EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

PhD thesis summary

Verb+Noun Multiword Expressions: A linguistic analysis for identification and translation

Uxoa Iñurrieta Urmeneta

2019

EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

Verb+Noun Multiword Expressions: A linguistic analysis for identification and translation

This is a shortened version of the Basque dissertation entitled *Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala*, written by Uxoa Iñurrieta under the supervision of Dr. Itziar Aduriz and Dr. Gorka Labaka. It also includes the papers published by the candidate on the research presented here.

September 2019

Acknowledgments

The Spanish Ministry of Economy and Competitiveness, who awarded me a predoctoral fellowship (BES-2013-066372) to conduct research within the SKATeR project (TIN2012-38584-C06-02).

Abstract

Multiword Expressions (MWEs) are combinations of words which exhibit some kind of idiosyncrasy. Due to their idiosyncratic nature, they pose several problems to Natural Language Processing (NLP). In this PhD, two of the most challenging tasks concerning MWE processing are addressed: the automatic identification of MWE occurrences in corpora and their translation in Machine Translation (MT).

On the one hand, to test whether the use of specific linguistic data was beneficial for MWE identification, an in-depth analysis of Spanish verb+noun MWEs was undertaken where lexical and morphosyntactic data were carefully considered. These data were used to identify occurrences of the studied MWEs, improving on results reported by related work. On the other hand, the Basque translations of the studied MWEs were also analysed along lexical and morphosyntactic dimensions. This additional information was then added into a rule-based MT system, and an improvement was observed concerning MT quality, both according to a manual evaluation and according to statistical measures. All the analysed linguistic data was collected in a publicly available database, which can be either queried online or fully downloaded to be used for NLP-related purposes.

Finally, to complete the analysis of Basque MWEs, verbal MWEs were annotated in a Basque corpus, which was then released along with annotated corpora in 19 more languages. Part of this multilingual corpus served as a basis for a subsequent study on literal occurrences of MWEs, carried out in five languages from different phylogenetic families, including Basque. Both the annotation and the study on literal occurrences are included in this PhD.

Contents

1 Introduction

Multiword Expressions (MWEs) are combinations of words which exhibit some kind of lexical, morphosyntactic, semantic, pragmatic or statistical idiosyncrasy (Baldwin and Kim, 2010). Several types of word combinations are comprised in the category of MWEs (Corpas Pastor, 1996; Urizar, 2012; Gurrutxaga and Alegria, 2013), two of which are considered in this work: idioms (example 1), which have a non-compositional meaning, and collocations, characterised by their frequent and restricted co-occurrence (example 2). The latter also include light verb constructions (example 3), in which the verb tends to be semantically bleached[1].

(1) *She always ends up* ***spilling the beans*** (lit. revealing the secret)

(2) *All students* ***passed*** *the* ***exam****.*

(3) *She is* ***giving*** *a* ***lecture*** *this afternoon.*

Due to their idiosyncratic nature, MWEs pose important challenges to Natural Language Processing (NLP), and sophisticated strategies are needed in order to process them correctly (Sag et al., 2002; Constant et al., 2017). The features of MWEs that prove most challenging for NLP can be summarised as follows.

- Arbitrarily prominent co-occurrence. Lexical co-occurrence is arbitrarily restricted in some kinds of MWEs, which causes problems in many NLP applications. One such application is Machine Translation (MT), since lexical choice is language-dependent, meaning that word-for-word translations are often inappropriate. For instance, in English (EN), Spanish (ES), Basque (EU) and French (FR), the noun *attention* is combined with different verbs to express the act of listening or observing something carefully (example 4).

  (4) EN: ***pay attention***
  EU: ***arreta jarri*** (lit. put attention)
  FR: ***faire attention*** (lit. make attention)
  ES: ***prestar atención*** (lit. lend attention)

- Non-compositionality. The meaning of some MWEs cannot be inferred from the separate meanings of their component words. The meaning of *kicking the bucket*, for example, is not related to the act of *kicking* nor to a *bucket*, but to *dying*. This is also problematic for MT, since this kind of expression is rarely translated word-for-word, especially when the source and target languages are of very different typology like Spanish and Basque (example 5).

[1]In the examples in this summary, the lexicalised components of MWEs are shown in bold, and non-MWEs or other words and morphemes to be marked have been underlined.

(5) ES: ***dormir a pierna suelta***
sleep to leg loose
'sleep soundly')
EU: ***lo seko egon***
asleep dry be
'sleep soundly')

- Ambiguity. Many word combinations can have both an idiomatic and a literal meaning. This is the case of *pull somebody's leg*, which can either be used as 'to kid/trick someone' (example 6) or literally (example 7).

(6) *She is not serious. She is just* ***pulling your leg****.*

(7) *Grab your knee, pulling your leg toward your chest.*

Automatically distinguishing when a given word combination is the occurrence of an MWE and when it has a literal meaning is a very complex task which can have an impact in several NLP applications. MT systems are among them, since a different translation is often needed for each meaning. Examples (8) and (9) show how *pull somebody's leg* would be translated into Spanish and Basque in each case.

(8) ES-id: *Solo te está* ***tomando el pelo****.*
only to-you is taking the hair
'She is just pulling your leg.'
ES-lit: *Tira de tu pierna hacia el pecho.*
pull of your leg toward the chest
'Pull your leg toward your chest.'

(9) EU-id: ***Adarra jotzen*** *ari_zaizu.*
horn-the playing is
'She is pulling your leg.'
EU-lit: *Erakarri hanka bularrerantz.*
pull leg-the chest-toward
'Pull your leg toward your chest.'

- Variability and discontiguity. MWEs can have multiple morphosyntactic variants, meaning that their component words can occur in several word forms and can be separated by other elements in a sentence (example 10). This flexibility often makes it hard to automatically identify MWE occurrences in corpora. For example, in the MWE *take steps*, the noun phrase (NP) can be either singular or plural, it can contain modifiers inside, an adverb or other elements can separate the verb and the NP, and word order can be reverted e.g. in passive or relative sentences. Thus, searching for a fixed sequence like 'take steps' or 'take a

step' would be insufficient to identify the diverse occurrences that this MWE can have.

(10) *It is important to **take steps** on the matter.*
*They **took** an important **step** on the matter.*
*The **steps** they **took** were vital.*
*Important **steps** are currently being **taken**.*

Two main tasks are involved in MWE processing: MWE discovery and MWE identification (Evert, 2009; Seretan, 2011; Ramisch, 2015). Discovery consists in extracting MWEs from corpora, usually with the aim to create or enlarge lexicons. Text corpora are used as input, and the output is a list of word combinations which are considered MWEs. Identification, in contrast, involves finding occurrences of previously known MWEs. As well as text corpora, a list of MWEs is also needed as an input for identification, and the output is a set of MWE annotations in the corpus.

These tasks can affect each other, as well as a number of NLP applications like parsing and MT. Figure 1 (adapted from Constant et al. 2017) shows how MWE processing relates to the applications mentioned: MWE lists extracted by discovery methods can be used as an input for MWE identification; morphosyntactic information obtained from parsers can equally be useful for MWE identification; and identifying MWEs is necessary to properly translate MWEs.

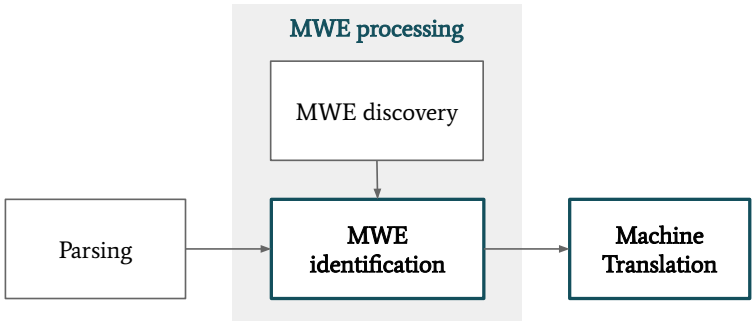


Figure 1 – MWE processing tasks in relation to parsing and MT, adapted from Constant et al. 2017.

Note that the inter-relation between these tasks and applications is more complex than what is shown in Figure 1. Only the relations relevant to this specific PhD are shown in it, but a more exhaustive scheme can be found in Constant *et al.*'s survey (2017). One of the tasks and one of the applications mentioned above are covered in this work, which are highlighted: MWE identification and MT. More details and explanations about them are given in the following subsections.

1.1 MWE identification

Automatically identifying MWE occurrences in corpora is a very complex task, especially because of three of their challenging features: ambiguity, discontiguity and variability. The last two are especially prominent in MWEs where the syntactic head is a verb (henceforth, verbal MWEs), on which this research work is focused.

For the identification of verbal MWEs, basic methods which try to match fixed word sequences against dictionary entries are too limited. For instance, let *make conclusions* and *take into account* be two entries in a dictionary. If this basic method was employed to identify occurrences of these entries in the sentences below, all occurrences would be ignored, because: the component words are separated by external elements in (11a)–(11c) and (12b); word forms in examples (11b), (11c) and (12a) are different from the ones in the entry; and word order is altered in example (11c).

(11) a. *They **made** a **conclusion**.*
b. *They **made** some simple but still interesting **conclusions**.*
c. *The **conclusions** they **make** are always interesting.*

(12) a. *Their advice should be **taken into account**.*
b. *You should **take** their advice **into account**.*

On the other hand, strategies where only the lemmas of the component words are searched for (within a given word distance) are not effective either, since these are, in their turn, too wide. These strategies would identify all of the occurrences in examples (11) and (12), but also the following ones and many others alike, which would be false positives:

(13) *They will make progress and will soon come to a conclusion.*

(14) *You should take the money and put it into your account.*

Therefore, it is necessary to develop identification methods which consider not only the morphosyntactic flexibility of verbal MWEs but also their possible restrictions. One of the main hypotheses behind this work is that very few word combinations occur in corpora both literally and idiomatically with the very same morphosyntactic features, that is, that most ambiguities concerning MWEs can be solved by looking at morphology and syntax. If we consider example (13), the fact that *conclusion* is not the direct object of *make* makes it evident that this is not an occurrence of the MWE *make conclusions*, but a coincidental occurrence of its component words. In example (14), on the other hand, two morphosyntactic features would be helpful to show that the underlined words are not part of the expression *take into account*: (1) the fact that the noun *account* is modified by a possessive determiner (*your*), which would be unacceptable if *take into account* was used

as an MWE, and (2) the fact that the head verb syntactically related to the prepositional phrase *into account* is not *take* but *put*.

Different methodologies have been employed in order to identify non-fixed MWEs. In some of them (Forcada et al. 2011; Padró and Stanilovsky 2012), a variable element is specified inside the MWE, and all possible inflected word forms are automatically generated and listed. In verbal MWEs, the verb is selected as the variable element, and the rest of the component words are only searched for exactly as they are in the lexicon entry, next to each other and with the same word form and order. For example, for *make conclusions*, *make/makes/made/making conclusions* would all be identified, but not *make a/the/one conclusion* and other variants alike. This is the methodology used by the Freeling parser, and the one we will use as a baseline in this work.

Some other approaches combine morphosyntactic information obtained from parsers and MWE-specific rules. This can be done either by applying general or category-based rules (Oflazer et al. 2004; Copestake et al. 2002; Ramisch et al. 2010), or by using MWE-specific lexicons which contain further linguistic information, i.e. the morphosyntactic restrictions of the entries (Hashimoto et al. 2006; Urizar 2012).

As well as rule-based methods, word sense disambiguation techniques have also been employed for MWE identification (Katz and Giesbrecht 2006; Cook et al. 2007; Hashimoto and Kawahara 2008; Sporleder and Li 2009; Tu 2012), where idiomatic and non-idiomatic occurrences are distinguished by looking at the surrounding words of a given occurrence. On the other hand, machine learning techniques have also proved to be useful for this task. Most of these only identify contiguous word combinations (Blunsom and Baldwin 2006; Constant and Sigogne 2011; Shigeto et al. 2013), although a few have developed more complex models in order to identify non-contiguous occurrences as well (Schneider et al. 2014).

The identification of MWEs being so challenging, two shared tasks were organised for that purpose by the PARSEME project (Savary et al. 2017; Ramisch et al. 2018). A multilingual corpus was released where verbal MWEs were annotated following universal guidelines (Savary et al. 2018), which was used as a basis for the shared task. The annotation of the Basque part is covered in this PhD, as well as a subsequent study on literal occurrences of MWEs based on it. More details about both the corpus, the shared task and the study of literal occurrences are given in the publications in the Appendix.

1.2 MWEs in Machine Translation

The second challenge considered in this PhD is the processing of MWEs within MT. As already mentioned, many MWEs cannot be translated word-for-word from one language to another, which makes their processing even more demanding when several languages are involved. In typologically different languages like Spanish and Basque, non-word-for-word translations are

especially prominent (examples 15–16).

(15) ES: ***meter ruido*** (lit. put noise in, 'make a noise')
EU: ***zarata atera*** (lit. take noise out, 'make a noise')

(16) EU: ***lan egin*** (lit. do work, 'to work')
ES: *trabajar* ('to work')

Some MT models, specifically statistical and neural systems, are based on corpora and use word co-occurrence as the main feature to do translations. Since these systems learn how words are combined within sentences, MWE-related problems are not as prominent as in rule-based ones, where every word is usually translated independently, by means of a bilingual lexicon and a set of linguistic rules. Namely, the *Matxin* MT system (Mayor et al., 2011) fails to translate many Spanish MWEs into Basque, due to three main reasons:

- The limits of the system's bilingual lexicon
- Its insufficient identification strategy
- Its insufficient transfer strategy

In order to properly translate MWEs, it is vital that the bilingual lexicon used by the MT system is MWE-aware, i.e. that it contains not only individual words but also MWEs along with their translations. Otherwise, whenever a given MWE is not included in the lexicon, the component words are treated regularly, and a literal word-for-word translation is thus assigned to them. This is the case of the Spanish MWE *contraer matrimonio* (lit. contract marriage 'get married'), which is not included in *Matxin*'s bilingual lexicon and, hence, the system ignores that the whole expression needs to be translated by a single verb, *ezkondu* ('marry') (example 17).

(17) ES: *La pareja* ***contrajo matrimonio***.
the couple contracted marriage
'The couple got married.'
MT: *Bikotea ezkontza uzkurtu zen.*
couple-the marriage shrink AUX
'The couple got shrunk marriage.'
EU: *Bikotea ezkondu zen.*
couple-the married AUX
'The couple got married.'

Apart from feeding the bilingual lexicon, two tasks must be carried out: on the one hand, the identification of MWEs in the source sentence, and on the other hand, their transfer into the target language. If a given MWE is

not properly identified, the system is usually unable to give it an appropriate translation, like in example (18). Since *Matxin* only searches for contiguous and almost completely fixed word combinations (see explanation about the Freeling parser in the previous section), the MT system fails to identify the MWE *tomar el pelo* (lit. take the hair 'pull sb's leg'), because its component words are not contiguous. Consequently, the verb and the noun are translated separately, and an incorrect output sentence is produced in Basque: instead of the MWE *adarra jo* (lit. play the horn 'pull sb's leg'), an erroneous literal translation is given.

(18) ES: *Nos* ***toma*** *siempre* ***el pelo***.
us takes always the hair
'He/She always pulls our leg.'
MT: *Beti* *hartzen* *digu* *ilea*.
always takes AUX hair-the
'He/She always takes our hair.'
EU: *Beti* ***jotzen*** *digu* ***adarra***.
always plays AUX horn-the
'He/she always pulls our leg.'

When an MWE is identified, the translation linked to it in the lexicon is selected by *Matxin*. However, it is not always evident to the system how this translation needs to be used in the target sentence, which can sometimes be the source of additional errors. In example (19), the MWE *buscarse la vida* (lit. search the living for oneself 'get by') is correctly identified in the source sentence, but the Basque translation produced by the MT system is ungrammatical: although an adequate lexical translation is selected, it is ignored that a transitive auxiliary verb should be used with it instead of an intransitive one.

(19) ES: *Ella* ***se busca la vida*** *como puede.*
she AUX searches the living as can
'She gets by as she can.'
MT: *Hura* ***bizimodua ateratzen*** *da ahal duen bezala.*
He/She living-the takes-out AUX.INTR power has as
'He/She gets by as he/she can' (grammatically incorrect)
EU: *Hark* ***bizimodua ateratzen*** *du ahal duen bezala.*
He/She living-the takes-out AUX.TR power has as
'He/She gets by as he/she can'

Most rule-based systems use basic MWE-processing methods which treat fixed word sequences as single words, both for the identification and for the translation tasks. When the source MWE is non-variable, this basic

method gets fairly good results (Barreiro 2008; Bouamor et al. 2012; Tan and Pal 2014); not, however, when morphosyntactically flexible MWEs occur in the source sentence. More complex strategies are needed in such cases, which some authors have developed by adding MWE-specific linguistic rules (Anastasiou 2008; Forcada et al. 2011; Monti et al. 2011).

Wehrli *et al.* (2009), for instance, identify MWEs by applying linguistic patterns after the analysis phase, and then translate them by using a syntax-based formal representation. Some other approaches code MWEs differently, by using *interlingua*s (Oepen et al. 2004; Monti et al. 2011), such as Lexical Functions (Heylen et al. 1994).

Concerning Spanish-Basque MT, apart from *Matxin*, other MT systems have also been created for Spanish-Basque, such as the EUSMT statistical system (Labaka 2010), a hybrid system which combines *Matxin* and EUSMT (Labaka et al. 2014), and a more recent neural system, MODELA (Etchegoyhen et al. 2018), which produces translations of significantly better quality than the rest. Among all of these systems, only *Matxin* includes an MWE-specific processing strategy. As already mentioned, this is the system on which this PhD is based. More details on our proposed MWE translation method will be given in the following sections and in the papers included in the Appendix.

2 General outline of the dissertation

This is a shortened version of the Basque-written PhD dissertation entitled *Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala*. The work was undertaken inside the Ixa research group, within the fields of Natural Language Processing and Computational Linguistics, and it specifically focuses on phraseology and Machine Translation.

The Basque dissertation is divided into eight chapters, the contents of which are mostly described in the seven publications in the Appendix. An outline of the work carried out in this PhD, as well as how the different parts are related to each other, is shown in Figure 2. The pieces of work referring to linguistic analyses are placed in the middle, with the resources created as an outcome to the left, and the related experiments to the right. Circled numbers refer to publication codes (see below), and arrows depict how each part is connected to others.

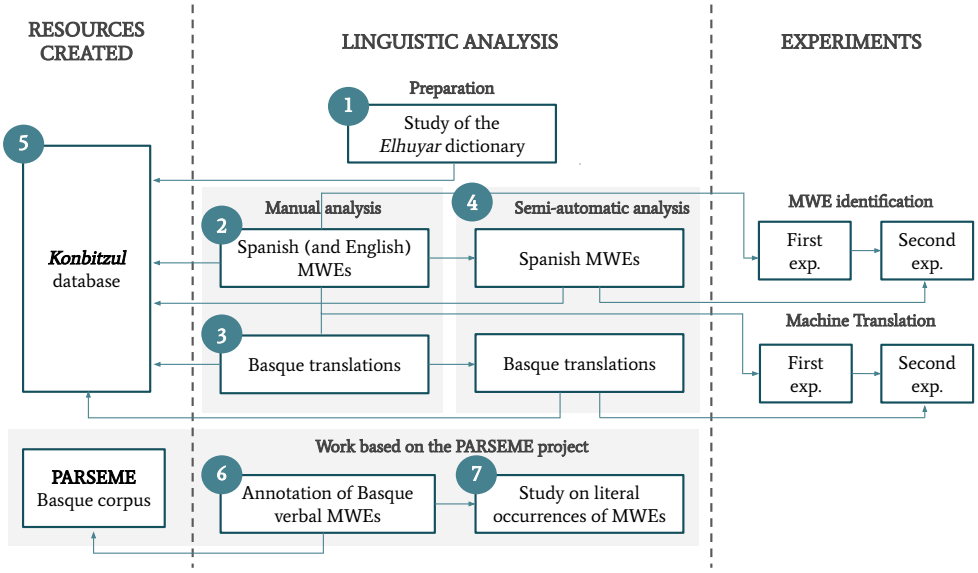


Figure 2 – Outline of the research done in this PhD.

For a proper understanding of the work carried out in this PhD, it is recommended to read this summary first, including the introduction (Section 1), the general outline in this section, and the hypotheses, conclusions and contributions in Section 3. Then, the publications in the Appendix should be read in order, having the outline in Figure 2 in mind, so as to better follow the link between the different parts[2].

[2] Please note that no changes were made to the publications before collecting them in the Appendix; they were added exactly as published (or in preprint format, when necessary). As a consequence, terminology might vary from one publication to another,

The seven publications are briefly described below. A code is given to each of them, in order to make it easier to connect them to the list in the Appendix (page 23).

[P1] A preliminary analysis of the verb+noun entries in the Elhuyar Spanish-Basque dictionary is described. A total of 2,343 Spanish combinations (along with 6,587 Basque translations) and 2,954 Basque combinations (along with 6,390 Spanish translations) were examined, with a special focus on lexical and morphosyntactic features. This study serves to strengthen the motivation behind this PhD, since it shows how rare word-for-word translations of MWEs are according to dictionaries.

[P2] A more in-depth analysis on Spanish verb+noun MWEs is described, as well as the first identification experiment undertaken. Out of the Spanish combinations from the Elhuyar dictionary, the 150 most frequent ones were selected and studied along lexical-semantic and morphosyntactic dimensions, and 117 were found to be suitable for the subsequent identification experiment. Morphosyntactic data was then used for MWE identification, and good results were obtained. The same analysis and experiment was also applied for English by using 173 frequent verb+noun entries from the Oxford Collocations Dictionary[3], obtaining comparable results.

[P3] Details on the first MT experiment are given. The 117 Spanish combinations used for identification in the previous paper were now studied for MT purposes, as well as 22 combinations from the DiCE dictionary. A Basque translation was selected or manually given to each MWE, and their lexical and morphosyntactic features were examined. This linguistic data was integrated into the *Matxin* MT system, and an improvement in translation quality was observed both according to statistical measures and according to human judgements.

[P4] An enhanced automatic analysis method is proposed, drawn from the premise that previous work (P2 and P3) had promising results but, at the same time, the use of a completely manual method posed a problem of scalability. Monolingual and parallel corpora were employed, and MWE-specific data for both MWE identification and MT were automatically gathered. The combinations used as a basis for this study and the following experiments were extracted from the DiCE dictionary and the PARSEME multilingual corpus of verbal MWEs. A total

either because of the publishers' requirements or, in a few cases, because the authors changed their terminological preferences.

[3]Note that the English part of this work was done during a research stay at the University of Sussex, under the supervision of Dr. John Carroll.

of 668 MWEs were analysed semi-automatically (i.e. automatic information was manually revised), and 440 additional MWEs completely automatically. According to the experiments carried out, the gathered data was found to be hugely beneficial for MWE identification, and helpful to a lower extent for MT.

[**P5**] The *Konbitzul* database is described, which collects all the linguistic information gathered from the previous analyses (P1 to P4). It is publicly available online, and NLP-applicable data can be fully downloaded. Please note that the database was updated since this paper was written, meaning that the data given is probably not up-to-date.

[**P6**] The annotation of verbal MWEs carried out on a Basque corpus is described. The PARSEME universal guidelines were followed, and an 11,158-sentence corpus containing 3,823 MWE annotations was added to the PARSEME multilingual corpus. Annotation issues concerning Basque are discussed in the paper, as well as some particularities of Basque verbal MWEs.

[**P7**] A thorough study on literal occurrences of MWEs is presented. Five languages included in the PARSEME multilingual corpus were considered for study: Basque, German, Greek, Polish and Portuguese. The MWE annotations in the PARSEME corpus were used, and candidate literal occurrences were automatically extracted, to be then manually analysed. Evidence was found that literal occurrences of MWEs are extremely rare in real texts, and that the vast majority can be distinguished from idiomatic occurrences by looking at morphological and syntactic features.

3 Hypotheses, conclusions and contributions

The main assumption behind the work in this PhD is that specific linguistic information is helpful for MWE processing. Based on this assumption, several hypotheses were proposed and tested through the work outlined in the previous section. These hypotheses are listed below (Section 3.1), and conclusions about them are also briefly explained. Then, the main contributions made through this research are summarised (Section 3.2), and ideas for future work are presented (Section 3.3).

3.1 Hypotheses and conclusions

Six hypotheses were covered in this PhD, four of which were mostly related to linguistic aspects of MWEs, and the resting two, to MWE processing. All six hypotheses are listed below, along with a summary of the main evidence found to support them.

[H1] Many MWEs are not translated word-for-word from one language to another.

Our data shows that this is true as far as Spanish and Basque are concerned. On the one hand, in paper 1, we showed that few verb+noun entries in the Elhuyar bilingual dictionary were given verb+noun translations, namely, half of the Spanish entries and only a third of the Basque ones. Besides, very few among them were literal translations: in all, only 11% of all verb+noun entries were translated word-for-word from Spanish into Basque, and only 7% from Basque into Spanish. On the other hand, half of the automatically extracted MWE translations were evaluated as incorrect when word-alignments were performed on parallel corpora (paper 4), which suggests that this kind of word combination is not usually translated word-for-word. The NLP-related problems arisen from phraseological differences between languages were also made evident throughout the whole PhD, notably in the work related to rule-based MT.

[H2] Verbal MWEs tend to be rather flexible concerning morphosyntax, although not completely, since they also have some restrictions.

Among all of the Spanish verb+noun MWEs we analysed (papers 2 and 4), none of them was found to be completely fixed. Out of the final set of Spanish MWEs studied, 36% were tagged as completely flexible, and the rest as semi-fixed, meaning that most of the MWEs in this dataset have some morphosyntactic restrictions. Furthermore, after having studied the literal occurrences of the verbal MWEs annotated in the PARSEME corpus (paper 7), one of the main conclusions was

that the vast majority of literal occurrences can be distinguished from their idiomatic counterparts by looking at morphosyntactic features. As a matter of fact, less than 1% of the Basque occurrences considered for study were literal cases which could not be discerned under morphosyntactic criteria, and this percentage was also very low in the rest of the languages considered, except for Polish.

[H3] Compared to many other languages which have been analysed from a phraseological perspective, light verb constructions are especially frequent in Basque.

This is confirmed in papers 1 and 6. In the dictionary-based study, it was observed that more than half of the noun+verb entries were formed by one of six common verbs which tend to be light inside MWEs: *egin* ('do'), *izan* ('be/have'), *eman* ('give'), *hartu* ('take'), *egon* ('be') and *jarri* ('put'). Additionally, after having annotated a Basque corpus following the PARSEME universal guidelines, comparisons were made with 19 languages. According to these corpora, only two languages (Farsi and Hindi) use more light verb constructions than Basque per sentence on average, and the frequency gap is especially noteworthy in relation to the three languages in closest contact with Basque: the extent of annotated light verb constructions is three times higher than in Spanish and French, and six times higher than in English.

[H4] Although many word combinations can be idiomatic or literal depending on the context, very few of them are actually used literally in real texts.

The study carried out using the PARSEME corpora (paper 7) supports this hypothesis. Five languages from different phylogenetic families were analysed: Basque, German, Greek, Polish and Portuguese. Based on the annotations in the PARSEME corpora, the non-annotated occurrences of the word combinations which form MWEs were examined, and the idiomaticity rate was observed to be very high in all five languages. Only 2% of the occurrences were considered literal.

[H5] Detailed morphosyntactic information is helpful for MWE identification.

Two of our experiments show that the quality of MWE identification increases when data from parsers and MWE-specific lexical and morphosyntactic information are combined (papers 2 and 4). Data about a small set of MWEs was analysed firstly, and an automatic analysis method was then proposed, aiming at reducing manual work and consequently increasing the number of MWEs covered. Results were very good, with an F score of 0.51 using data from the PARSEME shared task on automatic identification of verbal MWEs. This result

is 28 points higher than the average score and 13 points higher than the best result obtained for Spanish in edition 1.1. Besides, when the MWEs annotated in the Test part were also considered (and not only the ones in the Train and Development parts), F score increased up to 0.72.

[H6] MWE-specific linguistic information is beneficial for MT.

This was confirmed for a Spanish-Basque rule-based MT system (papers 3 and 4). According to a manual evaluation, 62–65% of the studied MWEs were better translated when specific linguistic information was added to the system, and only 8% got worse translations than the baseline. On the other hand, an improvement was also appreciated by using statistical measures, with an increase of 2.25% in BLEU.

3.2 Contributions

Apart from confirming the hypotheses in the previous section, a number of contributions were made through the work in this PhD. The main ones are listed below.

[C1] Comprehensive NLP-applicable study of verb+noun MWEs in Spanish and Basque.

Although there exist other studies on verb+noun MWEs in both languages, the one in this PhD differs from them in two main aspects. On the one hand, because it is NLP-oriented, unlike most of the phraseological analyses carried out for Spanish and Basque. On the other hand, because most of the data obtained from it is quantified, which is helpful to see the extent of the MWE-specific features under study. Furthermore, as will be shown in contribution 6, all data were made publicly available.

[C2] Analysis of the translation of verb+noun MWEs between Spanish and Basque.

Almost no research has been undertaken about phraseology in Spanish-Basque translation, and this work brings a contribution into the field. The verb+noun entries and translations in the Elhuyar dictionary were firstly analysed, and translations were automatically extracted from parallel corpora for further MWEs from other sources. In both analyses, lexical and morphosyntactic features were examined, to see how these change when MWEs are translated from one language to the other.

[C3] Proposal or application of methodologies which are adaptable to other languages.

The idea of replicability and reusability of our methods in several languages was present throughout the whole work. Firstly, in order to test whether the proposed analysis was applicable to other languages, the manual study of Spanish verb+noun MWEs was undertaken also in English. The output data was then used for an MWE identification experiment, where results were even better than the Spanish ones. Secondly, part of the method proposed to automatise the analysis of MWEs was reused on Basque corpora to gather translation-oriented information. Only a few modifications needed to be done, which means that it is easily adaptable to languages other than Spanish. Thirdly, the PARSEME universal guidelines were followed to annotate verbal MWEs in a Basque corpus, just like in 19 other languages. Fourthly, the study on literal occurrences of MWEs was also done in five languages of different phylogenetic families. And finally, it must be pointed out that the analysis method proposed in this PhD was recently reused within a study of MWEs in Catalan.

[C4] Improvement of the identification of verb+noun MWEs.

The identification method proposed in this PhD outperforms all results in the Spanish part of the PARSEME shared task edition 1.1. As a matter of fact, an F score of 0.51 was obtained, which is 13 points higher than the best-performing system in the Spanish task. Besides, it was made evident precisely what morphosyntactic features are helpful for identification.

[C5] Integration of MWE-specific linguistic data into MT.

Lexical and morphosyntactic information specific to a set of verb+noun MWEs was added to the *Matxin* rule-based MT system, and results were better than the basic system both according to a manual evaluation and according to statistical measures. As explained in hypothesis and conclusion 6, translations were better in 62–65% of the cases according to human evaluators, and 2.25% better as per BLEU.

[C6] Creation of a database collecting all MWEs and translations covered in this work, along with NLP-applicable linguistic data.

The *Konbitzul* database is publicly accessible online[4]. Its interface enables users to make queries according to several criteria and filters, and all NLP-applicable information can be fully downloaded[5]. In all, 1,927 Spanish MWEs (along with 4,043 translations) and 2,074 Basque MWEs (along with 3,022 translations) are collected in it, out of which

[4] `http://ixa2.si.ehu.es/konbitzul/`

[5] `http://ixa.eus/node/4484`

894 Spanish MWEs and their translations contain NLP-applicable information.

[C7] **Annotation of Spanish and (especially) Basque corpora from a phraseological perspective.**

The PARSEME multilingual corpus comprises texts of 20 different languages, including Spanish and Basque. Its annotation was carried out in two phases, and we contributed to both of them: in the first edition, as part of the Spanish annotation team; in the second one, by participating in the process of enhancing the guidelines and, more importantly, by creating the Basque corpus, which consists of 11,158 sentences (157,807 words) and 3,823 MWE annotations. Then, literal occurrences were also studied and annotated on five of the languages in the PARSEME corpus, including Basque. Both the original PARSEME corpus and the one including annotations about literal occurrences are publicly available online[6].

3.3 Future work

The work in this PhD can be extended in several ways.

- **Enhancement and expansion of the *Konbitzul* database.** The database will keep being fed with new MWEs and linguistic information. Basque MWEs and their Spanish translations will be considered in particular, since mostly Spanish into Basque translation was covered so far. Furthermore, the intention is to include MWEs other than the verb+noun type, such as the ones consisting of a verb and an adjective, or of a verb and an adverb.

- **Application of the analysed MWE-specific data for statistical and neural MT systems.** The data obtained from our linguistic analyses was only tested in a rule-based system in this PhD, although most of it is reusable in other kinds of tools. Continuing on MT research, it would be interesting to study what the effect of MWE-specific data is on the translation quality of statistical and neural systems.

- **Study of different strategies to solve semantic ambiguity.** One of the main hypotheses in this PhD is that morphosyntactic information is useful to disambiguate most ambiguities concerning MWEs. However, a few ambiguous cases cannot be clarified without further seman-

[6]The PARSEME multilingual corpus can be downloaded from `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842`. The corpus including annotations about literal occurrences can be found here: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2966`.

tic information, and there would be interest in examining these by using distributional methods and other similar strategies.

- **Use of specialised corpora.** A great number of MWEs being domain-specific, not only general but also specialised corpora should be exploited in order to compare results, as well as to have a broader perspective. Besides, specialised phraseology is an understudied field especially in Basque, meaning that the collection and analysis of domain-specific MWEs would play a part in filling this gap.

- **Closer look at translation patterns concerning MWEs.** When proposing a lexical-semantic classification of MWEs, hypotheses about their behaviour in translation were made (for example, that metaphoric idioms are more prone to receive literal translations than opaque idioms). Other work was prioritised at the time, and these hypotheses are still to be tested by looking at parallel corpora.

Bibliography

Anastasiou, D. (2008). Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. In *Proceedings of the 12th Conference of the European Association of Machine Translation (EAMT 2008)*, pages 12–20. Hanburg, Germany.

Baldwin, T. and Kim, S. N. (2010). Multiword Expressions. *Handbook of Natural Language Processing*, 2:267–292.

Barreiro, A. (2008). *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation*. PhD thesis, Universidade do Porto.

Blunsom, P. and Baldwin, T. (2006). Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 conference on Empirical Methods in Natural Language Processing*, pages 164–171. Sydney, Australia.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (at COLING 2012)*, pages 95–108. Mumbai, India.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword Expression processing: a survey. *Computational Linguistics*, 43(4):837–892.

Constant, M. and Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world*, pages 49–56. Portland, USA.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on Multiword Expressions (at ACL 2007)*, pages 41–48. Prague, Czech Republic.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., and Flickinger, D. (2002). Multiword expressions: Linguistic precision and reusability. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2002)*, pages 1941–1947. Canary Islands, Spain.

Corpas Pastor, G. (1996). *Manual de fraseología española*. Editorial Gredos.

Etchegoyhen, T., Martinez Garcia, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes Etxabe, I., Jauregi Carrera, A., Ellakuria Santos, I., Martin, M., and Calonge, E. (2018). Neural Machine Translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 139–148. Alicante, Spain.

Evert, S. (2009). Corpora and collocations. In *Corpus linguistics. An international handbook*, volume 2, pages 1212–1248. De Gruyter.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., ORegan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gurrutxaga, A. and Alegria, I. (2013). Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125. Atlanta, USA.

Hashimoto, C. and Kawahara, D. (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the conference on empirical methods in natural language processing*, pages 992–1001. Honolulu, Hawaii.

Hashimoto, C., Sato, S., and Utsuro, T. (2006). Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 353–360. Sydney, Australia.

Heylen, D., Maxwell, K. G., and Verhagen, M. (1994). Lexical functions and machine translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japan.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Sydney, Australia.

Labaka, G. (2010). *EUSMT: incorporating linguistic information to SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.* PhD thesis, University of the Basque Country, UPV/EHU.

Labaka, G., España-Bonet, C., Màrquez, L., and Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2):91–125.

Mayor, A., Alegria, I., De Ilarraza, A. D., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based Machine Translation system for Basque. *Machine Translation*, 25(1):53–82.

Monti, J., Barreiro, A., Elia, A., Marano, F., and Napoli, A. (2011). Taking on new challenges in multi-word unit processing for machine translation. In *Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 11–19. Barcelona, Spain.

Oepen, S., Dyvik, H., Lønning, J. T., Velldal, E., Beermann, D., Carroll, J., Flickinger, D., Hellan, L., Johannessen, J. B., and Meurer, P. (2004). Som å kapp-ete med trollet? Towards MRS-based Norwegian-English Machine Translation. In *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Alicante, Spain.

Oflazer, K., Say, B., et al. (2004). Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71. Barcelona.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, pages 2473–2479.

Ramisch, C. (2015). *Multiword Expressions acquisition: a generic and open framework*. Springer.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Mititelu, V. B., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngür, T., Hawwari, A., Iñurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebskind, C., Monti, J., Parra, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, pages 222–240. Santa Fe, USA.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). MWEtoolkit: a framework for Multiword Expression identification. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, volume 10, pages 662–669. Valletta, Malta.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: a pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., and Gompel, M. v. (2018). PARSEME multilingual corpus of Verbal Multiword Expressions. In *Multiword Expressions at length and in-depth: extended papers from the MWE 2017 workshop*, pages 87–147. Berlin Language Science Press.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., and Stoyanova, I. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (at EACL 2017)*, pages 121–126. Valencia, Spain.

Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Seretan, V. (2011). *Syntax-based collocation extraction*. Springer Science Business Media.

Shigeto, Y., Azuma, A., Hisamoto, S., Kondo, S., Kouse, T., Sakaguchi, K., Yoshimoto, A., Yung, F., and Matsumoto, Y. (2013). Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 139–144. Atlanta, USA.

Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Athens, Greece.

Tan, L. and Pal, S. (2014). Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (at ACL 2014)*, pages 201–206. Baltimore, USA.

Tu, Y. (2012). *English complex verb constructions: identification and inference*. PhD thesis, University of Illinois at Urbana-Champaign.

Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala*. PhD thesis, University of the Basque Country (UPV/EHU).

Wehrli, E., Seretan, V., Nerima, L., and Russo, L. (2009). Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *13th Annual Conference of the European Association for Machine Translation*, pages 128–135. Barcelona, Spain.

Appendix

This appendix includes a copy of the publications related to this dissertation, in the recommended reading order.

[**P1**] Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. (2018) **Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system**. In *Multiword Units in Machine Translation and Translation Technologies*, pages 41–60. John Benjamins Publishing Company.

[**P2**] Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K., Carroll J. (2016) **Using linguistic data for English and Spanish verb-noun combination identification**. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016): Technical Papers*, pages 857–867. Osaka, Japan.

[**P3**] Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. (2017) **Rule-based translation of Spanish verb-noun combinations into Basque**. *Proceedings of the 13th Workshop on Multiword Expressions (at EACL2017)*, pages 149–154. Valencia, Spain.

[**P4**] Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. (under review) **Learning about phraseology from corpora: A linguistically motivated approach for Multiword Expression identification and translation**. Sent to *PLOS ONE*.

[**P5**] Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. (2018) **Konbitzul: an MWE-specific database for Spanish-Basque**. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC2018)*, pages 2500–2504. Miyazaki, Japan.

[**P6**] Iñurrieta U., Aduriz I., Estarrona A., Gonzalez-Dios I., Gurrutxaga A., Urizar R., Alegria I. (2018) **Verbal Multiword Expressions in Basque corpora**. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, pages 86–95. Santa Fe, New Mexico, USA.

[**P7**] Savary A., Cordeiro S.R., Lichte T., Ramisch C., Iñurrieta U., Giouli V. (2019) **Literal occurrences of multiword expressions: rare birds that cause a stir**. In *The Prague Bulletin of Mathematical Linguistics*, pages 5–54.

Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system. In Mitkov R., Monti J., Corpas Pastor G., Seretan V. (eds). *Multiword Units in Machine Translation and Translation Technologies*, pp. 41–60. John Benjamins Publishing Company. 2018.
https://benjamins.com/catalog/cilt.341

Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system

Uxoa Inurrieta, Itziar Aduriz*, Arantza Diaz de Ilarraza, Gorka Labaka, Kepa Sarasola
Ixa NLP group, University of the Basque Country
∗Department of Linguistics, University of Barcelona
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu,
a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

Abstract

This paper describes an in-depth analysis of noun+verb combinations in Spanish-Basque translations. Firstly, we examined noun+verb constructions in the dictionary, and confirmed that this kind of MWU varies considerably from language to language, which justifies the need for their specific treatment in MT systems. Then, we searched for those combinations in a parallel corpus, and we selected the most frequently occuring ones to analyse them further and classify them according to their level of syntactic fixedness and semantic compositionality. We tested whether adding linguistic data relevant to MWUs improved the detection of Spanish combinations, and we found that, indeed, the number of MWUs identified increased by 30.30% with a precision of 97.61%. Finally, we also evaluated how an RBMT system translated the MWUs we analysed, and concluded that 42.96% needed to be corrected or improved.

1 Introduction

Multi Word Units (MWUs) are word combinations that pose difficulties to many research areas, as they do not usually follow the common grammatical and lexical rules of languages. Although they are made up of more than one lexeme, they are often used as a single unit in a sentence, and sometimes their meaning is not even transparent, which makes them particularly tricky for Natural Language Processing (NLP).

(1) a. *She always ends up **spilling the beans**.* → lit. giving information away
b. *They **buried the hatchet**.* → lit. stopped arguing

This kind of word combination is also highly variable cross-linguistically and, as it is a very common phenomenon in all types of texts, it presents an additional challenge to multilingual systems like Machine Translation (MT), especially if the source and target languages are from different language families.

(2) 'to kid/trick someone'
EN: ***pull someone's leg***
ES: ***tomar el pelo** [a alguien]*
take the hair [to someone]
take.INF ART.M.SG hair [PREP someone]

EU: *[norbait-i]* ***adar-a jo***
[someone-to] horn-the play
[someone-DAT] horn-ART.SG play.INF

The work presented in this chapter has been done within the framework of Computational Linguistics, and therefore, it involves both a linguistic analysis and an experiment aimed at improving a computer application. More specifically, our object is to analyse the translation of Spanish MWUs into Basque, in order to improve the existing MT system, which is based on linguistic rules and, up to now, has used a very basic method to process MWUs.

As we believe that linguistic data particular to MWUs is necessary in order to obtain good processing results, we undertook an in-depth analysis of a set of word combinations and their possible translations, with the aim of adding information both to the Spanish parser and to the Basque generation process.

In this paper, we will first give an overview of the challenges posed by MWUs to MT systems, and will discuss different techniques that have been used to meet those challenges (part 2). Secondly, we will give some information about our linguistic analysis: the features we focused on, a selection of statistical data, and the criteria we followed for the classification of the combinations (part 3). Finally, we will present our experiment and will show how it improves our system (part 4).

