idiomatic expressions but exclude collocations, pragmatic expressions and routine formulae. For example, expressions such as *force est de constater* ['it must be noted'], *il n'y a pas de quoi* ['you're welcome'] or *célibataire endurci* ['confirmed bachelor'] are not included in the PARSEME annotation project. However a major interest of the PARSEME project is the association with a syntactic annotation.

social web corpus with MWE annotations, which distinguishes between strong and weak MWEs, but does not provide a detailed typology.

3.1 Typology of Multiword Expressions and Annotation Process

Typology of MWEs. Our typology of MWEs (see Table 2), inspired by several linguistic typologies (Granger and Paquot 2008; Heid 2008; Tutin 2010; Mel'čuk 2012), is fairly broad and includes 8 types of MWEs, ranging from most compositional MWEs such as collocations (e.g., *heavy smoker*, *pay attention*) to fully frozen expressions such as *insofar as* or *black sheep* and oral expressions with pragmatic constraints (e.g., *see you later! Thank you. You're welcome*). We thought it was important to include more compositional expressions such as collocations because they are known to be crucial in NLP applications such as text generation (Lareau et al. 2011) or Machine Translation (Liu et al. 2010; Ramisch 2017).

The robustness of this typology has been evaluated. We computed the inter-annotator agreement on a sub-corpus (2,000 words in the literary text and 2,000 words in the scientific text) (Cf. (Tutin et al. 2015)). It provided acceptable results (Fleiss kappa of 0.683 for the literary text and 0.741 for the scientific report) with a good agreement on functional MWEs, performing slightly less well with collocations and idioms, which brought us to provide more detail and formal criteria in our annotation guidelines.

Table 2. Typology of multiword expressions in PolyCorp.

Multiword expressions	Examples
Idioms[a]: frozen multiword expressions	*cul de sac* (fr)/*dead end, prendre en compte* (fr)/*take into account*
Collocations: preferred binary association, including light verb constructions	*gros fumeur* (fr)/*heavy smoker; faire une promenade* (fr)/*to take a walk*
Functional Multiword Expressions: functional adverbs, prepositions, conjunctions, determiners, pronouns	*c'est pourquoi* (fr)/ *that is why; d'autre part* (fr)/*on the other hand; insofar as*
Pragmatic MWEs: Multiword expressions related to specific speech situations	*de rien* (fr)/*You're welcome; à plus tard* (fr.)/*see you later*
Proverbs	*Pierre qui roule n'amasse pas mousse* (fr)/*A rolling stone gathers no moss*
Complex terms	*Natural Language Processing*
Multiword named entities	*Université Grenoble Alpes; the European Union*
Routine formulae: routines generally associated with rhetorical functions	*force est de constater* (fr)/*it must be noted*

[a]Idioms include compounds in our typology.

Annotation Scheme and Annotation Process. The annotation scheme is a surface annotation, where each token included in a MWE is annotated with an identifier, the

grammatical category of the whole expression, the part of speech of every element of the MWE, and the type of MWE as a whole (idiom, collocation …). Cases of overlapping MWEs, i.e., when some parts of the MWEs belong to several MWEs, as in the example *We paid close attention,* where we can have two collocations (*close attention* and *pay attention),* are dealt with.

The annotation process has been performed semi-automatically with the help of an NLP tool, NooJ (Silberztein 2016), and a large dictionary of 5,000 frequent MWEs. This dictionary, which provides the morpho-syntactic and the type of each MWE, has been compiled from several resources, including the *Dictionnaire Électronique des Mots* (Dubois and Dubois Charlier 2010), *Wiktionary* or the *DELAC* (Courtois et al. 1997). About 35% up to 50% of MWEs are semi-automatically annotated. The automatic annotation has been completed and checked manually with the help of guidelines by at least two skilled annotators[7] (some texts have been annotated by three persons). Moreover, a Perl script has been used to ensure the consistence of the annotation. It allowed to spot and to correct some errors, when the various elements of 1 MWE had been labeled with different tags of part of speech.

Figure 3 provides an example of annotated text extracted from *Thérèse Raquin*.

