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.

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.