2 Definitions, challenges and treatment of MWUs in MT

Although authors usually agree when it comes to the most important features of MWUs, there are almost as many definitions as researchers in the field. The broadest definition is probably the one given by Sag et al. (2002), who define them as lexical items that can be decomposed into multiple lexemes and that display some kind of idiomaticity, which, according to Baldwin and Kim (2010), can be of several types: lexical (ad hoc), syntactic (by and large), semantic (kick the bucket), pragmatic (good morning), or statistical (immaculate performance, black and white).

In fact, idiomaticity is understood as a key factor of this kind of word combination by other authors too (Gurrutxaga & Alegria, 2011), and forms the basis of a number of classifications. Howarth (1998), for example, proposes a three-layer grouping in which the last layer corresponds to the division between idiomatic and non-idiomatic combinations (see Table 1).

Functional expressions		**Composite units**			
non-idiomatic	idiomatic	Grammatical composites		Lexical composites	
		non-idiomatic	idiomatic	non-idiomatic	idiomatic

Table 1: Howarth's classification of word combinations

Other classifications follow different criteria to sort MWUs, like the one created by Corpas Pastor (1996) for Spanish combinations, which has later been reused and adapted to other languages, including Basque (Urizar, 2012). Its main focus is upon two features of what she terms Phraseological Units: whether they are complete speech acts or not, and the nature of their fixedness (see Table 2).

Phraseological statements	**Collocations**	**Idioms**
fixed in speech	fixed in norms of usage	fixed in the system
complete speech acts	not complete speech acts	

Table 2: Corpas Pastor's classification of Phraseological Units

As regards the computational treatment of MWUs, however, it is essential to take into account their level of syntactic fixedness. While some approaches focus solely on word combinations that are indivisible, if we take a look at real texts, it soon becomes evident that a large number of them can be separated by other words, and sometimes even the word order can be changed. Therefore, this can be a determining feature in the adequate processing of a given combination.

Sag et al. (2002), for example, make a distinction between institutionalised and lexicalised phrases, and rank the latter as fixed, semi-fixed or syntactically free.

Institutionalised phrases	Lexicalised phrases		
	Fixed	Semi-fixed	Syntactically flexible

Table 3: Classification of Multiword Expressions by Sat *et al.* (2002)

These kinds of expressions are used very frequently both in oral and written texts, and are hence important linguistic phenomena to be borne in mind for NLP systems. Jackendoff (1997) estimates that the number of MWUs in an English speaker's lexicon is of the same order of magnitude as the number of single words, and, indeed, 41% of the entries in WordNet 1.7 (Fellbaum, 1998) are constituted of more than one word.

Thus, word combinations pose an important challenge to NLP in general (Sag et al., 2002; Villavicencio et al., 2005), but even bigger when the language to be processed has a rich morphology, as with Basque (Alegria et al., 2004). Furthermore, difficulties multiply when it comes to multilingual systems, as MWUs vary a great deal from one language to another, especially when the languages are very different. As stated in Baldwin and Kim (2010):

> "There is remarkable variation in MWEs across languages (...) There are of course many MWEs which have no direct translation equivalent in a second language. (...) Equally, there are terms which are realised as MWEs in one language but single-word lexemes in another."

As a matter of fact, Simova and Kordoni (2013) studied the translation of English phrasal verbs into Bulgarian, and found out that asymmetry is a major problem when translating word combinations:

> "MWEs constitute a major challenge, since it is very often the case that they do not receive exact translation equivalents. (...) In Bulgarian, phrasal verbs do not occur as multiword units, but are usually translated as single verbs."

On the other hand, regarding MT systems, there are two major issues to be addressed: (1) the identification of MWUs in the source language, and (2) their adequate transfer into and correct generation in the target language. Concerning the identification process, the most basic method is probably the words-with-spaces strategy, which consists in searching solely for sequential word combinations (Zhang et al., 2006; Alegria et al., 2004). Nonetheless, as previously mentioned, non-sequential combinations are as frequent as the sequential ones, and this approach does not allow us to find them.

It is important to use a flexible method which allows the detection of as many combinations as possible, but also to impose some restrictions, so that only real MWUs are detected. The tendency of recent years has been to combine computational methods, like association measures, with linguistic features (Dubremetz and Nivre, 2014; Pecina, 2008). For example, information obtained from deep parsers has been proved to be very helpful (Baldwin et al., 2004; Blunsom, 2007).

It must be noted, however, that, while a lot of detection and extraction work has been done, not that much research has been conducted on MWU integration into MT systems. Most reports explain experiments in which combinations are added to Statistical Machine Translation (SMT) systems (Bouamor et al., 2012; Tsvetkov and Wintner, 2012), all of which greatly improve translation quality. As is pointed out in Seretan (2013):

> "Phrase-based SMT systems already incorporate MWE/collocational knowledge as an effect of training their language and translation models on large (parallel) corpora. These systems are successful in dealing with local collocations, but are arguably ill-suited for handling collocations whose components are not in close proximity to one another."

Meanwhile, integration experiments on Rule-Based Machine Translation (RBMT) systems have also been confirmed to have a very positive effect. Wehrli et al. (2009), for instance, replaced the parsing strategy in an RBMT system with a new one which integrated collocation identification, and obtained much better results regarding MWU translation adequacy.

It must be mentioned that, according to studies, even the simplest treatment of MWUs improves translation quality, although, of course, more complex processing methods will obtain better results, especially concerning non-sequential word combinations (Copestake et al., 2002).

3 Linguistic analysis of Basque and Spanish noun+verb combinations

As previously mentioned, our aim is to study MWUs and their translations, in order to establish the linguistic grounds for their appropriate treatment in MT systems. So, we focused on several features of noun+verb combinations in Basque and Spanish, and we analysed how they were translated.

First of all, we gathered noun+verb combinations from bilingual dictionaries, and we looked at their morphological composition and some semantic features (see 3.1). Secondly, we searched for these combinations in a parallel corpus, so that we could check to what extent they were used in real texts and how they were translated (see 3.2). Thirdly, we chose the most frequent combinations in the corpus, and classified them according to their syntactic flexibility and their semantic compositionality (see 3.3).

All of our results are collected in a public database: Konbitzul[1]. It is now available online, and it allows users to search for the appropriate translation of a given combination, along with all the linguistic data we garnered from our in-depth analysis.

3.1 Noun+verb combinations in bilingual dictionaries

Although it was clear to us that parallel corpora were the most useful resource for extracting frequently-used word combinations, we decided to take a look at bilingual dictionaries first, in order to get a general idea of the translation challenges the combinations can pose. To that end, we used the Elhuyar dictionaries1 (Spanish into Basque and Basque into Spanish), from which we gathered 2,954 Basque combinations (along with 6,392 Spanish equivalents) and 2,650 Spanish combinations (along with 6,587 Basque equivalents).

All of the Basque combinations we analysed consisted of just a noun and a verb. However, it is important to note that Basque is an agglutinative language and, as such, constructs phrases by attaching elements, typically at the end of the phrase (Laka, 1996). This means that the nouns that are used in MWUs can also be marked by different grammatical cases and postpositions.

(3) a. ***lan egin***
work do
work.ABS do.INF
'to work'

b. ***deabru-a-k hartu***
devil-the takes
devil-ART.SG-ERG take.INF
'the devil take [someone/something]'

c. ***joko-a-n jarri***
game-the-in put
game-ART.SG-LOC put.INF
'to risk'

d. ***buru-tik egon***
head-the.from be
head-ART.SG.ABL be.INF
'to be crazy'

Spanish, on the other hand, uses prepositions instead of postpositions and grammatical cases, and determiners in Spanish are not morphemes attached to the phrases, but always separate words. Therefore, of the Spanish combinations we selected for this study, each one consisted of at least a verb and a noun, but many of them also contained prepositions and/or determiners in-between.

[1] http://ixa2.si.ehu.eus/konbitzul

(4) a. ***tener afecto***
have affection
have.INF affection
'to have affection'

b. ***ser una pena***
be a pity
be.INF ART.F.SG pity
'to be a pity'

c. ***saber de memoria***
know by memory
know.INF PREP memory
'to know by heart'

d. ***dejar a un lado***
leave to a side
leave.INF PREP ART.M.SG side
'to leave aside/to one side'

We focused on the combinations in each language separately first, without taking their translations into account (see 3.1.1). Then, we examined their translations (see 3.1.2), paying special attention to those combinations that are also translated by noun+verb constructions (see 3.1.3).

3.1.1 Basque and Spanish noun+verb combinations in the dictionary

To begin with our analysis, we focused on the morphological composition of the combinations in the Elhuyar dictionaries. As we mentioned earlier, the Basque combinations we chose for this project consisted of a noun and a verb (see example 3), while the Spanish combinations were of four types:

- verb + noun (example 4a)
- verb + determiner + noun (example 4b)
- verb + preposition + noun (example 4c)
- verb + preposition + determiner + noun (example 4d)

Concerning the Basque combinations in our list, we found many kinds of morphemes attached to the end of the nouns: three grammatical cases, and ten different postpositional marks. However, not all of them were used as often. As a matter of fact, 76.18% of the nouns were in the absolutive case, and the rest of the cases and postpositional marks were hardly used. On the other hand, there was no such difference among the Spanish structures, even though the combinations of the type verb + determiner + noun were slightly more common than the rest (37.70%).

It is also interesting to note that a large number of the verbs in the combinations are very common, both in Basque and in Spanish. In addition, the most frequent verbs in both languages are equivalent to each other: egin hacer ('do'), izan ser/estar/tener ('be/have'), eman dar ('give'), hartu tomar ('take') and so on. This is no surprise though, as light verb constructions are very frequent among MWUs (Sag et al., 2002; Butt, 2010).

3.1.2 Translations of noun+verb combinations in the dictionary

As a second step, we looked at the dictionary translations of the combinations we had extracted. When translating between languages from the same family, most word combinations in the source language were also word combinations in the target one. However, this is not the case in Spanish into Basque translations, where asymmetry is much more in evidence.

In fact, of the Spanish translations of Basque combinations we analysed, 58.07% were single verbs, while just 30.85% contained a noun and a verb. This was to be expected, given that in Basque, it is very common to use two-word verbs to represent some actions that are expressed with single verbs in most

European languages (see example 5). On the other hand, this asymmetry was slightly less prominent but still significant when Spanish was the source language, as fewer than half of the Basque equivalents (48.54%) were noun+verb combinations (see example 6).

(5) 'to work'

EU: *lan egin*
work do
work.ABS do.INF

ES: *trabajar*
work
work.INF

(6) 'to open one's eyes'

ES: *abrir los ojo-s*
open the eye-s
open.INF ART.M.PL eye-PL

EU: *begi-ak ireki*
eye-s open
eye-ART.PL.ABS open.INF

3.1.3 Equivalences of noun+verb constructions in translations

Before finishing our dictionary-based study, we considered it worth analysing syntactically-symmetrical translations further. So, we selected those noun+verb constructions that were also translated by other noun+verb constructions, and we found that there was a link between the morphological composition of the combinations in both languages.

As previously mentioned, the natural equivalents of Basque postpositions are prepositions in Spanish. Our study has found that, despite their high idiosyncrasy, MWUs are not always an exception to this rule, as most Spanish combinations containing a preposition in our list were translated by combinations with a postposition into Basque, and vice versa.

(7) 'to eat hungrily'

ES: *comer con apetito*
eat with appetite
eat.INF PREP appetite

EU: *gogo-z jan*
desire-with eat
desire-INS eat.INF

(8) 'to be a case in point'

EU: *hari-ra etorri*
string-to.the come
string-ART.SG.ALL come.INF

ES: *venir al caso*
come to.the case
come.INF PREP.ART.M.SG case

This symmetry, however, is not consistent when it comes to the (in)definiteness and singularity/plurality of noun phrases, which is usually highly irregular cross-linguistically. The only exceptions are indefinite Basque nouns, which mostly remain indefinite when the combinations are translated into Spanish (80.72%).

To conclude, we found it pertinent to make a comparison between the noun phrases and verbs in the source language and those in the target language. As we had expected, very few combinations were translated by substituting each component with an equivalent (see example 9). Most of the time, at least

one of the components was translated by a word that was not its equivalent in the dictionary (see example 10).

(9) 'to leave [somebody/something] alone'

ES: *dejar en paz*
leave in peace
leave.INF PREP peace

EU: *bake-a-n utzi*
peace-the-in leave
peace-ART.SG-LOC leave.INF

(10) 'to make noise'

ES: *armar bulla*
build racket
build.INF racket

EU: *zarata egin*
noise make
noise.ABS make.INF

3.2 Contrasting information with parallel corpora

The dictionary-based analysis provided us with a general view of the high complexity of MWU translation, but in order to learn about the actual use of these units, we needed to look at real texts. To do this, we used a parallel corpus of Spanish into Basque translations, constituted of 491,853 sentences from many different sources.

Out of the 2,650 combinations we had gathered from the dictionary, just 200 were found within the corpus. However, we did not search for whole word sequences, but for noun lemmas and verb lemmas only, accepting any preposition and/or determiner in-between. This allowed us to find many other variants of the combinations we had already analysed (see example 11), and, in addition, we also added new combinations that could be worth examining. These variants and extra combinations numbered 698 in all.

(11) Previously examined: *alzar la voz*
raise the voice
raise.INF ART.F.SG voice
'to raise the voice'

New variant 1: *alzar su voz*
raise his/her voice
raise.INF POS.3SG voice
'to raise his/her voice'

New variant 2: *alzar voces*
raise voices
raise.INF voice.PL
'to raise voices'

On the other hand, while the aforementioned 200 combinations had no more than 385 Basque equivalents in the dictionary, they were translated in as many as 1,641 different ways in the corpus, which enabled us to feed new translations into our database.

3.3 Classification of the Spanish MWUs

For the next study, we ranked all the combinations extracted from the corpus by their number of occurrences, and we selected the most frequently-used ones: a total of 150. Our aim this time was to analyse linguistic information that could be useful for MT systems, so we focused on two main features of the Spanish combinations: (1) their syntactic flexibility, and (2) their semantic compositionality.

3.3.1 Syntactic flexibility

In order to measure how flexible the combinations were, we asked the following questions about each of them:

- Was the noun phrase definite or indefinite? Was this consistent for every occurrence?
- Was the noun phrase singular or plural? Was this consistent?
- Could the noun phrase include a modifier? Adjectives, prepositional phrases and so on.
- Was it possible to add something between the noun phrase and the verb? An adverb, an extra phrase etc.
- Could the order of the components be changed? In passive sentences, for example.

As our judgement was that syntactic information was a key element for the adequate treatment of a given MWU, we used that information to sort the combinations into three groups, following Sag et al. (2002): fixed, semi-fixed and free (see Table 4).

Fixed expressions	0%
Semi-fixed expressions	30.67%
Syntactically free expressions	66.67%

Table 4: Syntactic classification of Spanish MWUs.

We call fixed expressions those word combinations that are always used together, using the same word forms (except for the verb, which can be inflected) and the same word order. Therefore, the MWUs in this group should be detected easily, simply by searching for the lemma of a given verb and the word sequence that follows it.

(12) *dar paso [a algo]* ('to give raise [to something]')

a. *Las elecciones dieron paso a un nuevo gobierno.*
'The elections gave raise to a new government.'

b. **El paso al nuevo gobierno lo dieron las elecciones.*
'The raise to the new government was given by the elections.'

Semi-fixed expressions, on the other hand, are more problematic regarding automatic detection tools. The components of these kinds of MWUs are often separated by other words (example 12), and even the word order can be changed, for example when the sentence is in the passive voice. They are not completely free though, as they have certain syntactic restrictions, such as that modifiers and/or determiners cannot be inserted. It is important to take those restrictions into account in order to detect only the combinations we are interested in, as in examples 13a and 13b, where the first one is an MWU whereas the second one is not.

(13) *hacer memoria* ('to try to remember' vs. 'to do a report')

a. *Haz memoria, Qué hiciste ayer?*
'Try to remember: what did you do yesterday?'

b. *Harán una memoria exhaustiva sobre su labor.*
'They will do a comprehensive report on their activities.'

The combinations we classified as free expressions do not seem to have any syntactic restriction. As a result, the MWUs in this group are probably the most difficult ones to detect.

(14) *fijar un plazo* ('to set a deadline')

a. *Hemos fijado el plazo de inscripción.*
'We have set the enrolment deadline.'

b. *El plazo de inscripción ha sido fijado.*
'The enrolment deadline has been set.'

c. *Cuál es el plazo de inscripción que se ha fijado?*
'What deadline has been set for enrolment?'

3.3.2 Semantic compositionalty

Apart from analysing the syntax of the combinations, we also considered it important to look at their meaning. We sorted the combinations into four groups, depending on their degree of semantic idiomaticity.

Non-compositional expressions	2%
Figurative expressions	10.67%
Semi-compositional expressions (collocations and light verb constructions)	52%
Compositional expressions (free)	35.33%

Table 5: Semantic classification of Spanish MWUs.

Non-compositional expressions are word combinations in which the meaning is not derivable from the separate meanings of their components. They are also called opaque expressions.

(15) *llevar a cabo*
take to ending
take.INF PREP ending
'to carry out', 'to do'

Figurative expressions, on the other hand, are combinations which can have a figurative sense in addition to the canonical one.

(16) *poner [algo] sobre la mesa*
put on the table
put PREP ART.F.SG table
'to put [something] on the table' or 'to draw attention [to something]'

In the case of semi-compositional expressions, one of the components keeps its literal meaning, while the other one adopts a new sense (in collocations) or is emptied of meaning to work as a supporting element for the other word (in light verb constructions). In verb+noun combinations, the component which keeps its original meaning is usually the noun.

(17) *cumplir su palabra*
fulfil his7her word
fulfil.INF POS.3SG word
'to keep his/her word'

(18) *tener dificultad [para algo]*
have difficulty
have.INF difficulty
'to have difficulty [doing something]', 'to find [something] difficult'

Finally, compositional expressions are completely regular in terms of semantics, as their meaning is made up of the separate meanings of the components. Hence, the constructions in this group are not semantically idiomatic, and most of them do not need any special computational treatment, as their literal translation is usually correct.

(19) *ir a un lugar*
go to a place
go.INF PREP ART.M.SG place
'to go to a place'

4 Evaluation of MWU detection and translation adequacy

As we mentioned earlier, the aim of our work is to establish the linguistic basis for the treatment of MWUs in MT systems. The experiment we will explain here was carried out with an RBMT system, namely Matxin1 (Mayor et al., 2011), which translates from Spanish into Basque.

Matxin works in three phases: (1) analysis, (2) transfer and (3) generation. In the first phase, it analyses the text in Spanish syntactically, based on the information given by Freeling 3.0 (Padr and Stanilovsky, 2012). Secondly, it transfers the structure of the sentences to be translated, as well as the lexicon, which is gathered from wide-coverage dictionaries. And in the third place, the words and phrases are re-ordered and the necessary morphological information is added to them.

Before we used our linguistic data, the system already had a MWU processing method, but it was based on the words-with-spaces approach (see section 2) and was thus unable to identify non-sequential word combinations (see example 13). The old MWU detection system was part of the analysis process, and searched only for the lemmas of the verbs and the forms of the rest of the words, which made it impossible to find combinations in which the components were non-adjacent and/or used a different order or word forms.

The new system, however, is based on all the data we acquired from the linguistic analysis presented in section 3.3. It is much more flexible, but, at the same time, it has many restrictions that prevent the identification of free combinations as MWUs. If a given combination is marked as a fixed expression, the system employs the old strategy, as this kind of MWU is always sequential and unchangeable (except for the verb inflection, which is also taken into account). If the unit is marked as semi-fixed, on the other hand, the system looks at the linguistic data we provided.

For the expression cambiar de tema ('change the topic'), for example, the system identifies those word combinations in which:

- The noun phrase is singular and definite, and preceded by the preposition de. According to this constraint, example 16a would be accepted, whereas 16b would not.

  (20) a. *Cambiemos de tema.*
  'Let's change the topic.'
  b. **Cambiemos de los temas.*
  'Let's change the topics.'

- The noun phrase has no modifier.

  (21) **Cambiemos de aburrido tema*.
  'Let's change the boring topic.'

- There may be more words between the verb and the prepositional phrase.

  (22) *Cambiemos inmediatamente de tema*.
  'Let's change the topic inmediately.'

- The word order cannot be changed (see example 19).

  (23) **De tema han cambiado*.
  'The topic, they changed.'

Here again, we used the linguistic analysis provided by Freeling 3.0, combined with the linguistic data we had manually analysed. This was very helpful for limiting, on the one hand, the number of words between the verb and the noun phrase that constitute the MWUs, and on the other hand, the modifiers that could be inside the noun phrase.

In the following sections, we will compare the results of the old and new detection systems, and will also evaluate the performance of our MT system when it comes to translating the MWUs we analysed.

4.1 Evaluation of MWU detection

To test whether or not our new detection method was useful, we used 15,182,385 sentences in Spanish, taken from the parallel English-Spanish corpus made public for the shared task in WMT workshop 2013[2]. Out of the 150 word combinations we analysed, we discarded those which were neither syntactically nor semantically idiomatic, that is, the ones classified as free and compositional expressions (see Section 3.3). In all, the set we used for the experiment consisted of 117 MWUs.

We did the detection experiment both with the old system and with the new one, and we found that, as we had expected, the method based on linguistic data was able to identify quite a large number of additional combinations. As a matter of fact, of the 433,092 MWUs detected by the new system, 27.80% was constituted of combinations that the old system did not manage to detect (see Table 6).

MWUs identified by both systems	311,966
MWUs identified by the new system only	120,362
MWUs identified by the old system only	764

Table 6: Comparison of the old and new MWU detection systems.

Our next step was to evaluate the combinations that were identified by just one of the systems, so that we could see (1) whether the 97,382 extra combinations detected by our method produced a real improvement, and (2) why we failed to identify 731 combinations that the old system did manage to detect. The evaluation was undertaken manually by linguists, on a representative set of sentences containing MWUs detected by one of the systems only.

Out of the evaluation set, all but one of the MWUs extracted with the words-with-spaces method were correct (99%), and the hit rate obtained by the new system was 95%. Assuming that the accuracy of the old system would still be 98% for the combinations detected by both methods, the total hit rate of our new system would be 98%[3], which would be a very satisfactory result. Therefore, this confirms that linguistic data specific to MWUs does improve the detection process, as the number of identified combinations increased by 27.80% with a very high degree of precision.

On the other hand, when evaluating the correct MWUs that were detected by the old system but not by the new one, we realised that most of them had parsing errors that prevented our method from working correctly. Thus, taking into consideration that the words-with-spaces method is extremely accurate, we decided to use both systems from now on: the old one first, in order to detect all sequential MWUs, and then the new one, which allows us to identify a large number of additional non-sequential combinations. More details about this experiment and its results can be found in (Inurrieta *et al.,* 2016).

4.2 Evaluation of MWU translation quality in an RBMT system

Apart from evaluating the detection quality, we also wanted to get a general picture of the improvement our data would make to Matxin, an RBMT system created by IXA NLP group. So, we translated all 117 MWUs (see section 4.1) using Matxin, and we also provided a manual translation for each of them. We sorted the results into three groups (see Table 7): correct, improvable and incorrect.

The MT is as good as the manual translation	57.02%
The MT is not correct, but the manual translatiton is better	9.91%
The MT is incorrect	33.05%

Table 7: Evaluation of MWU translations given by Matxin.

The results we obtained in this evaluation show that much improvement remains to be made concerning MWU translation in Matxin, as 42.96% of the MTs were incorrect or improvable. In addition, it must be considered that we undertook this test without any context, and this percentage would surely be much

[2] `http://www.statmt.org/wmt13/translation-task.html`

[3] (311,966*99/100 + 120,362*95/100)/(311,966+120,362)

higher if the combinations were used in the context of sentences, especially if they were separated by other words or if a non-canonical word order was used. In (Inurrieta *et al.,* 2017), it is further explained how MWU-specific linguistic data helps improving translation quality in Matxin.

5 Conclusions and future work

In order to establish the grounds for the computational treatment of MWUs in MT systems, we undertook an in-depth linguistic analysis of some word combinations and their translations. First of all, we extracted combinations containing nouns and verbs from bilingual dictionaries: Spanish into Basque, and Basque into Spanish. We examined the morphological and semantic features of both the combinations (5,604) and their translations (12,979) and, as we had expected, we confirmed that MWUs cannot usually be translated word for word and morpheme for morpheme, as this kind of expression varies considerably from language to language.

Secondly, we searched for the combinations analysed in a parallel corpus, which allowed us (1) to know to what extent each combination was used in real texts, and (2) to obtain a large number of additional translations that were not in the dictionary. All of our results were included in our database, Konbitzul1, which is now available for public use.

Then, we selected the 150 most frequent combinations in Spanish, we analysed them further and classified them according to their syntactic fixedness and their semantic compositionality, which helped us determine the kind of treatment that each MWU needed. As we wanted to carry out an experiment with a Spanish into Basque RBMT system, we did an detection test to establish whether the data we provided had a real effect on MWU identification, and we obtained very satisfactory results. On the one hand, the number of MWUs identified increased by 27.80% with our data, and, on the other hand, our method achieved a precision of 98% according to a manual evaluation undertaken by linguists.

Finally, we also evaluated the translations given by our RBMT system for the MWUs we analysed, and we concluded that at least 42.96% of them were either incorrect or improvable, which underscores the need for specific techniques to process MWUs in the systems.

We are currently working on semi-automatising the whole linguistic analysis explained here, so that this methodology can be applied to a larger number of word combinations more easily. In addition, there would be merit in analysing semantic data about MWUs, as we believe this information could make further improvement both to the detection and generation processes.

Acknowledgements

Uxoa Inurrieta's work is funded by a PhD scholarship from the Ministry of Economy and Competitiveness (BES-2013-066372). This research was undertaken as part of the SKATeR (TIN2012-38584-C06-02) and QTLeap (FP7-ICT-2013.4.1-610516) projects.

References

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., & Urizar, R. (2004, July). Representation and treatment of multiword expressions in Basque. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing* (pp. 48–55). Association for Computational Linguistics.

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., & Oepen, S. (2004, May). Road-testing the English Resource Grammar Over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. *Handbook of Natural Language Processing* (2nd ed.). Morgan and Claypool.

Blunsom, P. (2007). *Structured classification for multilingual natural language processing* (Doctoral dissertation, University of Melbourne, Melbourne, Australia).

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012, May). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *LREC 2012, Eigth International Conference on Language Resources and Evaluation*, (pp. 674679). Istanbul, Turkey.

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective* (pp. 4878).

Corpas Pastor, G. (1997). *Manual de Fraseologa Espaola*. Editorial Gredos.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., & Flickinger, D. (2002). *Multiword Expressions: Linguistic Precision and Reusability*. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002, (pp. 19411947). Las Palmas, Spain.

Dubremetz, M., & Nivre, J. (2014). Extraction of Nominal Multiword Expressions in French. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, (pp. 7276,). Gothenburg, Sweden.

Fellbaum, C. (1998). *WordNet*. Blackwell Publishing Ltd.

Gurrutxaga, A., & Alegria, I. (2011, June). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World* (pp. 27). Association for Computational Linguistics.

Heylen, D., & Maxwell, K. (1994). Lexical Functions and the Translation of Collocations. In *Proceedings of Euralex*.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied linguistics*, 19(1), 2444. doi: 10.1093/applin/19.1.24

Inurrieta, U., Aduriz, I., Diaz de Ilarraza, A., Labaka, G., Sarasola, K., & Carroll, J. (2016). Using linguistic data for English and Spanish verb-noun combination identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers* (pp. 857867).

Inurrieta, U., Aduriz, I., Diaz de Ilarraza, A., Labaka, G., & Sarasola, K. (2017). Rule-based trans- lation of Spanish verb-noun combinations into Basque. In *Proceedings of the 13th Workshop on Multiword Expressions, in EACL 2017* (pp. 149154).

Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.

Laka, I. (1996). *A brief grammar of Euskara, the Basque language*. Universidad del Pas Vasco.

Mayor, A., Alegria, I., De Ilarraza, A. D., Labaka, G., Lersundi, M., & Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 25(1), 5382. doi: 10.1007/s10590-011-9092-y

Padr, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey.

Pecina, P. (2008, June). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)* (pp. 5461).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing* (pp. 115). Springer Berlin Heidelberg.

Seretan, V. (2013, October). On collocations and their interaction with parsing and translation. In *Informatics* (Vol. 1, No. 1, pp. 1131). Multidisciplinary Digital Publishing Institute.

Simova, I., & Kordoni, V. (2013, September). Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice, France.

Tsvetkov, Y., & Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04), 549573. doi: 10.1017/S1351324912000101

Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala* (Doctoral dissertation, Faculty of Computer Science, University of the Basque Country).

Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (Eds.). (2005). *Computer Speech & Language (Special issue on Multiword Expressions)*, volume 19. Elsevier.

Wehrli, E., Seretan, V., Nerima, L., & Russo, L. (2009, May). Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation* (pp. 128135).

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006, July). Automated multiword expres- sion prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 3644). Association for Computational Linguistics. doi: 10.3115/1613692.1613700

Using Linguistic Data for English and Spanish Verb-Noun Combination Identification

Uxoa Iñurrieta, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA NLP group, University of the Basque Country
usoa.inurrieta|a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

Itziar Aduriz
Department of Linguistics, University of Barcelona
itziar.aduriz@ub.edu

John Carroll
Department of Informatics, University of Sussex
j.a.carroll@sussex.ac.uk

Abstract

We present a linguistic analysis of a set of English and Spanish verb+noun combinations (VNCs), and a method to use this information to improve VNC identification. Firstly, a sample of frequent VNCs are analysed in-depth and tagged along lexico-semantic and morphosyntactic dimensions, obtaining satisfactory inter-annotator agreement scores. Then, a VNC identification experiment is undertaken, where the analysed linguistic data is combined with chunking information and syntactic dependencies. A comparison between the results of the experiment and the results obtained by a basic detection method shows that VNC identification can be greatly improved by using linguistic information, as a large number of additional occurrences are detected with high precision.

1 Introduction

Multiword Expressions (MWEs) are recurrent combinations of two or more words expressing a single unit of meaning, this meaning not always derivable directly from the meanings of the component words (Sag et al., 2002). Therefore, Natural Language Processing (NLP) tasks that need to be sensitive to lexical meaning should treat MWEs as single units. However, this is a challenging problem since many MWEs can have multimple morphosyntactic variants, which makes them difficult to recognise or generate. Examples (1)-(3) below contain *take steps*; correct translation of this MWE into another language, for instance, requires it to be recognised as a single unit[1].

(1) The Government will *take* all the necessary *steps* to prepare.

(2) They set out five important *steps* the Minister needs to *take*.

(3) What were the *steps* that should have been *taken*?

Although the most straightforward method for recognising MWEs is to attempt to match word sequences against entries in a lexicon, this method does not work for combinations that can have multiple variants. This is often the situation for verb+noun combinations (VNCs), since this kind of MWE is usually morphosyntactically flexible.

In the case of Machine Translation (MT), there are two challenges that need to be addressed concerning VNCs: (1) the detection of a given combination in the source language, and (2) its translation into the target language. If the first part fails, the words that constitute the MWE will be translated separately,

[1]Google Translate English-French and English-Spanish https://translate.google.co.uk apparently detects *take steps* as an MWE in (1) but not in (2) or (3).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 857–867, Osaka, Japan, December 11-17 2016.

which will usually result in an incorrect translation. Then, for the second part, it is vital to have the necessary information to know what translation should be given to each VNC. A further problem arises here, since the morphosyntax of this kind of MWE varies a great deal from one language to another, meaning that it is not necessarily translated by another VNC into the target language. This problem is especially acute when the source and target languages are typologically different, as with English, Spanish and Basque[2]. This is what happens in example (4).

(4) English (EN): *get married* (V+V)
Spanish (ES): *contraer matrimonio* (V+N)
'contract marriage'
Basque (EU): *ezkondu* (V)
'(to) marry'

In this paper, we present a linguistic analysis undertaken with the aim of improving the detection of VNCs in *Matxin* (Mayor et al., 2011), a rule-based MT system which translates English and Spanish into Basque. Although we ground our study in this particular MT system, our methodology, analysis and conclusions are relevant to any kind of NLP task that needs to be sensitive to lexical meaning.

The paper is structured as follows. After discussing related work (Section 2), we present our linguistic analysis (Section 3) including: our procedure for VNC tagging, how we classify the combinations, and levels of inter-annotator agreement. In Section 4 we present a VNC detection experiment, and give the results obtained by combining linguistic information with chunking and dependency parsing. Finally, in Section 5, we draw conclusions and propose directions for future work.

2 Related Work

It is widely acknowledged that good MWE processing strategies are necessary for NLP systems to work effectively (Sag et al., 2002), since these kinds of word combinations are very frequent in both text and speech. It is estimated that the number of MWEs in an English speaker's vocabulary is of the same order of magnitude as that of single words (Jackendoff, 1997), and that at least one MWE is used per sentence on average (Sinclair, 1991).

Various classifications of MWEs have been proposed, employing different criteria to match the requirements of a particular kind of target application. Some researchers propose a binary categorisation of literal and non-literal word combinations (Birke and Sarkar, 2006; Cook et al., 2008), whereas others propose a grading containing several MWE types based on semantic idiomaticity, considered as a continuum (Wulff, 2008). Within the Meaning-Text Theory, collocations are sorted according to the notion of lexical functions (Mel'ćuk, 1998), that is, taking into account how the component words are semantically related. Furthermore, some experiments have investigated automatic methods—such as distributional similarity or word embeddings—for the task of classification, leading to fairly good results (Baldwin et al., 2003; McCarthy et al., 2003; Fazly et al., 2007; Rodríguez-Fernández et al., 2016).

In addition to MWE classification, a great deal of work has been undertaken over the last two decades on MWE acquisition (Ramisch, 2015) and identification (Li et al., 2003; Seretan and Wehrli, 2009; Sporleder and Li, 2009). Precise and detailed syntactic information is crucial for both tasks, and, at the same time, MWE identification can also help parsers obtain better results (Seretan, 2013). Moreover, accurate MWE detection is crucial for MT, since MWEs vary greatly from one language to another, and are not usually translated word for word. In the context of MT systems, Wehrli (2014) states "the non-identification of collocations dramatically affects the quality of the output".

3 Linguistic Analysis

The linguistic analysis we present here aims at improving MWE processing in MT. More specifically, we base our study on Matxin (Mayor et al., 2011), a rule-based MT system for English-Basque and

[2]Whereas English (Germanic) and Spanish (Romance) are Indo-European languages, Basque is a non-Indo-European language which moreover belongs to no known language family.

Spanish-Basque translation. One of the problems Matxin has concerning MWEs is that it currently fails to detect many instances of morphologically flexible word combinations, since it only searches for word sequences against entries in a lexicon.

As mentioned in Section 1, our study foccusses on one particular kind of MWE: verb+noun combinations (VNCs). As well as the principal constituents of a verb and a noun, we also allow for combinations containing a preposition and/or a determiner in between. Candidate combinations were first gathered from machine-readable dictionaries and were then searched for in corpora, the most frequent combinations being selected for detailed analysis.

More details about the procedure for selecting the combinations are given in the following subsections, as well as explanations of a manual tagging process, the criteria used to classify the combinations, and the overall results and conclusions drawn from this analysis. How this information is used for VNC identification is explained in Section 4.

3.1 Selection of Verb+Noun Combinations

The Spanish combinations for this study were extracted from the Elhuyar Spanish-Basque dictionary[3], and the corpus used to obtain frequency information was made up of 491,853 sentences taken from a Spanish-Basque parallel corpus containing a range of text genres. A total of 150 distinct VNCs were selected, each of which occurred more than five times as a word sequence in the corpus.

For English, our original intention was to extract combinations from the Elhuyar *English*-Basque dictionary, in part because the Basque translations would be useful for the translation process in the MT system. However, the dictionary contained too few combinations for this study, so instead we decided to use the Oxford Collocations Dictionary (Deuter, 2008). After extracting the combinations matching our grammatical pattern, we searched for them in the British National Corpus (Burnard, 2007). If the verb and the noun (and the preposition, when necessary) were found as main elements in adjacent chunks more than 500 times, the combination was selected. The final set consisted of 173 combinations in all.

3.2 Tagging Process

The combinations were tagged manually and classified along lexico-semantic and morphosyntactic dimensions, as discussed in the next sections. Although annotators looked at corpora to take decisions, the tagging was not done on instances in a corpus but on combinations out of sentential context. Therefore, each annotator gave each combination a single tag per task.

The lexico-semantic classification was done for two reasons: to determine which combinations were worth detecting and which ones should not be treated as MWEs, and because making groups depending on the combinations' idiomaticity was considered relevant for the later translation process. The morphosyntactic data, on the other hand, was analysed to be used for VNC detection (Section 4).

The tagging was performed by five linguists, all of whom are Spanish native speakers and fluent in English. Firstly, a 'super-annotator' tagged all the data, comprising a total of 323 distinct combinations in Spanish and English. Then, the data were split in four parts, and a further four annotators each tagged one of these parts, following the guidelines created for this purpose.

3.3 Lexico-Semantic Classification

The tags assigned by the annotators separated the combinations into four lexico-semantic groups, from less to more idiomatic: (1) free expressions, (2) collocations and light verb constructions, (3) metaphoric expressions, and (4) idioms. This was not an easy task, as the boundaries between one group and another are not always clearly defined. Idiomaticity is rather understood as a continuum (Wulff, 2008), and some combinations are very difficult to classify (we return to this point in Section 3.5).

Idioms (also called **opaque expressions**) are combinations in which the whole meaning cannot be understood by looking at the meanings of the words separately. Two clear examples of these would be the sentences in examples (5) and (6), which are impossible to interpret correctly without knowledge of

[3] http://hiztegiak.elhuyar.eus/

the figurative meaning of the expressions in italics.

(5) Do not believe her, she is just *pulling your leg*.
= Do not believe her, she is just *joking*.

(6) Ese chico *no se corta un pelo*, es un descarado.
'That boy *does not cut a hair*, he is shameless.'
= That boy *is never intimidated*, he is shameless.

Metaphoric expressions are not used in their literal sense either, but it is possible to understand their meaning in terms of a metaphor, as in examples (7) and (8).

(7) He did not come to the meeting and the boss *had a word* with him.
= He did not come to the meeting and the boss *spoke* with him.

(8) Las experiencias de ese tipo *dejan huella*.
'These kinds of experiences *leave (a) mark*.'
= These kinds of experiences have a very significant effect (on people's life).

Unlike the combinations in examples (5)–(8), those in examples (9) and (10) are easily understandable on the basis of their component words; they belong to the group of **collocations and light verb constructions**. Collocations are defined as lexically constrained and recurrent combinations of words which are in a given syntactic relation (Evert, 2008; Bartsch, 2004). When they are VNCs, the verb is often a very common word which is semantically bleached—meaning that it loses its usual sense to a certain extent (Butt, 2010). These kinds of combinations are called light verb constructions (LVCs). Examples (9) and (10) would be classified in this group.

(9) Volunteers *gave support* to disadvantaged children.

(10) La educación *tiene* vital *importancia* para los niños desaventajados.
'Education *has* vital *importance* for disadvantaged children.'

Finally, **free expressions** are groups of words that can be combined freely, that is, following the standard lexical and grammatical rules of a given language. These kinds of expressions are not idiomatic, and are thus not considered MWEs, as in examples (11) and (12). Therefore, the combinations sorted in this group by the annotators were excluded for the later detection experiment (Section 4).

(11) They *are using a* new *technique* now.

(12) Este año *iremos a un lugar* diferente.
'This year *we will go to a* different *place*.'

As mentioned in Section 3.2, we consider that classifying the VNCs is relevant for translation. Our hypothesis is that the kind of translation a VNC should be given is often dependent on its lexico-semantic class. For instance, the combinations we have analysed so far suggest that, although idioms are usually translated by other (morphosyntactically equivalent or non-equivalent) idioms into the target language, they are unlikely to receive a word-for-word translation (see example (13)). On the other hand, in collocations, the noun is very likely to receive a direct translation, whereas the verb is often given a translation other than the one expected when it is not part of the collocation (see example (14)).

(13) EN: *pull* (somebody)*'s leg*
ES: *tomar el pelo* (a alguien)
'take (somebody)'s hair'
EU: (norbaiti) *adarra jo*

'play (somebody) the horn'

(14) EN: *take steps*
ES: *dar pasos*
'give steps'
EU: *pausoak eman*
'give steps'

We will not focus on the correlation between VNC classes and their translation in this paper. However, we do consider it an interesting topic for future investigation.

3.4 Morphosyntactic Classification

As well as the lexico-semantic tagging described above, we examined morphosyntactic features of combinations to classify them into three groups: (1) fixed combinations, (2) semi-fixed combinations, and (3) morphosyntactically free combinations. The annotators had to consider five questions to determine how fixed the combinations were:

- Does the noun phrase (NP) have a determiner? (always/never/optional)
- Is the NP singular or plural? (singular/plural/optional)
- Can there be a modifier (i.e. an adjective) inside the NP? (yes/no)
- Can the verb and the NP be separated by other words? (yes/no)
- Can the order of the elements be altered? (yes/no)

A given VNC needed to be classified as completely free when: the determiner and the number of the NP were marked as optional; there could be a modifier inside the NP; the verb and the NP could be separated by other words; and the order of the elements in the expression was judged to be alterable. When some of the answers were different to these, the combination had to be marked as semi-fixed, and as completely fixed if all the answers were different (that is, when the syntactic variability of the VNC was completely restricted).

None of the combinations was tagged as **fixed** by both the super-annotator and the second annotator, but this was not surprising, as VNCs which do not accept any kind of morphosyntactic variation are extremely rare. Usually, they can undergo some alterations (**semi-fixed expressions** as in examples (15) and (16)), or they can even be completely flexible (**morphosyntactically free expressions** as in examples (17) and (18)).

(15) be in love; be always in love; *be in the love; *be in loves.

(16) dar paso (a algo); dar siempre paso (a algo); *dar pasos (a algo)
'give way (to sth); always give way (to sth); *give ways (to sth)'

(17) cause a problem; cause two important problems; the problem was caused

(18) hacer un favor; hacer un gran favor; hacer dos favores; el favor que se hizo
'do a favour; do a big favour; do two favours; the favour that was done'

As these features have a direct impact on the detection of the combinations, the answers to the above-mentioned questions were also specified by the super-annotator one by one, so that this information could later be used to improve detection (see Section 4).

	Lexico-semantics	Morphosyntax
Agreement	70.52%	84.39%
κ	0.55	0.55

Table 1: IAA for English VN combinations.

	Lexico-semantics	Morphosyntax
Agreement	76.00%	81.34%
κ	0.63	0.61

Table 2: IAA for Spanish VN combinations.

3.5 Inter-Annotator Agreement

Inter-annotator agreement (IAA) was measured in two ways: the percentage of combinations in which the annotators agreed, and Cohen's Kappa, κ (Cohen, 1960).

As shown in Tables 1 and 2, annotator agreement was 70% to 84% for all tagging tasks and for both languages. With κ scores between 0.55 and 0.63, we conclude that the task is coherent and that the tagging results are usable for further investigation. The lexico-semantic IAA for English is similar to the IAA obtained in previous related work (Fazly et al., 2007; Vincze, 2012), and for Spanish it is appreciably higher.

