Konbitzul: a database for Spanish-Basque verb+noun combination translation

(related to Working Groups 1, 2 and 3)

Uxoa Iñurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola

While Multiword Expressions (MWEs) are constantly used in both oral and written texts[1], they are very problematic for Machine Translation (MT) systems, which often fail to translate these kinds of word combinations correctly. There are two main challenges that must be addressed when facing these problems: on the one hand, the detection of MWEs in the source language, and on the other hand, their transfer into the target language.

Matxin[2] is an open-source Rule-Based Machine Translation (RBMT) system which translates Spanish into Basque. The method it currently employs to deal with MWEs is based on the *words-with-spaces* strategy, which consists in searching solely for adjacent word combinations and assigning a fixed equivalent to the whole expression. However, non-adjacent combinations are as frequent as the adjacent ones[3], and this approach does not allow us to find them and give them an adequate translation.

We undertook a comprehensive linguistic analysis of Spanish MWEs and the features that must be taken into account when translating them into Basque, and we created a public database with all the results: *Konbitzul*. We also carried out an experiment to test whether the detection of Spanish MWEs could be improved by combining linguistic data with chunking information and dependency parsing, which produced very positive results.

Linguistic analysis

The combinations we selected for our study were those consisting of a verb and a noun, as these kinds of expressions are usually very flexible when it comes to syntax and, therefore, very difficult to process. In the case of Spanish, many of the combinations also contained a preposition and/or a determiner in-between, and Basque nouns had many different cases and postpositions.

First of all, we extracted a list of verb+noun combinations from the Elhuyar dictionary[4]: 2,650 Spanish combinations (together with 6,587 Basque equivalents), and 2,954 Basque combinations (together with 6,392 Spanish equivalents). We analysed their **morphological features** to see to what extent their translation was irregular. As we had expected, we saw that the translation of Spanish verb+noun MWEs into Basque is indeed a very complex task, as only 48.54% of them were translated regularly, that is, substituting the verb with a verb, the noun with a noun, and the determiners and prepositions with the morphemes that are usually used for their translation (1). When it comes to Basque into Spanish translations, on the other hand, the irregularity was even more evident: 58.07% of the Basque combinations were not translated by a word combination but by a verb only (2), and 11.90% were combinations other than the verb+noun type.

(1) *hacer caso* (do attention) > *jaramon egin* (attention do), 'to pay attention'

(2) *lan egin* (work^{N} do) > *trabajar* (work^{V}), 'to work'

As a second step, we searched for the analysed verb+noun combinations in a Spanish-Basque bilingual corpus, and we selected the 150 most frequently-used ones to examine them further. We now looked at some relevant **syntactic features** of the Spanish MWEs, such as the

1 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. 1-15). Springer Berlin Heidelberg.

2 http://matxin.elhuyar.eus/

3 Vincze, O., & Ramos, M. A. (2013). Incorporating frequency information in a collocation dictionary: Establishing a methodology. *Procedia-Social and Behavioral Sciences*, *95*, 241-248.

4 http://hiztegiak.elhuyar.eus/

possibility of separating the elements of the combinations, changing the number and definiteness of the noun phrase, adding a modifier, or changing the order of the elements. Most of the combinations had a completely free syntax (3), and the major part of the rest also accepted some kind of variation (4), which, once again, justifies the need for a more sophisticated treatment of MWEs.

(3) Dar una clase (give a lecture), 'give a lecture'
dio clases, la clase que dio, dio una interesante clase...

(4) Llevar a cabo (take to cape), 'carry out, perform, conduct'
*llevarán algo a cabo, *a cabo algo llevarán...*

Improving MWE detection by using specific linguistic information

To test whether our linguistic data improved the verb+noun MWE detection process, we undertook an experiment in which we combined and compared three identification methods: (a) the old one, based on the *words-with-spaces* strategy, (b) a second one, based on our linguistic data and automatic chunking information, and (c) a third one, based on our linguistic data and automatic syntactic dependencies. A total of 433,092 MWEs were detected combining all three methods, and we evaluated the improvement made by (b) and (c) manually.

While method (a) was extremely precise (99.80%), it just searched for adjacent word combinations, and specific linguistic data allowed (b) and (c) methods to increase the number of identified expressions by 27.80%, with a slightly lower but still acceptable precision (see Table 1).

Method	MWEs detected	% of MWEs detected	Precision
B+C only	90,293	20.85%	96.60%
B only	17,828	4.12%	92.60%
C only	12,241	2.83%	83.20%

Table1: Results of the experiment on the detection of Spanish verb+noun MWEs

In conclusion, our linguistic data does help in the detection process of verb+noun MWEs, especially when it comes to non-adjacent word combinations.

The *Konbitzul* database

The *Konbitzul* database[5] collects all the information gathered from our linguistic analysis, and its 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, a list of combinations matching those criteria is shown, along with their possible translations and other linguistic data.

For example, if a user searches for combinations containing the verb *tender* (’tend’ or ’stretch’), three MWEs are found in the database: *tender la mano* (’reach out’), *tender las velas* (’stretch the sails’), and *tender puentes* (’build bridges’). Then, users can click on the [+] icon next to each translation to see additional linguistic information. In the case of *tender la mano > laguntza eman* ('give help'), for instance, the information in the database is the following: the Spanish combination is made up of a verb + determiner + noun (singular and definite); the Basque combination is made up of a noun (in the absolutive case) + verb; the nouns in the combinations (*tender* 'stretch' and *eman* 'give') are not equivalent to each other, and neither are the verbs (*mano* 'hand' and *laguntza* 'help').

5 http://ixa2.si.ehu.es/konbitzul