# Translation of Spanish Multiword Expressions into Basque: linguistic analysis and detection experiment*

## *Traducción de Expresiones Multipalabra del castellano al euskera: análisis lingüístico y experimento de detección*

**Uxoa Iñurrieta**
IXA NLP group, University of the Basque Country
usoa.inurrieta@ehu.eus

**Resumen:** Este artículo presenta un análisis lingüístico sobre la traducción de expresiones multipalabra entre castellano y euskera, y describe un experimento de detección basado en información lingüística de dichas expresiones en castellano. Todos nuestros resultados se encuentran disponibles en una base de datos pública: *Konbitzul*.
**Palabras clave:** Expresiones Multipalabra, Traducción Automática, euskera, castellano

**Abstract:** This paper describes a linguistic analysis of the translation of MWEs between Spanish and Basque, as well as a detection experiment of Spanish MWEs based on linguistic information. All of our results are available in a public database: *Konbitzul*.
**Keywords:** Multiword Expressions, Machine Translation, Basque, Spanish

## *1 Introduction*

While Multiword Expressions (MWEs) are constantly used in both oral and written texts, they do not usually follow the common grammatical and lexical rules of languages. Sometimes, the way they are formed is atypical; at other times, their usage in a sentence is non-standard; and sometimes, their meaning is not even what one would expect from the separate meanings of the words that form them. All those characteristics make this kind of word combination problematic for Natural Language Processing (NLP), especially when it comes to multilingual systems like Machine Translation (MT).

(1) English: *to kick the bucket*
= to die
Spanish: *estirar la pata*
'to stretch the leg'
= to die
Basque: *azken arnasa eman*
'to give one's last breath out'
= to die

The aim of our research work is to establish the linguistic basis for an appropriate treatment of MWEs in a Rule-Based Machine Translation (RBMT) system, which translates from Spanish into Basque. To that end, we undertook a linguistic analysis of Basque and Spanish noun+verb combinations, to examine both the features of the combinations and how they are translated (see sections 3.1 and 3.2). Then we did an experiment to test to what extent the linguistic data obtained from that analysis improved the identification of Spanish MWEs (see section 3.4), and we found that it is indeed very helpful.

## *2 Multiword Expressions in Machine Translation: related work*

MWEs are lexical items that can be decomposed into multiple lexemes and display some kind of idiomaticity (Sag et al., 2002). This particular feature makes them very challenging for NLP in general, but even more when the language to be processed has a rich morphology (Alegria et al., 2004).

Translating these kinds of word combinations is not an easy task either, as they are very asymmetric cross-linguistically (Simova

* Uxoa Iñurrieta's work within the SKATeR project (TIN2012-38584-C06-02) is funded by a PhD scholarship from the Ministry of Economy and Competitiveness (BES-2013-066372), and is supervised by Itziar Aduriz and Kepa Sarasola. We thank Arantza Díaz de Ilarraza and Gorka Labaka for their advice.

y Kordoni, 2013). Therefore, difficulties get bigger for multilingual systems like MT, in which it is essential to treat those constructions separately. Most research work focuses on Statistical Machine Translation (SMT) systems (Seretan, 2013), but other experiments have also been done to integrate MWE knowledge into RBMT systems (Wehrli et al., 2009).

It has been noted that even the simplest treatment of MWEs can improve MT systems (Copestake et al., 2002). However, more sophisticated strategies produce better results, and the tendency of recent years has been to combine computational methods with linguistic information (Dubremetz y Nivre, 2014; Seretan, 2013).

Concerning Basque, a few research has been conducted on MWE processing (Gurrutxaga y Alegria, 2011; Urizar, 2012), but it does not take multilingual systems into account. It is thus our objective to help handling this problem.

## 3 Noun+verb combinations in Matxin, a Spanish-Basque RBMT system

*Matxin* is an RBMT system which translates from Spanish into Basque (Mayor et al., 2011). The method it currently uses to treat MWEs is quite basic, and, although it works very well for some kinds of constructions, it has two important shortcomings: (1) on the one hand, it is restricted to a list of word combinations obtained from bilingual dictionaries, and (2) on the other hand, it works only when the words that constitute the MWE are sequential and use the same word form (except for the verb, which is detected even if it is inflected).