Consistent with previous work (Seretan, 2013), we found that in our selection of 323 of the most frequently occurring VNCs in Spanish and English, collocations and LVCs are the most common type of combination, and that opaque expressions (idioms) are very scarce.

We also found that the combinations that led to disagreements among annotators were not classified in random groups, but were almost always in classes lexico-semantically (and morphosyntactically) close to each other (see Tables 3 and 4). Indeed, only a few combinations were classified in two groups that were not directly adjacent on the idiomaticity continuum. This provides further evidence that MWEs form a continuum of idiomaticity with no clear boundaries between MWE types (McCarthy et al., 2003).

		Other annotators			
		Idiom	**Metaphoric**	**Colloc/LVC**	**Free**
Super-annotator	**Idiom**	**0**	0	0	0
	Metaphoric	1	**24**	0	1
	Colloc/LVC	0	12	**73**	22
	Free	0	2	13	**25**

Table 3: Confusion matrix for English showing lexico-semantic tag agreement between the annotators.

		Other annotators			
		Idiom	**Metaphoric**	**Colloc/LVC**	**Free**
Super-annotator	**Idiom**	**1**	0	1	0
	Metaphoric	0	**20**	2	1
	Colloc/LVC	0	8	**69**	15
	Free	0	1	8	**24**

Table 4: Confusion matrix for Spanish showing lexico-semantic tag agreement between the annotators.

4 Identification Experiment

To test whether the analysed morphosyntactic data (see Section 3.3) could improve MWE detection, we undertook an experiment where three **identification methods** were combined and compared: (A)

the old one, used by Matxin, which searched only for word sequences; (B) a second one, based on the analysed linguistic data and automatically-produced chunking information; and (C) a third one, based on the analysed linguistic data and automatically-produced syntactic dependencies. Depending on how morphosyntactically fixed a given combination was, more or less linguistic restrictions were applied to identify them.

The **experimental set** was made of the combinations presented in Section 3, excluding the ones tagged as completely free by the super-annotator (Section 3.3). The final set consisted of 117 combinations in Spanish and 133 in English.

4.1 Results of the English Experiment

The corpus used for the experiment on English VNCs was the British National Corpus (Burnard, 2007), and chunking and dependency information was computed by the Stanford parser (Manning et al., 2014). A total of 152,051 occurrences of the 133 VNCs were identified by combining all three methods, 78.92% of which were not detected by method A, currently used for English-Basque translation in Matxin. Figure 1 shows the percentages of all the instances detected by each of the methods.

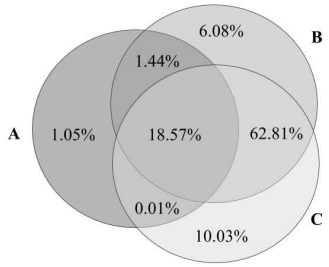


Figure 1: Percentages of English VNC occurrences identified by each method. (For clarity, areas are not drawn in scale with percentages)

We cannot calculate recall since our evaluation dataset contains only the occurrences identified collectively by the three methods, and it is almost certain that some occurrences of the VNCs under investigation were not detected. For future work, we would need to use MWE-tagged corpora to calculate recall, such as the Prague Czech-English Dependency Treebank (Uresova et al., 2013). In any case, the results obtained clearly show that the number of identified occurrences is increased considerably by using linguistic data specific to VNCs, as well as confirming that VNCs are commonly used in multiple morphosyntactic variations, as only 21.08% of the instances could be identified by searching for word sequences against entries in a lexicon.

To estimate the precision of VNC detection, we considered a representative sample of the full set, and evaluation was carried out manually by linguists. The precisions of methods B and C were not as good as that of method A. However, the evaluation on instances identified by both B and C methods reveals that detection quality is still very high when linguistic data specific to VNCs is combined with parsing (the second row of scores in Table 5).

	Additional VNCs %	**Precision**
Method A (in all)	21.08%	99%
Method B+C but not A	62.81%	96%
Method B only	6.08%	70%
Method C only	10.03%	79%

Table 5: Identification precision for the additional VNC occurrences detected in English

The least satisfactory results were those obtained by method B. When verifying the results, we noticed that the vast majority of false instances detected were light verb constructions (LVCs) containing verbs

that could also work as auxiliaries. In example (19), for instance, *have influences* is erroneously detected since *influence* is mis-analysed as the object of *have* rather than the subject of *have been likened*.

(19) These *influences have* also been likened to the forces effected by a millenarian journey to a new faith...

The overall improvement we obtained was substantial, as expected from previous work. Li et al. (2003) report an F-score improvement of 9 percentage points (86.9% to 95.6%) when using parsers and hand-crafted lexical patterns to identify phrasal verbs in English, as well as a precision improvement of 8 percentage points (90% to 98%). In our case, precision falls from 99% to 93% when combining all three methods, but the number of new instances detected suggests an appreciable increase in recall. As we already mentioned, MWE-annotated corpora would be needed to calculate recall and F-score and compare our results to those reported by other authors.

4.2 Results of the Spanish Experiment

For the experiment on Spanish, VNCs were searched in 15,182,385 sentences taken from the parallel English-Spanish corpus made public for the shared task in the ACL 2013 workshop on statistical MT[4], and the parser used was Freeling (Padró and Stanilovsky, 2012). A total of 433,092 occurrences were identified, 27.80% of which were not detected by method A (the percentages of the combinations identified by each method are shown in Figure 2). Consistent with the results obtained for English, this further reveals that the morphosyntactic data we took into account (Section 3.4) is very relevant for VNC identification.

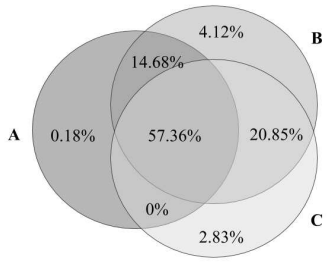


Figure 2: Percentages of Spanish VNC ocurrences identified by each method

Furthermore, as well as the quantity improving considerably, the manual evaluation reveals that the quality of our method is also very satisfactory. As is shown in Table 6, methods B and C, although not as precise as method A, got very good precision scores.

	Additional VNCs %	**Precision**
Method A (in all)	72.20%	99%
Method B+C but not A	20.85%	97%
Method B only	4.12%	93%
Method C only	2.83%	83%

Table 6: Identification precision for the additional VNCs detected in Spanish

As the corpora and parsers we used were different for English and Spanish, the experiments in both languages are not really comparable. However, it is evident that the improvement obtained for English was considerably higher than the one obtained for Spanish. Taking into account that the Freeling and Stanford parsers work in similar ways and that the manual tagging of the VNCs was done following the same criteria, this difference could suggest that syntactic variations of VNCs other than the canonical form are more common in English than in Spanish. One of the possible reasons for this could be the

[4] http://www.statmt.org/wmt13/translation-task.html

different word order inside NPs in both languages. In Spanish, adjectives can either precede or follow the head noun, whereas in English adjectives are almost never placed after the noun: *importantes pasos* or *pasos importantes* vs. *important steps* but not **steps important*. An exhaustive analysis would be needed to verify this hypothesis or identify other possible reasons.

5 Conclusions and Future Work

Morphosyntactically flexible MWEs constitute a problem for NLP systems, which often fail to process these kinds of word combinations correctly. In this paper, we presented a linguistic analysis undertaken with the aim of improving the identification of VNCs, as well as an experiment which shows how linguistic data can improve identification results greatly.

Firstly, we classified a selection of frequent VNCs in English and Spanish, following both lexico-semantic and morphosyntactic criteria. A total of 323 distinct combinations (173 in English and 150 in Spanish) were tagged by several annotators, with very reasonable inter-annotator agreement scores (κ 0.55 to 0.63). We noted moreover that the combinations that led to disagreements among annotators were always tagged in groups that were lexico-semantically and morphosyntactically close to each other, which gives further evidence that idiomaticity should be viewed as a continuum. More detailed morphosyntactic information was also specified for each combination, and this information was then used to improve VNC identification.

Our experiment confirmed that specific linguistic data about VNCs is useful for the identification of this kind of word combination, as it allows for the recognition of occurrences that do not match a combination's canonical form. Indeed, a large number of instances that were not identified by searching for fixed word sequences could be identified by combining linguistic data with chunking information and syntactic dependencies, with fairly good precision scores (79% to 97%).

Building on the satisfactory results obtained, we will test our methods in the context of MT, and we will keep analysing more VNCs. The next step will be to explore what kind of data is needed for an adequate translation of VNC combinations within MT systems. In addition, we intend to investigate how semantic information can be used within the translation process.

Acknowledgements

Uxoa Iñurrieta's doctoral research is funded by the Spanish Ministry of Economy and Competitiveness (BES-2013-066372). The work was carried out in the context of the SKATeR (TIN2012-38584-C06-02) and TADEEP (TIN2015-70214-P) projects. We thank Diana McCarthy for helpful advice, as well as Begoña Altuna, Nora Aranberri, Ainara Estarrona and Larraitz Uria for helping us with the tagging work.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 89–96.

Sabine Bartsch. 2004. *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Gunter Narr Verlag, Tübingen.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 329–336.

Lou Burnard (ed.) 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services.

Miriam Butt. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey (eds.), *Complex Predicates: Cross-linguistic Perspectives on Event Structure*. Cambridge University Press, 48–78.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 19–22.

Margaret Deuter (ed.) 2008. *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, IMS, University of Stuttgart.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, Berlin, 1212–1248.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL-SIGLEX Workshop on a Broader Perspective on Multiword Expressions*, 9–16.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Wei Li, Xiuhong Zhang, Cheng Niu, Yuankai Jiang, and Rohini Srihari. 2003. An expert lexicon approach to identifying English phrasal verbs. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics: Long Papers*, 513–520.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 25(1): 53–82.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 73–80.

Igor A. Mel'ćuk. 1998. Collocations and lexical functions. In Anthony P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford University Press, 23–53.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2473–2479.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer International Publishing, Switzerland.

Sara Rodríguez-Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, 499–505.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'02)*. Springer Berlin Heidelberg, 1–15.

Violeta Seretan and Eric Wehrli. 2009. Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, 43(1): 71–85.

Violeta Seretan. 2013. On collocations and their interaction with parsing and translation. *Informatics*, 1(1): 11–33.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Caroline Sporleder, and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 754–762.

Zdenka Uresova, Jana Sindlerova, Eva Fucikova, and Jan Hajic. 2013. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proceedings of the Ninth Workshop on Multiword Expressions (MWE 2013)*, 58–63.

Rule-Based Translation of Spanish Verb+Noun Combinations into Basque

Uxoa Iñurrieta, Itziar Aduriz*, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA NLP group, University of the Basque Country
University of Barcelona
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu
a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

Abstract

This paper presents a method to improve the translation of Verb-Noun Combinations (VNCs) in a rule-based Machine Translation (MT) system for Spanish-Basque. Linguistic information about a set of VNCs is gathered from the public database Konbitzul, and it is integrated into the MT system, leading to an improvement in BLEU, NIST and TER scores, as well as the results being significantly better according to human evaluators.

1 Introduction

Multiword Expressions (MWEs) constitute a challenging phraseological phenomenon for Natural Language Processing (NLP). They are formed by more than one word, but the whole expression has to be taken into account in order to understand its meaning (Sag et al., 2002). They are very frequent in natural language, but their processing is not straightforward, especially due to their morphosyntactic variability. Furthermore, difficulties multiply when it comes to Machine Translation (MT), since MWEs are not usually translated word for word and, hence, sophisticated processing methods are needed.

In this paper, we will deal with Verb-Noun Combinations (VNCs), and we will explain how MWE-specific linguistic information can be used to improve a rule-based MT system which translates Spanish into Basque, namely Matxin (Mayor et al., 2011). After discussing some related work (Section 2), a brief explanation about Matxin and the way it handles MWEs will be given (Section 3). Then, the experimental setup will be presented (Section 4), and results will be shown (Section 5).

2 Related Work

MWEs are word combinations that need to be treated as a whole in order to get good results in lexically-sensitive NLP tasks (Sag et al., 2002). Not all MWEs are morphosyntactically fixed –there are also semi-fixed and flexible combinations–, which makes their processing a complex task. Some kinds of MWEs, like VNCs, are specially tricky, as they are more likely to have multiple morphosyntactic variants.

Over the last decades, quite a lot of research has been done on MWE identification and extraction (Gurrutxaga and Alegria, 2011; Ramisch, 2015), which is relevant not only for NLP applications but also for other disciplines like Lexicography (Vincze et al., 2011). MWE-specific resources are being developed in a number of languages, as reported by Losnegaard *et al.* (2016) in a survey carried out within the PARSEME COST Action (IC1207).

However, not so much work has been undertaken concerning the multilingual aspects of this phraseological phenomenon, although challenges get bigger when multiple languages are involved. One of the reasons why this happens is that MWEs are not usually translated word for word from one language to another, especially when these languages are from very different typologies (Baldwin and Kim, 2010; Simova and Kordoni, 2013), as with Basque and Spanish[1].

Joint efforts are also being made towards improving Machine Translation systems, for example, within the european QTLeap project (Agirre et al., 2015). Although statistical MT systems already integrate some phraseological knowledge as a consequence of training their models on large

[1]Whereas Spanish is a romance language, Basque is a non-indoeuropean language which belongs to no known family. More details about the main differences between both languages are given in Section 3.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 149–154,
Valencia, Spain, April 4.

corpora (Ren et al., 2009; Bouamor et al., 2012; Kordoni and Simova, 2014), rule-based systems often get bad results when MWEs are involved, as they tend to translate each word separately. Thus, this kind of expression being so frequent in natural language, MT systems benefit greatly from including phraseological knowledge, and several studies have shown that even the simplest method to process MWEs makes a difference in the system's translation quality (Wehrli et al., 2009; Seretan, 2014).

3 Matxin: Rule-based MT from Spanish into Basque

Matxin (Mayor et al., 2011) is an MT system which translates Spanish into Basque, two long-distance families. As opposed to Spanish, which uses prepositions, Basque is a morphologically rich language where postpositions and cases are used and word order is free. The system is rule-based, mainly because of the scarcity of parallel corpora available in these languages.

Matxin's general architecture is divided into three phases:

1. **Analysis**. The source text is analysed using the FreeLing parser (Padró and Stanilovsky, 2012), which gives morphological information, chunking information, and determines the dependency relationship between words.

2. **Transfer**. The deep syntactic representation of the Spanish sentence is transferred into an equivalent representation in Basque. During this phase, on the one hand, the lexical components in the source language are replaced with their corresponding elements in the target language, and, on the other hand, the structure is also transferred. Specific modules for Spanish-Basque translation are included in this phase, like the one to change prepositions into postpositional information.

3. **Generation**. Firstly, the nodes in each chunk and the chunks themselves are reordered in the sentence from scratch, and postpositional information is added to the chunks when needed. Then, the forms of the words in Basque are created from the labelled lexical elements. The morphological processor used for this purpose is *Morfeus* (Alegría et al., 1996).

3.1 Current MWE handling

At the moment, Matxin uses a very simple method to process MWEs. When an entry in the system's bilingual dictionary is formed by more than one word, the whole expression is treated as a fixed sequence, that is, as if it was a single word. During the transfer phase, the Spanish MWE is replaced by its corresponding Basque word(s), as shown in example (1)[2].

(1) 'A vacancy was filled.'
ES: Se *cubrió_una_plaza*.
Refl covered a vacancy
MT: *Plaza_bat_bete* zen.
vacancy a fill AuxV

In the case of verbal MWEs (including VNCs), verb inflection is taken into account, but the rest of the words have to follow the verb exactly like they appear in the entry. This means that morphosyntactic variation is not processed correctly, neither when identifying the MWE in the source language, nor when translating it into the target language. More details about this are given in Sections 4.1 and 4.2.

(2) 'They filled all vacancies.'
ES: *Cubrieron* todas las *plazas*.
they-covered all the vacancies
MT: *Plaza* guztiak *estali* zituzten.
vacancy all.abs cover AuxV
CT: *Plaza* guztiak *bete* zituzten.
vacancy all.abs fill AuxV

(3) 'He doesn't pay me attention.'
ES: No me *hace_caso*.
not me.IndObj he-does attention
MT: Ez nau *kasu_egiten*.
not AuxV.DObj attention do
CT: Ez dit *kasu(rik) egiten*.
not AuxV.IndObj attention.part do

In example (2), the VNC *cubrir plazas* is not identified as a MWE and, as a consequence, the wrong lexical choice is done when translating it into Basque. In example (3), on the other hand, the VNC is identified well, but the grammatical information of its Basque translation is incorrect, because the system ignores that the Basque VNC needs an indirect object instead of a direct one.

[2] In examples, we use ES for the Spanish text to be translated, MT for the result of the MT system, and CT for the correct Basque translation.

4 Experimental setup

The VNC set used for the experiment consisted of 92 combinations taken from the Konbitzul database[3], where a number of Spanish VNCs and their Basque translations are collected along with linguistic data. The combinations in Konbitzul were gathered from several sources; the set we used here originally came from the Elhuyar Spanish-Basque dictionary[4] and was then analysed and tailored to meet the requirements of the database. According to the information in Konbitzul, 57 out of the 92 combinations were morphosyntactically semi-fixed, while the resting 26 were completely flexible.

Concerning the corpus, 4,991 sentences were selected from a bigger parallel corpus made of cross-domain texts collected by web-crawling and automatically aligned between Spanish and Basque. It was expressly crafted for this experiment, meaning that it did not consist of random sentences but of selected sentences containing: either instances of the Spanish VNCs in our set (Example 4), or both the verb and the noun of a given VNC in our set, but not being part of the VNC in this context (Example 5). This allowed us to test the performance of the MT system both when the VNC needed to be processed as a whole and when the verb and the noun needed to be translated separately.

(4) Iban *dando voces* por la calle.
they-went giving voices on the street
'They were shouting on the street.'

(5) Aquellas *voces* le *dieron* una pista.
those voices her.$_{\text{IndObj}}$ gave a clue
'Those voices gave her a clue.'

The information in Konbitzul was first used to help to identify instances of the VNCs when analysing the source text (Section 4.1), and then to transfer the source sentence into the target language (Section 4.2). Therefore, the identification of VNCs was done within the Analysis phase of the translation procedure, and their translation was done within the Transfer phase, the Generation phase not needing any special adaptation for MWE handling (Section 3).

[3] http://ixa2.si.ehu.eus/konbitzul
[4] http://hiztegiak.elhuyar.eus/

4.1 Identifying the Spanish VNCs

In Konbitzul, comprehensive linguistic information is specified for the VNC set we use here, including some features specifically analysed for NLP purposes. The morphosyntactic classification is first used, according to which the VNCs can be of three types: fixed, semi-fixed or flexible.

When a given VNC is classified as flexible, it means that, concerning morphosyntax, the noun and the verb work as any other noun and verb in the sentence, that is, they can have as many variants as any non-phraseological VNC.

(6) Me *da* muchísimo *miedo*.
me.$_{\text{IndObj}}$ gives very-much fear
'It scares me very much.'
¡Qué *miedo* me *da*!
what fear me.$_{\text{IndObj}}$ gives
'How scary (I find it)!'

On the other hand, when the VNC is classified as semi-fixed, some restrictions are needed in order to distinguish occurrences of the VNC from other sentences where the verb and the noun are present but should not be treated as an MWE.

(7) *Estoy* muy *de acuerdo*.
I-am very of agreement
'I agree very much.'
Estoy harta *del acuerdo*.
I-am fed-up of-the agreement
'I'm fed up with the agreement.'

In example (7), two sentences are shown, both of which contain the verb *estar* and the noun *acuerdo* preceded by the preposition *de*. In the first sentence, those words constitute a MWE (*estar de acuerdo*, 'agree'), but not in the second one, where the noun phrase (NP) has a determiner. By restricting determiners from the NP in the VNC, the system identifies a MWE in the first sentence but not in the second one[5].

For the identification task, we followed the same procedure as the one used in (Iñurrieta et al., 2016). First of all, the method currently used by Matxin is run, that is: word sequences are searched for against entries in the database, taking verb infletion into acount, but not considering the potential variability of the rest of the elements.

[5] All restrictions are collected and explained in (Iñurrieta et al., 2016).

Then, automatically-produced chunking information and syntactic dependencies are used, and morphosyntactic restrictions specified in Konbitzul are applied (Example 7).

4.2 Translating the VNCs into Basque

Concerning translation, Konbitzul classifies the Spanish VNCs according to what needs to be changed when translating them into Basque: lexicon, grammar, or both lexicon and grammar.

For the VNCs needing lexical treatment, Basque equivalents are specified for the verb and the noun in Spanish. This information is integrated into Matxin, so that, when a VNC is identified, the system does not translate it regularly (Example 8).

(8) 'The topic aroused interest.'
ES: El tema *despertó interés*.
the topic awakened interest
MT: Gaiak *interesa esnatu* zuen.
topic.$_{erg}$ interest awaken $_{AuxV}$
CT: Gaiak *interesa piztu* zuen.
topic.$_{erg}$ interest turn-on $_{AuxV}$

On the other hand, for the VNCs needing special grammatical treatment, the features that need to be taken into account are specified. For those cases, exceptional rules are added within the Transfer phase, so that the specified feature(s) is/are not translated regularly.

The features specified in the database are:

- Cases or postposition marks of the NPs
- Determiner irregularities
- Number and definiteness of the NPs
- Syntactic relations of the verbs and the NPs
- Postpositions of open slots

In example (9), for instance, the Basque NP needs a postposition other than the one automatically given as a translation of the Spanish preposition. Furthermore, it needs to be indefinite, but it would be translated as definite if no special rule was applied.

(9) 'She treats me with respect.'
ES: Me *trata con respeto*.
she-me.$_{DObj}$ treats with respect
MT: *Errespetuarekin tratatzen* nau.
respect.$_{soc}$ treat $_{AuxV}$
CT: *Errespetuz tratatzen* nau.
respect.$_{ins}$ treat $_{AuxV}$

When it comes to example (10), the noun in the Spanish VNC is preceded by a preposition, and this prepositional phrase works as a modifier of the verb. On the other hand, the combination has an object which works as an open slot, that is, an element which is always present but can be filled with any NP. In the Basque translation, the object of the verb in the VNC is actually the noun in the VNC, and the open slot is a postpositional phrase which works as a modifier. Therefore, both the syntactic relation and the postposition of the open slot need special rules to be processed correctly.

(10) 'They miss him.'
ES: Lo *echan en falta*.
him.$_{IndObj}$ throw in lack
MT: *Faltan botatzen* dute.
lack.$_{ine}$ throw $_{AuxV}$
CT: Haren *falta sumatzen* dute.
his lack.$_{abs}$ feel $_{AuxV}$

5 Results

After integrating all the linguistic information into Matxin, the system was evaluated using three automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and TER (Snover et al., 2006). Evaluation was carried out without casing, and two systems were compared: (a) the original one, Matxin, and (b) the same system with VNC-specific information.

System	BLEU	NIST	TER
Matxin	7.28	3.88	84.36
Matxin-VNC	7.50	3.90	84.27

Table 1: BLEU, NIST and TER scores obtained by Matxin with and without VNC-specific information

As shown in Table 1, all scores improve when VNC-specific information is used. The greatest improvement is obtained in BLEU score (0.22 points), and results are statistically significant according to paired bootstrap resampling ($p>0.05$). It must be noted that BLEU scores are low for Spanish-Basque, and this result means a relative increase of 3.02%.

5.1 Human evaluation

Apart from using automatic evaluation metrics, three human evaluators were also given a representative sample of the sentences translated differently by both systems and were asked to compare

them. All evaluators were Spanish and Basque native speakers: two of them (A and B) were linguists, whereas the third one (C) had no linguistic background.

System	A	B	C
Matxin-VNC	77.50%	77.50%	46.50%
Matxin	6.50%	8%	40.50%
No preference	16%	14.50%	13%

Table 2: Scores by three human evaluators

Although scores clearly show that the system with VNC-specific information gets better results, they also suggest that improvements are much more evident for linguists than for native speakers with no linguistic background (Table 2). In fact, 43.52% of the evaluation set led to disagreements among annotators, but 78.57% of these (33% of the whole set) were cases in which both linguists said the new system performed better while annotator C chose the other translation.

Taking into account that only a few combinations were tested and the corpus used was specifically prepared based on those combinations, it can be foreseen that the overall improvement this method would produce on large corpora would not be as significant. However, as the kind of linguistic information we chose is proved to have a positive effect on the system's output, we conclude that this methodology is relevant and useful for further investigation.

6 Conclusion

In the experiment presented in this paper, linguistic information was used to improve the translation of VNCs in Matxin, a rule-based MT system for Spanish-Basque. MWE-specific linguistic information was gathered from Konbitzul, a database collecting data about a list of VNCs, and this information was then used both for the identification of idiomatic VNCs in Spanish and for their translation into Basque.

After integrating information about 92 VNCs into Matxin, the system was evaluated on a 4,991-sentence cross-domain corpus, using three automatic metrics: BLEU, NIST and TER. The score that raised the most was BLEU, with an increase of 0.22 points (3.02%). A human evaluation was also carried out, where the improvement became even more evident, even if it also suggested that linguists are more likely to notice improvements than native speakers with no linguistic background.

It must also be noted that the corpus we used here was specifically crafted for this experiment, which means that the improvement would probably not be as significant in a bigger general corpus. However, results are positive as a start, and we intend to keep investigating how this methodology can be enhanced. The next step will be to add more VNCs and test them in bigger corpora, so that conclusions can be drawn at a greater scale.

Acknowledgments

Uxoa Iñurrieta's doctoral research is funded by the Spanish Ministry of Economy and Competitiveness (BES-2013-066372). The work was carried out in the context of the SKATeR (TIN2012-38584-C06-02), EXTRECM (TIN2013-46616-C2-1-R) and TADEEP (TIN2015-70214-P) projects.

References

Eneko Agirre, Iñaki Alegría, Nora Aranberri, Mikel Artetxe, Ander Barrena, António Branco, Arantza Díaz de Ilarraza, Koldo Gojenola, Gorka Labaka, Arantxa Otegi, et al. 2015. Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches. *Procesamiento del Lenguaje Natural*, 55:169–172.

Iñaki Alegría, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multiword expressions for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 674–679.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic extraction of nv expressions in basque: basic issues on cooccurrence techniques. In *Proceedings*

of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, pages 2–7. Association for Computational Linguistics.

Uxoa Iñurrieta, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola, Itziar Aduriz, and John Carroll. 2016. Using linguistic data for english and spanish verb-noun combination identification. In *The 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, pages 857–867.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1208–1211.

Gyri Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sacha Bargmann, and Johanna Monti. 2016. Parseme survey on MWE resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA)*.

Aingeru Mayor, Iñaki Alegría, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, 25(1):53–82.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2473–2479.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhug. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition*. Springer.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (ACL 2009*, pages 47–54. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Violeta Seretan. 2014. On collocations and their interaction with parsing and translation. In *Informatics*, volume 1, pages 11–31. Multidisciplinary Digital Publishing Institute.

Iliana Simova and Valia Kordoni. 2013. Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Orsolya Vincze, Estela Mosqueira, and Margarita Alonso Ramos. 2011. An online collocation dictionary of Spanish. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 275–286.

Eric Wehrli, Violeta Seretan, Luka Nerima, and Lorenza Russo. 2009. Collocations in a rule-based mt system: A case study evaluation of their translation adequacy.

Learning about phraseology from corpora: A linguistically motivated approach for Multiword Expression identification and translation

Uxoa Inurrieta[1], Itziar Aduriz[1,2], Arantza Diaz de Ilarraza[1], Gorka Labaka[1], Kepa Sarasola[1],

1 Ixa NLP group, University of the Basque Country, Spain
2 Department of Linguistics, University of Barcelona, Spain

* usoa.inurrieta@ehu.eus (UI)

Abstract

Multiword Expressions (MWEs) are idiosyncratic combinations of words which pose important challenges to Natural Language Processing. Some kinds of MWEs, such as verbal ones, are particularly hard to identify in corpora, due to their high degree of morphosyntactic flexibility. Besides, MWEs are not always translated word-for-word, which complicates their processing even more in multilingual tools like Machine Translation. This paper describes a linguistically motivated method to gather detailed linguistic information about verb+noun MWEs (VNMWEs) from corpora. Monolingual and parallel corpora are used, and data about both morphosyntactic variability of VNMWEs and their translation are extracted. Two experiments confirm that the method is useful to improve MWE identification and MT, with an F score of 0.72 in identification (which is considerably higher than related work) and an improvement of translation quality both according to human judgements and according to statistical measures like BLEU.

Introduction

Multiword Expressions (MWEs) are combinations of words which exhibit some kind of lexical, morphosyntactic, semantic, pragmatic or statistical idiosyncrasy [2]. Due to their idiosyncratic nature, they pose multiple challenges to Natural Language Processing (NLP), and sophisticated strategies are needed in order to process them correctly.

Several types of word combinations are comprised in the category of MWEs [7, 12], such as idioms (example 1), which have a non-compositional meaning, and collocations, where the lexical choice is restricted (example 2). The latter also include light verb constructions (example 3), where the verb tends to be semantically bleached. In the examples in this paper, lexicalised component words of MWEs [24] are bold, and other words or morphemes that need to be marked are underlined. When glosses are given, the Leipzig glossing rules and abbreviations are used.

(1) *She always ends up* ***spilling the beans*** (lit. revealing the secret)

(2) *All students* ***passed*** *the* ***exam****.*

(3) *She is* ***giving*** *a* ***lecture*** *this afternoon.*

Two of the most challenging features of MWEs are variability and discontiguity [6], that is, the fact that the component words of many MWEs can occur in several word forms, can be separated by other elements in a sentence, and can even have an altered word order. These variations are especially prominent in MWEs where the syntactic head is a verb, since combinations of these kinds tend to be rather flexible morphosyntactically (example 4). However, many such combinations are not completely flexible and have some restrictions (example 5).

(4) a. *They **made** a **conclusion**.*
b. *They **made** some simple but still interesting **conclusions**.*
c. *The **conclusions** they **make** are always interesting.*

(5) a. *Their advice should be **taken into account**.*
b. *You should **take** their advice **into account**.*
c. *The accounts into which their advice should be taken.

Therefore, for the identification of verbal MWEs, basic methods which try to match fixed word sequences against dictionary entries are too limited. Let *make conclusions* and *take into account* be two entries in a dictionary. If this basic method was employed to identify occurrences of these entries in the sentences in examples (4) and (5a)–(5b), all occurrences would be ignored, because: the component words are separated by external elements in (4a)–(4c) and (5b); word forms in examples (4b), (4c) and (5a) are different than the ones in the entry; and word order is altered in example (4c).

On the other hand, opposite strategies where only the lemmas of the component words are searched for (within a given word distance) are not effective either, since these are, in their turn, too wide. These strategies would identify all of the occurrences in examples (4) and (5a)–(5b), but also the following ones and many others alike, which would be false positives:

(6) *They will make progress and will soon come to a conclusion.*

(7) *You should take the money and put it into your account.*

Furthermore, many MWEs cannot be translated word-for-word from one language to another, which makes their processing even more demanding when several languages are involved. For instance, in English (EN), Spanish (ES), Basque (EU) and French (FR), the noun *attention* is combined with different verbs to express the act of listening or observing something carefully:

(8) EN: ***pay attention***
EU: ***arreta jarri*** (lit. attention put)
FR: ***faire attention*** (lit. make attention)
ES: ***prestar atención*** (lit. lend attention)

The need for specific translation strategies for MWEs becomes evident in Machine Translation (MT), especially in rule-based systems which tend to translate every word independently. Namely, the *Matxin* MT system [18] fails to translate the Spanish MWE *tomar el pelo* (lit. take the hair 'pull sb's leg') into Basque, since it translates the verb and the noun separately (example 9). Consequently, instead of producing a correct output sentence containing *adarra jo* (lit. play the horn 'pull sb's leg'), the system gives the MWE an erroneous literal translation.

(9) ES: *Nos **toma** siempre **el pelo**.*
us takes always the hair

'He/She always pulls our leg.'

MT: *Beti hartzen digu ilea.*
always takes AUX hair-the
'He/She always takes our hair.'

EU: *Beti* ***jotzen*** *digu* ***adarra***.
always plays AUX horn-the
'He/she always pulls our leg.'

The main assumption behind the work explained here is that MWE-specific morphosyntactic information is helpful for MWE processing. As a matter of fact, recent studies [25] have shown that very few word combinations occur in corpora both literally and idiomatically with the very same morphosyntactic features, suggesting that most ambiguities concerning MWEs can be solved by looking at morphology and syntax.

In previous work, in-depth lexical and morphosyntactic information about verb+noun MWEs (VNMWEs) was proven to have a positive impact both in identification [14] and in MT [15]. Detailed data was manually provided in these experiments, and the results obtained using a controlled set of sentences were promising. However, only a few VNMWEs were analysed and, the analysis process being completely manual, the method had a clear scalability problem.

This paper describes an improved method where detailed linguistic information about VNMWEs is automatically gathered from corpora, with an aim to reduce manual work and consequently increase the number of analysed VNMWEs. Data acquired by this method was tested on MWE identification and MT, and both the method and the experiments are explained here.

The paper is organised as follows. The resources and methodology are first described, with comprehensive explanations on each of the six steps taken, as well as on quality assessments of the gathered data (after Step 3, Step 4 and Step 6). It goes on to show how the gathered data was used for MWE identification and for MT, and results of both experiments are shown. Finally, some conclusions are drawn and ideas for future work are presented.

Resources and methodology

The information-gathering process is organised in six steps, half of which obtain data for VNMWE identification, and the other half, for MT. The languages selected for this study are Spanish (source) and Basque (target), two languages of very different typology between which MWE translation is highly complex [16]. Fig 1 shows the main steps followed, as well as the resources employed and what the data from each step was then used for.

All data acquired from this process is stored in the *Konbitzul* database [17], which is openly accessible at `http://ixa2.si.ehu.eus/konbitzul` and can be fully downloaded on CVS from `ixa.eus/node/4484?language=en`.

Fig 1. Outline of the general methodology.

As Fig 1 shows, a set of VNMWEs is first extracted from dictionaries, which is used as a basis for the whole process. For the experiments explained here, two dictionaries were used: the Elhuyar Spanish-Basque general dictionary (`https://hiztegiak.elhuyar.eus`) and the DiCE dictionary of Spanish collocations [1] (`www.dicesp.com`). Note, however, that this set can easily be extended

or modified in the future, either by collecting additional combinations from other dictionaries or by automatically extracting them from corpora [11, 22].

Due to the different nature of the source dictionaries, the extracted combinations were diverse. The ones coming from the Elhuyar dictionary all consisted of a verb and a noun, but sometimes included a preposition and/or a determiner in-between. The preposition was treated as lexicalised, but not the determiner, since determiners are usually variable elements of VNMWEs (see explanations on Step 1). In the DiCE dictionary, however, entries contain nouns only, and collocates are classified by grammatical category under each of the nouns. Therefore, in this case, we gathered the entries along with their verbal collocates, but no prepositions were included; this information was added later, by looking at corpora (see explanations on Step 1). The 500 most frequent collocations were selected [28], and the ones repeated in Elhuyar or the ones manually analysed in previous work [15] were discarded. In all, the set of combinations to be used as a basis for the information-gathering method consisted of 1,205 entries from the Elhuyar dictionary and 437 from DiCE.

Once this set ready, both parallel and monolingual corpora were used to obtain data about the VNMWEs. Explanations on each step are given below.

Step 1. Extraction of linguistic data from Spanish corpora

In the first step, a monolingual corpus was employed to see how the combinations in our set were used in text. The corpus selected for this purpose was the 15-million-sentence Spanish corpus released for the 2013 Workshop on Machine Translation (available at `https://www.statmt.org/wmt13/`), and the Freeling 3.0 parser [19] was used to analyse it. For each of the VNMWEs, the lemmas of the component words were searched for in cases where the noun (and the preposition, when necessary) was dependent on the verb.

Our aim at this stage was to obtain detailed morphosyntactic information about each of the occurrences. Taking into account the characteristic morphosyntactic aspects of Spanish VNMWEs [4, 20], we looked at the following features:

- Number of the noun phrase (NP): singular (Sing.) or plural (Pl.)
- Determiners in the NP (Det.)
- Definiteness of the NP, in case a determiner was present: definite (Def.) or indefinite (Ind.)
- Modifiers inside the NP (Mod.)
- Alterations in the order of the component words (Ord.)

The information was stored in percentages for all occurrences of each VNMWE candidate (an example will be shown later). However, many verb+noun (VN) combinations can constitute a VNMWE in some sentences but a free expression –or even a different VNMWE– in another one, even when the noun depends on the verb in both cases. This is the case of the verb *dar* 'give' and the noun *paso* 'step' in Spanish, which can be part of both the VNMWE *dar paso* 'give way' (example 10) and the VNMWE *dar pasos* 'take steps' (example 11), as well as having coincidental non-idiomatic occurrences such as the one in example (12).

(10) *No es posible* ***dar paso*** *a muchas preguntas hoy.*
no is-it possible give step/way to lots-of questions today
'It is not possible to give way to many questions today.'

(11) *Vamos a **dar** un **paso** transcendental.*
we-will to give one step vital
'We will take a vital step.'

(12) *Los pasos dieron media vuelta y se marcharon.*
the steps gave half return and REFL left
'The steps turned away and left.'

Based on previous investigations, our hypothesis was that some morphosyntactic features could be especially useful to distinguish between different meanings of the same VN pairs. These features were: the syntactic relation (example 13), the possibility to add determiners inside the NP (example 14), and the use of the pronominal form of the verb (example 15). It was thus decided that variants differing in these three aspects would be treated as separate candidates at this step. Information was stored separately for each of them.

(13) a. *Pueden **tomar parte** en los debates.* → obj.
they-can take part in the debates
'They can take part in the debates.'

b. *Cada parte tomará las medidas necesarias.* → subj.
each part will-take the measures necessary
'Each party will take the necessary measures.'

(14) a. *Esas cuestiones pueden **ser de interés** para los participantes.* → no det.
those questions can be of interest for the participants
'Those questions can be interesting for the participants.'

b. *Esto debería **ser del interés** del cliente.* → det.
this should be of-the interest of-the client
'This should be of the client's interest.'

(15) a. ***Nos damos cuenta** de lo ocurrido.* → pronominal
AUX give account of the happened
'We realise what happened.'

b. *Las autoridades deben **dar cuenta** de lo ocurrido.* → non-pronominal
the authorities must give account of the happened
'Authorities must report on what happened.'

As already mentioned, the prepositions in the combinations from Elhuyar were specified, but not the ones in the combinations from DiCE. This kind of division was also done for the VN pairs from the DiCE dictionary used with several prepositions in the corpus (example 16).

(16) a. *No lo **dejaremos a** un **lado**.*
no him we-will-leave to one side
'We will not let him aside.'

b. *No lo **dejaremos de lado**.*
no him leave of side
'We will not let him aside.'

The features listed above were counted in every occurrence, and a table was created collecting all the data about each candidate. Fig 2 shows how information was stored for the combinations in examples (13)–(16). Candidate keys are organised as follows: VERB, reflexive/pronominal use of the verb (pron|-), syntactic relation (subj|obj|ccomp|pred), PREPOSITION, determiners (*|?|-), NOUN.

Fig 2. Example of the way information is stored in Step 1. Data in percentages.

Many combinations did not occur in the corpus, and others were discarded because their frequency was too low (below 10) for the extracted information to be reliable. Finally, 435 candidates from Elhuyar and 544 from DiCE were collected along with linguistic data. Note that many VNMWEs from Elhuyar did not occur in the corpus, and that more than one candidate was stored per VNMWE from DiCE on average, as a result of the divisions explained above. This contrast in number is probably due to the different nature of the source dictionaries: DiCE is a dictionary of collocations, a kind of MWE which is very frequent; conversely, general dictionaries like Elhuyar contain mostly idioms, which usually occur less often in general corpora.

Step 2. Classification of candidates according to morphosyntactic patterns

The next step consisted in classifying the candidates into morphosyntactic patterns, based on the percentages obtained. The underlying idea here was that, if the very large majority of the occurrences of a given candidate had e.g. a singular NP with a definite article, the rest of the sentences were unlikely to be relevant to the VNMWE candidate (bearing in mind that divisions were made during Step 1 to help solve ambiguities). Half of the candidates were used for trials: 218 from Elhuyar and 272 from DiCE.

A threshold was established per feature, in order to indicate from what point on each of the features should be treated as determining. According to these thresholds, the values of features were classified in three groups: Y (yes, this feature is always present for the candidate), O (optional, this feature is sometimes present for the candidate), or N (no, this feature is never present for the candidate). Except for the syntactic relation, only morphological features were taken into account at this step. Information about intra-NP modifiers and alterations in word order was added later, for the identification experiment.

Due to the different nature of the source dictionaries, morphosyntactic patterns were created separately based on the source. Candidates were grouped according to their features and, after generalising the less productive ones, twelve patterns were established for the combinations from the Elhuyar dictionary. These patterns and their corresponding feature values are specified in Table 1.

The same patterns were then applied to the candidates extracted from DiCE as well, but it was observed that this set of combinations had less morphosyntactic restrictions and the patterns were thus too specific. Therefore, for the candidates from DiCE, we decided to reduce the specific patterns to five (Table 2).

As can be noticed, seven patterns were discarded in all, due to their lack of relevance to the candidates from DiCE. The discarded patterns were: (A) those collecting the combinations used in the pronominal form only and (B) those collecting the combinations used in the plural form only. DiCE being a dictionary of collocations only, the fact that these candidates have a higher morphosyntactic flexibility strengthens the idea that the level of lexical-semantic idiomaticity is somehow linked to the degree of morphosyntactic fixedness of a given combination [27].

Table 1. Morphosyntactic patterns for the candidates from the Elhuyar dictionary.

	Pron.	Sing.	Pl.	Det.	Def.	Ind.
FREE	N	O/N	O/N	Y/O/N	Y/O/N	Y/O/N
PL_NO-DET	N	N	Y	N	O/N	O/N
PL_DET_DEF	N	N	Y	Y	Y	N
PL	N	N	Y	Y	Y/O/N	Y/O/N
SING_NO-DET	N	Y	N	N	N	N
SING_DET_DEF	N	Y	N	Y	Y	N
SING_DET_IND	N	Y	N	Y	N	Y
SING	N	Y/O	N	Y/O	O/N	O/N
P_PL	Y	N	Y	Y/O	Y/O/N	Y/O/N
P_SING_NO-DET	Y	Y	N	N	N	N
P_SING_DET_DEF	Y	Y	N	Y	Y	N
P_SING	Y	Y	N	Y/O	Y/O/N	Y/O/N

Note that rules were applied from the most general to the most specific, so that combinations which fitted in more than one group were classified in the most specific pattern possible. A few extra rules were also created to discard some combinations which were typically non-MWEs, like those where the noun was the subject of a very common verb (*ser/estar* 'be', *hacer* 'do/make'...).

Table 2. Morphosyntactic patterns for the candidates from the DiCE dictionary.