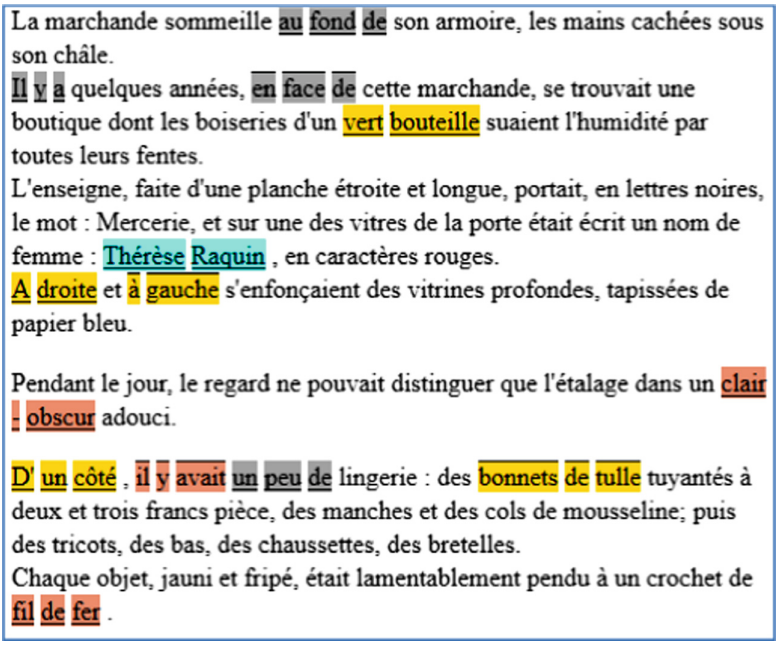

Fig. 3. An example of annotated text (*Thérèse Raquin*) with a stylesheet (collocations, idioms, named entities, functional words) (Color figure online)

[7] Master's degree students and scholars in computational linguistics.

3.2 Some Tricky Cases

The annotation process is not trivial. The annotators encounter two main kinds of problems: delimiting boundaries of MWEs and choosing the right type of MWE. Due to the lack of space, we will only address some examples.

Delimiting boundaries of MWEs is generally trivial for functional words (e.g. *in order to*) or nominal idioms (e.g. *point of view*). It is far more complicated for verbal MWEs. As a rule, we decided to exclude the annotation of grammatical words that are not essential in the MWE. For example, in the following example,

Il faut faire un effort [‘we have to make an effort]

we do not include the determiner in the MWE annotation, since it is highly variable (*faire l’effort, des efforts, cet effort*). Conversely, in some MWEs, the determiner is fixed and should be included:

Je ne veux pas passer l’ éponge là-dessus. [I don’t want to wipe the slate clean on that’]

Nor do we include it within the prepositions that are subcategorized by the MWE verbs (e.g.: *to give a talk to*). Including or not including the grammatical words strongly depends on the type of MWE, and often requires corpus queries in order to make the right decision.

Selecting the right type of MWE is also tricky. Some MWEs such as functional MWEs and Named entities are less prone to inter-annotator disagreement than other ones such as decision between collocations and frozen expressions (we will see that they are also the most tricky cases for statistical Machine Translation). Clear-cut decisions along this continuum of “frozenness” can be considered somewhat arbitrary. However, providing syntactic and semantic criteria in the guidelines can help to ensure consistency during the annotation process and enables us to refine the annotation process. For example, we decided to include expressions such as *developed country* as collocations. These expressions can be considered as kinds of “hyponymic collocations” (the collocation refers to a subtype of the head noun. A *developed country* is a kind of *country*). We also included binomials (e.g., *I work day and night for this project*) in the class of collocations.

4 A Corpus Annotated with Multiword Expressions: Some Results

4.1 The PolyCorp Corpus

Our annotated corpus with MWEs (see Table 3), the PolyCorp corpus, reaches the size of almost 70,000 tokens and includes several freely available corpora. Except EIIDA, all texts have an English equivalent, which enabled some studies on MT (see Sect. 4.4).

Table 3. Types of textual genres included in the PolyCorp corpus

Sub-corpus	Textual genre	Number of tokens
BAF Citi 1 (Baf corpus)	Scientific writing	14,500 tokens
Thérèse Raquin (Emile Zola)	Literary novel	7,260 tokens
News	"Journalese"	17,400 tokens
Film Le fabuleux destin d'Amélie Poulain	Film dialogues (subtitles)	9,900 tokens
TED talks	Talks in science popularization	8,160 tokens
EIIDA French corpus (scientific talks)	Scientific talks	12,630 tokens

The corpus is diverse since it includes 3 oral texts (film dialogues, scientific talks, TED talks) and 3 written texts (scientific report, news, literary novel) in several fields[8]. This diversity of textual genres is useful to better understand the linguistic properties of MWEs.

In this corpus, 5.560 MWEs have been annotated. Some interesting results could be computed from this annotated corpus.