(2) Spanish: *contraer matrimonio*
contract.INF marriage
'to get married'
MT: *ezkontza uzkurtu*
marriage.ABS shrink.INF
'to shrink marriage'
Correct: *ezkondu*

When we want to translate a given MWE but it is not included in the bilingual dictionary of the MT system, the latter translates each word with its most common equivalent, which usually leads to incorrect results (see example 2). At other times, the MWE we want to translate is included in the dictionary, but the system fails to detect it and thus produces a wrong translation. The combination *llevar a cabo*, for example, is well translated when there is nothing between the verb and the prepositional phrase (see example 3), but not when the combination is non-sequential (see example 4).

(3) Spanish: *Llevar a cabo algo.*
take.INF to cape something
'To conduct something.'
MT: *Zerbait burutzea.*
Correct: *Zerbait burutzea.*
something.ABS conduct.INF
'to conduct something'

(4) Spanish: *Llevar algo a cabo.*
take.INF something to cape
'To conduct something.'
MT: *Zerbait eramatea kabora.*
something.ABS take.INF corporal.ALL
'To take something to the corporal.'
Correct: *Zerbait burutzea.*
something.ABS conduct.INF
'to conduct something'

As previously mentioned, our aim is to provide the linguistic information necessary to process MWEs adequately. Therefore, we analysed a set of Basque and Spanish word combinations and their possible translations, and, after getting a general idea of the differences and similarities between the MWEs in both languages, we selected some of them to look at those features that could determine what kind of treatment they require (see sections 3.1 and 3.2). Then, we carried out an evaluation to see whether or not our information was useful (see section 3.4).

All of our results are collected in a public database, *Konbitzul*[1], which is available online and allows users to search for the combinations we analysed, along with their possible translations and other linguistic data (see section 3.3).

### 3.1 Dictionary-based linguistic analysis

Although it was evident to us that parallel corpora were the most adequate basis for a linguistic analysis of MWEs, we considered it worthwhile to take a look at bilingual

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

dictionaries first. We used the Elhuyar dictionaries[2] for that purpose, 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 the Basque combinations were constituted only of a noun and a verb. However, it is important to note that Basque is an agglutinative language (Laka, 1996) and thus it constructs phrases by attaching elements to the endings of phrases, which means that the nouns that are part of MWEs can also have different cases and postpositions (see example 5). Spanish combinations, on the other hand, were made of verbs and nouns, but could also contain determiners and/or prepositions in-between (see example 6).

(5) Absolutive: *denbora galdu*
time.ABS lose.INF
'to waste time'
Locative: *jokoan jarri*
game.LOC put.INF
'to put (st) at risk'
Allative: *aurrera egin*
front.ALL do.INF
'to go ahead'
Instrumental: *gogoz jan*
attitude.INS eat.INF
'to eat hungrily'
Ablative: *aurretik erabaki*
front.ABL decide.INF
'to predetermine'

(6) V+D+N: *hacer un favor*
do.INF a favor
'to do a favor'
V+N: *tener afecto*
have.INF affection
'to have affection'
V+P+N: *saber de memoria*
know.INF by memory
'to know by heart'
V+P+D+N: *dejar a un lado*
leave.INF to a side
'to leave aside'

First of all, we examined the MWEs in both languages separately, without taking their translations into account. Although their morphological features were very different cross-linguistically, it was interesting to note that a large number of the verbs that were part of the MWEs were very common in both languages, and that the most frequent ones were equivalent to each other: *egin – hacer* ('do'), *izan – ser/estar/tener* ('be/have'), *eman – dar* ('give'), *hartu – tomar* ('take') and so on.

Secondly, we analysed how those combinations were translated, and we verified that asymmetry is a very common phenomenon when translating MWEs between Basque and Spanish. As a matter of fact, among the Spanish translations of the Basque combinations, 58.07% were single verbs (see examples 7, 8 and 9).