	Pron.	Sing.	Pl.	Det.	Def.	Ind.
FREE	N	O/N	Y/O/N	Y/O/N	Y/O/N	Y/O/N
SING_NO-DET	N	Y	N	N	N	N
SING_DET_DEF	N	Y	N	Y	Y	N
SING_DET_IND	N	Y	N	Y	N	Y
SING	N	Y/O	N	Y/O	O/N	O/N

By way of example of the kind of classification done at Step 2, Fig 3 shows how the data on Fig 2 evolved, and how the candidates in examples (13)–(16) were classified.

Fig 3. Example of the pattern-based classification in Step 2. The main features which determine why a given combination is classified in its pattern are marked in blue.

Step 3. Adjustments in pattern assignations by using parallel corpora

As explained in Step 1, the occurrences of some combinations were divided into more than one candidate, and morphosyntactic information was stored separately for each of them, leading to different pattern-based classifications. In order to verify if the candidates should really be treated as different combinations or if they should be merged into a single one, a parallel corpus was used. The English-Spanish parallel corpus matching the 15-million Spanish corpus used in Step 1 was chosen for this purpose.

The assumption behind this stage was that, if two candidates containing the same word lemmas were usually translated similarly, they were probably morphosyntactic

variants of the same combination; however, if they were translated in very different ways, it was likely that they had different meanings, and they should remain separate.

N-gram alignments were used to extract possible translations for every candidate. It was counted how many of the translations were shared and, if the percentage was higher than a given (manually set) threshold, the candidates were merged into a single one. Then, the new merged candidate was re-classified, according to the new combined information.

This is what happened to the combination pairs in examples (14) and (15), classified separately in Step 2 (Fig 3). After looking at parallel corpora, it was observed that the amount of translations shared by both candidates in each pair was the following: 15.78% for *dar cuenta* and *darse cuenta* and 64.52% for *ser de interés* and *ser de(l) interés*. Unlike the first pair of candidates, the second pair passed the threshold of shared translations, which was manually set on 35% after some trial. Thus, the information of this pair of candidates was combined, and the merged candidate was re-classified (Fig 4).

Fig 4. Example of the second pattern-based classification in Step 3.

Quality assessment of Steps 1–3

As pointed out previously, trials were made on half of the candidates, and the rest of the candidates (217 from Elhuyar and 272 from DiCE) were used to assess the quality of the data gathered from Steps 1–3. Note, however, that the real impact of this data on MWE identification will be shown later.

The candidates on the test set were manually assigned morphosyntactic patterns, and these were compared to the automatically assigned ones. Since the number of considered patterns was different for the candidates from Elhuyar and the ones from DiCE, evaluation was carried out separately for each source. Similarly, both the classification in Step 2 and the one in Step 3 were evaluated, in order to see what the impact of each step was. Results were calculated both in percentages and according to Cohen's κ [5] (Table 3).

Table 3. Quality assessment of the data gathered from Steps 1–3.

		Elhuyar		DiCE	
		Candidates	**%**	**Candidates**	**%**
Step 2	✓	118	54.38	148	54.42
	×	99	45.62	124	45.59
	κ	0.45		0.39	
Step 3	✓	127	58.53	161	59.19
	×	90	41.47	111	40.81
	κ	0.50		0.45	

General results were fairly good. More than half of the candidates were correctly classified on the first round, and an improvement of around 4 percentage points and 5 to 6 κ points was obtained on the second round.

Although both sets obtain very similar percentages, the candidates from DiCE have a lower κ score. This was to be expected, considering that Cohen's κ takes into account the probability of the method to be right by chance. Since less morphosyntactic patterns were used for the DiCE candidates, this probability was bigger, which led to a decrease in the score.

On the other hand, it is worth mentioning that a high amount of the erroneously classified candidates were not MWEs, but free counterparts of MWEs. This is the case

of examples (17b) and (18b), which are erroneous counterparts of the VNMWEs in (17a) and (18a), created as a consequence of the divisions made during Step 1.

(17) a. ***aplicar pena*** (lit. apply penalty) → obj.
b. *aplicar a pena* (lit. apply to penalty) → ccomp.

(18) a. ***expresar esperanza*** (lit. express hope) → obj.
b. *expresar esperanza* (lit. express hope) → subj.

Division of candidates was still positively valued as a disambiguation strategy. Nevertheless, for future work, it would be convenient to develop a refined division procedure, perhaps including a more sophisticated way of discarding redundant candidates, since this overgeneration can affect the quality of the whole method negatively.

After quality assessment and manual corrections, 282 VNMWEs from Elhuyar and 264 VNMWEs from DiCE were collected along with detailed linguistic data. This final set was then used as a starting point for the translation-related steps of the method.

Step 4. Extraction of translation candidates from Spanish-Basque parallel corpora

In order to obtain Basque translations of the VNMWEs analysed in the previous steps, word alignments were automatically generated by using the mGIZA tool [10] on a 7-million-sentence parallel Spanish-Basque corpus. Since these translations would later be integrated into an MT system, a few adjustments were made on the word-aligning process, in order to adapt candidate translations to our purpose as much as possible.

As a matter of fact, one of the conclusions from previous work was that morphosyntactically regular translations were usually better handled by the *Matxin* MT system. Therefore, whenever more than one possible alignment existed for a word combination, some morphological structures were prioritised. The priority scale we used was the following:

1. noun+verb combinations
2. adjective+verb or adverb+verb combinations
3. single verb or verb+one or more components of any other category
4. other morphological structures

A list of candidate translations was created as an output of this process. Some clean-up was undertaken next, in order to discard translations where no verb was included, and the most frequent translation candidate was then selected per VNMWE. For instance, if the Spanish combination *generar confianza* (lit. generate confidence 'build confidence') occurred 255 times in the parallel corpus and 202 of the occurrences were aligned with the Basque combination *konfiantza sortu* (lit. confidence create 'create confidence'), this translation would be assigned a value of 79.21%. Its frequency being higher than the rest, this translation candidate would be selected (Fig 5).

Fig 5. Example of the selection of translation candidates in Step 4.

Quality assessment of Step 4

Automatically generated translations were evaluated manually. Some VNMWEs did not get any translations from the corpus, either because they were absent from it or because the obtained translations were discarded because they were not suitable (i.e. because they did not contain any verb at all). This group accounted for 70 VNMWEs from Elhuyar and 52 from DiCE.

Excluding these, for each Spanish VNMWE, it was specified if the Basque translation was correct or incorrect, and incorrect cases were split in two: those which needed lexical adjustments only, and those which (also or exclusively) needed grammatical adjustments. Results are collected in Table 4.

Table 4. Quality assessment of the translations gathered automatically in Step 4.

	Elhuyar		**DiCE**	
	Translations	**%**	**Translations**	**%**
✓	98	46.23	102	48.11
× **lex.**	100	47.17	101	47.64
× **gr.**	14	6.60	9	4.25

Nearly half of the translations were correct, and most of the incorrect ones needed lexical adjustments. Although these results are not brilliant, it must be born in mind that the size and nature of the corpus greatly affects the quality of alignments. The 7-million-sentence corpus collecting general texts was perhaps insufficient for this purpose, especially taking into account that 122 VNMWEs did not get any valid translation. Besides, there was a clear tendency for the most frequent VNMWEs in the parallel corpus to obtain correct translations, as opposed to the less frequent ones which were often evaluated as incorrect.

After manual adjustments were made, 535 different Basque translations were collected for the 546 Spanish VNMWEs (as some VNMWEs shared translations with other VNMWEs). These were used as a basis for the following step.

Step 5. Extraction of linguistic information about translations from Basque corpora

Once all translations were collected, information about their usage was extracted from corpora. Since the same information-gathering methodology from Step 1 was reused here, translations were firstly divided into two groups: those consisting of a verb and a noun, and the rest. Then, the combinations consisting of components other than a verb and a noun (a total of 40) were momentarily left aside. Only the categories of their component lemmas were specified, like in examples (19) and (20).

(19) ES: ***ser*** *un* ***consuelo*** ('be a consolation')
EU: *kontsolagarri izan* ('be consoling') → adj.+verb

(20) ES: ***dar alivio*** ('give relief')
EU: *lasaitu* ('relieve') → verb

Meanwhile, more data about the noun+verb (NV) translations was gathered from a Basque corpus, by reusing the methodology from Step 1. The features we looked at were the same as the ones considered in Step 1: number and definiteness of the NP, determiners and modifiers inside the NP and alterations in the order of the component words.

A few modifications had to be made in order to adapt the methodology to a different language. Basque being an agglutinative language, the main adjustment consisted in changing the way definiteness was looked at. In fact, in Spanish, information on definiteness is given by the determiner, whereas in Basque, definite articles are not independent words and are typically attached to the last element in an NP [13, pp. 135] (example 21).

(21) ES: ***dar*** *el* ***paso; dar*** *el gran* ***paso***
give the setp; give the big step
'take the step; take the big step'

EU: ***pausoa eman; pauso*** *handia* ***eman***
step-the give; step big-the give
'take the step; take the big step'

Therefore, this information was obtained by looking at the head noun in the NP first; then, if the head noun did not contain any information about definiteness, the rest of the elements in the NP were examined in descending order according to the dependency tree. An example of the output of Step 5 is shown in Fig 6.

Fig 6. Example of the output of Step 5. Basque keys stand for the following word combinations: *airean egon* (lit. be in the air 'be up in the air'), *aukera aprobetxatu* (lit. profit the opportunity 'take the opportunity), *eskuak garbitu* (lit. wash the hands 'wash one's hands'), *hitz eman* (lit. give word 'give one's word'), *pikutara bidali* (lit. go for figs 'tell sb to get lost')

Step 6. Classification of Basque translations according to morphosyntactic patterns

Finally, translations were also classified according to morphosyntactic patterns. The scheme from Step 2 was followed: half of the translations (269) were used to set the patterns and for trials, and the rest (268) were employed for evaluation.

As before, percentages collected in the previous step were converted into Y, O and N values by applying thresholds, and patterns were created by grouping translations with similar features. Six different morphosyntactic patterns were created, which are collected and described on Table 5. An example of the output of this Step is shown in Fig 7

Table 5. Morphosyntactic patterns for the Basque translations.

	Sing.	Pl.	Det.	Def.	Ind.
FREE	O/N	O/N	Y/O/N	O/N	Y/O/N
IND	N	N	N	N	Y
SING_DEF	Y	N	N	Y	N
SING	Y	N	O/Y	Y/O/N	N
PL_DEF	N	Y	N	Y	N
PL	N	Y	Y/O	Y/O/N	N

Fig 7. Example of the output of Step 6.

On the other hand, three more patterns were created for the 40 translations which were momentarily left aside in the previous step, i.e. for the translations which were

non-NV combinations. These three patterns had no other feature than the grammatical category of the component words, that is: AdjV (adjective+verb), AdvV (adverb+verb) and V (verb).

Quality assessment of Steps 5–6

Evaluation of the pattern assignations for Basque translations was carried out both including the 40 non-NV translations and excluding them, and results were calculated in percentages and by using Cohen κ (Table 6).

Table 6. Quality assessment of the pattern assignations in Step 6. Note that 62 VNMWEs did not occur in the corpus with a frequency of 10 or higher, so these were not counted for evaluation.

	Complete set		**Only NV set**	
	Translations	**%**	**Translations**	**%**
✓	150	72.82	110	66.27
×	56	27.18	56	33.73
κ	0.62		0.53	

As was to be expected, the results obtained using the complete set of translations were considerably higher, since the classification of the 40 non-NV translations was straight-forward. Nevertheless, even when only the NV combinations were counted, two thirds of the set were correctly classified, which is a fairly good result.

Manual adjustments were made on the patterns, and the final datasets were prepared to be used in the following experiments. More details will be given in the next sections, and the real impact of the whole method on MWE identification and MT will be shown.

Application of the data for MWE identification

In order to test whether the information gathered by the method is useful to improve MWE identification, an experiment was undertaken. The Spanish part of the PARSEME multilingual corpus [24] was used, so that our results could be compared to the results from the edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal MWEs [23].

However, it must be born in mind that the criteria followed to annotate this corpus were not completely compatible with ours. More precisely, the MWEs annotated in the PARSEME corpus differ from the MWEs considered in our previous information-gathering process in two main aspects:

- MWEs composed by all kinds of grammatical categories are annotated, such as verb+adjective and verb+preposition, not only VNMWEs.
- Collocations which are not Light Verb Constructions (LVCs) are excluded.

With a view to carrying out an experiment as comparable as possible to the PARSEME Shared Task, a few adjustments had to be made both on the VNMWE dataset and on the corpus. Since the PARSEME multilingual corpus was released with not only MWE tags but also morphosyntactic information (based on Universal Dependencies), the preparation of the corpus was quite simple. Only the MWE tags in which the component words were a verb and a noun (sometimes also with a preposition or determiner) were considered, omitting the rest and creating an adapted corpus of 5,515 sentences and 662 MWE tags, distributed as follows: 355 in the Train corpus, 136 in the Development corpus, and 171 in the Test corpus.

On the other hand, so as to avoid problems coming from disparities in the conception of MWEs according to PARSEME and according to the source dictionaries we used, only the word combinations annotated in the corpus were looked for. The information-gathering method was applied to all 662 VNMWE tags in the corpus, and a new dataset was built. Some of the tags (273 in all) could not be classified by the method because of their low frequency; the resting 389 were found to be occurrences of 156 VNMWEs, and were automatically classified according to patterns, following Steps 1–3 described in the previous section.

Then, the overall identification strategy was the following:

1. The component lemmas of the automatically classified VNMWEs were searched for, and the restrictions corresponding to their morphosyntactic patterns were applied in order to discard non-MWE occurrences. Both dependency-based information and automatically generated chunks were used to apply restrictions, like in [14].
2. The VNMWEs which could not be automatically classified were treated in two different ways:
   - **A.** As fixed word sequences where the only variation possible was verb inflection. The verb lemma was looked for, but the rest of the components needed to be in the same word form as in the corpus tag. All component words needed to be contiguous and respect the order in the corpus tag.
   - **B.** As completely flexible combinations. The lemmas of the component words were looked for, in any order and word form.

In addition, two different datasets were employed for the experiment: the full dataset of VNMWEs extracted from the corpus (henceforth, *full*), including the VNMWEs from the Test part, and a reduced dataset which contained only the VNMWEs in the Train and Development parts (henceforth, *train+dev*). Results are discussed below.

Results of the identification experiment

All results are collected in Table 7. As can be noticed, scores are very good, with an F score of 0.51 when only the *train+dev* dataset is used and of 0.72 when the *full* dataset is used. Additionally, results show that, when no morphosyntactic pattern is available for some VNMWEs, it is better to treat those VNMWEs as fixed word sequences, accepting verb inflection as the only possible variation.

	P	R	F
A-train+dev	0.74	0.29	**0.51**
B-train+dev	0.60	0.31	0.46
A-full	0.84	0.60	**0.72**
B-full	0.76	0.67	0.71

Table 7. Results of the identification experiment

Recall was pretty low when only the *train+dev* dataset was used, since a lot of VNMWEs in the Test corpus did not occur in the Train and Development parts. This was to be expected, especially taking into account that this was quite general among the 17 systems which participated in the Shared Task [23] (Table 8).

Although our recall is a bit under the average recall in Spanish, our precision is much higher, which led us to achieve an F score of 13 points higher than the best-performing system in the Spanish part of the Shared Task [3].

	P	R	F
ES	0.19 (0.00-0.32)	0.33 (0.00-0.49)	0.23 (0.00-0.38)
All languages	0.36 (0.00-0.68)	0.29 (0.01-0.53)	0.31 (0.00-0.54)

Table 8. Results of the PARSEME 1.1 edition, in Spanish and in all 20 languages

While it is true that these results are not completely comparable to ours, since verbal MWEs other than VNMWEs are also considered in the Shared Task, category-based results suggest that VNMWEs are precisely one of the most difficult kind of MWE to identify. As a matter of fact, MWEs composed of a verb and a noun fall into three of the MWE categories considered in the PARSEME corpus (LVC.full, LVC.cause and VID), and these are the categories where the lowest F scores were obtained (Table 9).

	P	R	F
LVC.cause	0.13 (0-100)	0.02 (0-0.21)	0.03 (0-0.30)
LVC.full	0.17 (0-0.48)	0.14 (0-0.41)	0.13 (0-0.33)
VID	0.14 (0-0.46)	0.08 (0-0.23)	0.10 (0-0.31)

Table 9. Spanish results of the PARSEME 1.1 edition by MWE category

Therefore, it can be concluded that our results are better than the ones obtained in Spanish in the Shared Task edition 1.1, which confirms the usefulness of the information-gathering method for the identification of MWEs. Additionally, it must be taken into account that the dataset used here was the output of a completely automatic analysis process, with no manual adjustment on the morphosyntactic patterns obtained. If manual checks were performed, results would probably be even better.

Application of the data for Machine Translation

In the second experiment, the data used was not only identification-oriented (Steps 1–3) but also translation-oriented (Steps 4–6). The *Matxin* MT system was selected as a basis for the experiment, an open-source Spanish-Basque rule-based system [18]. Note that, while there exists a neural system [9] which outperforms *Matxin*, rule-based MT was considered a better choice here, since detailed morphosyntactic data can be integrated in a more straight-forward way, and its effect on translation quality is more appreciable.

Like any other rule-based system, *Matxin* works in three main phases: analysis, transfer and generation. Our approach was to carry out MWE processing between the first and the second phase, so that the output of the analysis of the source sentence could be employed. Let us explain this process with two cases in point (examples 23 and 22).

(22) *Quiere* ***mantener*** *el* ***equilibrio****.*
he/she-wants maintain the balance
'He/She wants to maintain balance.'

(23) *La pareja* ***contrajo matrimonio****.*
the couple contracted marriage
'The couple got married.'

After the source sentence is analysed, the MWEs *mantener el equilibrio* (lit. maintain the balance 'maintain the balance') and *contraer matrimonio* (lit. contract

marriage 'get married') are identified as explained in the previous section, by using the restrictions specified by their corresponding patterns (SING in both cases). Then, the lemmas of their translations are specified before the transfer phase, and morphosyntactic information is modified when necessary. Figs 8 and 9 show the output of the analysis and transfer phases of both the original MT system (*Matxin*) and the one which integrates MWE-specific information (*Matxin*-MWE). The generation phase is carried out normally.

Fig 8. Output of the analysis and transfer phases of *Matxin* and *Matxin*-MWE when translating *mantener el equilibrio* ('maintain the balance').

Fig 9. Output of the analysis and transfer phases of *Matxin* and *Matxin*-MWE when translating *contraer matrimonio* (lit. contract marriage 'get married').

As can be seen in Fig 8, when translating the sentence in example (22), modifications are made both on the lexicon and on morphosyntactic features. On the one hand, the Spanish verb is translated by a verb different than the usual one: *eutsi* ('hold') instead of *mantendu* ('mantain'). On the other hand, the noun *oreka* ('balance') is no longer treated as the direct object of the verb, and the dative case is assigned to it instead of the absolutive. Since the pattern corresponding to the translation *oreka-(DAT) eutsi* (lit. hold balance 'mantain balance') is SING, no additional morphosyntactic changes can be appreciated, because the literal translation given by *Matxin* is also singular. Otherwise, changes on number and definiteness would also be marked at this stage.

The final translation given by *Matxin* is shown in example (24), and the one given by *Matxin*-MWE, in example (25).

(24) *Oreka mantendu nahi du.*
balance.ABS mantain desire has
'He/She wants to maintain balance.'

(25) ***Orekari eutsi*** *nahi dio.*
balance.DAT hold desire has
'He/She wants to maintain balance.'

Concerning the sentence in example (23), the MWE *contraer matrimonio* (lit. contract marriage 'get married') is usually translated by a single verb into Basque: *ezkondu* ('marry'). In such cases, as Fig 9 shows, the verb and the noun are not translated separately by *Matxin*-MWE, but the whole MWE is given a single translation instead, consequently deleting the node corresponding to the noun. The output sentence given by *Matxin* is shown in example (26), and the output of *Matxin*-MWE, in example (27).

(26) **Bikotea ezkontza uzkurtu zen.*
couple-the marriage contract AUX
'The couple got contracted marriage.'

(27) *Bikotea* ***ezkondu*** *zen.*
couple-the married AUX
'The couple got married.'

In previous work [15], manually analysed detailed information was integrated into *Matxin*, and a major conclusion was that too many grammatical modifications on MWE transfer usually led the system to do errors. Therefore, apart from selecting translations composed by a verb and a noun when possible (see Step 4), three additional rules were created, by generalising three linguistic features which used to be applied MWE by MWE in the previous approach. These rules were the following:

- When the syntactic relation between the verb and the NP is not of direct object in Spanish but it is in Basque, if there is a direct object in the Spanish sentence, its equivalent node in Basque is treated as the indirect object. This helps solve many errors that used to arise from syntactic disparity between languages, but without needing to add one rule per VNMWE as in the previous approach.

  (28) ES: ***Castigar*** *a alguien (dobj.)* ***con*** *una* ***pena*** *(ccomp)*
  punish to somebody with one penalty
  'Punish somebody with a penalty'
  EU: *Norbaiti (iobj.)* ***zigorra*** *(dobj.)* ***ezarri***.
  somebody.DAT punishment-the establish
  'Set a punishment for somebody'

- In negative sentences, the partitive postposition is usually attached to the NP in Basque, and there are only a few VNMWEs which are an exception to this. *Matxin* currently follows the general rule of always applying the partitive postposition to negated NPs. However, an exception was added here, which concerns the Basque VNMWEs where the NP is always indeterminate and in the absolutive case, and where the verb is *izan* ('be') or *ukan* ('have'), two of the most common verbs which typically form light verb constructions.

  (29) a. *Ez* ***dut nahi***. (vs **nahirik*)
  no AUX desire (vs desire.PART)
  'I do not want to'.
  b. *Ez* ***naiz bizi***. (vs **bizirik*)
  no AUX life (vs life.PART)
  'I do not live'.

- The third rule is the most general one, since it does not only affect VNMWEs but also any other source sentence. It consists in adapting the impersonal use of Basque verbs when translating Spanish reflexive verbs. Nowadays, *Matxin* gives pronouns one of a set of tags during the analysis process, based on the *Freeling* parser [19]. Some errors of the system come from an erroneous selection of such tags, which affects the choice of verbal forms in the translation: some reflexive verbs are translated as impersonal when they should be transitive (example 30) and viceversa (example 31).

  (30) ES: *Ella* ***se busca la vida***.
  She REFL searches the life
  'She gets by'.
  MT: *Hura* ***bizimodua ateratzen*** *da*.
  He/she leaving-the comes-out AUX-INTR
  'The leaving comes out'.
  EU: *Hark* ***bizimodua ateratzen*** *du*.
  He/She leaving-the takes-out AUX-TR
  'He/She gets by'.

(31) ES: *Se busca trabajo.*
REFL search job
'Job wanted'.
MT: *Lana bilatzen du.*
job search AUX-TR
'(He/She) searches for a job'.
EU: *Lana bilatzen da.*
job search AUX-INTR
'Job wanted'.

The rule proposed here is simple but effective to solve some of these kinds of problems: when both the subject and the object of a given sentence are present, it means the reflexive pronoun is not linked to the impersonal use of the verb, and the auxiliary verb is thus transferred as transitive into Basque.

Once these rules were integrated, an evaluation was carried out to see how VNMWE information and these three rules affected translation quality. For a more exhaustive analysis of the results, the VNMWE datasets were separated in several groups according to the way they were analysed.

- MWEmanual: 668 VNMWEs analysed either manually or semi automatically. The 133 completely manual VNMWEs were the same as the ones used in previous work [15], and the resting 535 were the ones used during the creation and quality assessment of the information-gathering method explained in this paper.
- MWEfiltered: 226 VNMWEs which were automatically analysed, but which occurred in the parallel corpus (Step 4) at least 10 times. These were collected from DiCE and from the PARSEME corpus.
- MWEall: 214 VNMWEs which were automatically analysed and did not have a frequency of 10 or higher in Step 4. These were also collected from DiCE and from the PARSEME corpus.

Results of the MT experiment

Evaluation was twofold: automatic and manual. Firstly, three automatic evaluation metrics were used: BLEU [21], NIST [8] and TER [26]. A 21,786-sentence parallel corpus was employed as a reference, which was specifically crafted for this experiment, meaning that all sentences contained the component lemmas of at least one VNMWE of the sets mentioned above. Several versions of the MT system were compared:

- *Matxin*: the original system
- *Matxin+*: the original system with the third general rule explained in this section
- *Matxin*-MWEmanual: the *Matxin+* version with the 668 manually or semi-automatically analysed VNMWEs
- *Matxin*-MWEfiltered: the previous version plus information about 226 VNMWEs which were automatically analysed but which had a frequency filter
- *Matxin*-MWEall: the previous version plus information about 214 VNMWEs which were automatically analysed with no frequency filter

As Table 10 shows, all versions led to an improvement in translation quality according to the evaluation metrics, except for the last one, which had almost no impact. However, the statistical improvement VNMWE-specific information brings to *Matxin* is not very big. This is mostly due to the fact that these measures, although useful to check the general quality of MT systems, are not too meaningful for specifically evaluating the quality of MWE translation [6].

	BLEU	**NIST**	**TER**
Matxin	7,08	4,04	85,90
Matxin+	7,17	4,05	85,44
Matxin-MWEmanual	7,23	4,07	85,34
Matxin-MWEfiltered	7,24	4,07	85,31
Matxin-MWEall	7,24	4,08	85,30

Table 10. Results of the MT experiment according to BLEU, NIST and TER

On account of the inadequacy of statistical measures for our purpose, a second evaluation was also undertaken, this time manually. Three experts participated in it, and the evaluation set was split in three parts: firstly, the *Matxin+* system was compared with *Matxin*-MWEmanual (150 sentences); secondly, *Matxin*-MWEmanual with *Matxin*-MWEfiltered (150 sentences); and finally, *Matxin*-MWEfiltered with *Matxin*-MWEall (50 sentences).

As can be seen in Table 11, the manually analysed VNMWE set and the automatically analysed but filtered VNMWEs have a very positive impact on translation quality, since 62–65% of the VNMWEs are better translated by them, and only 7–8% are translated worse. The exception is the non-filtered set of VNMWEs, which makes the system do more errors than improvements.

	✓	=	×	**Disagreement**
Matxin-MWEmanual	% 62	% 11	% 8	% 19
Matxin-MWEfiltered	% 65	% 12	% 7	% 16
Matxin-MWEall	% 14	% 14	% 56	% 16

Table 11. Results of the MT experiment according to the manual evaluation

Therefore, it can be concluded that VNMWE-specific information gathered by the method proposed in this paper, if manually supervised or filtered by frequency, is beneficial for MT.

Conclusion

A six-step method to automatically gather VNMWE-specific linguistic information from corpora was described and tested in this paper. Morphosyntactic features were especially looked at, and the resulting information is now stored in a publicly available database, *Konbitzul* (`http://ixa2.si.ehu.eus/konbitzul/?lang=en`). Two experiments were undertaken to see what the effect of the automatically or semi-automatically gathered data was on MWE identification and MT, and the main conclusion was that its impact is positive for both tasks.

In MWE identification, an F score of 0.51 was obtained using the Spanish part of the PARSEME corpus, released for the PARSEME shared task on automatic identification

of verbal MWEs. This score is 17 points higher than the best-performing system (among 17) in edition 1.1 of the shared task. Besides, if all VNMWEs to identify were listed in a lexicon (and not only the ones in the Train and Development parts of the corpus), F score would be of 0.72 by using the automatically gathered linguistic information. Likewise, the analysed linguistic information has a positive effect on the translation quality of a rule-based MT system. A modest increase is appreciated in BLEU, NIST and TER scores, and an improvement of 62–65% according to human judgements.

For future work, the method would benefit from some improvements, especially concerning the extraction of VNMWE translations from parallel corpora. The use of different corpora (i.e. of specialised text) would also be helpful to see how the use of MWEs changes between domains. Finally, it would be interesting to keep developing the method by including MWEs other than VN combinations, as well as by adapting it to languages other than Spanish and Basque.

References

1. Alonso Ramos M., Nishikawa A. & Vincze O. DiCE in the web: an online Spanish collocation dictionary. Proceedings of eLex 2009, e-lexicography in the 21st century: new challenges, new applications. 2010: 369–374.

2. Baldwin T. & Kim SN. Multiword Expressions. Handbook of Natural Language Processing, 2. 2010: 267–292.

3. Boros T. & Burtica R. GBD-NER at PARSEME Shared Task 2018: Multiword Expression detection using bidirectional long-short-term memory networks and graph-based decoding. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018). 2018: 254–260.

4. Buckingham L. Las construcciones con verbo soporte en un corpus de especialidad. Peter Lang. 2009.

5. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1). 1960: 37–46.

6. Constant M., Eryiğit G., Monti J., Van Der Plas L., Ramisch C., Rosner M. & Todirascu A. Multiword Expression processing: a survey. Computational Linguistics, 43(4). 2017: 837–892.

7. Corpas Pastor G. Manual de fraseología española. Editorial Gredos. 1996.

8. Doddington G. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. Proceedings of the second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. 2002: 138–145.

9. Etchegoyhen T., Martinez Garcia E., Azpeitia A., Labaka G., Alegria I., Cortes Etxabe I., Jauregi Carrera A., Ellakuria Santos I., Martin M. & Calonge E. Neural Machine Translation of Basque. Proceedings of the 21st Annual Conference of the European Association for Machine Translation. 2018: 139–148.

10. Gao Q. & Vogel S. Parallel implementations of word alignment tool. Software engineering, testing, and quality assurance for Natural Language Processing. 2008: 49–57.

11. Gurrutxaga A. & Alegria I. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world (at ACL 2011). 2011: 2–7.

12. Gurrutxaga A. & Alegria I. Combining different features of idiomaticity for the automatic classification of noun+ verb expressions in Basque. Proceedings of the 9th Workshop on Multiword Expressions. 2013: 116–125.

13. Hualde JI., Oyharçabal B. & Ortiz de Urbina J. Verbs. A grammar of Basque. De Gruyter. 2003: 155–198.

14. Inurrieta U., Diaz de Ilarraza A., Labaka G., Sarasola K., Aduriz I. & Carroll J. Using linguistic data for English and Spanish verb-noun combination identification. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers. 2016: 857–867.

15. Inurrieta U., Aduriz I., Diaz de Ilarraza A., Labaka G. & Sarasola K. Rule-based translation of Spanish Verb-Noun combinations into Basque. Proceedings of the 13th Workshop on Multiword Expressions (at EACL 2017). 2017: 149–154.

16. Inurrieta U., Aduriz I., Diaz de Ilarraza A., Labaka G. & Sarasola K. Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system. Mitkov R., Monti J., Corpas Pastor G. & Seretan V. (eds.). Multiword Units in Machine Translation and Translation Technology. John Benjamins Publishing Company. 2018: 41–59.

17. Inurrieta U., Aduriz I., Diaz de Ilarraza A., Labaka G. & Sarasola K. Konbitzul: an MWE-specific database for Spanish-Basque. Proceedings of LREC 2018, the 11th Language Resources and Evaluation Conference. 2018: 2500–2504.

18. Mayor A., Alegria I., Diaz De Ilarraza A., Labaka G., Lersundi M. & Sarasola K. Matxin, an open-source rule-based Machine Translation system for Basque. Machine Translation, 25(1). 2011: 53–82.

19. Padró L. & Stanilovsky E. FreeLing 3.0: Towards Wider Multilinguality. Proceedings of LREC 2012, the Language Resources and Evaluation Conference. 2012: 2473–2479.

20. Parra Escartín C., Nevado Llopis A. & Sánchez Martínez E. Spanish Multiword Expressions: looking for a taxonomy. Sailer M. & Markantonatou S. (eds.). Multiword expressions, insights from a multilingual perspective. 2018: 271–323.

21. Papineni K., Roukos S., Ward T., & Zhu WJ. BLEU: a method for automatic evaluation of Machine Translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311–318.

22. Ramisch C. Multiword Expressions acquisition. A generic and open framework. Springer. 2015.

23. Ramisch C., Cordeiro SR., Savary A., Vincze V. et al. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018). 2018: 222–240.

24. Savary A., Candito M., Mititelu VB., Bejček E., Cap F. et al. PARSEME multilingual corpus of Verbal Multiword Expressions. Markantonatou S., Ramisch C., Savary A., Vincze V. (eds.), Multiword Expressions at length and in depth: extended papers from the MWE 2017 workshop. 2018: 87–147.

25. Savary A., Cordeiro SR., Lichte T., Ramisch C., Inurrieta U. & Giouli V. Literal Occurrences of Multiword Expressions: Rare Birds that Cause a Stir. The Prague Bulletin of Mathematical Linguistics, 112(1). 2019: 5–54.

26. Snover M., Dorr B., Schwartz R., Micciulla L., & Makhoul J. A study of translation edit rate with targeted human annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. 2006: 223–231.

27. Timofeeva L. Los principios definitorios de las Unidades Fraseológicas: nuevos enfoques para viejos problemas. ELUA, Estudios de Lingüística de la Universidad de Alicante, 22. 2008: 243–261.

28. Vincze O. & Alonso M. Incorporating frequency information in a collocation dictionary: Establishing a methodology. Procedia, Social and Behavioral Sciences, 95. 2013: 241–248.

Fig 1

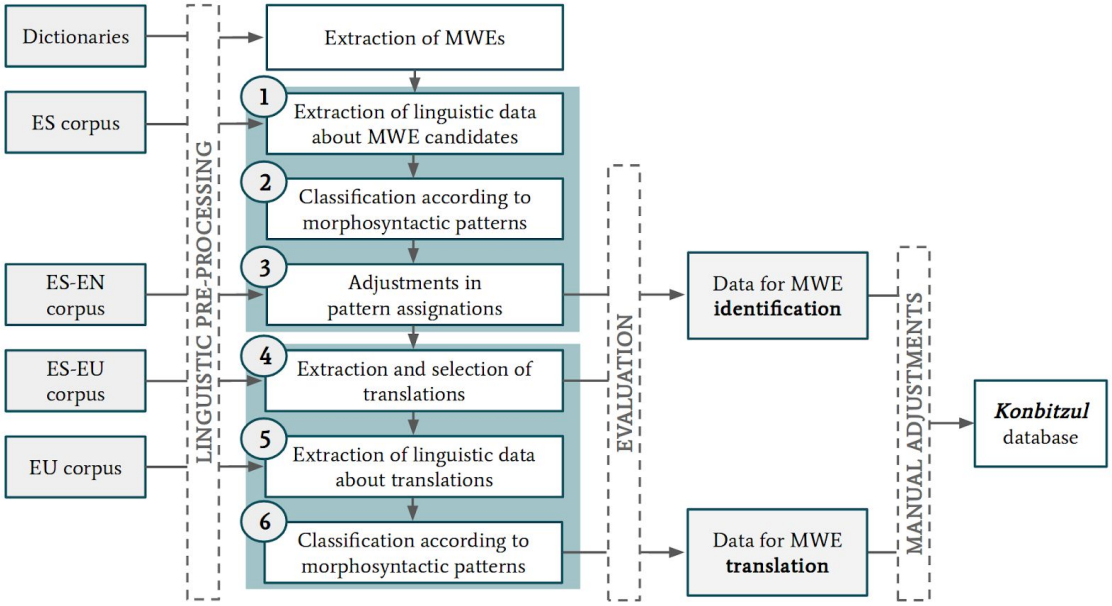


Fig 2

	NP number			Definiteness			
	Pron.	Sing.	Pl.	Det.	Def.	Ind.	Mod.	Ord.
TOMAR - subj - - **PARTE**	0	100	0	0	0	0	3.55	89.94
TOMAR - obj - - **PARTE**	0	100	0	0	0	0	4.85	78.97
SER - ccomp **DE** - **INTERÉS**	0	100	0	0	0	0	47.21	6.61
SER - ccomp **DE** * **INTERÉS**	0	83.98	16.02	95.39	86.65	8.74	43.69	5.83
DAR pron obj - - **CUENTA**	100	100	0	0	0	0	85.71	2.38
DAR - obj - - **CUENTA**	0	100	0	0	0	0	97.37	25.66
DEJAR - ccomp **A** * **LADO**	0	92.86	7.14	100	16.67	83.33	9.52	2.38
DEJAR - ccomp **DE** * **LADO**	0	100	0	100	30	70	20	10

Fig 3

	NP number			Definiteness		
	Pron.	Sing.	Pl.	Det.	Def.	Ind.	
TOMAR - subj - - **PARTE**	N	Y	N	N	N	N	DISCARD
TOMAR - obj - - **PARTE**	N	Y	N	N	N	N	SING_NO-DET
SER - ccomp **DE** - **INTERÉS**	N	Y	N	N	N	N	SING_NO-DET
SER - ccomp **DE** * **INTERÉS**	N	O	O	O	O	O	FREE
DAR pron obj - - **CUENTA**	Y	Y	N	N	N	N	P_SING_NO-DET
DAR - obj - - **CUENTA**	N	Y	N	N	N	N	SING_NO-DET
DEJAR - ccomp **A** * **LADO**	N	Y	N	Y	O	O	SING
DEJAR - ccomp **DE** * **LADO**	N	Y	N	Y	O	O	SING

Fig 4

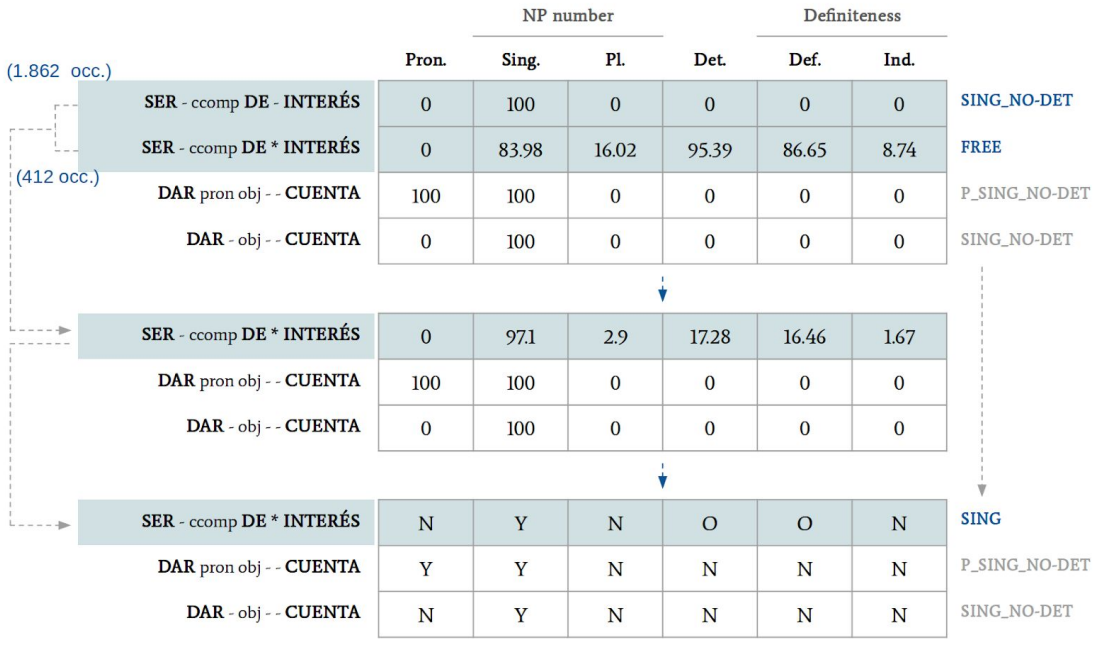


Fig 5

GENERAR - *obj* - - **CONFIANZA**	konfiantza (abs) sortu 79.21%
	konfiantza (abs) sorrarazi 6.67%
	konfiantza (abs) galdetu 3.53%
	konfiantza (abs) areagotu 3.53%
	konfiantza (abs) eman 3.53%
	konfiantza 3,53%

Fig 6

	NP number		Definiteness				
	Sing.	**Pl.**	**Det.**	**Def.**	**Ind.**	**Mod.**	**Ord.**
AIRE * **ine** ccomp **EGON**	100	0	0	100	0	17.78	4.44
AUKERA * **abs** obj **APROBETXATU**	42.23	44.44	52.22	86.67	13.33	32.22	28.89
ESKU * **abs** obj **GARBITU**	0	100	0	100	0	33.33	44.44
HITZ * **abs** obj **EMAN**	39.77	1.14	3.41	40.91	59.09	19.32	12.50
PIKU - **ala** ccomp **BIDALI**	0	0	0	0	100	0	0

Fig 7

	NP number		Definiteness			
	Sing.	Pl.	Det.	Def.	Ind.	
AIRE * **ine** ccomp **EGON**	Y	N	N	Y	N	SING_DEF
AUKERA * **abs** obj **APROBETXATU**	O	O	O	O	O	FREE
ESKU * **abs** obj **GARBITU**	N	Y	N	Y	N	PL_DEF
HITZ * **abs** obj **EMAN**	O	N	N	O	O	FREE
PIKU - **ala** ccomp **BIDALI**	N	N	N	N	Y	IND

Fig 8

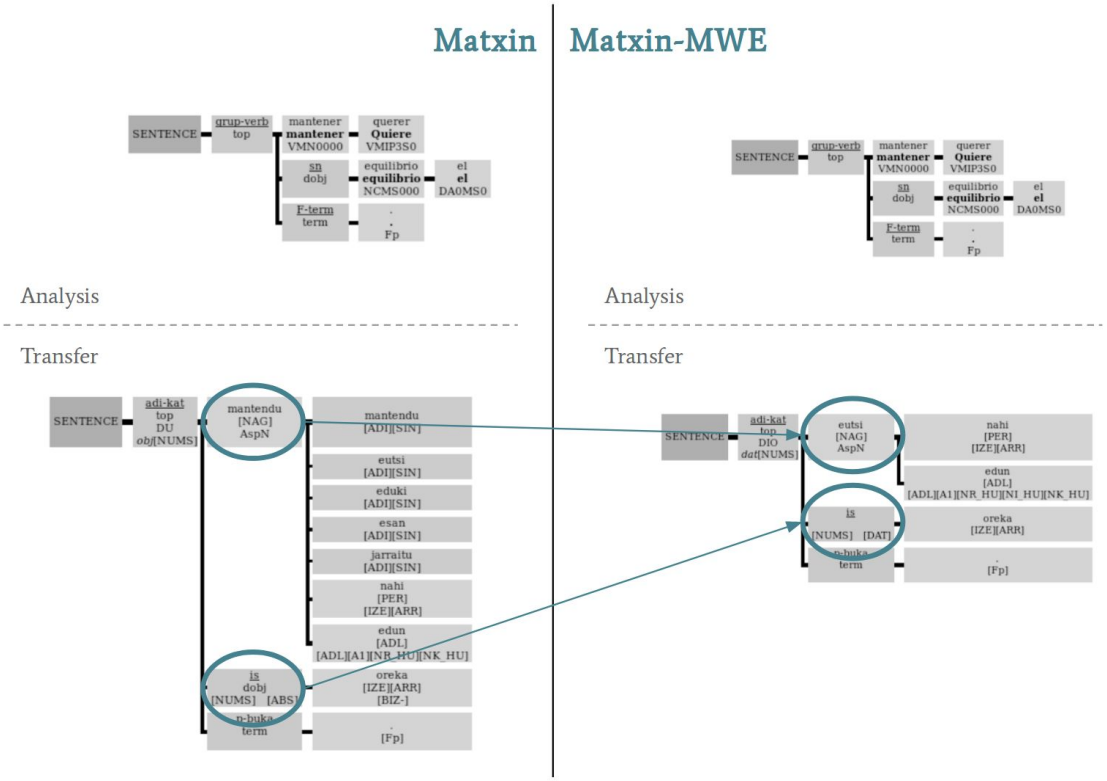
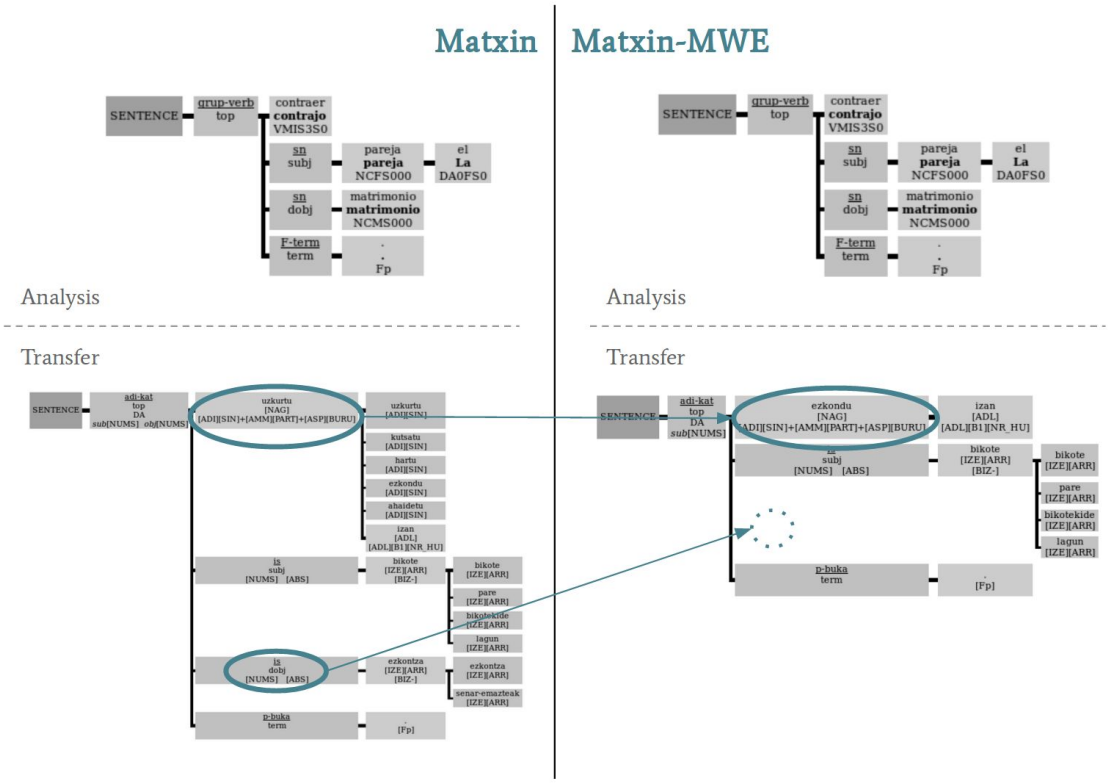


Fig 9


Matxin
Matxin-MWE
SENTENCE
grup-verb
top
contraer
contrajo
VMIS3S0
sn
subj
pareja
pareja
NCFS000
el
La
DA0FS0
sn
dobj
matrimonio
matrimonio
NCMS000
F-term
term
.
.
Fp
Analysis
Transfer
adi-kat
top
DA
sub[NUMS] obj[NUMS]
uzkurtu
[NAG]
[ADI][SIN]+[AMM][PART]+[ASP][BURU]
uzkurtu
[ADI][SIN]
kutsatu
[ADI][SIN]
hartu
[ADI][SIN]
ezkondu
[ADI][SIN]
ahaidetu
[ADI][SIN]
izan
[ADL]
[ADL][B1][NR_HU]
is
subj
[NUMS] [ABS]
bikote
[IZE][ARR]
[BIZ-]
bikote
[IZE][ARR]
pare
[IZE][ARR]
bikotekide
[IZE][ARR]
lagun
[IZE][ARR]
is
dobj
[NUMS] [ABS]
ezkontza
[IZE][ARR]
[BIZ-]
ezkontza
[IZE][ARR]
senar-emazteak
[IZE][ARR]
p-buka
term
.
[Fp]
ezkondu
[NAG]
[ADI][SIN]+[AMM][PART]+[ASP][BURU]
izan
[ADL]
[ADL][B1][NR_HU]
[NUMS] [ABS]
bikote
[IZE][ARR]
[BIZ-]
bikote
[IZE][ARR]
pare
[IZE][ARR]
bikotekide
[IZE][ARR]
lagun
[IZE][ARR]
p-buka
term
.
[Fp]

Konbitzul: an MWE-specific database for Spanish-Basque

Uxoa Iñurrieta, *Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA NLP group, University of the Basque Country
*University of Barcelona
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu, {a.diazdeilarraza, gorka.labaka, kepa.sarasola}@ehu.eus

Abstract

This paper presents Konbitzul, an online database of verb+noun MWEs in Spanish and Basque. It collects a list of MWEs with their translations, as well as linguistic information which is NLP-applicable: it helps to identify occurrences of MWEs in multiple morphosyntactic variants, and it is also useful for improving translation quality in rule-based MT. In addition to this, its user-friendly interface makes it possible to simply search for MWEs along with translations, just as in any bilingual phraseological dictionary.