4.2 Phraseological Density

The phraseological density is the ratio of tokens included in MWEs. For example, in the following sentence

Nous avons pris ce problème en compte.
[we have taken this problem into account]

our ratio is 3/8. The phraseological density for the whole corpus is 20.45%, which is quite close to that of Schneider et al. (2014). These results show that although MWEs are widespread, they are far from being as frequent as simple words. In French, phraseological density would greatly increase if we included compound tenses (e.g. *j'ai mangé* 'I have eaten') and discontinuous negation, which can be considered as a kind of MWE (e.g. *je ne mange pas*).

With regard to the kind of text, we can observe that phraseological density is quite stable from each textual genre to another (between 17.40% and 27.88%). The differences between the genres are not as important as we could have expected, but it seems that oral transcriptions, especially scientific talks, contain more MWEs than the other written genres (e.g. journalese and scientific writing) (Fig. 4).

[8] The criteria for selecting the texts were the following ones: diversity of genres and availability of aligned bilingual corpora.

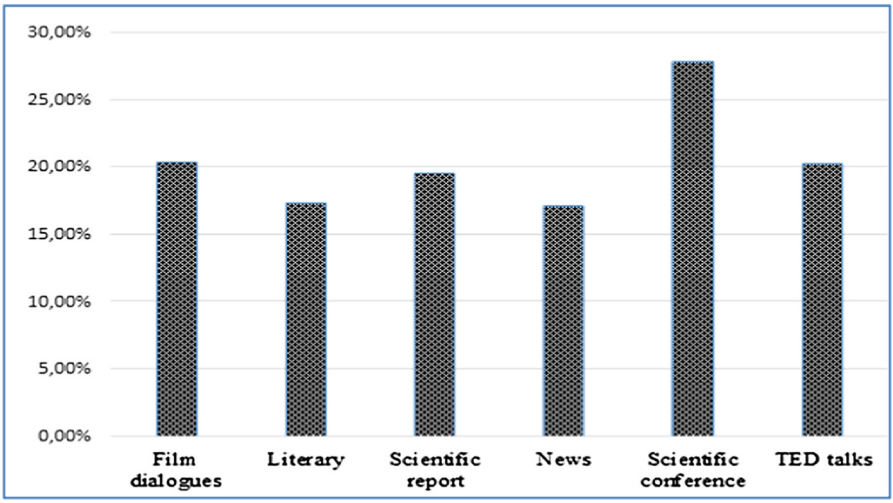

Fig. 4. Phraseological density across genres

4.3 Distribution of MWEs Across Textual Genres

More interesting is the distribution of MWEs across textual genres (Fig. 5). We notice that some expressions are stable across all kinds of texts: collocations, idioms and functional words. As is often claimed in literature, collocations are more frequent than

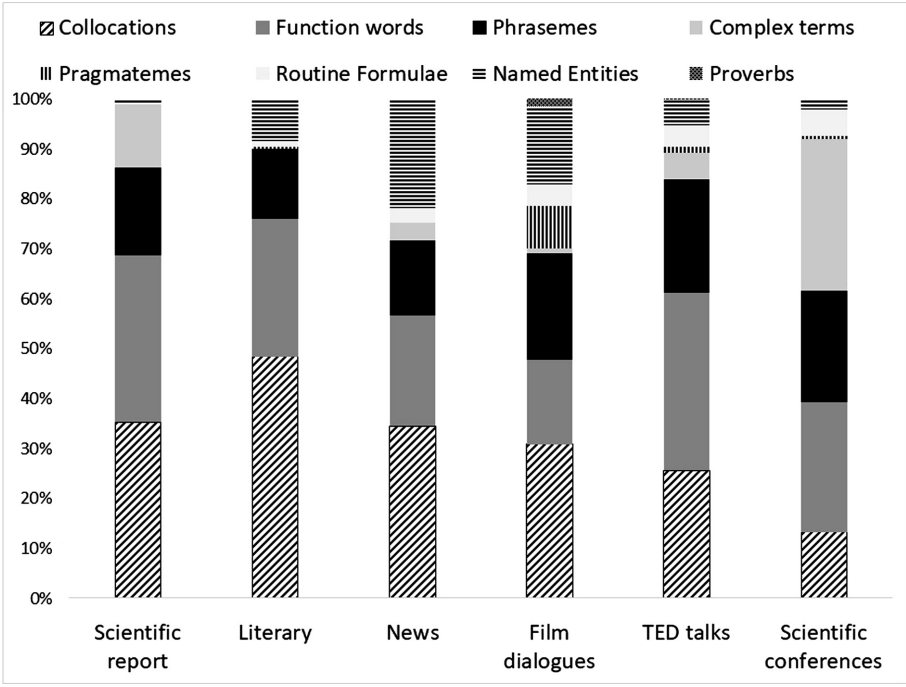

Fig. 5. Distribution of MWEs across textual genres

idioms in all genres, except scientific talks. Among the most frequent MWEs across textual genres, we obviously find functional words such as *un peu* ('a little bit'), *au cours de* ('over the course of'), …

We also observe strong dissimilarities between textual genres: complex terms in scientific reports and talks, named entities in journalese and films, and pragmatic MWEs in dialogues.

