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

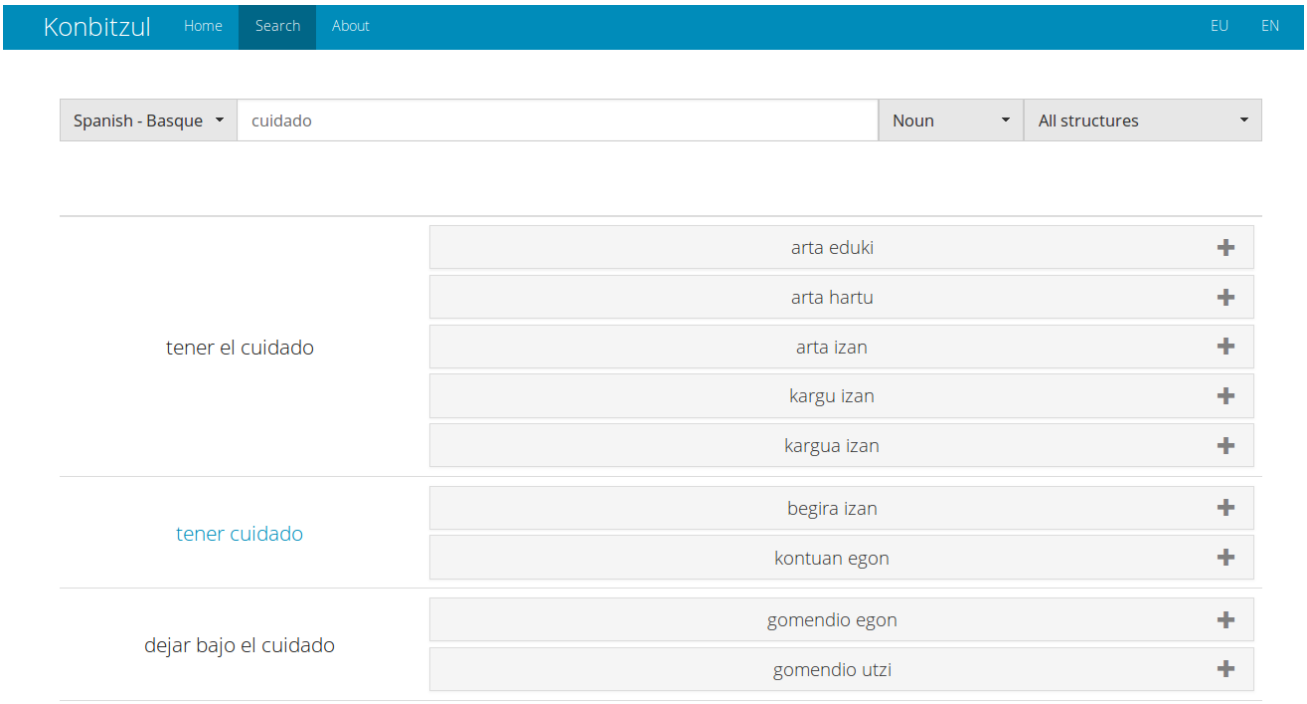

Figure 1: The Konbitzul database's interface. The noun *cuidado* (care) is searched, and three combinations are shown along with their possible translations: *tener el cuidado* (lit. have the care, 'be careful'), *tener cuidado* (lit. have care, 'be careful') and *dejar bajo el cuidado* (lit. leave under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

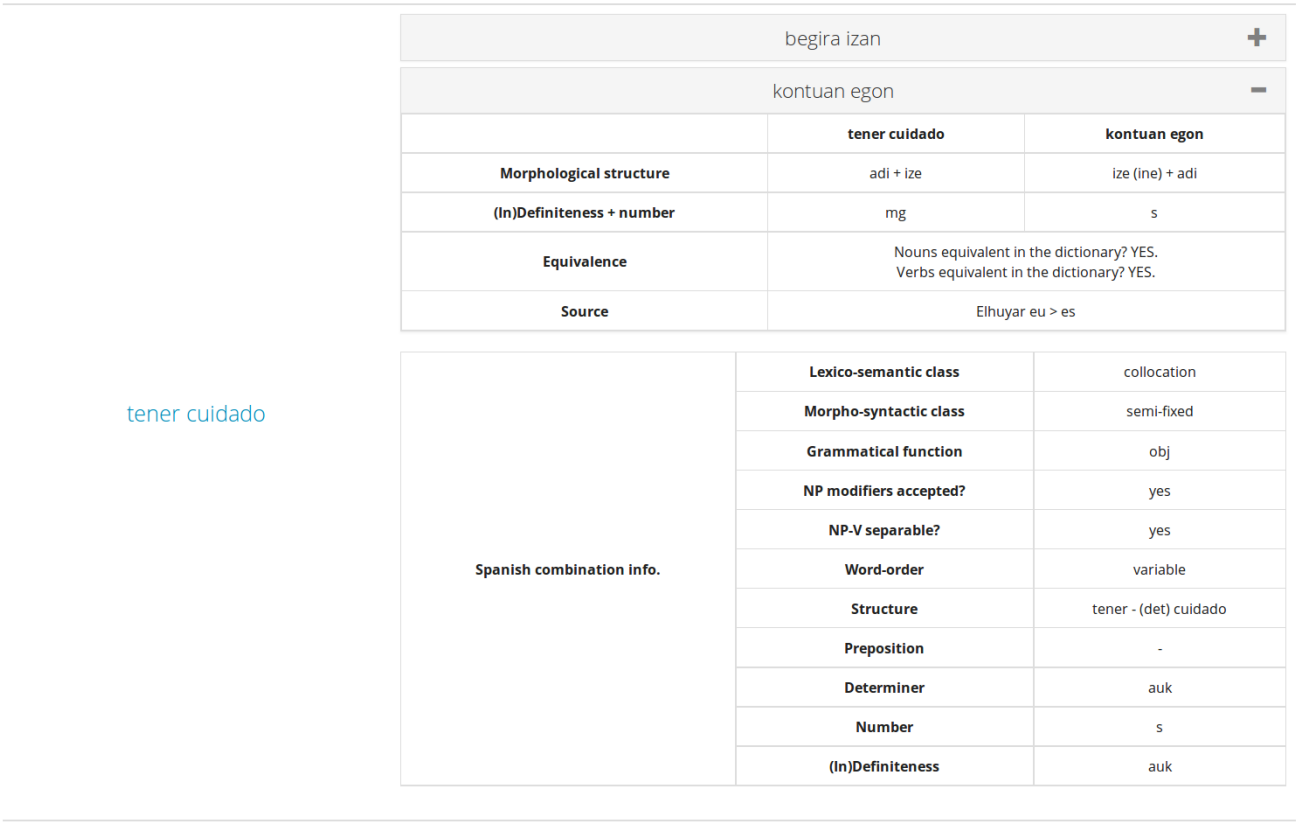

Figure 2: An example of how linguistic information is shown. Two tables are opened after clicking both on the entry *tener cuidado* and on the plus button besides the translation *kontuan egon* (lit. care-in be, 'be careful).

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.