1. Introduction

Multiword Expressions (MWEs), also called Phraseological Units (PUs), are combinations of words which together express a single meaning (Sag et al., 2002). They often have irregular lexical-semantic and/or morphosyntactic features, and they are not always translated word-for-word (Examples 1-3). This means they cause challenges in various disciplines, such as Lexicography, Translation and Natural Language Processing (NLP).

(1) EN: *pull sb's leg*
ES: *tomar el pelo (a)* lit. *take sb's hair*
EU: *adarra jo* lit. *play the horn (to sb)*

(2) EN: *take steps*
ES: *dar pasos* lit. *give steps*
EU: *urratsak egin* lit. *steps do*

(3) EN: *take off*
ES: *alzar el vuelo* lit. *raise the flight*
EU: *aireratu* lit. *go-to-the-air*

Although interest in Phraseology has a longer history, studies on MWEs have multiplied considerably over the last two decades (Baldwin and Kim, 2010; Savary et al., 2015). Most of the work undertaken within the field of NLP focuses on MWE candidate extraction (Ramisch, 2015) – mainly for lexicographic purposes– or identification of MWE occurrences in corpora (Savary et al., 2017). However, some research has also been conducted into improving Machine Translation (MT) quality by enhancing MWE processing (Kordoni and Simova, 2014; Seretan, 2014). Meanwhile, a considerable amount of resources have been created for several languages, including MWE lists, lexicons and MWE-annotated treebanks (Losnegaard et al., 2016).

Concerning Basque phraseology, research has been done both to describe some linguistic phenomena and to develop NLP tools (Alegria et al., 2004; Gurrutxaga and Alegria, 2012), but researchers have had an almost exclusively monolingual perspective. Thus, our aim is, on the one hand, to analyse how MWEs are translated, and, on the other hand, to propose a method to improve their computational treatment in bilingual tools.

In this paper, we will present Konbitzul, a database of verb+noun MWEs in Spanish and Basque. As well as working as a bilingual phraseological dictionary, the database contains linguistic information which is useful for NLP-related tasks, notably for Parsing and MT.

We will start by introducing the database, including: the verb+noun MWEs collected (Section 2.1.), how linguistic information is included in the database (Section 2.2.), and how the interface is structured (Section 2.3.). We will then go on to explain what the database can be used for: as a helpful tool for MWE identification (Section 3.1.), or as a resource to improve MT quality (Section 3.2.). Finally, we will discuss some conclusions and ongoing and future work.

2. The database

Konbitzul is a database which can be publicly accessed online (Section 2.3.). It currently comprises 3,195 Spanish verb+noun MWEs (along with 7,132 translations) and 2,954 Basque noun+verb MWEs (along with 6,392 translations).

The MWEs in the database were gathered from two main sources: the Elhuyar Spanish-Basque and Basque-Spanish dictionaries[1] and the DiCE dictionary of Spanish collocations[2] (Vincze et al., 2011). However, the detabase being part of an ongoing project, additional sources will probably be used in the future, such as a list of Basque MWEs extracted from corpora by using Gurrutxaga *et al.*'s method (Gurrutxaga and Alegria, 2011). NLP-applicable linguistic information was added afterwards. As this was done in several phases, the amount of linguistic data provided varies from one MWE to another. More information about the analysis will be given in the following paragraphs.

2.1. Verb+Noun MWEs in Spanish and Basque

Whereas Spanish is a romance language, Basque is a non-indoeuropean language which does not belong to any known family. Their typological features are very different:

- Spanish is SVO-ordered, head-initial, fusional, and uses prepositions

[1] http://hiztegiak.elhuyar.eus
[2] www.dicesp.com

- Basque is canonically SOV-ordered[3], head-final, agglutinative, and uses postpositions

Thus, given that they are so dissimilar in such fundamental aspects, it is not surprising that both languages differ considerably in phraseology as well, as typological features directly affect the way in which languages combine words. The MWEs collected in Konbitzul are all made up of a verb and a noun. The Spanish ones can have a preposition and/or a determiner in-between (Example 4), and similarly, Basque noun phrases can have case markers or postpositions attached (Example 5).

(4) A. *tener afecto* (V+N)
lit. *have affection* 'have affection'
B. *hacer un favor* (V+D+N)
lit. *do a favour* 'do a favour'
C. *saber de memoria* (V+P+N)
lit. *know of memory* 'know by heart'
D. *dejar a un lado* (V+P+D+N)
lit. *leave to one side* 'leave aside'

(5) A. *denbora galdu* (N.abs+V)
lit. *time lose* 'waste time'
B. *sutan egon* (N.loc+V)
lit. *fire-in be* 'be very angry'
C. *aurrera egin* (N.alla+V)
lit. *front-to do* 'move forward'
D. *hutsetik hasi* (N.abl+V)
lit. *zero-from start* 'start from scratch'

In previous work, we showed that it is rare for a verb+noun MWE to be translated literally between Spanish and Basque (Example 6). As a matter of fact, out of the Spanish verb+noun combinations in a general bilingual dictionary, only 48.54% had a noun+verb translation in Basque, and only 10.58% were translated word-for-word.

(6) ES: *poner en libertad* (V+P+N)
lit. *put in liberty*
EU: *aske utzi* (Adv+V) / *askatu* (V)
lit. *free leave* / *(to) free*
EN: '(to) release'

As for Basque into Spanish (Example 7), the gap was even bigger: only 30.85% of the noun+verb combinations were translated by a verb and a noun, and only 8.64% of the translations were literal.

(7) EU: *zin egin* (N.abs+V)
lit. *oath do*
ES: *jurar* (V)
lit. *swear*
EN: 'swear'

[3]Note that, although Basque is classified as an SOV language, it is often said to be free-ordered, as word order can be freely altered for emphasis.

2.2. Methodology for analysing linguistic data

As we have already mentioned, most of the linguistic information in Konbitzul is analysed and structured so that it can later be used in NLP tools. The collection and analysis of the MWEs was done in five phases: during the first three, the annotation was mainly manual; the last two are the result of our attempt to automatize the previous manual work. We will now briefly explain the phases one by one.

Phase 1. All the entries consisting of a verb and a noun were gathered from the Elhuyar Spanish-Basque and Basque-Spanish dictionaries. Basic information about them was analysed semi-automatically: morphological structure, number and definiteness of the noun phrases (NPs), and whether the nouns and the verbs in both languages were regular translations or not. This information was used to make some preliminary estimations about the irregularities which occur when translating MWEs between Spanish and Basque (Inurrieta et al., in print).

Phase 2. After having looked at the frequencies of the MWEs analysed in Phase 1, the 150 most common combinations in Spanish were selected for more in-depth study, which would then be used for an identification experiment (Section 3.1.). The combinations were classified into lexical-semantic and morphosyntactic groups, and further morphosyntactic data was examined, such as: possible determiners inside the NPs, variations in number and definiteness, possibility of altering word order, etc. Detailed information about this can be found in (Inurrieta et al., 2016).

Phase 3. A Basque translation was manually given to each of the combinations analysed in Phase 2, and information about this translation was examined: lexical components, whether the number and/or definiteness needed changing between one language and the other, cases in which the translation was not made up of a noun and a verb, etc. The data obtained from this phase was later tested and evaluated in an MT system (Inurrieta et al., 2017).

Phase 4. Once having seen that the analysed information was helpful for MWE identification, the next step was to semi-automatize the linguistic analysis, so that our method could be useful on a bigger scale. We used both the list of Spanish verb+noun combinations from the Elhuyar Spanish-Basque dictionary and a new one obtained from the DiCE collocation dictionary (Vincze et al., 2011). Some data about the features analysed in Phase 2 was automatically extracted from both monolingual and parallel corpora, and this information was employed to group the MWEs according to fifteen morphosyntactic patterns: those never occurring with a determiner, those only used in the plural form, those where the pronominal form of the verb is especial, those which can be freely altered just like any other word combination, etc. We are now in the process of testing this information in MWE identification within parsing.

Phase 5. Parallel corpora were used to obtain translation candidates for the MWEs, by word and n-gram alignment. For each MWE, one of the translations was chosen as the most suitable for MT (usually the most common one requiring less grammatical changes when transferring it from the

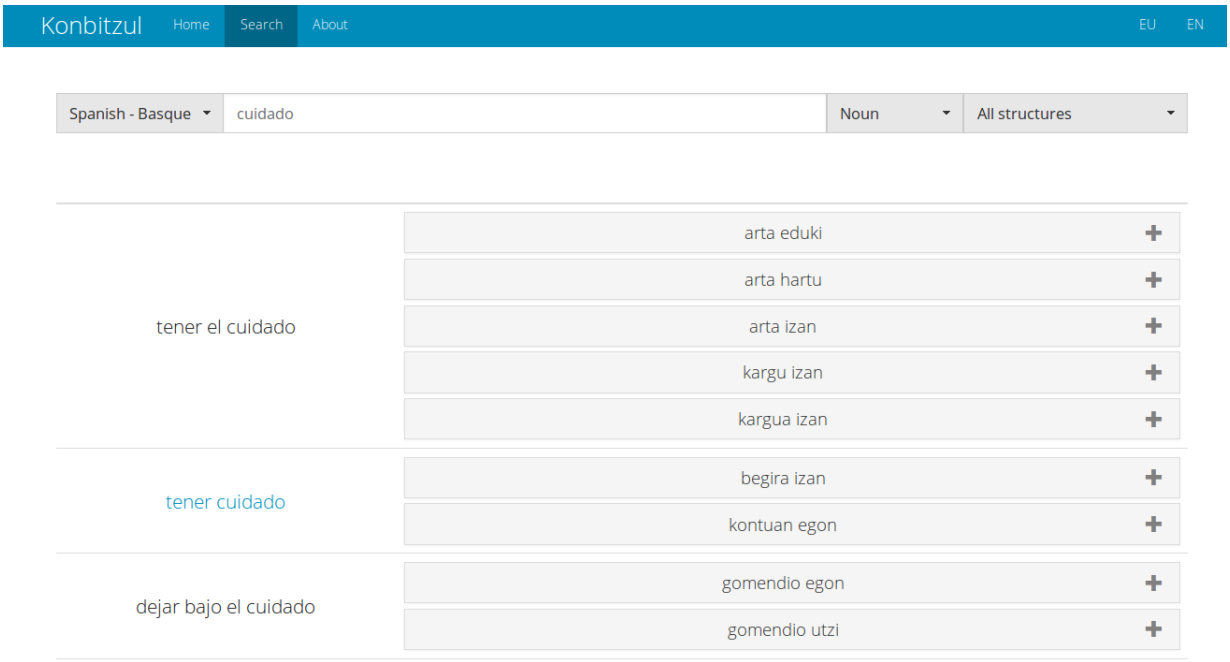


Figure 1: The Konbitzul database's interface. The noun *cuidado* (care) is searched, and three combinations are shown along with their possible translations: *tener el cuidado* (lit. have the care, 'be careful'), *tener cuidado* (lit. have care, 'be careful') and *dejar bajo el cuidado* (lit. leave under the care, 'leave in charge of').

Figure 2: An example of how linguistic information is shown. Two tables are opened after clicking both on the entry *tener cuidado* and on the plus button besides the translation *kontuan egon* (lit. care-in be, 'be careful).

source to the target language). Then, lexical and grammatical information was added. This information is yet to be tested in MT.

2.3. The interface

The database can be publicly accessed at `http://ixa2.si.ehu.eus/konbitzul`. Combinations can be found by typing verb or noun lemmas or full combinations, and morphosyntactic structures can be filtered as well. The results matching the query are listed along with one or more possible translations (Figure 1).

By clicking on the plus button beside each translation, the basic information analysed in Phase 1 can be seen (top table in Figure 2). When the Spanish entry is in a different colour, it means that the combination was (either manually or semi-automatically) analysed in Phase 2 or 4; this information can be seen by clicking on the entry (bottom table in Figure 2). Finally, when one of the translations is also differently coloured and clickable, it means that this translation was marked as the most appropriate for MT (Phases 3 and 5).

3. Applications

Konbitzul was originally created as an NLP-applicable resource. However, with its user-friendly interface, it can simply be used as a phraseological dictionary as well.

In Sections 3.1. and 3.2., two past experiments will be explained, to show the potential impact of the analysed linguistic data on MWE identification and MT.

3.1. MWE identification

Concerning identification, one of the major problems of MWEs is their morphosyntactic variability (Example 8). The most straightforward means of identification is to try to match word sequences against dictionary entries; however, this method falls short in most cases, especially when it comes to verbal MWEs, which tend to have multiple morphosyntactic variants (Savary et al., 2017).

(8) *dar clase* lit. *give lecture*
dar una clase lit. *give one lecture*
dar clases lit. *give lectures*
la clase dada lit. *the lecture given*

In previous work (Inurrieta et al., 2016), the linguistic information in Konbitzul was used to help identify occurrences of a list of verb+noun MWEs in corpora. To be precise, the MWEs were the same ones studied during the second phase of the analysis presented in Section 2.2..

Two identification methods were compared: (A) that used by the Freeling parser (Padró and Stanilovsky, 2012), which only searches for non-separable occurrences of MWEs, and (B) a new one combining the linguistic data in Konbitzul with the chunking and dependency information provided by the parser. The results clearly showed that method B was considerably better, as it identified 28% more occurrences than method A, with a precision score as high as 98% (as opposed to 99%).

3.2. Machine Translation

Likewise, another experiment was undertaken to see whether the information in Konbitzul could improve MT quality. Matxin was used for this study, a rule-based system for Spanish-Basque (Mayor et al., 2011).

As with any rule-based system, Matxin works in three phases: analysis, transfer and generation. The data gathered from Konbitzul was added both to the analysis and transfer phases. Firstly, identification of MWEs was carried out as explained in Section 3.1., and then, lexical and grammatical information about the translation of each MWE (analysed in Phase 3 of Section 2.2.) was used.

The experiment resulted in an increase of 3% in BLEU score (Papineni et al., 2002). In addition, a manual evaluation by three experts was carried out in a controlled corpus, and it was concluded that the new translation was better than the old one in 78.6% of the cases (Inurrieta et al., 2017). Once again, this proves that the kind of linguistic information in the database is helpful for NLP purposes.

4. Conclusion

Konbitzul is an open-source online database of verb+noun MWEs in Spanish and Basque. It currently comprises 6,149 entries in all, which all have one or more translation and rich NLP-applicable linguistic information. Part was added manually, and the reminder is the result of a semi-automatic analysis.

Experiments have confirmed that the information in the database is helpful for NLP tools. Due to the large amount of MWEs requiring a non-regular translation, the database is of special interest for the area of MT, as well as being a useful resource to help identifying multiple morphosyntactic variants of MWEs in text.

As this is an ongoing project, the database is constantly being updated with further MWEs, translations and linguistic information. At the same time, new experiments are being undertaken both to semi-automatize the linguistic analysis and to test the automatic information in NLP tools.

5. Acknowledgements

Uxoa Iñurrieta's doctoral research is funded by the Spanish Ministry of Economy and Competitiveness (BES-2013-066372). This work was carried out in the context of the SKATeR (TIN2012-38584-C06-02) and TADEEP (TIN2015-70214-P) projects.

6. Bibliographical References

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and treatment of multiword expressions in basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Gurrutxaga, A. and Alegria, I. (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 2–7. Association for Computational Linguistics.

Gurrutxaga, A. and Alegria, I. (2012). Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. In *LREC*, pages 2389–2394.

Inurrieta, U., Díaz de Ilarraza, A., Labaka, G., Sarasola, K., Aduriz, I., and Carroll, J. A. (2016). Using linguistic data for English and Spanish verb-noun combination identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, pages 857–867.

Inurrieta, U., Aduriz, I., Díaz de Ilarraza, A., Labaka, G., and Sarasola, K. (2017). Rule-based translation of Spanish verb+noun combinations into Basque. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 149–154.

Inurrieta, U., Aduriz, I., Díaz de Ilarraza, A., Labaka, G., and Sarasola, K. (in print). Analysing linguistic information about word combinations for a Spanish-Basque rule-based machine translation system. In Ruslan Mitkov, et al., editors, *Multiword Units in Machine Translation and Translation Technologies*, pages 41–59. John Benjamins Publishing Company.

Kordoni, V. and Simova, I. (2014). Multiword expressions in machine translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1208–1211.

Losnegaard, G., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S., and Monti, J. (2016). Parseme survey on MWE resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France*, pages 2299–2306. European Language Resources Association (ELRA).

Mayor, A., Alegría, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, 25(1):53–82.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2473–2479.

Papineni, K., Roukos, S., Ward, T., and Zhug, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Ramisch, C. (2015). *Multiword expressions acquisition*. Springer.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard Smørdal, G., et al. (2015). PARSEME: PARSing and Multiword Expressions within a European multilingual network. In *Proceedings of the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., Qasemizadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., et al. (2017). The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47.

Seretan, V. (2014). On collocations and their interaction with parsing and translation. In *Informatics*, volume 1, pages 11–31. Multidisciplinary Digital Publishing Institute.

Vincze, O., Mosqueira, E., and Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 275–286.

Verbal Multiword Expressions in Basque Corpora

Uxoa Iñurrieta, Itziar Aduriz*, Ainara Estarrona,
Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, Iñaki Alegria**
IXA NLP group, University of the Basque Country
*IXA NLP group, University of Barcelona
**Elhuyar Foundation
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu,
ainara.estarrona@ehu.eus, itziar.gonzalezd@ehu.eus,
a.gurrutxaga@elhuyar.eus, ruben.urizar@ehu.eus, i.alegria@ehu.eus

Abstract

This paper presents a Basque corpus where Verbal Multiword Expressions (VMWEs) were annotated following universal guidelines. Information on the annotation is given, and some ideas for discussion upon the guidelines are also proposed. The corpus is useful not only for NLP-related research, but also to draw conclusions on Basque phraseology in comparison with other languages.

1 Introduction

For Natural Language Processing (NLP) tools to produce good-quality results, it is necessary to detect which words need to be treated together (Sag et al., 2002; Savary et al., 2015). However, identifying Multiword Expressions (MWEs) is a challenging task for NLP, and current tools still struggle to do this properly. This is mainly due to the multiple morphosyntactic variants that these kinds of word combinations can have, especially when their syntactic head is a verb.

(1) *They **made** a **decision**.*

(2) *They **made** some difficult **decisions**.*

(3) *The **decisions** they **made** were correct.*

In order to promote research on this topic, the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (VMWEs) was organised (Savary et al., 2017), which holds its second edition this year. One of the outcomes of this initiative is an MWE-annotated corpus including 20 languages. Along with other relevant resources (Losnegaard et al., 2016), this kind of corpus can be helpful to tackle the problems posed by MWEs to NLP. The present paper aims at describing the Basque annotation carried out for this Shared Task (ST), Basque being one of the novel languages included in the new edition.

Comprehensive work has been done on Basque MWEs, not only from a linguistic perspective (Zabala, 2004), but also concerning identification within parsing (Alegria et al., 2004), extraction of VMWEs for lexicographical purposes (Gurrutxaga and Alegria, 2011) and translation (Inurrieta et al., 2017). Nevertheless, this is the first corpus where these kinds of expressions are manually annotated[1].

The paper starts by introducing what resources are used (Section 2), and it goes on to briefly describe how the annotation process was done overall (Section 3). Then, the main confusing issues concerning Basque VMWEs are commented on (Section 4), and a few questions about the guidelines are proposed for future discussion (Section 5). Some remarks about Basque VMWEs are also made based on the annotated corpus (Section 6), and finally, conclusions are drawn (Section 7).

[1] Annotation of Verb+Noun MWEs in Basque was carried out by Gurrutxaga and Alegria (2011), but note that this was not done on corpora but on automatically extracted out-of-context word combinations.

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 86–95
Santa Fe, New Mexico, USA, August 25-26, 2018.

2 Resources and setup

For the annotation described in this paper, a **Basque corpus** was created by collecting texts from two different sources: (A) 6,621 sentences from the Universal Dependencies treebank for Basque (Aranzabe et al., 2015), that is, the whole UD treebank, and (B) 4,537 sentences taken from the Elhuyar Web Corpora[2]. Thus, in all, the Basque corpus consists of 11,158 sentences (157,807 words).

The UD subcorpus comprises news from Basque media, whereas the Elhuyar subcorpus consists of texts which were automatically extracted from the web. Although only good-quality sources were selected and a cleanup was done before performing the annotation, a few strange sentences can still be found in this part due to automatic extraction (such as sentences missing some words or a few words in languages other than Basque). Scripts made available by the ST organisers[3] were used to prepare the corpus before and after annotation.

Likewise, the **annotation guidelines**[4] created specifically for the ST edition 1.1 were used. The guidelines are intended to be universal and were the result of thoughtful discussions among experts from many different languages (Savary et al., 2018). Six different categories of VMWEs are included in the guidelines, but only two of them are applicable to Basque: Verbal Idioms (VID) and Light Verb Constructions (LVCs), the latter being divided into two subcategories, LVC.full and LVC.cause. All of them are universal categories.

Detailed information about each of the categories can be found in the guidelines, as well as decision trees and specific tests provided in order to make it easier to decide whether/how a given combination should be annotated. As a brief explanation to better follow the content of this paper, categories can be broadly defined as follows.

- **VID:** combinations of a verb and at least another lexicalised component whose meaning is not derivable from the separate meanings of the component words.

  (4) ***adarra jo***[5]
  horn-the.ABS play
  '(to) trick, (to) pull somebody's leg'

- **LVC.full:** combinations of a verb and a noun phrase (sometimes introduced or followed by an adposition) where the noun denotes an event or state and the verb adds only morphological features but no meaning.

  (5) ***proba egin***
  test.BARE do
  '(to) try'

- **LVC.cause:** combinations of a verb and a noun phrase (sometimes introduced or followed by an adposition) where the noun denotes an event or state and the verb is causative.

  (6) ***berri izan***
  news.BARE have
  '(to) know (about), (to) have heard (of)'

As for the **annotation platform**, FLAT[6] was used, which has a very user-friendly interface and greatly simplifies the task of adding, deleting or modifying tags.

[2] http://webcorpusak.elhuyar.eus/

[3] https://gitlab.com/parseme/utilities/tree/master/1.1

[4] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=home

[5] Explanations for glosses in examples: ABS → absolutive case; ADV → adverb; AUX → auxiliary verb; BARE → bare noun; FUT → future; LOC → locative postposition; 1PS/3PS → 1st/3rd person singular; 3PP → 3rd person plural.

[6] http://flat.readthedocs.io/en/latest/

3 The annotation process

The annotation process had several phases. First of all, a few training sessions were organised with a dual objective: on the one hand, to help participants get familiarised with the guidelines and the annotation platform; on the other hand, to identify tricky issues that might arise from annotating Basque VMWEs in corpora. Some decisions were made on problematic cases, which were then collected in an internal document to be used as a reference tool along with the guidelines.

Six experts took part in this annotation task: five linguists and a lexicographer, most of which have broad experience in the field of phraseology. The training sessions will now be briefly described (Section 3.1), and some more details on the final annotated corpus will be given (Section 3.2).

3.1 Training sessions

After receiving explanations about the guidelines and the annotation platform, all participants were asked to annotate the same part of the corpus: 500 sentences in all. At this first attempt, the degree of disagreement was considerably high among annotators, whose number of tags varied from 85 to 170 for the same sentences. The main reason for this was that two oposed positions were adopted: whereas some participants marked everything which showed any kind of similarity with VMWEs, others opted for annotating only the cases they were completely sure of.

All examples which caused disagreements were collected and classified, and three more sessions were organised, where participants tried to reach an agreement on the main problematic cases. A lot of the differently-annotated sentences were quite easy to decide on, as they were due to misunderstandings on basic concepts, either related to general language or to the guidelines. The rest of the cases, however, required further discussion. Decisions made on these cases were collected in an internal document for Basque annotators, so that they knew what criteria they should follow. Details about this document will be given in Section 4.

3.2 Final annotation and Inter-Annotator Agreement

After disagreements were discussed and decided on, each annotator was assigned some texts, and a small part of the corpus was double-annotated as a basis to calculate Inter-Annotator Agreement (IAA). This subcorpus was fully annotated by one participant, and was then split into two parts, so that two more annotators would work on one part each. Following the measurements of the first edition of the ST, the final IAA scores for Basque are sumed up in Table 1[7].

sent	inst-file1	inst-file2	mwe-fscore	kappa	kappa-cat
871	327	355	0.86	0.82	0.86

Table 1: IAA scores

As it can be noticed, scores are noteworthily high for all three measures. This is presumably an outcome of, on the one hand, the clarity of the guidelines and the specific tests provided, and on the other hand, the effectiveness of the training sessions held before starting the real annotation. Additionally, as a further step towards ensuring the unity of all annotations, consistency checks were performed once the main annotations were finished. Considering that before such checks these IAA scores were already much higher than average (comparing to the rest of the languages included in the ST), the good quality of this resource becomes evident beyond doubt.

The final annotated corpus comprises 3,823 VMWE tags of three categories in a total of 11,158 sentences. General data about the annotations is collected in Table 2, and further comments on them will be made in Section 6.

[7]Meaning of the table columns: sent = sentence; inst-file1 = instances annotated by one of the annotators; inst-file2 = instances annotated by the other two annotators; mwe-fscore = F score for MWEs; kappa = kappa score for VMWEs annotated; kappa-cat = kappa score for VMWE categories. More details on how scores were calculated are given in (Savary et al., 2018).

sentences	tokens	MWEs	LVC.cause	LVC.full	VID
11,158	157,807	3,823	183	2,866	774

Table 2: Data about the final Basque VMWE corpus

4 Difficult language-dependent cases

As pointed out previously, all the conclusions drawn from the training sessions were collected in an internal document for annotators. The main issues found during the annotation of Basque VMWEs will now be commented on, and the decisions made for each of the issues will be explained. Note that only general questions will be brought here. Individual cases which led to disagreements among annotators will not be included in this section, although a few examples of this kind were also collected.

4.1 Morphological variation of the nouns inside LVCs

In Basque, noun phrases almost always have a determiner, and there are hardly any instances of "bare" nouns (Laka, 1996), that is, nouns with no determiner at all. However, the presence of this kind of noun followed by a (usually light) verb seems to be a common characteristic among VMWEs. More specifically, it is frequent in VMWEs which denote very common actions, usually expressed by single verbs in other languages.

(7) ***lo egin***
sleep.BARE do
'(to) sleep', (ES) 'dormir', (FR) 'dormir'

(8) ***hitz egin***
word.BARE do
'(to) speak', (ES) 'hablar', (FR) 'parler'

While some of these VMWEs accept almost no morphological modification in the noun phrase, others are also used with determiners and modifiers, as the one shown in Examples (9)-(10). In these cases, the VMWEs display a canonical morphosyntactic variation.

(9) ***lan egin***
work.BARE do
'(to) work'

(10) ***lana egin***
work-the.ABS do
'(to) work, (to) do some work'

Morphological variants of this kind of LVC caused some trouble to annotators at the beginning, probably because only variants where the noun is "bare" are currently considered MWEs by Basque parsers (Alegria et al., 2004). Although it has sometimes been argued that instances with a determiner should not be treated as VMWEs, they pass all the LVC tests in the guidelines. Thus, our decision was to annotate these kinds of combinations both when they have some determiner and when they do not.

4.2 The future time in LVCs containing the verb *izan*

Izan 'have/be' is one of the most common verbs inside Basque LVCs, but it is also an auxiliary verb, which can be confusing for annotators sometimes. The usage of this verb is somewhat peculiar concerning the future form of LVCs. When we want to express that a given action will happen in the future, the verb participle is inflected by taking the morpheme *-ko/-go* at the end. However, this morpheme does not

always follow the verb when an LVC with *izan* is used: in many cases, it can also be attached to the noun inside the VMWE, eliding the verb.

(11) ***behar dut***
need.BARE have.1PS.PR
'I need'

(12) ***behar izango*** *dut*
need.BARE have-FUT AUX.1PS
'I will need'

(13) ***beharko*** *dut*
need-FUT AUX.1PS
'I will need'

Example (11) shows the VMWE *behar izan* '(to) need' in its present form, while the other examples show two variants of the future form. In Example (12), the *-go* morpheme is attached to the verb as usual, while in Example (13) the verb is elided, and the morpheme *-ko* is added to the noun *behar* instead[8]. Whereas the first two cases must be annotated, there is no VMWE in the third one, as only one lexicalised component is present, *behar*.

The fact that *izan* is also an auxiliary verb makes it easy to mistakenly think that the auxiliary after a word like *beharko* is a lexicalised component of the VMWE. However, this difference is an important detail annotators should always bear in mind. To see this difference, it can be helpful to use a morphological analyzer like Morfeus (Alegria et al., 1996), as it analyses *beharko* as an inflected form of *behar_izan*.

4.3 The blurred limit between adjectives and nouns in Basque VMWEs

All languages have words which can belong to more than one different part of speech. In some Basque VMWEs, it is not always clear if the non-verbal element is a noun or an adjective, and many parsers struggle to get the right tag. For instance, the word *gose* 'hunger/hungry' can be either one or the other depending on the context, even though its usage as an adjective is quite marginal nowadays. In Examples (14)-(15), two VMWEs containing this word and the verb *izan* 'be/have' are shown. Although intuition indicates us that *gose* is an adjective in Example (14) but a noun in (15), it is very common for parsers to tag both instances as nouns.

(14) ***gose naiz***
hungry/hunger.BARE be.1PS.PR
'I am hungry.'

(15) ***gosea dut***[9]
hunger-the.ABS have.1PS.PR
'I am hungry.'

Besides, sometimes, the usage of a word which always holds one category may even suggest that it belongs to a different part of speech within a VMWE. For instance, the first element in the expression *nahi izan* (wish.BARE have → '(to) want') can take the comparative suffix *-ago*, which is used to grade adjectives and adverbs: *nahiago izan* (wish-more have → '(to) prefer'). This usage may suggest that *nahi* is used as an adjective in this expression, even if it is always used as a noun out of it.

For coherence, it was concluded that these kinds of examples should all be grouped equally, and they were classified in the LVC categories. Given that the non-verbal element is sometimes closer to adjectives

[8]Note that *-ko* and *-go* are allomorphs of the same morpheme (due to phonemic context).

[9]Example (15) is probably a loan translation, as this is the way the idea of *being hungry* is expressed in Spanish and French, the main languages sharing territory with Basque. This usage is more recent and, according to some speakers, it is not as 'proper' as the first one. However, it is more and more common in real corpora and, thus, it must be considered.

than to nouns, it could be pertinent to add a note in the guidelines along with the one about Hindi, which states "the noun can be replaced by an adjective which is morphologically identical to an eventive noun". Exactly the same could be applied to Basque as well.

(16) ***bizi izan***
live/life be
'(to) live'

In fact, as the adjectives of this kind have identical nouns, combinations like the one in Example (16) pass LVC tests with no difficulty, and thus, this is the category they were assigned, regardless of their adjectival nature.

4.4 (Apparently) cranberry words inside LVCs

Some VMWEs which have reached us from a former stage of the language may present some idiosyncrasies from a diachronic perspective, e.g. the lack of determiners in noun phrases (see Section 4.1). They may also contain words which are only used within the context of a given verbal expression. For example, the word *merezi* is almost exclusively used as part of the VMWE *merezi izan* 'to deserve'.

Something similar occurs with *ari* in the verbal expression *ari izan*, which is categorised as a complex aspectual verb in Basque grammars (Etxepare, 2003). It is used in phrases such as *lanean ari izan* 'to be at work' and becomes grammaticalised when used to make the continuous forms of verbs, as in *jaten ari izan* 'to be eating'.

For the vast majority of Basque speakers, it is not a straight-forward assumption that these words are nouns. Nevertheless, if we take a look at the *Orotariko Euskal Hiztegia* (Mitxelena, 1987), the reference historical dictionary created by the Royal Academy of the Basque language, Euskaltzaindia[10], we realise that these words have an entry by themselves and are actually classified as nouns. Futhermore, while speakers might first think that these expressions do not pass test LVC.5, that is, that the verb can be ommitted when a possessive is added to the noun, some examples[11] of this kind can be found in the dictionary:

(17) *Eman diote (...)* ***bere merezia***.
give AUX.3PP (...) his/her deserved-the.ABS
'They gave him what he deserved.'

(18) *Ez zuen utzi* ***bere aria***.
not AUX.3PS leave his/her practice-the.ABS
'He did not stop doing what he was doing.'

To sum up, although some non-verbal elements in VMWEs might look like cranberry words, it is important to contrast information with reference material, especially when the verb is accompanied by a light verb. For the examples mentioned here, it was clear to us that LVC.full was the category where they fitted best.

5 Discussion on some conceptions in the guidelines

Overall, it is a remarkable point that the most controversial issues during the training sessions were all related to LVCs. This is probably an effect of the very high frequency of this type of VMWE in Basque corpora (more details will be given in Section 6), but it should also be considered that, as far as LVCs are concerned, there are notable differences between the guidelines and the rest of the literature on Basque (and Spanish) phraseology. Therefore, it is very likely that this fact has also conditioned the doubts arisen to participants.

It is an enormous challenge to create universal guidelines in a field like phraseology, where boundaries are never as definite as NLP tools would need. The guidelines created for both PARSEME Shared

[10] www.euskaltzaindia.eus

[11] For clarity, examples were re-written following current ortographical rules.

Tasks are a really important step towards unifying different conceptions about MWEs, and the clarity of tests simplifies the annotation task greatly. However, some points might still benefit from further consideration, which will be briefly noted here. If these points were problematic in other languages as well, the ideas presented in this section could be used as a starting point for future discussion.

Two main notions will be mentioned here related to the gap existent between the guidelines and our previous conceptions about phraseology: on the one hand, the understanding of collocations as a phenomenon separate from MWEs (Section 5.1), and on the other hand, the fact that LVCs are defined as combinations of a verb and a noun phrase only (Section 5.2).

5.1 Collocations as non-VMWEs

LVCs are usually understood as a subcategory of collocations in the reference literature about Basque phraseology (Urizar, 2012; Gurrutxaga and Alegria, 2013), as well as in that about Spanish phraseology (Corpas Pastor, 1997). However, in the guidelines, collocations are defined as a mere statistical phenomenon, and they are discriminated not only from LVCs but also from VMWEs in general. The line separating ones and others was not always clear, and despite the comprehensive tests, annotators sometimes found it hard not to annotate some instances which, according to them, were clearly related to phraseology somehow.

(19) ***deia egin***
call-the.ABS make
'(to) make a call'

(20) ***deia jaso***
call-the.ABS receive
'(to) receive a call'

For instance, the guidelines say that, whereas the combination in Example (19) must be annotated, the one in Example (20) must not. The fact that one passes all tests and the other one does not made it relatively easy to let the second example apart. However, it is still not that evident to us that it should not be treated as a VMWE at all, since the noun *deia* 'call' always chooses the verb *jaso* 'receive' to express that meaning. As a matter of fact, it is extremely rare to see it accompanied by other verbs which could equally express that meaning, such as *eduki* 'have'. Similar examples were found quite often in the corpus, so it might be worth examining those cases further for future editions.

5.2 LVCs accepting only noun phrases

On the other hand, according to the guidelines, LVCs can only be composed of a light verb and a noun phrase (except for Hindi, as it is pointed out in Section 4.3). This noun phrases can be preceded by prepositions or followed by postpositions. According to this, VMWEs like the one in Example (21) should not be annotated as LVC.full, as *korrika* is an adverb.

(21) ***korrika egin***
running.ADV do
'(to) run'

By definition, LVCs are VMWEs where the verb is void of meaning and the other component carries the whole semantic weight about the event or state the combination denotes. In Basque, many events can be expressed by adverbs, and this definition could equally be applied to constructions of adverbs and light verbs like the one in Example (21).

Furthermore, many of these adverbs are created by attaching a suffix to a noun, often *-ka*, such as *hazka* 'scratching', which comes from *hatz* 'finger' and forms part of the VMWE *hazka egin* (scratching do → '(to) scratch'). Thus, the LVC.full and LVC.cause categories would probably be more coherent if they had a wider scope and this kind of combination was also considered.

6 Information about Basque VMWEs inferred from annotations

As already mentioned, VMWEs from three different categories were annotated in Basque: VID, LVC.full and LVC.cause. Table 2 shows how many tags there are in the corpus, where the number of VMWEs annotated as LVC.full clearly stands out from the rest: 75% of all tags belong to this category. If we add the instances in the LVC.cause group to this number, the whole group of LVCs amounts to almost 80% of all annotations.

This is not surprising, since, as it is pointed out in Section 4.1, it is not strange that very common actions expressed by single verbs in some other languages are denoted by an LVC in Basque. Thus, it was to be expected that the number of instances in this category would be higher in our corpus than in other languages.

Table 3 makes this fact obvious. It collects the ratio of LVCs and VMWEs per sentence in the Basque corpus, as well as the average ratio of the whole ST corpus (20 languages in all) and the ratios for Spanish, French and English corpora[12], the three languages which affect Basque the most. In order to make comparisons properly, only the three universal categories were taken into account, even if all except Basque include other categories as well. From the languages included in the ST, only Farsi and Hindi have a higher number of LVCs per 100 sentences (95 and 40 respectively).

	VMWEs per 100 sentences	**LVCs per 100 sentences**
Basque	34	27
Average	18	11
French	20	9
Spanish	15	9
English	6	4

Table 3: Average frequencies of tags in Basque, Spanish, French and English

On the other hand, the number of instances annotated as LVC.cause is very low (less than 5% of all tags), and this seems to be quite a common tendency also in other languages. Considering only annotations from the three universal categories, the average percentage of VMWEs classified in this group is only 3% (taking all 20 languages into account). This might be a sign that either: (A) the LVC.cause category would be better merged with the LVC.full one, or (B) maybe it would be a good idea to broaden this category so that it includes combinations that are not yet annotated, such as collocations.