(7) Basque: *lan egin*
work.ABS do.INF
'to work'
Spanish: *trabajar*
'to work'

(8) Basque: *hitz egin*
word.ABS do.INF
'to talk'
Spanish: *hablar*
'to talk'

(9) Basque: *atseden hartu*
rest.ABS take.INF
'to take a rest'
Spanish: *descansar*
'to rest'

In addition, we examined further the *noun + verb* constructions which were translated with *verb + (determiner) + (preposition) + noun* constructions and vice versa. We observed whether the nouns and the verbs in the combinations were equivalent to each other, and we found that there were just a few cases in which both the noun and the verb in one language were equivalent to the noun and the verb in the other (28.01% of the Basque into Spanish translations, and 21.79% of the Spanish into Basque ones).

### 3.2 Corpus-based analysis

The second step of our linguistic analysis was to contrast the obtained results with comparable corpora. The corpus we used was made up of 491,853 sentences taken from Spanish-into-Basque translations, within which we found only 200 out of the 2,650 Spanish combinations we had previously analysed (see section 3.1).

[2] `http://hiztegiak.elhuyar.org`

However, as we searched only for the noun and verb lemmas of the MWEs (accepting any preposition and/or determiner in-between), we could also obtain new variants of the 200 Spanish MWEs (see example 10), as well as a few additional combinations. Furthermore, while those combinations had only 385 Basque equivalents in the dictionary, they were translated in as many as 1,641 ways in the corpus, which made it possible for us to add new information to our database.

(10) Combination: *fijar un plazo*
'to fix a deadline'
Variant 1: *fijar el plazo*
'to fix the deadline'
Variant 2: *fijar plazos*
'to fix deadlines'

Then, as we believe that linguistic data specific to MWEs is essential for their processing, we decided to select the most frequent combinations and look at some features that could be important for their computational treatment: prepositions, determiners, definiteness and number of the nominal phrases, variability of the nominal phrases, order changeability, etc.

We examined the 121 combinations that were repeated more than five times in the corpus, and we sorted them into three groups:

- Fixed combinations, which were always sequential and used the same word form. There could not be any element between the components of the MWE, and the order could not be changed. For example: *dar paso* ('give way').
- Semi-fixed combinations, which had some kind of lexical or grammatical restriction but did not always use the same word form and order. Moreover, they often accepted modifiers or other elements between the words that constituted the MWE. For instance: *impartir clases* ('give classes').
- Free combinations, which did not seem to have any special restriction to take into account: *cubrir de agua* ('cover with water').

### 3.3 The *Konbitzul* database

As mentioned at the beginning of section 3, all of the results obtained from our linguistic analysis are available in a public database: *Konbitzul*. The interface gives users the possibility to search for a given combination according to various criteria: the verb, the noun, the structure of the combination etc. As a result, the database shows a list of combinations that match those criteria, along with their possible translations and other linguistic data.

For example, figure 1 shows the result of searching for verb+noun combinations that contain the verb *tender* ('tend' or 'stretch'). Three combinations are found in the database: *tender la mano* ('reach out'), *tender las velas* ('stretch the sails'), and *tender puentes* ('build bridges').

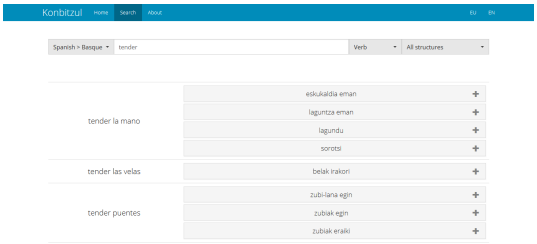


Figure 1: Result of the search for *tender*

On the other hand, each translation has a [+] icon on which users can click to see additional linguistic information. In the case of *laguntza eman* ('to help'), the information in the database (see figure 2) is the following:

- The Spanish combination is made up of a verb, a determiner and a noun, which is singular and definite.
- The Basque combination is made up of a noun, which is in the absolutive case, and a verb.
- The nouns and the combinations (*tender* and *eman*) are not equivalent to each other, and neither are the verbs (*mano* and *laguntza*).