Therefore, there seem to be two kinds of MWEs: the first ones fall within the category of "core MWE lexicon" used in any type of text. The second ones are specific to given genres and pragmatic contexts. These first results should obviously be confirmed on larger corpora.

4.4 Using the PolyCorp Corpus in the Framework of Machine Translation Studies

An interesting application in NLP has been performed. The corpus obtained has already been used in two translation-quality evaluation studies. The first one done by Esperança-Rodier and Didier (2016), using the scientific writing sub-corpus BAF Citi 1, demonstrated that most (actually 80%) of MWEs in French were translated into equivalent MWEs in English. This interesting result shows that there is a high structural and lexical similarity between two close languages such as English and French. Furthermore, the study showed that when an idiom, a collocation or a functional multiword expression was well translated, it was translated by its attested translation in more than 95% of the occurrences. Nevertheless, even if there is consistency when translated well, those 3 types behave in a different way when mistranslated. When looking at the errors, if we consider the case in which MWEs were not translated by their attested translation but by a translation which did not keep the source meaning, it mainly happened for roughly 41% of functional MWEs against almost 19% of idioms, while the percentage rose to approximately 32,7% for collocations. Moreover, when MWEs were not translated by their attested translation, but by a translation which did not keep the source meaning, we could see that it mainly occurred in collocations (13%) while the percentage dropped to 8% for idioms and to 2% for functional multiword expressions. These figures demonstrate that when collocations are mistranslated, the meaning is not kept while it is the contrary for functional multiword expressions.

The second work done by Esperança-Rodier (2019, to be published), focuses on the evaluation of the quality of the MWE translation produced by a state-of-the-art Statistical Machine Translation (SMT) system on the same sub-corpus. This study has shown that Named Entities are always well translated (100%) by the in-house SMT system, named hereafter Lig-Moses, which is logical as SMT systems provide the original words when they do not know the translation of a word in the target language. Then Technical terms (T) and Functional words (F) are partially well translated (respectively 15% of T were annotated with an error type and 8% of F.). Finally, Collocations and idioms were the ones for which most of the errors occurred (respectively 23% and 24% were annotated with an error type), which is consistent with the difficulties encountered by annotators. Considering the mistranslation, when the MWEs were not translated correctly it was mainly due to the fact that the MWE translation was missing in the target language (missing word, filler words), and then the fact that the MWEs were incorrectly translated (incorrect word sense, wrong lexical choice or incorrect disambiguation).

5 Conclusion and Perspectives

These experiments on French MWEs show that experts can reach a consensus and that an annotation task with a panel of multiword expressions can be performed satisfactorily, as long as the typology is not overly complex and clear guidelines are provided. The annotation of various texts provides interesting results concerning MWEs, which are prevalent but not as numerous as often claimed. The phraseological density is quite stable across textual genres. The lexicon of MWEs falls into 2 broad types: (a) core MWE lexicon including function words, collocations and idioms; (b) specific MWE lexicon including named entities, complex terms, pragmatic MWEs, and to a lesser extent, routines formulae and proverbs. The annotated corpus has been used to evaluate statistical machine translation, showing interesting differences among multiword expressions considered in the study.

This work needs to be extended to bilingual annotation (including the English version of this parallel corpus) and extended to other corpora in order to confirm these first results and better understand the nature and diversity of multiword expressions.

Acknowledgements. Internships for the corpus annotation have been funded by the Pôle Cognition (http://www.grenoblecognition.fr/) (Université Grenoble Alpes). Special thanks to the interns and students who contributed greatly to this project: Doriane Simonnet, Pauline Soutrenon, Manolo Iborra, Justine Reverdy, Zied Elloumi, Johan Didier.