Concerning morphology, the VMWEs in the Basque corpus are mostly combinations of a verb and a noun (94%)[13], which was easy to anticipate considering that LVCs can only be of this kind according to the guidelines. Consistent with other work about VMWEs in dictionaries (Inurrieta et al., 2017), such nouns are mainly found in the absolutive case (85%) in the corpus, and among the rest, the locative is the most frequent postposition, as in Example (22).

(22) ***jolasean ibili***
game-the.LOC be
'(to) be playing, (to) play'

Something comparable probably happens in other languages as well. In the Spanish corpus, for example, out of the VMWEs where the main constituents are a verb and a noun, only 23% include a preposition.

7 Conclusion

VMWEs were annotated in a 11,158-sentence Basque corpus, following the universal guidelines of edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In

[12]Corpora for all languages can be accessed here: `https://gitlab.com/parseme/sharedtask-data/tree/master/1.1`

[13]When calculating this number, non-verbal elements of LVCs which could be either a noun or an adjective (see Section 4.3) were counted as nouns.

all, 3,823 instances were annotated and classified into two main categories: Verbal Idioms and Light Verb Constructions. High Inter-Annotator Agreement scores make it evident that this is a very good-quality resource, which can be useful not only for NLP-related research, but also for future studies on Basque phraseology.

After explaining how the annotation process was organised, the main doubts arisen to Basque annotators while performing this task were commented on in this paper. The decisions taken on language-dependent issues were presented, and some ideas for discussion on the universal guidelines were also proposed. If these ideas are shared by annotators from other languages, it could be interesting to take a further look at them for future editions.

References

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Díaz de Ilarraza, Koldo Gojenola and Larraitz Uria. 2015. Automatic conversion of the Basque dependency treebank to universal dependencies. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT 2015)*, 233–241.

Iñaki Alegria, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. In *Literary and Linguistic Computing*, 11(4):193–203.

Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola and Ruben Urizar. 2004. Representation and treatment of Multiword Expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 48–55. Association for Computational Linguistics.

Gloria Corpas Pastor. 1997. Manual de fraseología española. Editorial Gredos.

Ricardo Etxepare. 2003. Valency and argument structure in the Basque verb. In Jose Ignacio Hualde and Jon Ortiz de Urbina (eds.) *A grammar of Basque*. Mouton de Gruyter.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world*, 2–7. Association for Computational Linguistics.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, 116–125. University of the Basque Country.

Uxoa Inurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka and Kepa Sarasola. 2017. Rule-based translation of Spanish Verb-Noun combinations into Basque. In *Proceedings of the 13th Workshop on Multiword Expressions, in EACL 2017*, 149–154. Association for Computational Linguistics.

Uxoa Inurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka and Kepa Sarasola. 2018 (in print). Analysing linguistic information about word combinations for a Spanish-Basque rule-based machine translation system. In Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor and Violeta Seretan (eds.), *Multiword Units in Machine Translation and Translation Technologies*, 39–60. John Benjamins publishing company.

Koldo Mitxelena. 1987. *Orotariko Euskal Hiztegia*. Euskaltzaindia, the Royal Academy of the Basque language.

Itziar Laka Mugarza. 1996. *A brief grammar of Euskera, the Basque language*. University of the Basque Country.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann and Johanna Monti. 2016. PARSEME survey on MWE resources. In *9th International Conference on Language Resources and Evaluation (LREC 2016)*, 2299–2306. European Association for Language Resources.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: a pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, 1–15. Springer.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, and others. 2015. PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova and others. 2017. The PARSEME Shared Task on automatic identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions, in EACL 2017*, 31–47. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Veronika Vincze and others. 2018. Edition 1.1 of the PARSEME Shared Task on automatic identification of Verbal Multiword Expressions. In *Proceedings of the 14th Workshop on Multiword Expressions, in COLING 2018*. Association for Computational Linguistics.

Ruben Urizar. 2012. *Euskal lokuzioen tratamendu konputazionala*. University of the Basque Country.

Igone Zabala Unzalu. 2004. Los predicados complejos en vasco. In *Las fronteras de la composicin en lenguas romnicas y en vasco*, 445–534. Universidad de Deusto.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 112 APRIL 2019 5–54

Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir

Agata Savary,[a] Silvio Ricardo Cordeiro,[b] Timm Lichte,[c] Carlos Ramisch,[d] Uxoa Iñurrieta,[e] Voula Giouli[f]

[a] University of Tours, France
[b] Paris-Diderot University, France
[c] University of Tübingen, Germany
[d] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[e] University of the Basque Country, Spain
[f] Athena Research Center, Greece

Abstract

Multiword expressions can have both idiomatic and literal occurrences. For instance *pulling strings* can be understood either as making use of one's influence, or literally. Distinguishing these two cases has been addressed in linguistics and psycholinguistics studies, and is also considered one of the major challenges in MWE processing. We suggest that literal occurrences should be considered in both semantic and syntactic terms, which motivates their study in a treebank. We propose heuristics to automatically pre-identify candidate sentences that might contain literal occurrences of verbal VMWEs, and we apply them to existing treebanks in five typologically different languages: Basque, German, Greek, Polish and Portuguese. We also perform a linguistic study of the literal occurrences extracted by the different heuristics. The results suggest that literal occurrences constitute a rare phenomenon. We also identify some properties that may distinguish them from their idiomatic counterparts. This article is a largely extended version of Savary and Cordeiro (2018).

1. Introduction

A multiword expression (MWE) is a combination of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies (Baldwin and Kim, 2010). MWEs encompass diverse linguistic objects such as idioms (*to **pull** the*

 Corresponding author: agata.savary@univ-tours.fr
Cite as: Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, Voula Giouli. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. The Prague Bulletin of Mathematical Linguistics No. 112, 2019, pp. 5–54. doi: 10.2478/pralin-2019-0001.

strings 'make use of one's influence to gain an advantage'), compounds (*a* ***hot dog***), light-verb constructions (*to* ***pay*** *a* ***visit***), rhetorical figures (***as busy as a bee***), institutionalized phrases (***traffic light***) and multiword named entities (***European Central Bank***). A prominent feature of many MWEs, especially of verbal idioms such as *to* ***pull*** *the* ***strings***, is their non-compositional semantics, that is, the fact that their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a way deemed regular for the given language. For this reason, MWEs pose special challenges both to linguistic modeling (e.g. as linguistic objects crossing boundaries between lexicon and grammar) and to natural language processing (NLP) applications, especially to those which rely on semantic interpretation of text (e.g. information retrieval, information extraction or machine translation).

Another outstanding property of many MWEs, as illustrated in Example (1), is that we can encounter their literally understood counterparts, as in (2).

(1) The boss was **pulling** the **strings** from prison. (EN)
'The boss was making use of his influence while in prison.'

(2) You control the marionette by pulling the strings. (EN)

This phenomenon, also called *literal-idiomatic ambiguity* (Savary et al., 2018), has been addressed in linguistic and psycholinguistic literature, and is considered a major challenge in MWE-oriented NLP tasks (Constant et al., 2017), as will be discussed in Section 10. Despite this considerable attention received from the scientific community, the notion of literal occurrence has rarely been formally defined. It is, thus, often unclear whether uses such as the following should be regarded as literal occurrences:

- "Coincidental" co-occurrences of components of a given MWE or of their homographs, as in Examples (3) and (4) respectively,[1]

  (3) As an effect of pulling, the strings broke. (EN)

  (4) He strings paper lanterns on trees without pulling the table. (EN)

- Variants, like (5), (6), (7) and (8), which change the syntactic dependencies between the components, as compared to (1),

  (5) Determine the maximum force you can pull on the string so that the string does not break. (EN)

  (6) My husband says no **strings** were **pulled** for him. (EN)

  (7) She moved Bill by **pulling** wires and **strings**. (EN)

[1] See below for an explanation of the different styles of highlighting and underlining used in this article.

(8) The article addresses the **strings** which the journalist claimed that the senator **pulled**. (EN)

- Co-occurrences exhibiting substantial changes in semantic roles, as in (9),

(9) The strings pulled the bridge. (EN)

- Uses like (10), where idiomatic and literal meanings are wittingly combined.

(10) He was there, **pulling** the **strings**, literally and metaphorically. (EN)

In this article, we put forward a definition of a literal occurrence which is not only semantically but also syntactically motivated. Intuitively, for a given MWE e with components $e_1, \ldots, e_n$, we conceive a *literal occurrence* (LO) of e as a co-occurrence e' of words $e'_1, \ldots, e'_n$ fulfilling the following conditions:

1. $e'_1, \ldots, e'_n$ can be attributed the same lemmas and parts of speech as $e_1, \ldots, e_n$.
2. The syntactic dependencies between $e'_1, \ldots, e'_n$ are the same or equivalent to those between $e_1, \ldots, e_n$ in a canonical form of e.[2]
3. e' is not an idiomatic occurrence of a MWE

When Conditions 1 and 3 are fulfilled but Condition 2 is not, we will speak of a *co-incidental occurrence* (CO) of e. Formal definitions of these conditions and notions will be provided in Section 2. What we eventually want to capture is that only Example (2) above is considered an LO. Examples (3), (5) and (9) are COs since they do not fulfill Condition 2. Examples (1), (6), (7), (8) and (10) do not fulfill Condition 3, since they are *idiomatic occurrences* (IOs). Finally, Example (4) is considered out of scope (not an IO, an LO or a CO), since it involves a lemma (*string*) with a different part of speech than the the MWE e, and therefore does not fulfill Condition 1. Because of Condition 2, the study of literal occurrences of MWEs is best carried out when explicit syntactic annotation is available, that is, in a treebank.

Assuming the above understanding of LOs as opposed to IOs and COs, this article focuses on verbal MWEs (VMWEs), which exhibit particularly frequent discontinuity, as well as syntactic ambiguity and flexibility (Savary et al., 2018). Henceforth, we use wavy and dashed underlining for LOs and COs, respectively. Straight underlining denotes emphasis. Lexicalized components of MWEs are shown in **bold**. Section 2.4 provides more details on the notation of examples used in this article.

We propose to study two main research questions. Firstly, we wish to quantify the LO phenomenon, that is, to estimate the relative frequency of LOs with respect to IOs

[2] As formally defined in Section 2, a canonical form of a VMWE is one of its least marked syntactic forms preserving the idiomatic meaning. A form with a finite verb is less marked than one with an infinitive or a participle, the active voice is less marked than the passive, etc. For instance, a canonical form of (1) is *the boss **pulled strings***. Dependencies are equivalent if the syntactic variation can be neutralized while preserving the overall meaning. For instance, (8) can be reformulated into *The journalist claimed that the senator **pulled** the **strings**, and this article addresses them.*

and COs, as well as the distribution of this frequency across different VMWE types and categories. Secondly, we are interested in cross-lingual aspects of LOs. To this aim, we focus on five languages from different language genera:[3] Basque (Basque genus), German (Germanic genus), Greek (Greek genus), Polish (Slavic genus) and Portuguese (Romance genus). We try to discover possible cross-lingual reasons that may favour the use of LOs, and, conversely, those reasons which are language specific.

The contributions of these efforts are manifold. We provide a normalized and cross-lingual terminology concerning the LO phenomenon. We pave the way towards a better understanding of the nature of ambiguity in VMWEs. We show that ambiguity between an idiomatic and a literal occurrence of a sequence is a challenge in MWE processing which is qualitatively major but quantitatively minor. We put forward recommendations for linguistically informed methods to automatically discover LOs in text. Last but not least, we provide an annotated corpus of positive and negative examples of LOs in five languages. It is distributed under open licenses and should be useful for linguistic studies, for example, on idiom transparency or figurativeness, as well as for data-driven NLP methods, for example, on MWE identification (Savary et al., 2017; Ramisch et al., 2018) or compositionality prediction (Cordeiro et al., 2019).

The article is organized as follows. We provide the necessary definitions, and in particular we formalize the notions of LOs and COs (Section 2). We exploit an existing multilingual corpus in which VMWE annotations are accompanied by morphological and dependency annotations, but literal occurrences are not tagged (Section 3). We propose heuristics to automatically detect possible LOs of known, that is, manually annotated, VMWEs (Section 4). We manually categorize the resulting occurrences using a typology which accounts for true and false positives, as well as for linguistic properties of LOs as opposed to those of IOs (Section 5). We report on the results in the five languages under study (Section 6), discussing characteristics of LOs (Section 7), of COs (Section 8) and of erroneous occurrences (Section 9). Finally, we present related work (Section 10), draw conclusions and discuss future work (Section 11).

This work is a considerably extended version of Savary and Cordeiro (2018). Compared to the previous article, we expanded our scope to five languages instead of one (Polish). We enhanced and formalized the definition of LOs. We enlarged the annotation typology and designed unified annotation guidelines, which were then used by native annotators to tag LOs, COs and annotation errors in their native languages. Finally, we produced results of both the automatic and the manual annotation for the five languages under study. Thanks to these extensions, the conclusions have a broader significance than in our previous work.

[3]The genus for each language is indicated according to the WALS (Dryer and Haspelmath, 2013).

2. Definitions and notations

In this section we formalize the nomenclature related to sequences and dependency graphs, and we summarize basic definitions concerning VMWEs and their components, adopted from previous work. We also formally define the central notions which are required in this work: VMWE tokens, variants and types, as well as idiomatic, literal and coincidental occurrences. Finally, we explain the notational conventions used throughout this article to gloss and translate multilingual examples.

2.1. Sequences, subsequences, graphs, subgraphs and coarse syntactic structures

Each *sequence* of word forms is a function $s : \{1, 2, \ldots, |s|\} \to W$, where the domain contains all integers between 1 and $|s|$, and W is the set of all possible word forms (including punctuation). A sequence s can be noted as $s := \{s_1, s_2, \ldots, s_{|s|}\}$, where $s_i := (i, w_i)$ is a single *token*. In other words, a sequence can be denoted as a set of pairs: $s = \{(1, w_1), (2, w_2), \ldots, (|s|, w_{|s|})\}$. For example, the sentence in Example (6), whose morphosyntactic annotation is shown in Figure 1(b), can be represented as a sequence $s = \{(1, \text{My}), (2, \text{husband}), (3, \text{says}), \ldots, (9, \text{him}), (10, .)\}$. Sequences can be seen as perfectly tokenized sentences, because they ignore orthographic conventions regarding spaces between word forms (e.g. before commas), compounding (e.g. *snowman* counts as two word forms), contractions (e.g. *don't* counts as two word forms), etc.

A sentence is a particular sequence of word forms for which the corpus used in our study provides lemmas, morphological features, dependency relations and VMWE annotations. For a given token $s_i = (i, w_i)$, let $\text{surface}(s_i)$, $\text{lemma}(s_i)$ and $\text{pos}(s_i)$ be its surface form, lemma and part of speech.[4] Consider Figure 1, which shows simplified morphosyntactic annotations of Examples (1), (6) and (7) from page 6. In Figure 1(a), $\text{surface}(s_6) = \textit{strings}$ and $\text{lemma}(s_6) = \textit{string}$.

A *dependency graph* for a sentence s is a tuple $\langle V_s, E_s \rangle$, where $V_s = \{\langle 1, \text{surface}(s_1), \text{lemma}(s_1), \text{pos}(s_1)\rangle, \ldots, \langle |s|, \text{surface}(s_{|s|}), \text{lemma}(s_{|s|}), \text{pos}(s_{|s|})\rangle\}$ and E_s is the set of labeled edges connecting nodes in V_s. For instance, Figure 1(a) shows a graphical representation of the dependency graph of sentence (1). Each token s_i of s is associated in the dependency graph with its parent, denoted as $\text{parent}(s_i)$, through a syntactic label, denoted as $\text{label}(s_i)$. Some tokens may have parent nil (and label root). In Figure 1(a), $\text{label}(s_2) = \text{nsubj}$, $\text{parent}(s_2) = s_4$, $\text{label}(s_4) = \text{root}$, and $\text{parent}(s_4) = \text{nil}$.

Given two sequences p and q over the same word forms, p is a *subsequence* of q iff there is an injection $\text{sub}_p^q : \{1, 2, \ldots, |p|\} \to \{1, 2, \ldots, |q|\}$, such that: (i) word forms are preserved, that is, for $i \in \{1, 2, \ldots |p|\}$, the condition $p(i) = q(\text{sub}_p^q(i))$ holds; and (ii) order is preserved, that is, for $i, j \in \{1, 2, \ldots |p|\}$, if $i < j$, then $\text{sub}_p^q(i) < \text{sub}_p^q(j)$. Thus, every subsequence is a sequence, and the definitions of lemmas, parts of speech and

[4] Morphological features are not used in our formalization of LOs and are further ignored, although they could be useful to improve our treatment of agglutinative languages like Basque in the future.

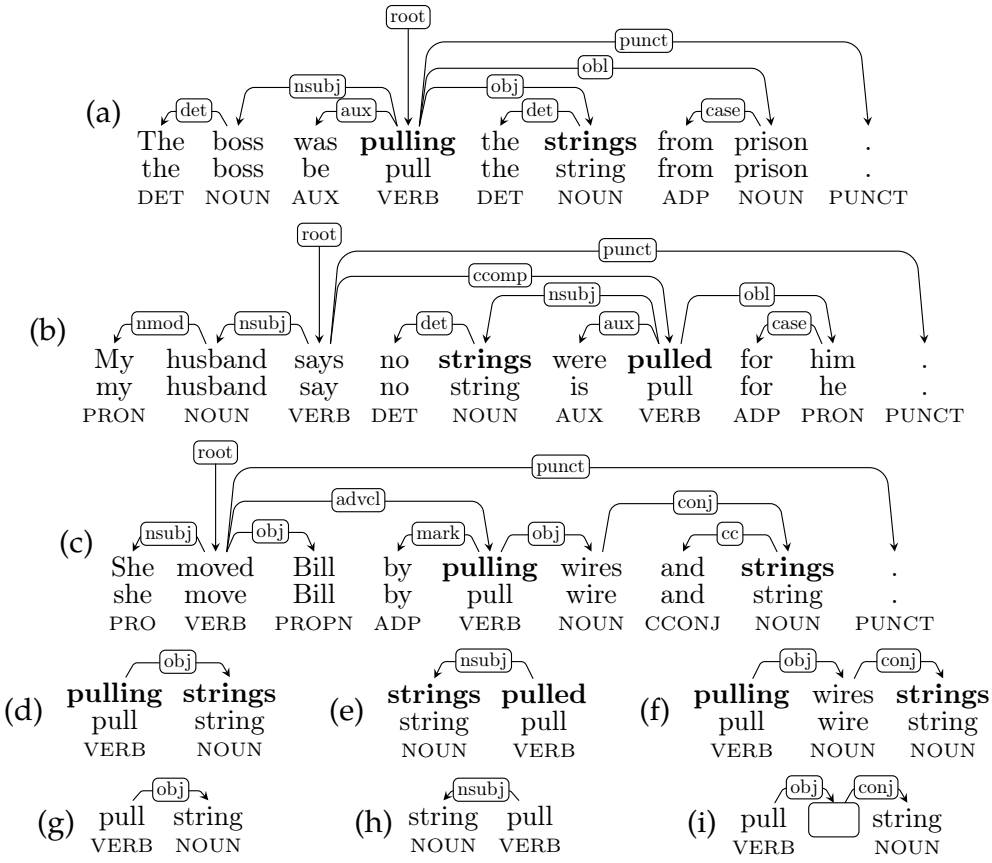


Figure 1. Dependency graphs (a-b-c) for the sentences in Examples (1), (6) and (7), the dependency subgraphs (d-e-f) corresponding to the VMWE tokens in bold, and the coarse syntactic structures (g-h-i) of these tokens. All examples use Universal Dependencies v2.

surface forms of sequence tokens apply straightforwardly to subsequence tokens. For instance, in Figure 1(a), the subsequence corresponding to the tokens in bold can be formalized as $p = \{p_1, p_2\} = \{(1, \text{pulling}), (2, \text{strings})\}$ and $\text{sub}^s_p(1) = 4, \text{sub}^s_p(2) = 6$. We also have $\text{lemma}(p_2) = \text{lemma}((\text{sub}^s_p(2), \text{strings})) = \text{lemma}(s_6) = \textit{string}$, etc.

A subsequence p of a sentence s defines a *dependency subgraph* $\langle V_p, E_p \rangle$ as a minimal weakly connected graph[5] containing at least the nodes corresponding to the tokens in p. In other words, only those edges from $\langle V_s, E_s \rangle$ are kept in $\langle V_p, E_p \rangle$ which appear in the dependency chains connecting the elements of p. If nodes not belonging to p appear in these chains, they are kept in the dependency subgraph for the sake of connectivity. Such nodes are called *intervening nodes*. For instance, Figures 1(d-e-f) show

[5]A directed graph is weakly connected if there is a path between every pair of vertices when the directions of edges are disregarded.

the dependency subgraphs corresponding to two-token subsequences (highlighted in bold) from the sentence graphs from Figures 1(a-b-c). Note that Figure 1(f) corresponds to a subsequence with words *pulling* and *strings* only but its subgraph also contains the intervening node for *wires*.

In a dependency subgraph of a subsequence p we can further abstract away from surface forms and their positions in the sentence, as well as from intervening nodes. In this way, we obtain the *coarse syntactic structure (CSS)* of p. Formally, if p contains k intervening nodes, then $\mathrm{css}(p) = \langle V_{\mathrm{css}(p)}, E_{\mathrm{css}(p)} \rangle$ is a directed graph where $V_{\mathrm{css}(p)} = \{\langle \_, \_, \mathrm{lemma}(p_1), \mathrm{pos}(p_1) \rangle, \ldots, \langle \_, \_, \mathrm{lemma}(p_{|p|})), \mathrm{pos}(p_p)) \rangle\}_{ms} \cup \{dummy_1, \ldots, dummy_k\}$, ms denotes a multiset, and $dummy_i$ are dummy nodes replacing the intervening words.[6] All dependency arcs from E_p are reproduced in $E_{\mathrm{css}(p)}$. Figures 1 (g-h-i) show the CSSes of the subsequences highlighted in bold in Figures 1 (a-b-c).

In a subsequence p, the definition of a parent still relies on the dependencies in the underlying sentence s, but is restricted to the tokens in p. Formally, for a given $1 \leqslant i \leqslant |p|$ and $k = \mathrm{sub}_p^s(i)$, if there exists $1 \leqslant j \leqslant |p|$ and $l = \mathrm{sub}_p^s(j)$ such that $\mathrm{parent}(s_k) = s_l$, then $\mathrm{parent}_p^s(p_i) := p_j$. Otherwise $\mathrm{parent}_p^s(p_i) :=$ nil. For instance, in Figure 1(a), if we take $p = \{p_1, p_2\} = \{(1, \mathrm{pulling}), (2, \mathrm{strings})\}$ and $\mathrm{sub}_p^s(1) = 4, \mathrm{sub}_p^s(2) = 6$, then $\mathrm{parent}_p^s(p_1) =$ nil and $\mathrm{parent}_p^s(p_2) = p_1$.

Note that, in Figure 1(c), where the subsequence *pulling strings* forms a non connected graph, the parents of both components are nil, that is, taking $\mathrm{sub}_p^s(1) = 5$ and $\mathrm{sub}_p^s(2) = 8$, we have $\mathrm{parent}_p^s(p_1) = \mathrm{parent}_p^s(p_2) =$ nil, although *strings* is dominated by *wires* in the dependency subgraph in Figure 1(f).

2.2. VMWE occurrences, variants and types

Concerning VMWEs, we adapt and extend the PARSEME corpus definitions from (Savary et al., 2018). Namely, if a sentence s is a sequence of syntactic words (i.e., elementary units linked through syntactic relations), then a *VMWE occurrence* (*VMWE token*) e in s is a subsequence of s (in the sense defined in Section 2.1) of length higher than one[7] which fulfills four conditions.

First, all components $e_1, \ldots, e_n$ of e must be *lexicalized*, that is, replacing them by semantically related words usually results in a meaning shift which goes beyond what is expected from the replacement. For instance, replacing *pulling* or *strings* in Example (1) by their synonyms *yanking* or *ropes*, respectively, leads to the loss of the idiomatic meaning: the sentence no longer alludes to using one's influence. Conversely, the determiner *the* can be interchanged with *some*, *many*, etc. with no harm to the idiomatic meaning. Therefore, *pulling* and *string* are lexicalized in (1) but *the* is not.

[6] The first two empty slots denote unspecified positions and surface forms.

[7] The PARSEME guidelines assume the existence of multiword tokens, some of which can be VMWEs, e.g. (DE) *aus-machen* 'out-make'⇒'open'. They consist of at least two words which occur as single tokens due to imperfect tokenization. Our definition of sequences excludes multiword tokens.

Second, the head of each of e's *canonical forms* must be a verb v. A canonical form of a VMWE is one of its least marked syntactic forms preserving the idiomatic meaning. A form with a finite verb is less marked than one with an infinitive or a participle, a non-negated form is less marked than a negated one, the active voice is less marked than the passive, a form with an extraction is more marked than without, etc. For most VMWEs, the canonical forms are equivalent to the so-called *prototypical verbal phrases*, that is, minimal sentences in which the head verb v occurs in a finite non-negated form and all its arguments are in singular and realized with no extraction. For some VMWEs, however, the prototypical verbal phrase does not preserve the idiomatic meaning, and then the canonical forms can be, for example, with nominal arguments in plural. This is the case in Example (11), which shows a canonical form of the VMWE occurrences from Examples (1), (6) and (7)[8], with a direct object in plural (for brevity, subjects are replaced by *he*).

(11) he **pulled** the **strings** (EN)

Other examples of canonical forms which are not prototypical verbal phrases include passivized phrases, as in (EN) ***the die is cast*** 'the point of no retreat has been passed' vs. (EN) *someone cast the die*.

Third, all lexicalized components other than v in a canonical form of e must form phrases which are syntactically directly dependent on v. In other words, $e_1, \ldots, e_n$ and the dependency arcs which connect them in s must form a weakly connected graph. This condition heavily depends on a particular view on syntax and, more specifically, on representing dependency relations. In this article, we follow the conventions established by the Universal Dependencies (UD) initiative (Nivre et al., 2016), which assume, in particular, that syntactic relations hold between content words, and function words depend on the content words which they specify. One of the consequences of this stance is that inherently adpositional verbs, composed of a verb and a selected preposition such as *rely on*, do not form connected graphs (the preposition is a *case* marker of the verb's object). Therefore, they are not considered VMWEs.

Finally, e in s must have an idiomatic meaning, that is, a meaning which cannot be deduced from the meanings of its components in a way deemed regular for the given language.[9] Semantic idiomaticity is hard to estimate directly, but has been approximated by lexical and syntactic tests defined in the PARSEME annotation guidelines (version 1.1).[10] These tests are applied to a canonical form of any VMWE candidate.

[8] As well as from Examples (8) and (10), which are further neglected.

[9] Morphological and/or syntactic idiomaticity of MWEs is also mentioned by some works. However, it implies semantic idiomaticity, because regular rules concern regular structures only. Thus, if an MWE is morphologically or syntactically irregular, its meaning cannot be derived by regular rules.

[10] `http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/`

Recall that a VMWE token e is a subsequence of a sentence s and is associated with a CSS $\text{css}(e) = \langle V_{\text{css}(e)}, E_{\text{css}(e)} \rangle$, as shown in Figures 1 (g-h-i).[11] We define a VMWE *syntactic variant*, or *variant* for short, ν as a set of all VMWE occurrences having the same CSS and the same meaning. Formally, let $\sigma_{\text{ID}}(e)$ be the idiomatic meaning contributed by the VMWE token e in sentence s. Then, the VMWE variant associated with e is defined as $\nu(e) := \{ e' \mid \text{css}(e') = \text{css}(e), \sigma_{\text{ID}}(e') = \sigma_{\text{ID}}(e) \}$. Note that VMWE variants as such are not ambiguous: they always come with one meaning. What can be ambiguous, however, is their CSS. For instance, the CSS in Figure 1(g) can have both the idiomatic meaning conveyed in Example (1) and a literal meaning, present in Example (2). Different VMWE occurrences may correspond to the same variant. For instance, the VMWE token from Example (1) and its canonical form in (11) correspond to the variant whose CSS is shown in Figure 1(g).

Finally, collections of VMWE variants form *VMWE types*. Formally, a *VMWE type*, or a *VMWE* for short, is an *equivalence class* of all VMWE variants having the same component lemmas and parts of speech, and the same idiomatic meaning. For each such equivalence class, its *canonical variant* is the variant stemming from its canonical forms, as defined above. The CSS of this canonical representative is called the *canonical structure* of the VMWE. For instance, Figure 1(g) contains the canonical structure of the VMWE type whose occurrences are highlighted in bold in Figures 1(a-c).

2.3. Idiomatic, literal and coincidental occurrences

Given the definitions from the previous section, consider a VMWE type t with n components and $|t|$ variants. Formally, $t = \{\langle \text{css}_1, \sigma_{\text{ID}} \rangle, \langle \text{css}_2, \sigma_{\text{ID}} \rangle, \ldots, \langle \text{css}_{|t|}, \sigma_{\text{ID}} \rangle\}$, and $\text{css}_i = \langle V, E_i \rangle$, where $V = \{\langle \_, \_, \text{lemma}_1, \text{pos}_1 \rangle, \ldots, \langle \_, \_, \text{lemma}_n, \text{pos}_n \rangle\}_{\text{ms}}$. Let s be a sentence of length $|s|$. A *potential occurrence* p of t in s is defined as a subsequence of s whose lemmas and parts of speech are those in (any of the CSSes of) t. Formally, p is a subsequence of length n of s (in the sense of the definitions in Section 2.1) and $\{\langle \_, \_, \text{lemma}(p_1), \text{pos}(p_1) \rangle, \ldots, \langle \_, \_, \text{lemma}(p_n), \text{pos}(p_n) \rangle\}_{\text{ms}} = V$.

Then, we assume the following definitions:

- p is an *idiomatic reading occurrence*, or *idiomatic occurrence* (IO) for short, of t iff
  - The CSS of p is identical to one of the CSSes in t.
  - p occurs with the meaning σ_{ID}, or with any other idiomatic meaning[12].
- p is a *literal reading occurrence*, or *literal occurrence* (LO) for short, of t iff

[11] Since $\text{css}(e)$ only specifies the lemmas of e's components, it might lack morphosyntactic constraints associated with e, e.g., the nominal object must be plural in *pull strings*. This motivates the annotation categories LITERAL-MORPH and LITERAL-SYNT presented in Section 5.

[12] This alternative condition covers cases of VMWE variants with the same CSS but different idiomatic meanings, for instance (EN) *to* ***take in*** 'to make a piece of clothing tighter', (EN) *to* ***take in*** 'to include something', (EN) *to* ***take in*** 'to remember something that you hear', etc. Note that, in this case, even if p is an idiomatic occurrence of t, it does not belong to any of t's variants, because of its different meaning. In other words, an IO of t is not necessarily an occurrence of t. It is rather an IO of t's CSS.

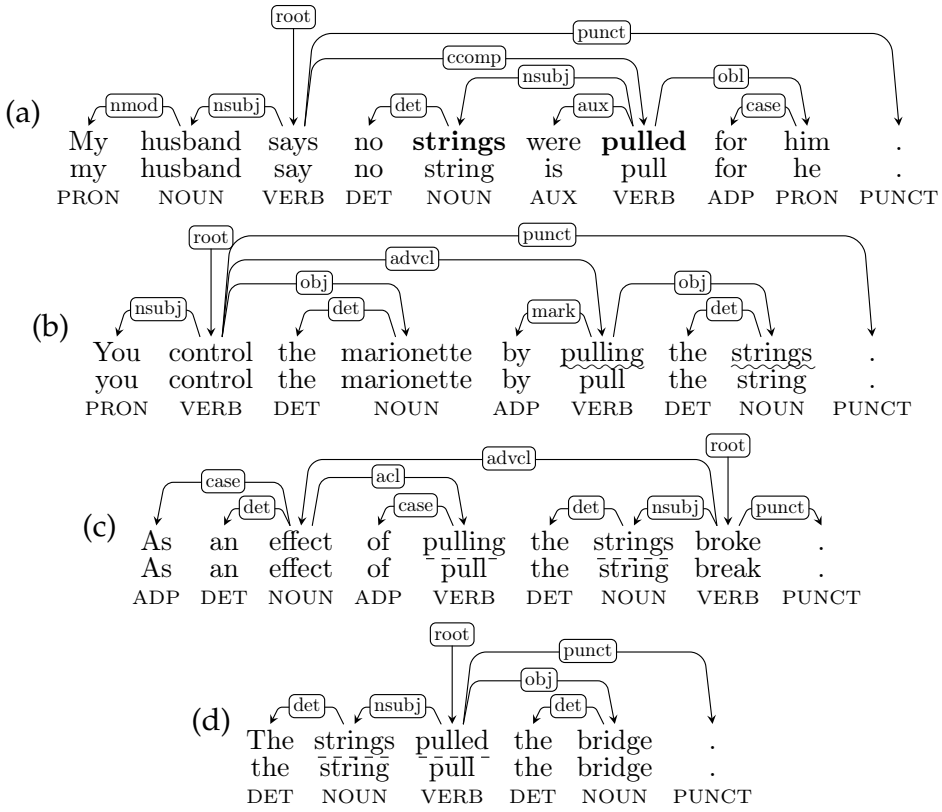


*Figure 2. Morphosyntactic annotations (disregarding morphological features) for occurrence contexts of the VMWE (EN) **pull strings**: (a) idiomatic occurrence, (b) literal occurrence, (c–d) coincidental occurrences.*

  - There is a rephrasing s' of s (possibly identical) such that: (i) s' is synonymous with s, (ii) there is a subsequence p' in s' such that the CSSes of p and p' have identical sets of vertexes ($V_{css(p)} = V_{css(p')}$), (iii) the CSS of p' is equal to the canonical structure of t.
  - p occurs with no idiomatic meaning (i.e not with the meaning σ_{ID} in particular), or it is a proper subsequence of a longer VMWE occurrence[13].
- p is a *coincidental occurrence* (CO) of t iff
  - there is no rephrasing s' of s which fulfills conditions (i–iii) describing an LO above.

For instance, consider the VMWE type t with the three variants whose CSSes are shown in Figure 1(g-h-i), and whose meaning is σ_{ID} = 'to make use of one's influ-

[13]This alternative condition covers cases like (EN) *He **pulled the string*** 'In baseball, he threw a pitch that broke sharply', which has one more lexicalized component (*the*) than the VMWE tokens in Figures 1(a-b-c).

ence'. Then, t occurs idiomatically, literally and coincidentally in the sentences from Figure 2(a), (b) and (c–d), respectively. In particular, the CO in Figure 2(d) has the same CSS as the IO in Figure 2(a). Still, the former is not an LO, since it cannot be rephrased in such a way that *strings* becomes the direct object of *pulling*, which is required in the canonical structure of t.

2.4. Notations for multilingual examples

Multilingual aspects of VMWEs addressed in this article are illustrated with examples which follow the notational conventions put forward in Markantonatou et al. (2018). A numbered example like (12) contains a sample VMWE in the original script followed by an ISO 639-1 language code,[14] a transcription (if any), a gloss, as well as a literal and an idiomatic translation. The inline version of the same example is: (EL) κάτι τέτοιο θα **ανοίξει την πόρτα** σ τη διαφθορά (kati tetio tha anixi tin porta s ti diaphthora) 'this will open the door to corruption'⇒'this will enable corruption'. The transliteration and the literal or idiomatic translations may sometimes be omitted for the sake of brevity or focus, as in (EL) κάτι τέτοιο θα **ανοίξει την πόρτα** σ τη διαφθορά 'this will open the door to corruption'.

(12) Κάτι τέτοιο θα **ανοίξει την πόρτα** στη διαφθορά. (EL)
Kati tetio tha anixi tin porta sti diafthora.
something such will open the door to-the corruption
This will open the door to corruption. 'This will enable corruption.'

These conventions also determine that segmentable morphemes are separated by a hyphen, as in the detachable verb-particle construction *ab-gesteckt* 'off-stuck' in Example (13), while one-to-many correspondences between the example and the gloss are marked by dots, as for *vom* 'by.the.DAT' in the same example.

(13) Der Rahmen für diese Verhandlungen soll vom Minister-rat
The framework for these negotiations should by.the.DAT Minister-council
ab-gesteckt werden. (DE)
off-stuck be.
The framework for these negotiations should be stuck off by the Council of Ministers. 'The framework for these negotiations should be set by the Council of Ministers.'

3. Corpus

We use the openly available PARSEME corpus, annotated for VMWEs in 19 languages (Savary et al., 2018; Ramisch et al., 2018).[15] Among its five major VMWE cat-

[14] DE for German, EL for Greek, EU for Basque, PL for Polish and PT for Portuguese

[15] Downloadable from the LINDAT/CLARIN infrastructure at: `http://hdl.handle.net/11372/LRT-2842`

egories, four are relevant to this study, dedicated to Basque, German, Greek, Polish and Portuguese:

- *Inherently reflexive verbs* (IRV) are pervasive in Romance and Slavic languages, present in German, but absent or rare in English or Greek. An IRV is a combination of a verb V and a reflexive clitic RCLI,[16] such that one of the 3 non-compositionality conditions holds: (i) V never occurs without RCLI, as is the case for the VMWE in (14); (ii) RCLI distinctly changes the meaning of V, like in (15); (iii) RCLI changes the subcategorization frame of V, like in (16) as opposed to (17). IRVs are semantically non-compositional in the sense that the RCLI does not correspond to any semantic role of V's dependents.

  (14) O aluno **se** **queixa** do professor. (PT)
  The student RCLI complains of.the teacher.
  'The student complains about the teacher.'

  (15) O jogador **se** **encontra** em campo. (PT)
  The player RCLI finds/meets on field.
  The player finds/meets himself on the field. 'The player is on the field.'

  (16) Eu **me** **esqueci** do nome dele. (PT)
  I RCLI forgot of.the name of.him.
  I forgot myself of his name. 'I forgot his name.'

  (17) Eu esqueci o nome dele. (PT)
  I forgot the name of.him.
  'I forgot his name.'

- *Light-verb constructions* (LVCs) are VERB(-ADP)(-DET)-NOUN[17] combinations in which the verb V is semantically void or bleached, and the noun N is a predicate expressing an event or a state. Two subtypes are defined:
  - *LVC.full* are those LVCs in which the subject of the verb is a semantic (i.e. compulsory) argument of the noun, as in Example (18),
  - *LVC.cause* are those in which the subject of the verb is the cause of the noun (but is not its semantic argument), as in (19).

  The idiomatic nature of LVCs lies in the fact that the verb may be lexically constrained and contributes no (or little) meaning to the whole expression.

[16] Some languages, e.g. German and Polish, use the term *reflexive pronoun* instead of *reflexive clitic*.

[17] Parentheses indicate optional elements. ADP stands for adposition, i.e. either a preposition or a postposition, spelled separately or together with the noun. The order of components may vary depending on the language, and intervening words (gaps) may occur.

(18) Ikasle hori-k ez **du interes**-ik ikasgai-a-n. (EU)
Student this-ERG no has interest-PART subject-the-LOC
This student has no interest in the subject. 'This student is not interested in the subject.'

(19) Kolpe-a-k **min eman** dio. (EU)
punch-the-ERG pain.BARE give AUX
The punch gave him/her pain. 'The punch hurt him/her.'

- *Verbal idioms* (VIDs) are verb phrases of various syntactic structures (except those of IRVs and VPCs), mostly characterized by metaphorical meaning, as in (20).

(20) Dawno już powinien był **wyciągnąć nogi**. (PL)
long.ago already should.3SG was stretch legs
He should have stretched his legs long ago. 'He should have died long ago.'

- *Verb-particle constructions* (VPC), pervasive in Germanic languages but virtually absent in Romance or Slavic ones, are semantically non-compositional combinations of a verb V and a particle PRT. Two subtypes are defined:
  - *VPC.full* in which the V without the PRT cannot refer to the same event as V with the PRT, as in Example (21),
  - *VPC.semi* in which the verb keeps its original meaning but the particle is not spacial, as in (22).

(21) Ein Angebot von Dinamo Zagreb hat Kovac bereits **aus-geschlagen**. (DE)
an offer of Dinamo Zagreb has Kovac already knocked-out
Kovac has already knocked out an offer from Dinamo Zagreb. 'Kovac has already refused an offer from Dinamo Zagreb.'

(22) Ende März **wertete** eine unabhängige Jury die Bilder **aus**. (DE)
end March evaluated an independent jury the paintings off
Late March, an independent jury evaluated the paintings off. 'Late March, an independent jury evaluated the paintings'

For all languages in the PARSEME corpus, the VMWE annotation layer is accompanied by morphological and syntactic layers, as shown in Figure 3. In the morphological layer, a lemma, a part of speech and morphological features are assigned to each token. The syntactic layer includes syntactic dependencies between tokens. For

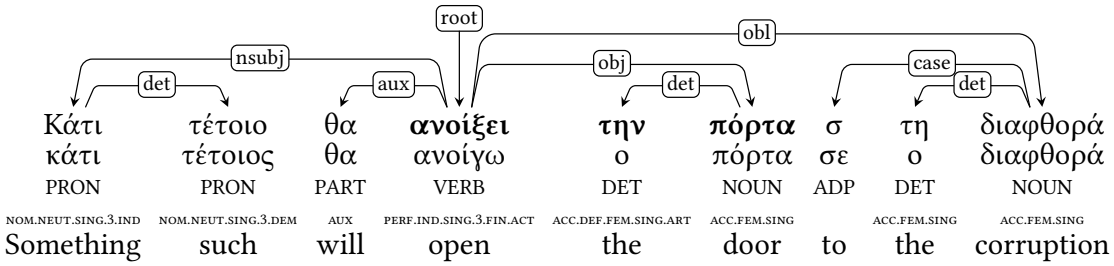


Figure 3. Morphosyntactic annotation for an occurrence context of the VMWE (EL) ***ανοίξει την πόρτα*** *(anixi tin porta) 'open the door'⇒'enable'.*

Language	Sentences	Tokens	VMWEs	Morphological layer		Syntactic layer	
				Tagset	Annotation	Tagset	Annotation
Basque	11,158	157,807	3,823	UD	partly manual	UD	partly manual
German	8,996	173,293	3,823	UD	automatic	UD	automatic
Greek	8,250	224,762	2,405	UD	automatic	UD	automatic
Polish	16,121	274,318	5,152	UD	partly manual	UD	partly manual
Portuguese	27,904	638,002	5,536	UD	partly manual	UD	partly manual

Table 1. Statistics of the PARSEME corpora used to extract LO candidates.

each language, this study combined the training, development and test sets into a single corpus whose sizes, tagsets and annotation methods are shown in Table 1.[18]

While the PARSEME corpus is manually annotated and categorized for IOs of VMWEs, it is not annotated for their LOs. Therefore, we developed several heuristics which allow us to identify them automatically, as discussed in the following section.

4. Automatic pre-identification of literal occurrences

We now consider the task of automatically identifying candidates for LOs in the corpora described in the previous section. In this work, we do not use any external resources. This allows us to compare all languages in a similar manner, but it also means that we can only automatically identify LO candidates for VMWEs which were annotated at least once in the corpus.