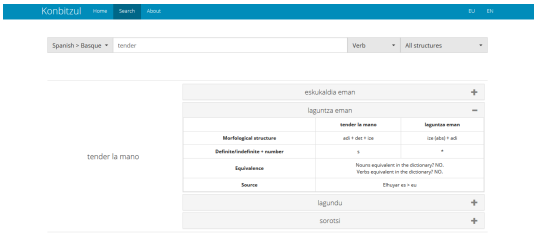


Figure 2: Linguistic information in *Konbitzul*

### 3.4 Spanish MWE detection based on linguistic information

After having added linguistic information to the Spanish combinations, we wanted to test whether that data really did improve the detection process. We left free combinations aside, and we used a 15,182,385 sentence corpus for our experiment, in which we compared two identification methods: the old one, used by *Matxin* (see section 3), and ours, based on linguistic information.

As table 1 shows, most of the MWEs were detected by both methods. However, the new one was able to identify 78,174 more combinations, increasing the number of detected MWEs by almost 25%.

| | |
|---|---|
| **Both methods** | 242,851 |
| **Old method only** | 517 |
| **New method only** | 78,174 |

Table 1: Detection experiment results

When evaluating the results manually, we found that, while the old method was slightly more accurate, the new one was also very precise (99.88% as opposed to 98.05%). This confirms that linguistic data specific to MWEs improves the detection of Spanish combinations, as it makes it possible to identify more combinations with very high accuracy.

## *4 Conclusions and future work*

The research work presented in this paper aims to analyse the translation of MWEs, in order to establish the linguistic grounds for their treatment in multilingual systems. First of all, we analysed verb+noun constructions in Spanish-Basque bilingual dictionaries linguistically, and confirmed that MWE translation is very asymmetric and thus needs a special treatment in NLP mboxapplications.

Secondly, we searched for those combinations in a parallel corpus. We selected the most frequent combinations in Spanish, and added linguistic data which could be important for their detection. We did an experiment to evaluate whether that information improved MWE identification, and found that the number of detected combinations rose by nearly 25% with very high accuracy (%98.05).

Our next purpose is to integrate that information into an MT system and explore what additional data is needed for an adequate translation into Basque. In addition, we would like to examine whether another kind of information could also improve MWE detection quality, like deep syntactic information or semantic data.

All the linguistic information achieved from this study is collected in a database, which is now available online: `http://ixa2.si.ehu.es/konbitzul`.

## *Bibliography*


Alegria, Inaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, y Ruben Urizar. 2004. Representation and treatment of multiword expressions in basque. En *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, páginas 48–55. Association for Computational Linguistics.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag, y Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability.

Dubremetz, Marie y Joakim Nivre. 2014. Extraction of nominal multiword expressions in french. *EACL 2014*, página 72.

Gurrutxaga, Antton y Inaki Alegria. 2011. Automatic extraction of nv expressions in basque: basic issues on cooccurrence techniques. En *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, páginas 2–7. Association for Computational Linguistics.

Laka, Itziar. 1996. A brief grammar of euskara, the basque language.

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

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

Seretan, Violeta. 2013. On collocations and their interaction with parsing and translation. En *Informatics*, volumen 1, páginas 11–31. Multidisciplinary Digital Publishing Institute.

Simova, Iliana y Valia Kordoni. 2013. Improving english-bulgarian statistical machine translation by phrasal verb treatment. En *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, France*.

Urizar, Ruben. 2012. Euskal lokuzioen tratamendu konputazionala. *Doktoregotesia, Informatika Fakultatea, UPV/EHU, Donostia*.

Wehrli, Eric, Violeta Seretan, Luka Nerima, y Lorenza Russo. 2009. Collocations in a rule-based mt system: A case study evaluation of their translation adequacy. En *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, páginas 128–135.