References

Abeillé, A., Clément, L., Toussenel, F.: Building a treebank for French. In: Abeillé, A. (ed.) Treebanks. Text, Speech and Language Technology, vol. 20, pp. 165–187. Springer, Dordrecht (2003). https://doi.org/10.1007/978-94-010-0201-1_10

Altenberg, B.: On the phraseology of spoken English: the evidence of recurrent word-combinations. In: Cowie, A.P. (ed.) Phraseology: Theory, Analysis and Applications, pp. 101–122. Clarendon Press, Oxford (1998)

Candito, M., et al.: Annotation d'expressions polylexicales verbales en français. In: 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN), pp. 1–9 (2017)

Courtois, B., et al.: Dictionnaire électronique DELAC: les mots composés binaires. Technical report 56, University Paris 7, LADL (1997)

Cowie, A.P. (ed.): Phraseology: Theory, Analysis, and Applications. OUP, Oxford (1998)

Dubois, J., Dubois-Charlier, F.: La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. Langages **179**(3), 3–56 (2010)

Esperanca-Rodier, E., Didier, J.: Translation Quality Evaluation of MWEs from French into English Using an SMT System (2016). Translating and the Computer 38 © AsLing, The International Association for Advancement in Language Technology (2015). www.asling.org

Esperança-Rodier, E.: Analyse de la qualité des traductions automatiques du français vers l'anglais, d'Expressions Poly-Lexicales (EPL) à partir d'un corpus parallèle – Quelles sont les erreurs les plus fréquentes par type d'EPL? In: LTT 2018 (2019, to be published)

Fraser, B.: Idioms within a transformational grammar. Found. Lang. 2242 (1970)

Granger, S., Paquot, M.: Disentangling the phraseological web. In: Granger, S., Meunier, F. (eds.) Phraseology: An Interdisciplinary Perspective, pp. 27–49. John Benjamins, Amsterdam (2008)

Gross, G.: Les expressions figées en français: noms composés et autres locutions. Ophrys, Paris (1996)

Hausmann, F.J.: Tout est idiomatique dans les langues. In: Martins-Baltar, M. (ed.) La locution entre langue et usages, pp. 19–52. ENS Editions, Fontenay-Saint-Cloud (1997)

Heid, U.: Computational phraseology. An overview. In: Granger, S., Meunier, F.: Phraseology. An Interdisciplinary Perspective, pp. 337–360. Benjamins, Amsterdam (2008)

Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 177–180 (2007)

Laporte, E., Nakamura, T., Voyatzi, S.: A French corpus annotated for multiword nouns. In: Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions, pp. 27–30 (2008a)

Laporte, E., Nakamura, T., Voyatzi, S.: A French corpus annotated for multiword expressions with adverbial function. In: Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop, pp. 48–51 (2008b)

Lareau, F., Dras, M., Börschinger, B., Dale, R.: Collocations in multilingual natural language generation: lexical functions meet lexical functional grammar. In Proceedings of ALTA 2011, pp. 95–104 (2011)

Liu, Z., Wang, H., Wu, H., Li, S.: Improving statistical machine translation with monolingual collocation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 825–833 (2010)

Mel'čuk, I., et al.: Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques 1, II, III, IV, Les Presses de l'Université de Montréal, Montréal (1984, 1988, 1992, 1998)

Mel'čuk, I.: Tout ce que nous voulions savoir sur les phrasèmes, mais… Cahiers de lexicologie **102**, 129–149 (2013)

Mel'čuk, I.: Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology **3**(1), 31–56 (2012)

Moon, R.: Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford University Press, Oxford (1998)

Polguère, A.: Principles of lexical network systemic modeling (Principes de modélisation systémique des réseaux lexicaux). In: Proceedings of TALN 2014 (Volume 1: Long Papers), pp. 79–90 (2014). (in French)

Potet, M., Esperança-Rodier, E., Besacier, L., Blanchon, H.: Collection of a large database of French-English SMT output corrections. In: LREC 2012, Istambul, 21–27 May 2012, pp. 4043–4048 (2012)

Ramisch, C.: Putting the horses before the cart: identifying multiword expressions before translation. In: Computational and Corpus-Based Phraseology - Second International Conference, Europhras 2017, London, UK, pp. 69–84 (2017)

Schneider, N., et al.: Comprehensive annotation of multiword expressions in a social web corpus. In: Proceedings of LREC, Reykjavík, Iceland, pp. 455–46 (2014)

Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley, London (2016)

Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991)

Tutin, A., Esperança-Rodier, E., Iborra, M., Reverdy, J.: Annotation of multiword expressions in French. In: European Society of Phraseology Conference (EUROPHRAS 2015), pp. 60–67 (2015)

Wray, A.: Identifying formulaic language: persistent challenges and new opportunities. In: Corrigan, G., Moravcsik, E., Ouali, H. (eds.) Formulaic Language, vol. I. Benjamins, Amsterdam (2009)