Moreover, in order to reliably perform the identification of LOs, we need to ensure that conditions 1, 2 and 3 from page 7 hold. To this aim, we may benefit from the

[18]UD stands for the Universal Dependencies tagset (`http://universaldependencies.org/guidelines.html`). For Basque, the PARSEME corpus uses both the UD tagset and a Basque-specific tagset. For this study, we unified the Basque corpus so that only the UD tagset is used.

morphological, syntactic and VMWE annotation layers present in the corpus. While checking Condition 1, we can rely on the underlying morphological annotation, which contains lemmas and parts of speech. However, as shown in Table 1, most of this annotation was performed automatically, and the risk of errors is relatively high. Therefore, the heuristics defined below rely only on lemmas but not on POS.[19] Condition 2 is closely linked to the syntactic annotations, but checking it fully reliably can be hindered by at least two factors. First, some dependencies can be incorrect, especially if determined automatically. Second, defining conditions under which two sets of dependency relations are equivalent is challenging and highly language-dependent because it requires establishing an exhaustive catalog of all CSSes for a VMWE type. Such a catalogue can be huge, or even potentially infinite, due to long-distance dependencies in recursively embedded relative clauses, as illustrated in Example (8) p. 7. Therefore, the heuristics defined below approximate VMWE types by abstracting away either from the dependency relations or from their directions and/or labels. Finally, Condition 3 can be automatically fulfilled by discarding all LO candidates that coincide with annotated VMWEs. Nonetheless, even if performed manually, VMWE annotations may still contain errors.

In order to cope with these obstacles, we design four *heuristics* which should cover a large part of LOs in complementary ways, while keeping the amount of false positives relatively low (i.e., the heuristics are skewed towards high recall). In the preprocessing step, we extract each occurrence of an annotated VMWE in a sentence s as a subsequence $e = \{e_1, e_2, \ldots, e_{|e|}\}$. For each VMWE e extracted in this way, and for each sentence $s' = \{s'_1, s'_2, \ldots, s'_{|s'|}\}$, we then look for relaxed non-idiomatic occurrences of e in s'. A relaxed non-idiomatic occurrence is a relaxed version of a potential occurrence (cf. Section 2.3), which applies to a VMWE occurrence rather than type, neglects POS and letter case, and is robust to missing lemmas. We first extend the definitions from Section 2 so as to account for missing or erroneous annotations. Namely, for a token s_i in sentence s, we define $\mathrm{lemmasurface}(s_i)$ as $\mathrm{lemma}(s_i)$, if available, and as $\mathrm{surface}(s_i)$ otherwise. Additionally, for any string x, $\mathrm{cf}(x)$ denotes its case-folded version. For instance, in Figure 1(a), $\mathrm{cf}(\mathrm{surface}(s_1)) = \mathrm{the}$. Finally, we say that r is a *relaxed non-idiomatic occurrence* (RNO) of e in s', if r is a subsequence of s' (cf. Section 2.1), $|r| = |e|$, and there is a bijection $\mathrm{rno}^r_e : \{1, 2, \ldots, |e|\} \to \{1, 2, \ldots, |e|\}$, such that: (i) for $i \in \{1, 2, \ldots, |e|\}$ and $j = \mathrm{rno}^r_e(i)$, we have $\mathrm{cf}(\mathrm{lemmasurface}(e_i)) \in \{\mathrm{cf}(\mathrm{lemma}(r_j)), \mathrm{cf}(\mathrm{surface}(r_j))\}$; and (ii) r has not been annotated as a VMWE. For instance, for the VMWE occurrence $e = \{(1, s_5), (2, s_7)\}$ from Figure 2 (a), we obtain the following RNO in sentence s' from Figure 2 (b): $r = \{(1, s'_6), (2, s'_8)\}$, with $\mathrm{rno}^r_e(1) = 2$ and $\mathrm{rno}^r_e(2) = 1$. Note that we do not require the POS tags in r to be the same as in e. In this way, we avoid sensitivity of the heuristics to tagging errors.

[19] Automatically determined lemmas may also be erroneous but we have to rely on them if LOs of previously seen VMWEs are to be found.

The set of such occurrences can be huge, and include a large number of false positives (that is, coincidental occurrences of e's components). Therefore, we restrain the set of *LO candidates* to the RNOs with the following criteria.

- **WindowGap**: Under this criterion, all matched tokens must fit into a sliding window with no more than g external elements (gaps). Formally, let J be the set of all matched indexes in sentence s', that is, $J = \{\, j \mid \mathrm{sub}_r^{s'}(i) = j \,\}$. Then r is only considered to match if $\max(J) - \min(J) + 1 \leqslant g + |e|$. For the subsequences e in Figure 2(a) and the RNO r in Figure 2(b), we have $J = \{6, 8\}$ and $|e| = 2$. Thus, the RNO *pulling strings* would be proposed as an LO candidate only if $g \geqslant 1$. The RNO in Figure 2(c) would also be proposed if $g \geqslant 1$. In the case of Figure 1(a), if this VMWE had not been annotated, it could also be proposed as an LO candidate with $g \geqslant 1$, while the occurrence in Figure 1(c) would require $g \geqslant 2$. In this article, WindowGap uses $g = 2$ unless otherwise specified.
- **BagOfDeps**: Under this criterion, an RNO must correspond to a weakly connected unlabeled subgraph with no dummy nodes, that is, the directions and the labels of the dependencies are ignored. For the VMWE in Figure 2(a), the RNO from Figure 2(b) would be proposed, as it consists of a connected graph of the lemmas *pull* and *string*, but the RNO in Figure 2(c) would not be suggested, as the tokens *pulling* and *strings* correspond to a subgraph with a dummy node.
- **UnlabeledDeps**: Under this criterion, an RNO r must correspond to a connected unlabeled graph with no dummy nodes, that is, the dependency labels are ignored but the parent relations are preserved. Formally, this criterion adds a restriction to BagOfDeps: r must be such that, if $\mathrm{parent}_e^s(e_k) = e_l$, $\mathrm{rno}_e^r(k) = i$, and $\mathrm{rno}_e^r(l) = j$, then $\mathrm{parent}_r^{s'}(r_i) = r_j$. For the VMWE in Figure 2(a), the RNO *pulling strings* in Figure 2(b) would be proposed, as it defines a connected subgraph with an arc between the lemmas *pull* and *string*.
- **LabeledDeps**: Under this criterion, an RNO must be a connected labeled graph with no dummy nodes, in which both the parent relations and the dependency labels are preserved. Formally, this criterion adds a restriction to UnlabeledDeps: For every $e_k \in e \setminus \{e_{root}\}$, if $\mathrm{rno}_e^r(k) = i$ then $\mathrm{label}(e_k) = \mathrm{label}(r_i)$. For the VMWE in Figure 2(a), differently from the heuristic UnlabeledDeps, the RNO *pulling strings* in Figure 2(b) would not be proposed because the label of the arc going from *pulled* to *strings* is not the same in both cases (*obj* vs. *nsubj*).

The heuristics defined by these criteria are language independent and were applied uniformly in the five languages: every RNO covered by at least one of the four heuristics was proposed as an LO candidate.

5. Manual annotation of literal occurrences

The sets of LO candidates extracted automatically were manually validated by native annotators. To this aim, we designed a set of guidelines which formalize the

methodology proposed for Polish in Savary and Cordeiro (2018), with some adaptations. We do not annotate the full corpus, but only the LO candidates retrieved by one of the heuristics, to save time and help annotators focus on potential LOs. As part of the morphological and syntactic layers in our corpora are automatically generated by parsers (Table 1), annotation decisions are taken based on ideal lemmas, POS tags and dependency relations (regardless of the actual dependency graphs in the corpora).

5.1. Annotation labels

We use the labels below for a fine-grained annotation of the phenomena. Each LO candidate is assigned a single label. The label set covers not only the target phenomena (LOs and COs of VMWEs) but also errors due to the original annotation or to the automatic candidate extraction methodology:[20]

- *Errors* can stem from the corpus or from the candidate extraction method.
  1. ERR-FALSE-IDIOMATIC: LO candidates that should not have been retrieved, but have been found due to a spurious VMWE annotation in the original corpus (error in the corpus, false positive):
     - *She […] brought back a branch of dill.* is retrieved as a candidate because *bring back* was wrongly annotated as an IO in ***bringing*** *the predator* ***back*** *to its former home*.
  2. ERR-SKIPPED-IDIOMATIC: LO candidates that should have been initially annotated as IOs in the corpus, but were not (error in the corpus, false negative).
     - *Bring down* was inadvertently forgotten in *Any insult […]* ***brings*** *us all* ***down***, although it is an IO.
  3. NONVERBAL-IDIOMATIC: LO candidates that are MWEs, but not verbal, and are thus out of scope (not an error, but a corpus/study limitation).
     - *Kill-off* functions as a NOUN in *After the major* ***kill-offs****, wolves […].*
  4. MISSING-CONTEXT: more context (e.g. previous/next sentences) would be required to annotate the LO candidate (genuinely ambiguous).
     - Without extra context, *blow up* is ambiguous in *Enron is blowing up.*
  5. WRONG-LEXEMES: The LO candidate should not have been extracted, because the lemmas or POS are not the same as in an IO (errors in the corpus' morphosyntactic annotation, or in the candidate extraction method).
     - The lexemes of *take place* do not occur in *Then take your finger and place it under their belly* because *place* is a VERB rather than a NOUN.
- *Coincidental* and *literal* occurrences are our focus. In the latter case, we also wish to check if an LO might be automatically distinguished from an IO, given additional information provided e.g., in VMWE lexicons.
  6. COINCIDENTAL: the LO candidate contains the correct lexemes (i.e., lemmas and POS), but the dependencies are not the same as in the IO.

[20] Although English is not part of this study, examples were taken from the PARSEME 1.1 English corpus.

  - The lexemes of *to* ***do the job*** 'to achieve the required result' co-occur incidentally in *[…] why you like the job and do a little bit of […]*, but they do not form and are not rephrasable to a connected dependency tree.
7. LITERAL-MORPH: the LO candidate is indeed an LO that could be automatically distinguished from an IO by checking morphological constraints.
   - The VMWE ***get going*** 'continue' requires a gerund *going*, which does not occur in *At least you get to go to Florida […]*
8. LITERAL-SYNT: the LO candidate is indeed an LO that could be automatically distinguished from an IO by checking syntactic constraints.
   - The VMWE *to* ***have*** *something* ***to do*** *with something* selects the preposition *with*, which does not occur in *[…] we have better things to do*.[21]
9. LITERAL-OTHER: the LO candidate is indeed an LO that could be automatically distinguished from an IO only by checking more elaborate constraints (e.g. semantic, contextual, extra-linguistic constraints).
   - *[…] we've come out of it quite good friends* is an LO of the VMWE *to* ***come of it*** 'to result', but it is unclear what kind of syntactic or morphological constraint could be defined to distinguish this LO from an IO.

5.2. Decision trees

Annotators label each automatically identified LO candidate using the decision tree below. Let $e = \{e_1, e_2, \ldots, e_{|e|}\}$ be a VMWE occurrence annotated in a sentence s and cs the canonical structure of e's type. Let $c = \{c_1, c_2, \ldots, c_{|c|}\}$ be e's LO candidate, i.e. an RNO extracted by one of the 4 heuristics from Section 4 in sentence s'.

Phase 1 – initial checks The automatic candidate extraction from Section 4 tries to maximize recall at the expense of precision, retrieving many false positives (e.g., annotation errors or wrong lexemes). Also, sometimes more context is needed to classify c. In this phase, we perform initial checks to discard such cases.

Test 1. [**FALSE**] Should e have been annotated as an IO of an MWE at all?
- NO → annotate c as ERR-FALSE-IDIOMATIC
- YES → go to the next test

Test 2. [**SKIP**] Is c actually an IO of an MWE that annotators forgot/ignored?
- YES, it is a verbal MWE → annotate c as ERR-SKIPPED-IDIOMATIC
- YES, but a non-verbal MWE → annotate c as NONVERBAL-IDIOMATIC
- UNSURE, not enough context → annotate c as MISSING-CONTEXT
- NO → go to the next test

Test 3. [**LEXEMES**] Do c's components have the same lemma and POS as cs's? That is, is c a potential occurrence (as defined in Section 2.3) of e?

[21] Here, the outcome depends on the PARSEME annotation conventions, in which selected prepositions are not considered as lexicalized components of VMWEs.

- NO → annotate c as WRONG-LEXEMES
- YES → go to the next test

Phase 2 – classification Once we have ensured that it is worth looking at the LO candidate c, we will (a) try to determine whether it is a CO or an LO, and (b) if it is the latter, then try to determine what kind of information would be required for an automatic system to distinguish an LO from an IO.

Test 4. [**COINCIDENCE**] Are the syntactic dependencies in c *equivalent* to those in cs? As defined in Section 2.3, dependencies are considered *equivalent* if a rephrasing (possibly an identity) of c is possible, keeping its original sense and producing dependencies identical to those in cs.[22]
- NO → annotate c as COINCIDENTAL
- YES → go to the next test

Test 4. [**MORPH**] Could the knowledge of morphological constraints allow us to automatically classify c as an LO?
- YES → annotate c as LITERAL-MORPH
- NO or UNSURE → go to the next test

Test 4. [**SYNT**] Could the knowledge of syntactic constraints allow us to automatically classify c as an LO?
- YES → annotate c as LITERAL-SYNT
- NO or UNSURE → annotate c as LITERAL-OTHER

5.3. Known limitations

As mentioned above, a precise definition of an LO, as proposed here, can only be done with respect to a particular syntactic framework. This is because we require the syntactic relations within an LO to be equivalent to those occurring in the canonical structure of a VMWE's type. The equivalence of the syntactic relations heavily depends on the annotation conventions of the underlying treebank. Here, we adopt UD, designed mainly to homogenize syntactic annotations across languages.

Suppose that the LVC in *the **presentation** was **made*** is annotated as an IO and that the heuristics propose the LO candidates (a) *his presentation made a good impression* and (b) *we made a surprise at her presentation*. In both LO candidates, the words *make* and *presentation* have a direct syntactic link, so we must base our decision on the relation's label. For Example (a), we cannot compare the labels between the LO candidate and the IO directly (both are `nsubj`), but we must first find the canonical structure of the IO (in which the label is `obj`) to conclude that this candidate is a CO rather than an LO. For candidate (b), the relation is `obl` and cannot be rephrased as `obj`, so this should

[22] Notice that we always compare the dependencies of c (or its rephrasing) with those in a canonical structure cs, never with those in an idiomatic occurrence e.

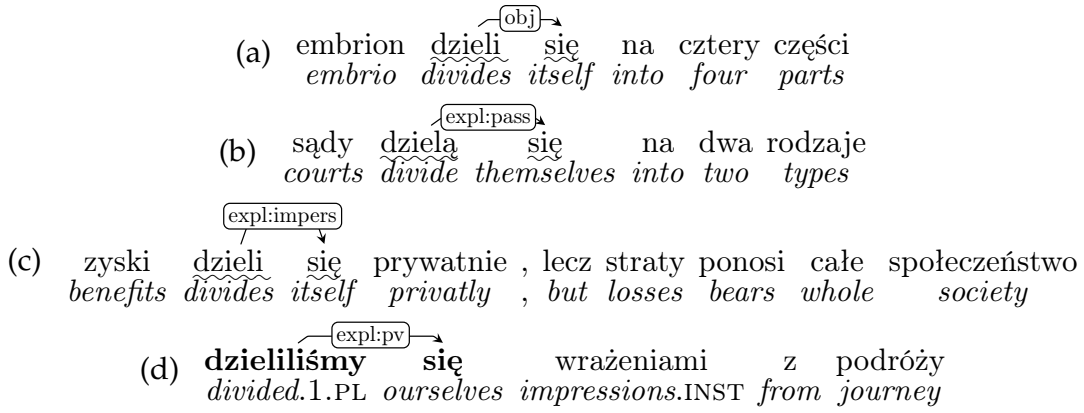


Figure 4. Four UD relations between a verb and a RCLI*. Translations: (a) 'the embryo splits into 4 parts', (b) 'there are 2 types of courts', (c) 'one shares benefits privately but loses are incurred by the whole society', (d) 'we shared our impressions from the journey'*

also be annotated as a CO. Notice that the outcomes could have been different in other syntactic frameworks, e.g., if `obj` and `obl` complements were treated uniformly.

The UD conventions are sometimes incompatible with our intentions. A notable example are verbs with reflexive clitics RCLI. According to UD, each RCLI should be annotated as *obj*, *iobj*, or as an expletive,[23] with one of its subrelations: *expl:pass*, *expl:impers* or *expl:pv* (Patejuk and Przepiórkowski, 2018), as shown in Figure 4. This means that the (semantic) ambiguity between the uses of the RCLI is supposed to be solved in the syntactic layer. Therefore, we ignore the (mostly language specific and often unstable) UD subrelations, so that the uses in Figure 4(b) and (c) are considered LOs of the IO in Figure 4(d). However, the use in Figure 4(a) has to be considered a CO, as we strictly cross our definition of an LO with this UD convention. Still, our intuition is that the (a) vs. (d) opposition in Figure 4 is one of the most challenging types of LOs and should be annotated as such. We postulate a future unification of the UD guidelines at this point, so that all examples in Figures 4(a-b-c-d) are annotated with the same dependency relation in the future. We argue that the distinction between purely reflexive and other uses of the RCLI should be avoided in the syntactic layer and be delegated to the semantic layer instead.

6. Results

In this section, we analyze the distribution of annotations across languages, and the suitability of heuristics (described in Section 4) to find genuine LOs.

[23] `http://universaldependencies.org/u/dep/expl.html#reflexives`

		DE	EL	EU	PL	PT
	Annotated IOs	3,823	2,405	3,823	4,843	5,536
	LO candidates	926	451	2,618	332	1,997
Distribution of labels	ERR-FALSE-IDIOMATIC	21.5% (199)	12.0% (54)	9.4% (246)	0.0% (0)	3.8% (76)
	ERR-SKIPPED-IDIOMATIC	27.0% (250)	47.5% (214)	17.3% (453)	5.4% (18)	10.7% (213)
	NONVERBAL-IDIOMATIC	0.0% (0)	0.0% (0)	0.2% (6)	0.0% (0)	0.5% (9)
	MISSING-CONTEXT	0.3% (3)	0.2% (1)	0.5% (12)	2.1% (7)	0.7% (13)
	WRONG-LEXEMES	40.1% (371)	0.9% (4)	26.7% (700)	1.8% (6)	38.1% (760)
	COINCIDENTAL (COs)	**2.6%** (24)	**27.9%** (126)	**42.4%** (1110)	**61.1%** (203)	**33.5%** (668)
	LITERAL (LOs)	**8.5%** (79)	**11.5%** (52)	**3.5%** (91)	**29.5%** (98)	**12.9%** (258)
	↪ LITERAL-MORPH	0.8% (7)	5.5% (25)	1.9% (51)	1.2% (4)	3.7% (73)
	↪ LITERAL-SYNT	1.5% (14)	2.0% (9)	0.7% (19)	8.1% (27)	2.2% (44)
	↪ LITERAL-OTHER	6.3% (58)	4.0% (18)	0.8% (21)	20.2% (67)	7.1% (141)
	Idiomaticity rate	**98%**	**98%**	**98%**	**98%**	**96%**

Table 2. General statistics of the annotation results. The idiomaticity rate is (#IOs)/(#IOs+#LOs), and #IOs include skipped idiomatic, e.g. $\frac{3823+250}{3823+250+79}$ *for DE.*

6.1. Annotation results

The general statistics of the (openly available) annotation results are shown in Table 2.[24] The VMWE annotations from the original corpus contained between 2.4 (EL) and 5.5 (PT) thousand annotated IOs of VMWEs (row 2).[25] The heuristics from Section 4 were then applied to these VMWEs to find LO candidates. An LO candidate was retained if it was extracted by at least one heuristic. The number of the resulting LO candidates (row 3) varies greatly from language to language, mainly due to language-specific reasons discussed in Sections 7–9. All LO candidates were annotated by expert native speakers (authors of this article) using the guidelines described in Section 5. The next rows (4–13) represent the distribution of annotation labels, documented in section 5.1, among the annotated candidates, across the five languages.

In most languages, a considerable fraction of the candidates turned out to be a result of incorrect annotations in the original corpus. These candidates may be false positives (row 4), or instances of false negatives (row 5).[26] In German, Basque and

[24] The annotated corpus is openly available at `http://hdl.handle.net/11372/LRT-2966`.

[25] In Polish, the reported number of annotated VMWEs is lower in Table 2 (4,843) than in Table 1 (5,152) because the former excludes VMWEs of the IAV (inherently adpositional verb) category, which were annotated only experimentally, and were disregarded in the present study.

[26] A point of satisfaction is that the number of errors of this kind dropped for Polish with respect to our previous work in (Savary and Cordeiro, 2018), performed on edition 1.0 of the PARSEME corpus. This indicates a better quality of the corpus in version 1.1.

	DE					EL				EU			PL				PT			
	IRV	LVC	VID	VPC	All	LVC	VID	VPC	All	LVC	VID	All	IRV	LVC	VID	All	IRV	LVC	VID	All
IdRate	99	100	99	97	**98**	99	95	100	**98**	99	93	**98**	98	99	96	**98**	93	99	88	**96**
EIR	99	100	97	97	**98**	94	92	100	**94**	86	58	**78**	95	94	90	**94**	85	92	73	**86**
ECR	0.6	0.3	1	.1	**.6**	5	3	0	**5**	14	37	**20**	3	5	7	**4**	9	7	18	**10**
ELR	1	0	1	3	**2**	1	5	0	**2**	1	5	**2**	2	1	3	**2**	6	1	10	**4**

Table 3. Extended idiomaticity (EIR), coincidentality (ECR) and literality (ELR). The numbers indicate percentages.

Portuguese, many of the incorrect candidates are also due to wrong lexemes, which results from two factors: (i) the fact that the heuristics rely on lemmas but not on parts of speech (Section 4), and (ii) incorrect lemmas in the underlying morphological layer.

The fraction of actual LOs among the extracted LO candidates (row 10) ranges from 3.5% (EU) to 29.5% (PL). This contrasts with a considerably higher number of COs (row 9) in almost all languages, with the exception of German. This might be partially explained by the fact that 30% of all German candidates stem from annotated multiword-token VPCs, e.g., (DE) *ab-geben* 'submit', which cannot have COs. The distribution of LITERAL-MORPH, LITERAL-SYNT and LITERAL-OTHER (rows 11–13) is addressed in sections 7–9.

The overall quantitative relevance of LOs can be estimated by measuring the *idiomaticity rate* (row 14), that is, the ratio of a VMWE's idiomatic occurrences (initially annotated IOs in the corpus or LO candidates annotated as ERR-SKIPPED-IDIOMATIC) to the sum of its idiomatic and literal occurrences in a corpus (El Maarouf and Oakes, 2015). If the overall idiomaticity rate is relatively low, distinguishing IOs and LOs becomes, indeed, a major challenge, as claimed by Fazly et al. (2009). However, as shown at the bottom of Table 2, the idiomaticity rate is very high (at least 96%) in all languages. In other words, whenever the morphosyntactic conditions for an idiomatic reading are fulfilled, this reading almost always occurs. This is one of the major findings of this work, especially from the point of view of linguistic considerations, given that most VMWEs could potentially be used literally.

From the point of view of NLP, however, more interesting is the proportion of IOs, COs and LOs with respect to the sum of these 3 types of occurrences. This is because a major MWE-oriented task is the automatic identification of MWEs in running text, where COs may play a confounding role. We call these the *extended idiomaticity rate* (EIR), *extended coincidentality rate* (ECR), and *extended literality rate* (ELR), respectively. Rows 4–6 in Table 3 show these three rates across languages and VMWE categories. EIR varies from language to language. In German, Greek and Polish, with total EIR over 94%, our heuristics become a powerful tool for identifying occurrences of previously seen VMWEs. In Basque and Portuguese, the proportion of IOs is much lower, notably due to language-specific CO-prone phenomena, discussed in Section 8. If

	DE		EL		EU		PL		PT	
	tokens	**types**	**tokens**	**types**	**tokens**	**types**	**tokens**	**types**	**tokens**	**types**
IOs	4 073	2 094	2 619	1 270	4 276	856	4 861	1 690	5 749	2 118
COs	24	0.9% (19)	126	5.5% (75)	1 110	18.0% (196)	203	4.7% (85)	668	10.7% (264)
LOs	79	2.4% (51)	52	2.0% (27)	91	3.6% (39)	98	2.6% (48)	258	3.2% (78)

Table 4. Distribution of IOs, LOs and COs across VMWE tokens and types. IO counts are updated to include err-skipped-idiomatic *cases.*

	IOs				COs				LOs			
	IRVs	**LVCs**	**VIDs**	**VPCs**	**IRVs**	**LVCs**	**VIDs**	**VPCs**	**IRVs**	**LVCs**	**VIDs**	**VPCs**
DE	9	8	34	49	8	4	79	8	4	0	27	70
EL	0	72	26	2	0	82	18	0	0	31	69	0
EU	0	79	21	0	0	50	50	0	0	24	76	0
PL	47	43	10	0	33	49	18	0	59	21	19	0
PT	16	64	21	0	14	43	43	0	25	15	60	0

Table 5. Distribution of IOs, LOs and COs, across VMWE categories (values are reported as percentages, adding up to 100 except for rounding).

those phenomena were treated as special cases (e.g., imposing additional morphological constraints) then the heuristics would also be effective for identifying previously seen VMWEs in these languages.

We also looked at the distribution of LOs and COs across VMWE types. Table 4 shows the number of IO, LO and CO tokens and types updated with respect to the initial VMWE annotation statistics, still considering ERR-SKIPPED-IDIOMATIC cases as IOs. Row 4 shows that the proportion of VMWE types which exhibit COs varies greatly among languages: from 0.9% in German to 10.7% in Portuguese and 18.0% in Basque. In Section 8, we further analyze the reasons for these particularities. Row 5 shows that the percentage of VMWE types with LOs is much more uniform, ranging from 2.0% for Greek to 3.6% for Basque. These LOs have a Zipfian distribution, as demonstrated by Figure 5: very few VMWEs have an LO frequency over 5, whereas a large majority of them has only one LO. The top-10 VMWE types with the highest individual LO frequency cover between 39% (in German) and 66% (in Greek) of all LOs. The appendix further shows the 10 VMWE types with the highest ELR and the 10 VMWE types with the highest frequency of LOs in each language. More in-depth language-specific studies might help understand why these precise VMWEs are particularly LO-prone.

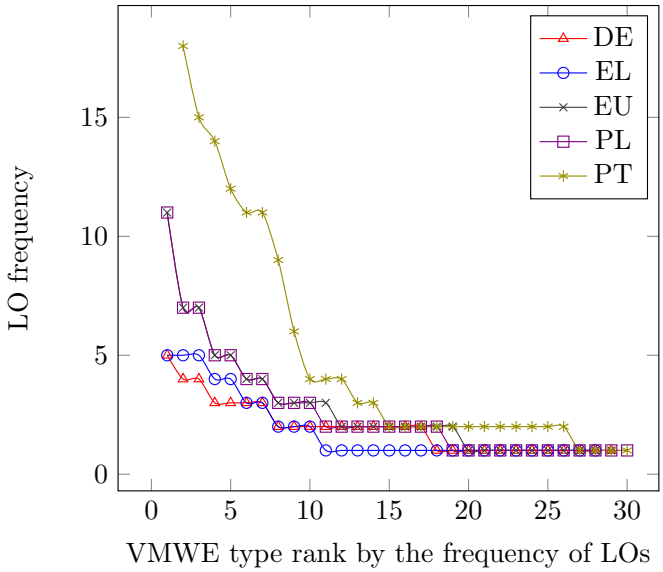


Figure 5. Frequency of LOs of the top-30 VMWE types per language. The VID (PT) ***já era*** *'already was.3SG.IPRF'⇒'it is over' (68 LOs) exceeds the vertical axis and is not shown.*

Table 5 shows the distribution of IOs, COs and LOs across VMWE categories. German has VMWEs of all 4 categories (with almost half of them being VPCs), while the other four languages are missing either IRVs or VPCs (or both). The distribution of COs and LOs across categories varies greatly across languages. The proportion of IOs to COs (excluding the cases of 0 occurrences) varies from 0.43 for German VIDs to 2 for German LVCs, except for German VPCs, with many IOs and LOs but few COs (probably due to the high percentage of mutiword tokens, as mentioned above). We also notice a pattern between LVCs and VIDs in Greek, Basque and Portuguese: LVCs are 2.8 to 3.8 times more frequent than VIDs, but their LOs exhibit roughly the inverse proportions. Interestingly, German seems to have no LOs for LVCs; while in Polish, most LOs stem from IRVs, with other occurrences almost evenly distributed between LVCs and VIDs.

6.2. Results of the heuristics in the task of finding literal occurrences

Once the candidates have been manually annotated, we can verify how well the four heuristics from section 4 solve the task of automatically identifying LOs of previously seen candidates. Table 6 presents precision (P), recall (R) and F-measure (F) in this task for each individual heuristic.

The precision represents the fraction of candidates that were then labeled as LITERAL. As expected, the most restrictive heuristic, LabeledDeps, obtains the highest precision, as its candidates are the ones that resemble the most the morphosyntactic

structure of the annotated VMWEs. In this work we were particularly interested in high recall, since the extracted candidates were further manually validated. The recall is the fraction of all candidates that were retrieved by a given heuristic. This definition of recall does not account for all of the LOs that could possibly have been found, but only for those which have been predicted by at least one heuristic, yielding a recall of 1.00 when the union of all heuristics is considered. We previously showed for Polish that this approximation proves accurate: these heuristics did not miss a single LO in the first 1,000 sentences of the corpus (Savary and Cordeiro, 2018).[27]

The recall for WindowGap is often quite high (91%–98%), suggesting that $g = 2$ is a good number of gaps in the common case, except for German (78%) and Greek (87%). This is consistent with Savary et al. (2018), in which German is an outlier concerning the average gap length within VMWEs (2.96), notably due to the frequency of long-distance dependencies in VPCs, which also occur in LOs, as in (DE) *Mutter Jasmin hielt ihn in letzter Sekunde fest* 'Mother Jasmin held him firmly till the last second'. Similarly, long-distance dependencies (i.e. those exceeding $g = 2$), due notably to the relatively free word order, especially in LVCs, may account for the 13% of LOs not found in Greek, as in (EL) έχει πολλές σπάνιες και αξιόλογες εικόνες (echi poles spanies ke aksiologes ikones) 'has many rare and valuable pictures'.

Through recall, we can attest that the heuristics are complementary, in the sense that no single heuristic is able to predict all of the LOs. For example, for German, WindowGap has R=78%, thus the other 22% of LOs were predicted through BagOfDeps (and possibly the other two more restrictive heuristics as well). Similarly, BagOfDeps has R=90%, implying that the other 10% were predicted only by WindowGap. This means that only 68% (i.e., $100\% - (22\% + 10\%)$) of the actual LOs were predicted by the intersection of both heuristics. Similar numbers are found for other languages, ranging from an intersection of 60% for Portuguese to 80% for Basque.

As expected, the recall of the BagOfDeps is systematically higher than the recall of UnlabeledDeps, which in turn is systematically higher than the recall of LabeledDeps (since these heuristics rely on increasing degrees of syntactic constraints). These constraints are often valuable in filtering out false literal candidates, which is why the precision of these 3 methods mostly shows an inverse behavior.

7. Characteristics of literal occurrences

This section provides a qualitative analysis of LOs. The goal is to identify both cross-lingual and language-specific reasons for LOs to occur. Additionally, we show examples of morphosyntactic constraints which, if known in advance, e.g., from MWE lexicons (Przepiórkowski et al., 2017), may help automatically distinguish LOs from IOs in the VMWE identification task. Because the morphosyntactic behavior varies

[27] It might be worth repeating the same experiment for German, where long-distance dependencies in LOs are more pervasive.

Language	WindowGap			BagOfDeps			UnlabeledDeps			LabeledDeps			All (union)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Basque	3	**91**	7	6	89	**11**	5	58	9	**6**	22	10	3	100	7
German	8	78	14	12	**90**	22	13	**90**	22	**14**	77	**23**	9	100	16
Greek	11	87	20	15	**90**	26	**16**	83	**27**	16	52	24	12	100	21
Polish	33	**96**	49	43	81	56	49	73	**59**	**52**	23	32	30	100	46
Portuguese	14	**98**	25	17	62	27	20	59	30	**34**	37	**36**	13	100	23

Table 6. Precision, recall and F-measure of the heuristics (all reported as percentages).

greatly across VMWE categories, this analysis is performed separately for each category.

7.1. IRVs

IRVs exhibit LOs due to homography with compositional VERB + RCLI combinations with true reflexive, reciprocal, impersonal and middle-passive uses. Recall from Section 5.3 and Figure 4 that these uses of RCLIs are supposed to be syntactically distinguished in UD via subrelations. However, due to their language-specific definition and inconsistent usage, subrelations are ignored in our annotation. Thus, examples like (23) are considered middle passive counterparts of the IRVs in (15), page 16.

(23) Nesse rio se encontraram muitos tipos de peixe. (PT)
In.this river RCLI found/met many kinds of fish.
'Many kinds of fish were found in this river.'

This large potential for LOs is displayed mainly in Portuguese and Polish (Table 5). Most of these LOs were annotated as LITERAL-OTHER, i.e., no explicit morphosyntactic hints can help automatically distinguish them from IOs, notably because the RCLI has a weak and infrequent inflection. Still, some LOs were labeled LITERAL-SYNT because they differ from the corresponding IOs by their valency frames. For instance, the IRV in Example (24) requires a genitive object, while the LO in (25) occurs with an accusative object.

(24) Polityk **dopuszczał się** bezprawia. (PL)
Politician allowed RCLI crime.GEN.
The politician allowed himself crime. 'The politician perpetrated crimes'

(25) Dopuszcza się inną działalność niż gastronomiczna. (PL)
Allows RCLI another activity.ACC than gastronomic.
'Activities other than gastronomic are allowed.'

7.2. LVCs

LVCs are mostly semantically compositional, in the sense that the light verb only contributes a bleached meaning (mostly stemming from morphological features, such as tense and aspect) to the whole expression. Therefore, the notion of an LO is less intuitively motivated for them. An LO of an LVC should be understood as a co-occurrence of the LVC's lexemes that does not have all the required LVC properties. This occurs, for instance, when a noun has both a predicative and a non-predicative meaning, i.e., it does or does not express an event or state. In Examples (26) and (27), the noun *zezwolenie* 'permission' means either the fact of being allowed to do something, or a concrete document certifying this fact (i.e. a permit), which yields an LVC and its LOs.

(26) Nie **mają** wymaganego **zezwolenia** na pracę. (PL)
Not have.3rd.PL required permission for work.
'They have no permission to work.'

(27) Kierowcy mieli sfałszowane zezwolenia. (PL)
Drivers had falsified permissions.
'The drivers had falsified permissions.'

The LVC in (26), like most other LVCs, exhibit a totally regular morhosyntactic behavior, therefore their LOs are usually classified as LITERAL-OTHER. Still, a few frequent LVCs do impose morphosyntactic constraints, like the LVC in (28), which prohibits modification of its direct object *miejsce* 'place'. Conversely, in the LO in (29), the same noun receives a nominal modifier, which makes it fall into the LITERAL-SYNT class.

(28) Zdarzenie **miało miejsce** w minioną sobotę. (PL)
Event had place in last Saturday.
'The event took place last Saturday.'

(29) Łódż miała stałe miejsce postoju na przystani. (PL)
Boat had permanent place of.parking on harbor.
'The boat had its permanent parking lot in the harbor.'

7.2.1. Polish-specific phenomena

Polish additionally exhibits a particular syntactic phenomenon which triggers a number of LOs. Namely, given the existential *być* 'to be' in present tense, e.g., in *są powody* 'are reasons.NOM'⇒'there are reasons', its negation is realized by the verb *mieć* 'to have' with the subject shifted to the object position, e.g., *nie ma powodów* 'not has reasons.ACC'⇒'there are no reasons'. Thus, an LVC occurring in present tense under the scope of negation, as in (30), is homonymic with a negated existential construction, as in (31).

(30) (Klient) nie **ma powodów** do satysfakcji. (PL)
Client not has reasons for satisfaction.
'(The client) has no reasons to be satisfied'

(31) Nie ma powodów do satysfakcji. (PL)
Not has reasons for satisfaction.
'There are no reasons to be satisfied'

Since Polish is a pro-drop language, the subject in (30) can be skipped, which makes both occurrences look identical. This clearly implies their labelling as LITERAL-OTHER.

7.2.2. Portuguese-specific phenomena

The Portuguese verb *ter* 'to have' exhibits two interesting language-specific phenomena which trigger LOs of LVCs: resultatives and secondary predication.

The structure of resultative constructions, illustrated by Example (32), may be very similar to some LVCs, as in (33). In both cases, the noun is the direct object of the verb *ter* 'to have' and it governs a participle. Because of the well known ambiguity of participles, in (32) the participle *renovada* 'renewed' depends on the noun via the *acl* relation, while in (33) *equilibrada* 'balanced' it is a plain adjectival modifier (one cannot specify the agent of *balance*).

(32) Ele tem sua força renovada quando descansa. (PT)
He has his strength renewed when rests.
'His strength gets renewed when he rests.'

(33) A criança **tem** uma **alimentação** equilibrada. (PT)
The child has a diet balanced.
'The child has a balanced diet.'

This subtle syntactic constraint might make (32) fall into the LITERAL-SYNT class, but it is unclear whether the presence of an outgoing *acl* relation is sufficient to distinguish an IO from an LO. Therefore, cases of this kind were labeled LITERAL-OTHER.

Secondary predication is illustrated in Example (34). There, the verb *ter* 'to have' has both a direct object (*obj*) and an indirect object (*iobj*) introduced by *como/por* 'as/by', the latter being a predicative of the former.

(34) João tem [seu irmão]$_{obj}$ [como um demônio]$_{iobj}$. (PT)
John has his brother as a demon.
'João considers his brother a demon.'

The indirect object can contain an abstract predicative noun, in which case its combination with *ter* 'have' is annotated as LVC.full, as in (35) and (36).

(35) Ela **tem [como objetivo]**$_{\text{iobj}}$ [a difusão de informações]$_{\text{obj}}$. (PT)
she has as goal the dissemination of information.
'Her goal is the dissemination of information.'

(36) Eles **tem** [essa atividade]$_{\text{obj}}$ [**como** uma **opção**]$_{\text{iobj}}$. (PT)
they have this activity as an option.
'This activity is a possible option for them.'

However, the opposite may also happen, that is, a predicative noun may appear in the *obj* position, as in (36). In this case, *tem atividade* 'has activity' is not an LVC.full, as it does not pass the V-REDUC test from the PARSEME guidelines.[28] Since the underlying CSS is identical to the canonical structure of this VMWE, this occurrence is annotation as LIT-OTHER.

7.3. VIDs

The origin of many VIDs lies in the metaphorical interpretation of semantically compositional constructions. Such VIDs are figurative (their literal meaning is easy to imagine) and naturally have a potential of LOs, as exemplified in (37)–(38).

(37) Gaixo dago eta ez **da** joateko **gauza**. (EU)
Sick is and no is going thing
He/She is sick and is no thing to go. 'He/She is sick and is unable to go.'

(38) Horiek beste garai bat-eko gauza-k dira. (EU)
These other time one-GEN thing-PL AUX
These are things from the past. 'These things belong to the past.'

Many of such cases, especially in Basque, Greek and Portuguese, can be distinguished by checking morphological or syntactic constraints (i.e. they are labelled LITERAL-MORPH or LITERAL-SYNT). Unlike in (37), the noun *gauza* 'thing' is in plural in (38). Since the noun inside the VID *gauza izan* 'be able (to)' is never used in the plural form, this feature indicates that the occurrence is literal.

Some LOs, however, fall into the LITERAL-OTHER class, notably when they are strong collocations or domain-specific terms. For instance, the LO in (40) is an institutionalized term, and has the same, both incoming and outgoing, syntactic dependencies as its corresponding IO in (39).

(39) Służenie nam **mają** **we krwi**. (PL)
serving us have.3rd.PL in blood
They have serving us in blood. 'Serving us is their innate ability.'

[28] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=lvc#test-lvc4

(40) Miał we krwi ponad 1,5 promila alkoholu (PL)
had.3rd.SING in blood over 1.5 per-mille alcohol
'His blood alcohol level was 1.5.'

7.3.1. Basque-specific phenomena

Basque, unlike the four other languages, is both postpositional and agglutinative, meaning that adpositions (which are separate words in the other four languages) are suffix-like (Inurrieta et al., 2018). Words decorated with different postpositions lemmatize to bare forms in which the postpositions are omitted. For instance, *kontu-a-n* 'account-ART-LOC' in Example (41) and *kontu-tik* 'account-ABL' in (42) both lemmatize to *kontu* 'account'. Additionally, the dependencies between these components and *hartu* 'take' are the same. Recall from Section 2.3 that the status of a candidate as an IO/LO/CO is based on comparing its CSS with the canonical structure of an IO. CSSes contain lemmas of the lexicalized components, which means that (suffix-like) adpositions in Basque are ignored in this comparison. This is why Example (42) counts as an LO of (41), despite the different adpositions *-n* 'LOC' and *-tik* 'ABL'.

(41) **Kontu-a-n** **hartu** du lagun-a-ren iritzi-a. (EU)
account-ART-LOC take AUX friend-ART-GEN opinion-ART.ABS
Took into account the opinion of his/her friend. 'He/She took his/her friend's opinion into account.'

(42) Diru-a hartu du kontu-tik. (EU)
money-ART.ABS take AUX account-ABL
Took money from the account. 'He/She withdraw money from the account.'

This behavior and modeling of adpositions is in sharp contrast with languages using prepositions on the one hand, and those using adverbial prefixes on the other. Prepositions are standalone words and can constitute independent lexicalized components of VMWEs. For instance, given the VID (EN) ***take** money **into account***, the occurrence (EN) *take money from my account* cannot be an LO/CO candidate because one lexicalized component (*into*) is missing. Conversely, adverbial prefixes, pervasive in Slavic languages, are inherent parts of the verb's lemma, i.e., they do not vanish in the process of lemmatization.[29] Therefore, given an IRV (PL) ***wy-nosić się*** 'out-carry oneself'⇒'to go away', an occurrence with a different prefix, like *pod-nosić się* 'lift oneself'⇒'stand up', can never be considered an LO/CO candidate.

7.3.2. German-specific phenomena

VIDs give raise to 27% of LOs in German (Table 5). Few of those (unlike in Basque, Greek and Portuguese) fall into the LITERAL-MORPH class (Table 2). The main reason is

[29]They resemble German VPCs as (DE) *auf-nehmen* 'up-take'⇒'to take up', but they are not separable.

that most of them stem from VIDs containing, along with the head verb, a functional word like an expletive pronoun or an adverb. The morphological range for the IO-LO distinction is therefore drastically reduced. Example (43) shows a VMWE with an expletive pronoun, and (44) a corresponding LO.

(43) **Es gilt** Hemmungen zu überwinden und zu lernen mit dem Lampenfieber
it holds inhibitions to overcome and to learn with the stage-fright
umzugehen. (DE)
to.deal

'You have to overcome inhibitions and learn how to deal with stage fright.'

(44) Es gilt der Grundsatz der Gleichbehandlung, erklärt die Sprecherin. (DE)
it holds the principle of equal-treatment says the speaker

'The principle of equal treatment applies, says the speaker.'

Besides the clear semantic contrast (the VMWE in (43) does not imply a legal provision), the two uses of *es gilt* 'it applies'⇒'one should' also differ with respect to their syntax: the VMWE in (43) governs a *zu*-infinitive, whereas the LO instance in (44) governs a noun phrase. Since the governed category is essential for the different readings to emerge, we have annotated the LO as LITERAL-SYNT.

In our German corpus, there is no common lemmatization for personal pronouns. *Es* 'it' is lemmatized as *es*, *er* 'he' as *er*, etc. Therefore, Example (45) cannot be suggested as an LO of (43) by the heuristics, even though this would be perfectly justified.

(45) Er gilt als russischer Mark Zuckerberg: [...] (DE)
he holds as Russian Mark Zuckerberg

'He is considered a Russian Mark Zuckerberg.'

7.3.3. Greek-specific phenomena

Like in German, many LOs of VIDs in Greek contain functional words, mainly pronouns, but in contrast to German, these LOs could be classified as LITERAL-MORPH. This is due to the diversity in how pronouns are modeled in both languages. In German, as just mentioned, each personal pronoun has its own lemma, e.g., *es* 'it' and *sie* 'they' are different lexemes. In Greek, pronouns are seen as exhibiting inflection for person, gender, number and case. Thus, e.g., *το* 'it' and *αυτούς* 'they' are inflected forms of the same lemma *εγώ* 'I'. This yields a large number of LOs. For instance, the VID in (46) comprises a clitic (i.e., a weak form of the personal pronoun) followed by a verb. The clitic *τα* 'them' is fixed with respect to the gender, number and case and does not co-refer with another nominal phrase.

(46) Ο Γιάννης **τα πήρε** με τα παιδιά. (EL)
Ο Gianis ta pire me ta pedia.
the John them took with the kids
John took them with the kids. 'John was very angry at the kids.'

The same clitic-verb combinations can occur in an LO, yet the morphosyntactic features of the clitic are not fixed, as in (47), which makes the LO fall into the LITERAL-MORPH category. It may also happen that the clitic in the LO has precisely the same morphology as in the VMWE, in which case the occurrence is labeled LITERAL-OTHER. Further ambiguity stems from clitic doubling (i.e., a construction in which a clitic co-occurs with a full noun phrase in argument position forming a discontinuous constituent with it), as illustrated in (48).

(47) Ο Γιάννης την πήρε με το αυτοκίνητο. (EL)
Ο Gianis tin pire me to aftokinito.
the John took her with the car
John took her in his car. 'John gave her a lift'

(48) Η κοπέλα τα πήρε τα έγγραφα (EL)
i kopela ta pire ta egrafa
the girl them took the documents
'The girl took the documents.'

As shown in Table 2, the LITERAL-MORPH class is the most frequent among Greek LOs. The rate of LITERAL-SYNT cases is lower, probably because when syntactic constraints can help solve the IO vs. LO ambiguity, morphosyntactic constraints also apply. In most LITERAL-SYNT cases, IOs either allow only for restricted modification of their elements, or no modification at all, as shown in (49), where the noun *χέρι* 'hand' allows no modifier.

(49) ο δημοσιογράφος τον **κρατάει στο χέρι** (EL)
ο dimosiografos ton kratai sto cheri
the journalist him holds in-the hand
The journalist holds him in the hand. 'The journalist has power over him.'

Conversely, LOs allow for modification, and can be identified on the grounds of syntactic features, as shown in (50), where the two modifiers of the noun are underlined.

(50) Στο δεξί του χέρι κρατάει το κουτί (EL)
sto dexi tu cheri kratai to kuti
in-the right his hand holds the box
'He holds the box in his right hand.'

Borderline cases between metaphors and VIDs were also identified, as shown in (51). Their corresponding LOs, like in (52), were marked as LITERAL-OTHER.

(51) Κάλεσε τους πολίτες να **βγουν** **στους** **δρόμους**. (EL)
kalese tus polites na vjun stus dromus
asked,03.SG the citizens to get-out.3PL to-the streets.
He asked citizens to get out to the streets. 'He asked the citizens to protest'

(52) Οι ποντικοί βγήκαν στους δρόμους του Παρισιού εξαιτίας […] (EL)
i pontiki vjikan stus dromus tu Parisiu eksetias […]
the rats went-out to-the streets of-the Paris because-of […]
'The rats appeared in the streets of Paris because of […]'

7.4. VPCs

Among our five languages of study, VPCs are mainly exhibited in German. LOs of a VPC occur whenever the verb is used literally and the particle is spacial. Thus, Example (53) is an LO of the VPC from Example (21) on page 17.

(53) Dem Michael wurden beide Schneidezähne aus-geschlagen (DE)
the.DAT Michael were both incisors out-knocked
'Michael's both incisors were knocked out.'

Despite their potential for LOs illustrated in Example (53), for many VPCs it is difficult to even imagine an LO. Trivially, this is the case where the verb is only used together with the particle, for example the verb *statten* in *aus-statten* 'equip'. But also VPCs such as *auf-geben* 'give up' are concerned, where it is rather the combination of verb and particle which is idiomatic. In the case of *auf-geben*, one might expect the availability of a literal meaning 'give upward', but this meaning is only available with the particle *hinauf*. Since both cases are particularly common in German VPCs (*aus-statten* and *auf-geben* alone occur 5 and 7 times in the corpus), this positively biases the idiomaticity rate.

Nevertheless, the few LOs which do occur in German are still dominated by VPCs 70%), probably due to their dominance also in the IOs (Table 5). Recall also from Table 2 that the majority of LITERAL annotations in the VPC category are classified as LITERAL-OTHER. The justification is similar to the one proposed in Section 7.3.2: since the particle has no inflection at all, VPCs and their LOs can hardly be distinguished in German based on the morphology of their components.

8. Characteristics of coincidental occurrences

Since LOs are contrasted in this work with IOs on the one hand and with COs on the other hand, it is interesting to also understand generic and language-specific

reasons for COs to arise. Recall that the heuristics described in Section 4 include WindowGap, which looks for a co-occurrence of the lexicalized components of a known VMWE within a window containing at most 2 gaps (external words). This leaves room for a large potential of COs and, indeed, those extracted only by the WindowGap method are 1.2 to 2.3 times more numerous than those yielded by BagOfDeps. Such candidates, e.g., (55) which is a CO of (54), in which the words in focus are not linked by direct syntactic dependencies, are of little general interest, except when language-specific studies cause their proliferation (see below).

(54) Es **kommt** auf die Qualität insgesamt **an**. (DE)
It comes on the quality totally on.
'It depends totally on the quality.'

(55) Union rannte an, kam zum Ausgleich … (DE)
Union ran on, came to deuce …
'Union attacked, came to a deuce …'

In the COs extracted with BagOfDeps, the syntactic dependencies are usually different from those occurring in the corresponding IOs. For instance, in (56) the dependency between the verb and the noun is of type *nmod*, while it is *obj* in the corresponding LVC in Example (28). Similarly, in (57), the verb *δίνω* 'give' is linked to the noun *απάντηση* 'answer' with the *subj* relation, while the *obj* relation occurs in the LVC ***δίνω απάντηση*** 'give an answer'.

(56) Teraz nie mam nikogo innego na jego miejsce. (PL)
now not have.1st.SING no-one else on his place
'Now, I have no one else to replace him.'

(57) Η απάντησή του μου δίνει αφορμή για […] (EL)
I apantisi tu mu dini aformi jia […]
the answer his me gives chance for […]
'His answer triggers […].'

Recall, however, from Figure 2 and Section 2.3 that sharing the same dependencies with an IO does not necessarily give an occurrence the status of an LO. It is, instead, the canonical structure of an IO's type which counts for evaluating the equivalence of syntactic relations.

8.1. Basque-specific phenomena

Basque has, by far, the highest number of COs, as attested in Table 2. It also has the highest extended coincidentality rate, especially in VIDs, as seen in Table 3. Many of the COs in Basque include nouns with adpositions, which vanish in the process of

lemmatization, as discussed in Section 7.3.1. For instance, in the VID from Example (58) the noun *aurre* 'front' is bare, and it is the direct object of the verb *egin* 'do'. Occurrences (59) and (60) contain the same noun but with adpositions, which is why their dependency to the verb is of different nature and they are COs rather than LOs.

(58) Arazo-e-i **aurre** **egin** zien. (EU)
problems-ART-DAT front.BARE do AUX
Did front to the problems. 'He/She faced the problems.'

(59) Irakasle-a-ren aurre-a-n egin zuen ariketa. (EU)
teacher-ART-GEN front-ART-LOC do AUX exercise.ART.ABS
'He/She did the exercise in front of the teacher.'

(60) Joan aurre-tik egin zuen ariketa. (EU)
leave front-ART.ABL did AUX exercise.ART.ABS
Did the exercise from front leaving. 'He/She did the exercise before leaving.'

Note that this example is quite analogous to (56) vs. (28), where the preposition does not vanish but is dependent on the noun, and therefore does not intervene in the comparison of the CSSes. It is therefore unclear why precisely the COs of this type are so much more frequent in Basque than in other languages exhibiting prepositions. Possible reasons are lemmatization errors in some corpora, or the fact that verbs in VMWE often govern functional words rather than nouns (e.g. in German VPCs, in German and Greek VIDs, and in Polish IRVs), which mostly excludes the use of prepositions.

8.2. Portuguese-specific phenomena

Portuguese has the second highest number of COs and ICR (Tables 2 and 3), especially in VIDs, like Basque, but also in IRVs. This is notably due to complex attachment mechanisms in reflexive clitics. They are adjacent to verbs in Portuguese, occurring immediately before (e.g., *me lavei* 'RCLI.1SG washed'⇒'I washed myself'), immediately after (e.g., *lavei-me* 'washed-RCLI.1SG') or, in some rare cases, in the middle of the verb, between its root and its suffix (e.g., *lavar-me-ei* 'wash-RCLI.1SG-FUT.1SG'⇒'I will wash myself'). A set of (more or less deterministic) rules allow choosing one of the three alternatives (e.g., a sentence cannot start with a reflexive clitic).

While the attachment of the clitic to its directly adjacent verb is mostly unambiguous, the interaction between reflexive clitics and verbal chains (e.g., auxiliary, modal, and controlled verbs) can be complex.[30] For instance, consider the verb *dever* 'to owe',

[30] In Brazilian Portuguese, a reflexive clitic is always adjacent to its verb (e.g., *vai se lavar* 'will RCLI wash'). European Portuguese has different rules, however, with auxiliary and modal verbs interposed between the clitic and the main verb (e.g., *se vai lavar* 'RCLI will wash'). We focus on Brazilian Portuguese only.

which is also used as a modal verb to express obligatoriness ('must'). In Example (61), the verb is combined with a reflexive clitic forming an IRV ***se deve*** *a* 'RCLI owe to'⇒'results from'. Examples (62) and (63), however, are not IOs of this VMWE, but candidates that must be annotated as a CO and an LO respectively.

(61) A demora **se** **deve** à burocracia. (PT)
the delay RCLI owe to.the bureaucracy
'The delay is due to the bureaucracy.'

(62) Os interessados devem se inscrever. (PT)
the interested.PL must RCLI register
'Those who are interested must register.'

(63) Deve se utilizar roupa ventilada. (PT)
must RCLI use clothes ventilated
'One must use ventilated clothes.'

The choice here depends on whether the clitic is attached to the main verb (CO) or to the modal verb (LO). In (63), the clitic marks an impersonal/middle reading of the whole verbal chain, hence the candidate is annotated as an LO (LITERAL-SYNT). Example (62), however, does not have this interpretation, as the clitic marks the reflexive object of the main verb *inscrever* 'register'. Therefore, it is annotated as a CO.

This distinction is tricky, but negation can be used as a test. One of the rules used to choose the clitic's position with respect to the verb is that negation "attracts" the clitic. The negation of Example (63) becomes *Não se deve utilizar* 'Not RCLI must use', indicating that the clitic is attached to the modal verb *dever* 'must'. In Example (62), negation does not change word order and fails to "attract" the clitic: *não devem se inscrever* 'not must RCLI register', indicating that the clitic attaches to the main verb.

8.3. Polish-specific phenomena

A similar ambiguity in the attachment of reflexive clitics occurs in Polish. It is less frequent but sometimes harder to solve, since *się* 'RCLI' benefits from the relatively free word order in this language and can often be separated from its governing verb. For instance the IRV in (64) triggers a CO in (65), where the reflexive clitic appears closer to the modal *ma* 'should' than to the infinitive *zmienić* 'change' which it depends on. One must therefore be extremely careful while annotating such cases. A possible test is to skip the modal and check if the clitic remains with the main verb as in *wszystko się zmieni* 'everything RCLI change.FUT'⇒'everything will change'.

(64) **Miał się** dobrze. (PL)
had RCLI well.
He had himself well. 'He was fine.'

(65) Teraz ma się wszystko zmienić. (PL)
Now has.to/should RCLI everything change.
'Now everything should change.'

9. Characteristics of erroneous occurrences

In this section, we are interested in the candidates labeled WRONG-LEXEMES, i.e., those which were extracted by the heuristics but do not respect Condition 1 from page 7. In other words, they have either different lemmas or different POS than the lexicalized components of an attested VMWE. Recall from Section 4 that the heuristics check the lemma but not the POS, so as to maximize recall even in presence of errors in morphosyntactic annotation.

As shown in Table 2, WRONG-LEXEMES are very frequent in German, Basque and Portuguese. In each case, this is due to the existence of homographs (understood here as words with the same lemma but different POS). One common case is the ambiguity of some common verbs between a main verb and an auxiliary. For instance, in (66), the auxiliary *tem* 'has' is ambiguous with the light verb appearing in the LVC ***tem força*** 'has strength'.

(66) O time tem mostrado força para reverter resultados. (PT)
the team has shown strength to revert results.
'The team has shown the strength to turn the results around.'

Other dominating classes of homographs are language-specific.

9.1. Basque-specific phenomena

Some Basque nouns (like some Hindi nouns[31]), such as the one in the LVC in Example (67), look identical to adjectives. This happens in (68), which triggers a candidate with a wrong lexeme.

(67) Plan-a-ren **berri** **eman** ziguten. (EU)
plan-ART-GEN news.BARE give AUX
Gave us news of the plan. 'They informed us about the plan.'

(68) Plan berri-a eman ziguten. (EU)
plan new-ART give AUX
'They gave us the new plan.'

Correct lemmatization can also be hindered by adpositions. Namely, several adverbs, such as *berriz* 'again' in Example (69), were formed by adding a postposition

[31] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=lvc

(here: *-z* 'INST') to a noun or an adjective (here: *berri* 'new'). Lemmatization of such adverbs is error-prone, therefore the occurrence in (69) was extracted on the basis of the LVC from Example (67).

(69) Plan-a berriz eman ziguten. (EU)
plan-ART again gave AUX
'They gave us the plan again.'

9.2. German-specific phenomena

Cases labeled WRONG-LEXEMES in German can be attributed to a large extent to particles in VPCs, which often have homographs with a different POS tag such as prepositions (e.g. *an* 'on'), the indefinite article *ein* 'a' and the infinitive marker *zu* (similar to *to* in English). For instance, in Example (70), the preposition *an* 'on' is wrongly confused with the particle appearing in the VPC from Example (54) in page 38.

(70) Beide Teams kamen an die free-throw-line. (DE)
both teams came on the free-throw-line.
'Both teams came up to the penalty line.'

9.3. Portuguese-specific phenomena

In Portuguese, one of the most frequent types of WRONG-LEXEMES stems from the fact that the conjunction *if* and the 3rd-person reflexive pronoun are homographs: *se*. Thus, a conditional sentence such as (71) is extracted on the basis of the IRV *perguntar-se* 'ask-RCLI'⇒'wonder'.

(71) Pergunta se sua mulher poderá vir. (PT)
asks if his wife can-3S-FUT come-INF.
'He asks if his wife will be able to come.'

Another common ambiguity is due to the fact that the subjunctive form *desse* of the verb *dar* 'to give' is a homograph of the contraction *desse* = *d-esse* 'of.this'. While, in this case, the lemmatized forms should have been different, errors in the underlying morphological annotation led to candidates such as the one in (72), extracted on the basis of the VID ***dar jeito*** 'give way'⇒'to find a workaround' .

(72) Foi bom porque vencemos e desse jeito. (PT)
was good because won-1PL and of.this way.
'It was a good thing, because we won, and in such manner.'

Other spurious candidates were proposed due to errors in lemmatization. For example, the verbs *ser* 'to.be' and *ir* 'to.go' have identical surface forms in some tenses

(e.g., *ele foi* 'he was / he went'). In the set of annotated expressions, there are cases in which ***foi bem*** 'went well'⇒'succeeded' and ***se foi*** 'RCLI went'⇒'left' had the word *foi* lemmatized as *ser*. This gave rise to the proposition of the spurious candidates *ser bem* 'be well' and *se ser* 'RCLI be'.

10. Related Work

Literal interpretation of utterances has been an important topic of debate in the philosophy of language. For instance, Recanati (1995) addresses the "standard model" by Grice (1989), which stipulates that "the interpretation of non-literal utterances proceeds in two stages: [a] the hearer computes the proposition literally expressed by the utterance; [b] on the basis of this proposition and general conversational principles, he or she infers what the speaker really means". Recanati (1995) further refutes the Gricean model by showing that, while non-literal interpretations presuppose literal ones, the latter are not necessarily processed before the former. This work does not explicitly address MWEs (i.e. expressions in which non-literal interpretations are conventionalized) but the proposed models of utterance interpretation (the *accessibility-based serial model*, in which only the most accessible interpretation is processed, and the *parallel model*, in which several sufficiently accessible interpretations are processed in parallel) seem applicable to MWEs, too.

Literal occurrences of MWEs, often called their literal readings or literal meanings, have also received a considerable attention from both linguistic and computational communities. From the psycholinguistic viewpoint, Cacciari and Corradini (2015) put special interest on the interplay between literal and idiomatic readings, as well as their distributional and statistical properties, when discovering how idioms are stored and processed in the human mind. Popiel and McRae (1988) collect ratings of frequency and familiarity for literal and figurative interpretations of 30 different idiomatic expressions in English. They find out that figurative interpretations obtain higher rankings in both aspects than literal interpretations. These results are further corroborated by Geeraert et al. (2018), who study the acceptability of lexical variation in VMWEs through rating and eye-tracking experiments. Judges are presented with sentences containing LOs and IOs of a VMWE with more or less variation. They judge the acceptability of the sentences, and at the same time the fixation duration is measured by eye tracking. The results show, in particular, that sentences with LOs are less acceptable than those with IOs, although the fixation duration for the former is shorter than for the latter. Overall, speakers do not feel comfortable with LOs. These results seem consistent with our quantitative analysis showing that LO are rare in our corpora across typologically different languages.

As to linguistic modelling, links between LOs and IOs are used by Sheinfux et al. (2019) to propose a novel typology of verbal idioms. It relies on figuration (the degree to which the idiom can be assigned a literal meaning) and transparency (the relationship between the literal and idiomatic reading). In *transparent figurative* idioms, the

relationship between the literal and the idiomatic reading is easy to recover (*to saw logs* 'snore'). In *opaque figurative* idioms, the literal picture is easy to imagine but its relationship to the idiomatic reading is unclear (*to shoot the breeze* 'chat'). Finally, in *opaque non-figurative* idioms, no comprehensible literal meaning is available, notably due to cranberry words which have no status as individual lexical units (*to take umbrage* 'to feel offended'). Their study also argues that the links between LOs and IOs can indicate which morphosyntactic variations are allowed or prohibited for some idioms.[32] Namely, transparent figurative idioms exhibit more flexibility than opaque figurative ones, because, in the former, the speakers can more easily relate to individual components and transpose their literal properties to the metaphoric level.

LOs and IOs were also addressed in the context of syntactic modelling by formal grammars. The challenge is to account for the difference between LOs and IOs when their syntax is identical. Abeillé and Schabes (1989) show how this problem can be elegantly solved by Lexicalized Tree-Adjoining Grammars containing a finite set of elementary (initial or auxiliary) trees, each of which has at least one lexicalized element. MWEs are represented as special kinds of elementary trees in which heads are made out of several lexical items that need not be contiguous. During parsing, a sentence can be derived by combining elementary trees via substitution (inserting an elementary tree at a non-terminal leaf) or adjunction (inserting an elementary tree at a non-terminal internal node), which yields a derived tree (the syntactic structure of the sentence) and a derivation tree (showing which elementary trees have been combined and how). While parsing ambiguous expressions (e.g., *he* ***kicked the bucket***), the idiomatic and the literal occurrences obtain the same derived trees, but the derivation trees differ. Accordingly, the idiomatic semantics stems from direct attachment of lexical items in the elementary trees, while the literal compositional semantics is a product of substitution (of non-terminal nodes with lexicon items). Lichte and Kallmeyer (2016) go even further and show how LTAGs combined with frame semantics can be used to model the LO-IO ambiguity only in the semantics. Here, derived trees and derivation trees remain identical across readings.

The LO-IO ambiguity is also considered a major challenge in computational processing of MWEs (Constant et al., 2017). This survey notably offers a state of the art in MWE identification, which is modelled by some approaches as a word sense disambiguation (WSD) problem: candidate expressions are extracted beforehand and then they are to be classified as literal or idiomatic. For example, Hashimoto and Kawahara (2008) deal with the ambiguity between literal and idiomatic interpretations of Japanese MWEs in a supervised WSD framework. The features, fed to a binary SVM classifier, account mainly for the morphosyntactic properties of the candidate MWEs, as well as for the lemmas, POS and domains of the words surrounding the them.

Fazly et al. (2009) use unsupervised MWE identification based on statistical measures of lexical and syntactic flexibility of MWEs. They draw upon the assumption

[32]Similar conclusions are drawn by Pausé (2017) from a corpus study of French VMWEs.

that usages in the canonical forms for a potential idiom are more likely to be IOs, and those in other forms are more likely to be LOs. There, the notion of an LO seems to have a much larger scope than in our approach: it notably includes variants stemming from replacement of lexicalized components by automatically extracted similar words, e.g., *spill corn* vs. ***spill** the **beans***. The test data is restricted to the 28 most frequent verb-object pairs and their manually validated IOs and LOs, i.e., COs are excluded from performance measures (unlike in our approach). Their precision and recall in LO identification range from 0.18 to 0.86 and from 0.11 to 0.61, respectively. These results are hard to compare to ours (Table 6), due to the very different understanding of the task and its experimental settings.

Peng et al. (2014) propose another approach to automatically classify LOs and IOs based on bag-of-words topic representations for 1–3 paragraphs containing the candidate phrase. Peng and Feldman (2016) further show how the same problem can be addressed via distributional semantics, where the semantics of a candidate expression, and of its component words, can be represented by their context vectors. In the same vein, Köper and Schulte im Walde (2016) automatically classify German particle verbs into literal or idiomatic by relying, notably, on distributional vectors (e.g. *aus-klingen* 'out-sound'⇒'end') and of their base verbs (e.g. *klingen* 'sound'). Other features, like abstractness of the context words, draw upon the hypothesis that idiomatic particle verbs are more likely to occur with abstract subjects or complements.

Distributional semantics also proves useful in the related task of predicting the semantic compositionality of an expression. Note that subtle links exist between idiomaticity and semantic non-compositionality. On the one hand, the LO-IO opposition is a dychotomy, and as such it did not seem problematic to apply in our corpus annotation experiments. On the other hand, idiomaticity usually stems from non-compositional semantics but this non-compositionality is known to be a matter of scale rather than a binary phenomenon. Estimating the *degree of (non-)compositionality* in MWEs is a convincing showcase for distributional semantics, where it is modelled via the degree of (non-)compositionality of the context vectors of their component words (see e.g., Katz and Giesbrecht 2006).

We are aware of only two previous works, our own, where the LO phenomenon was assessed in quantitative terms. In Waszczuk et al. (2016), we estimate the idiomaticity rate of Polish verbal, nominal, adjectival, and adverbial MWEs at 0.95, which confirms our current results also with respect to non-verbal VMWE categories. More importantly, this work also shows that the high idiomaticity rate can speed up parsing, if appropriately taken into account by a parser's architecture. Further, in Savary and Cordeiro (2018) we pave the way towards this article, by making the first attempt towards defining the notion of LO, and by estimating the idiomaticity rate of Polish VMWEs (at 0.98) on a smaller corpus.

Several datasets containing IO/LO annotations of MWEs were developed in the past. The dataset of Polish IOs and LOs created by us for the Savary and Cordeiro

(2018) publication, is openly available[33] and contains over 3,000 IOs, 72 LOs and 344 COs. The dataset of Tu and Roth (2011) consists of 2,162 sentences from the British National Corpus in which verb-object pairs formed with *do*, *get*, *give*, *have*, *make*, and *take* are marked as positive and negative examples of LVCs. Tu and Roth (2012) built a crowdsourced corpus in which VPCs are manually distinguished from compositional verb-preposition combinations, again for six selected verbs. Cook et al. (2008) present the VNC Tokens dataset, containing almost 3,000 occurrences of 53 Verb+Noun combinations in direct object relation, annotated as literal or idiomatic. In all, only 18% of all combinations were annotated as literal, which is roughly consistent with our study. Hashimoto and Kawahara (2008) offer a Japanese counterpart of these resources, with 146 idioms and over 102,000 example sentences. Sentences were automatically preselected in a corpus if they contained occurrences of the components of a reference MWE, and if the dependencies between those components were "canonical". This probably means that syntactic variability in LOs is underrepresented in this dataset. The authors mention that "some idioms are short of examples", which is corroborates our high idiomaticity rate results in another, typologically different, language. Our resource, described in this article, has a larger scope than these previous datasets: we address 5 languages from 5 language genera, and we cover VMWEs of unrestricted syntactic structures and lexical choices. The corpus is available under open licenses.

Let us finally mention datasets which provide human annotation of IO/LO candidates in a finer framework where semantic compositionality is estimated on a multivalued scale. Bott et al. (2016) offer such a resource for German VPCs, and Ramisch et al. (2016) for English, French and Portuguese Noun-Noun and Adjective-Noun compounds. A review of such datasets can be found in Cordeiro et al. (2019).

11. Conclusions and future work

This article offers an in-depth study of the phenomenon of literal occurrences of verbal multiword expressions, as well as of their interactions with two closely related phenomena: idiomatic occurrences on the one hand, and coincidental occurrences on the other. We firstly propose formal definitions of these three bordering notions, which were missing in the literature so far. The definitions stipulate that LOs, and consequently also COs, should be understood not only in semantic but also in syntactic terms, which motivates their study in treebanks. We then propose a thorough methodology to quantitatively and qualitatively estimate the importance of LOs. It consists in: (i) heuristics for automatic extraction of LOs tuned towards high recall with reasonable precision, (ii) a VMWE-annotated reference corpus in 5 typologically different languages, and (iii) manual annotation based on detailed annotation guidelines designed as decision trees. The results of this annotation are openly available.[34]

[33] http://clip.ipipan.waw.pl/MweLitRead

[34] http://hdl.handle.net/11372/LRT-2966

They constitute a novel resource, given that previous datasets with IO-and-LO annotation were mostly dedicated to a selected language and MWE category.

We claim to have shown that LOs are ***rare birds*** 'exceptional individuals' in our corpus, both among VMWE tokens and types, in all five languages under study. When syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96%–98% of the cases, as formalized via the IdRate. These results are only slightly less consistent across VMWE types, and range from 90% in Basque VIDs to 100% in Greek LVCs. This is an important finding from the linguistic viewpoint, because most VMWE could potentially be used literally, but they are rarely so in our corpus. This fact is somehow surprising since local ambiguity is inherent to natural language and humans generally deal with it very efficiently. For instance, numerous single words exhibit both rich polysemy and high frequency, and listeners easily disambiguate them based on context. IO-LO ambiguity can also be easily solved by context in most cases, and yet LOs occur surprisingly infrequently. We put forward the explanation of this fact as an interesting research question.

Given the instances of LOs found in the corpus, we also perform their qualitative analysis. Namely, we explain the conditions under which LOs occur in various VMWE categories, whether cross-lingually or in a language-specific manner. We show examples of morphosyntactic constraints which VMWE impose and which, if known in advance, e.g., from VMWE lexicons, might help automatically distinguish IOs from LOs. These observation might help tune various MWE processing tools (e.g., via fine-grained feature engineering). We additionally point at correlations that exist between the syntactic structure of VMWEs and their capacity to exhibit LOs. For example, many LOs are triggered by those VMWEs in which a head verb governs a functional word only (IRVs, VPCs and VID with expletive pronouns or adverbs). As future work, we wish to further examine these interactions.

We also provide quantitative analyses of LOs from the viewpoint of NLP, where automatic MWE identification is a major challenge for semantically-oriented downstream applications. There, IOs are to be opposed not only to LOs but also to COs (in which the lexemes in focus do occur, but not in the right syntactic configuration). We show that the predominance of IOs in this case is strong for German, Greek and Polish, but weaker for Basque and Portuguese. We show examples of language-specific phenomena which contribute to this fact. We also briefly account for some types of lexical ambiguity which challenge automatic IO/LO/CO extraction methods, and make them highly dependent on the quality of the underlying morphosyntactic annotation.

To conclude, in spite of being rare birds, LOs do *cause a stir* 'incite trouble or excitement'. Firstly, the IO-LO opposition provides a stimulating background for psycholinguistics and language-modeling considerations, which yields interesting insights into human language. Second, the IO-LO ambiguity is considered one of the major challenges in the NLP and has attracted much attention from the community, given that it relates to tasks such as MWE identification. Thirdly, even if we have shown that the LO phenomenon is quantitatively much more modest than expected,

it is still important due to both cross-lingually valid and language-specific phenomena, which are both interesting and not trivial to capture.

Let us finally stress that this is one of the first and few attempts to approach the naturally occurring IO-LO ambiguity on a larger scale in a cross-linguistic setting. We hope that this will inspire subsequent work in a variety of topics, be it in theoretical linguistics, psycholinguistics or computational linguistics.

Acknowledgements

This work was supported by the IC1207 PARSEME (PARSing and Multi-word Expressions) COST action[35], by the French PARSEME-FR project (ANR-14-CERA-0001)[36] and by German CRC 991 grant from Deutsche Forschungsgemeinschaft (DFG)[37].

Bibliography

Abeillé, Anne and Yves Schabes. Parsing Idioms in Lexicalized TAGs. In Somers, Harold L. and Mary McGee Wood, editors, *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pages 1–9. The Association for Computer Linguistics, 1989. URL http://dblp.uni-trier.de/db/conf/eacl/eacl1989.html#AbeilleS89.

Baldwin, Timothy and Su Nam Kim. Multiword Expressions. In Indurkhya, Nitin and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition, 2010. ISBN 978-1-4200-8592-1.

Bott, Stefan, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. $G_h$ost-PV: A Representative Gold Standard of German Particle Verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*, Osaka, Japan, 2016.

Cacciari, Cristina and Paola Corradini. Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology*, 27(7):797–811, 2015. doi: 10.1080/20445911.2015.1049178. URL http://dx.doi.org/10.1080/20445911.2015.1049178.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Multiword Expression Processing: A Survey. *Computational Linguistics*, to appear, 2017.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. The VNC-Tokens Dataset. In *Proceedings of the Workshop on Multiword Expressions*, 2008.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 2019. doi: 10.1162/COLI_a_00341. (to appear).

[35] http://www.parseme.eu

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

[37] https://frames.phil.uni-duesseldorf.de/

Dryer, Matthew S. and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/.

El Maarouf, Ismail and Michael Oakes. Statistical Measures for Characterising MWEs. In *IC1207 COST PARSEME 5th general meeting*, 2015. URL http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103, 2009. doi: 10.1162/coli.08-010-R1-07-048. URL https://doi.org/10.1162/coli.08-010-R1-07-048.

Geeraert, Kristina, R. Harald Baayen, and John Newman. "Spilling the bag" on idiomatic variation. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 1–33. Language Science Press., Berlin, 2018. doi: 10.5281/zenodo.1469551.

Grice, Herbert Paul. *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass., 1989.

Hashimoto, Chikara and Daisuke Kawahara. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001. Association for Computational Linguistics, 2008. URL http://aclweb.org/anthology/D08-1104.

Inurrieta, Uxoa, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, and Inaki Alegria. Verbal Multiword Expressions in Basque Corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 86–95, 2018.

Katz, Graham and Eugenie Giesbrecht. Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July 2006. URL http://www.aclweb.org/anthology/W/W06/W06-1203.

Köper, Maximilian and Sabine Schulte im Walde. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California, 2016. URL http://www.aclweb.org/anthology/N16-1039.

Lichte, Timm and Laura Kallmeyer. Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In Piñón, Christopher, editor, *Empirical Issues in Syntax and Semantics 11*, pages 111–140, 2016. URL http://www.cssp.cnrs.fr/eiss11/.

Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze. Preface. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin, 2018. ISBN 978-3-96110-123-8. doi: 10.5281/zenodo.1469527.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection.

In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation* , LREC 2016, pages 1659–1666. European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1. 23-28 May, 2016.

Patejuk, Agnieszka and Adam Przepiórkowski. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2018. (263 pages).

Pausé, Marie-Sophie. *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*. PhD thesis, Université de Lorraine, Nancy, France, 2017.

Peng, Jing and Anna Feldman. Automatic Idiom Recognition with Word Embeddings. In *SIMBig (Revised Selected Papers)*, volume 656 of *Communications in Computer and Information Science*, pages 17–29. Springer, 2016.

Peng, Jing, Anna Feldman, and Ekaterina Vylomova. Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1216.

Popiel, Stephen J. and Ken McRae. The figurative and literal senses of idioms, or all idioms are not used equally. *Journal of Psycholinguistic Research*, 17(6):475–487, Nov 1988. ISSN 1573-6555. doi: 10.1007/BF01067912. URL https://doi.org/10.1007/BF01067912.

Przepiórkowski, Adam, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. Phraseology in two Slavic Valency Dictionaries: Limitations and Perspectives. *International Journal of Lexicography*, 30(1):1–38, 2017.

Ramisch, Carlos, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany, 2016. ACL. doi: 10.18653/v1/P16-2026. CORE2018 rank: A*. https://aclweb.org/anthology/P16-2026.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-4925.

Recanati, François. The alleged priority of literal interpretation. *Cognitive Science*, 19:207–232, 1995. URL https://jeannicod.ccsd.cnrs.fr/ijn_00000181.

Savary, Agata and Silvio Cordeiro. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16), Jan 2018, Prague, Czech Republic*, pages 64 – 72, Prague, Czech Republic, Jan. 2018.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the EACL'17 Workshop on Multiword Expressions*, 2017.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Sla vomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Lie bes kind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Fe derico Sangati, Ivelina Stoyanova, and Veronika Vincze. PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin, 2018. ISBN 978-3-96110-123-8. doi: 10.5281/zenodo.1469527.

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. Verbal MWEs: Idiomaticity and flexibility. In Parmentier, Yannick and Jakub Waszczuk, editors, *Representation and Parsing of Multiword Expressions*, pages 5–38. Language Science Press, Berlin, 2019.

Tu, Yuancheng and Dan Roth. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, pages 31–39. Association for Computational Linguistics, June 2011. URL `http://www.aclweb.org/anthology/W11-0807`.

Tu, Yuancheng and Dan Roth. Sorting out the Most Confusing English Phrasal Verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 65–69. Association for Computational Linguistics, 2012. URL `http://dl.acm.org/citation.cfm?id=2387636.2387648`.

Waszczuk, Jakub, Agata Savary, and Yannick Parmentier. Promoting multiword expressions in A* TAG parsing. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 429–439, 2016. URL `http://aclweb.org/anthology/C/C16/C16-1042.pdf`.

Appendix: VMWEs with the highest extended literality rate and frequency of literal occurrences

VMWE	ELR	VMWE	Freq.
ausbauen 'dismount'⇒'enlarge'	0.8	*abgeben* 'give away'⇒'loose'	5
abwehren 'repel'⇒'repel'	0.67	*der heissen* 'its name is'⇒'it means that'	4
ansteigen 'increase'⇒'increase'	0.67	*ausbauen* 'dismount'⇒'enlarge'	4
einleiten 'lead in'⇒'initiate'	0.67	*umstellen* 'surround'⇒'rearrange'	3
sehen an 'watch'⇒'consider'	0.67	*gewachsen sein* 'be grown'⇒'withstand'	3
abgeben 'give away'⇒'loose'	0.625	*gehen weiter* 'go further'⇒'continue'	3
abgegeben (part.) 'give away'⇒'loose'	0.6	*abgegeben (part.)* 'give away'⇒'loose'	3
gewachsen sein 'be grown'⇒'withstand'	0.6	*sehen an* 'watch'⇒'consider'	2
umstellen 'surround'⇒'rearrange'	0.6	*recht haben* 'have the right'⇒'be right'	2
abgestellen (part.) 'park'⇒'switch off'	0.5	*nehmen ab* 'take off'⇒'decrease'	2

Table 7. VMWEs with the highest ELR and LO frequency in German

VMWE	ELR	VMWE	Freq.
τα βάζω 'them put'⇒'to be against'	0.83	*τα ρίχνω* 'them pour'⇒'to blame'	5
εκδίδω ανακοίνωση 'issue announcement' ⇒ 'to announce'	0.83	*εκδίδω ανακοίνωση* 'issue announcement' ⇒ 'to announce'	5
τα ρίχνω 'them throw'⇒'to blame'	0.83	*τα ρίχνω* 'them throw'⇒'to blame'	5
έχω στο χέρι 'have in the hand' ⇒ 'to have control over'	0.75	*τα παίρνω* 'them take'⇒'to become furious'	4
ανοίγω την πόρτα 'open the door'⇒'to allow'	0.67	*το ίδιο κάνει* 'does the same'⇒'never mind'	4
βρίσκομαι σε θέση 'be in position'⇒'to be able to'	0.6	*έχω στο χέρι* 'have in the hand'⇒'to have control over'	3
το ίδιο κάνει 'does the same'⇒'never mind'	0.57	*βρίσκομαι σε θέση* 'be in position'⇒'to be able to'	3
τα παίρνω 'them take'⇒'become furious'	0.5	*ανοίγω την πόρτα* 'open the door'⇒'to allow'	2
δίνω δύναμη 'give power'⇒'to empower'	0.5	*έχω υποχρέωση* 'have obligation'⇒'to be obliged'	2
κρατώ στο χέρι μου 'keep in the hand' ⇒ 'to have control over'	0.5	*παίρνω θέση* 'take seat'⇒'to express my opinion'	2

Table 8. VMWEs with the highest ELR and LO frequency in Greek

VMWE	ELR	VMWE	Freq.
ate ireki 'open door'⇒'to open sth up to sth'	0.75	*berdin izan* 'be equal'⇒'not to mind'	11
atzetik ibili 'walk behind'⇒'to be behind'	0.67	*alde izan* 'be side'⇒'to be in favour'	7
forma hartu 'take form'⇒'to take shape'	0.67	*gauza izan* 'be thing'⇒'to be able'	7
berdin izan 'be equal'⇒'not to mind'	0.55	*balio izan* 'have value'⇒'to be useful'	5
adar jo 'play horn'⇒'to be kidding'	0.5	*jokoan izan* 'be in game'⇒'to be at stake'	5
ate zabaldu 'open door'⇒'to open sth up to sth'	0.5	*laguntza eman* 'give help'⇒'to help'	4
hitz hartu 'take word'⇒'to take sb at sb's word'	0.5	*nabari izan* 'be evident'⇒'to show'	4
kantu egin 'do song'⇒'to sing'	0.5	*ate ireki* 'open door'⇒'to open st up to st'	3
nabari izan 'be evident'⇒'to show'	0.5	*behar izan* 'have need'⇒'to need'	3
pisu ukan 'have weight'⇒'to have an influence'	0.5	*buru ukan* 'have head'⇒'to be intelligent'	3

Table 9. VMWEs with the highest ELR and LO frequency in Basque

VMWE	ELR	VMWE	Freq.
mieć we krwi 'to have in blood'	0.8	*być w stanie* 'be in state'⇒'be able'	11
zerwać się 'break RCLI'⇒'get up abruptly' ⇒ 'have sth as an innate capacity'	0.8	*mieścić się* 'hold RFLI'⇒'fit'	7
dzielić się 'divide RCLI'⇒'share'	0.78	*znaleźć się* 'find RCLI'⇒'be'	5
oprzeć się 'lean RCLI'⇒'resist'	0.71	*oprzeć się* 'lean RCLI'⇒'resist'	5
dopuszczać się 'allow RCLI'⇒'perpetrate'	0.67	*zerwać się* 'break RCLI'⇒'get up abruptly'	4
prosić się 'ask RCLI'⇒'call for'	0.67	*mieć we krwi* 'have in blood' ⇒ 'have sth as an innate capacity'	4
doprowadzić do zatrzymania 'lead to arresting' ⇒ 'cause arresting'	0.5	*przedstawiać się* 'present RCLI'⇒'look'	3
mieć pewność 'have certainly'⇒'be sure'	0.5	*mieć udział* 'have share'⇒'take part'	3
mieć udział 'have share'⇒'take part'	0.5	*mieć się* 'have RCLI'⇒'be'	3
mieć wynik 'have result'	0.5	*znać się* 'know RCLI'⇒'be an expert'	2

Table 10. VMWEs with the highest ELR and LO frequency in Polish

VMWE	ELR	VMWE	Freq.
formar se 'form RCLI'⇒'graduate'	0.8	*já era* 'already was.3SG.IPRF'⇒'it is over'	68
ver se 'see RCLI'⇒'find oneself (in a situation)'	0.79	*dever se* 'owe RCLI'⇒'be due to'	18
posicionar se 'position RCLI'⇒'express an opinion'	0.67	*ter filho* 'have child'⇒'give birth'	15
quero ver 'want.1SG.PRS to.see'⇒'I doubt / I dare'	0.64	*ser a vez* 'be the time'⇒'be someone's turn'	14
ter filho 'have son'⇒'to have a son'	0.62	*ver se* 'see RCLI'⇒'find oneself (in a situation)'	11
fazer cobertura 'make news.coverage'⇒'cover (news)'	0.5	*dizer se* 'say RCLI'⇒'claim to be'	11
fazer placar 'make scoreboard'⇒'score goals'	0.5	*querer.1PS.PRS ver* 'I.want to.see'⇒'I doubt'	9
ganhar números 'gain numbers'⇒'increase in numbers'	0.5	*ir.IMP lá* 'go there'⇒'come on!'	6
morrer em a praia 'die on the beach'⇒'fail at the last stage'	0.5	*querer dizer* 'want to.say'⇒'mean'	4

Table 11. VMWEs with the highest ELR and LO frequency in Portuguese

Address for correspondence:
Agata Savary
agata.savary@univ-tours.fr
University of Tours, IUT of Blois, 3 place Jean-Jaurès, 41000 Blois, France