COMPUTATIONAL TREATMENT OF BASQUE MULTIWORD EXPRESSIONS

(WG1, WG2, WG3)
Ruben urizar
University of the Basque Country

This paper describes the representation of Basque multiword expressions (MWE) in a lexical database and their automatic processing. Being aware of the importance of recognizing and analyzing MWEs if we want any NLP tool to perform accurately, we took the challenge of developing a system for automatic processing of MWEs in Basque.
The definition of the term 'multiword expression' and the types of such MWEs to be treated in NLP may vary considerably depending on the purposes or "the depth of processing being undertaken" (Copestake *et al.,* 2002).

MWE sources

In our case, when deciding which Basque MWEs to treat, we mostly relied on lexicographers' expertise since we considered *lexicalized phrases* have a top priority for both lemmatizing and syntactic purposes. We mainly resorted to two sources. On the one hand, we made use of the *Statistical Corpus of 20th Century Basque*[1], a 4,7-million-word balanced corpus manually tagged. Among the MWEs lemmatized in the corpus, we selected those that occurred 10 times or more in it. On the other hand, we also collected all multiword subentries in the 2000 edition of the dictionary *Hiztegi Batua* (Euskaltzaindia, 2000). The final list amounted to 2,200 MWEs.

Formal description of multiword expressions

Then, to describe all these MWEs in the Lexical Database for Basque, EDBL (Aldezabal *et al.*, 2001), we worked out a single representation covering all types of MWEs ranging from fixed expressions to those of highest morphosyntactic flexibility. The purpose of the description is to formally encode all the possible surface realizations of each MWE.
The description of MWEs within EDBL includes, at least, three aspects:

1. their *composition*, i.e. which the components of the MWE are, whether each of them can be inflected or not, and which one-word lexical unit conveys the morphosyntactic information to the whole MWE
2. what we call the *surface realization*, that is, the order in which the components may occur in the text, the mandatory or optional contiguousness of components, and the inflectional restrictions applicable to each one of the components. According to these features, we use a formal description where different realization patterns may be defined for each MWE
3. their possible ambiguity, i.e. whether the sequence of words matching a given surface realization pattern (SRP) must be unambiguously analyzed as an instance of the MWE or, on the contrary, may be analyzed as separate words in some contexts.

In total, we used 177 different realization patterns and 145 inflection restrictions to describe the 2,200 expressions described in the lexical database.

Processing MWEs with HABIL

For the processing of the Basque MWEs we implemented HABIL, a tool that detects and analyzes MWEs, or candidate MWEs, based on the features previously described in the lexical database. The most important features of HABIL are the following:

- It deals with both contiguous and split MWEs.
- It takes into account all the possible orders of the components (SRP).
- It checks that inflectional restrictions are complied with.
- It generates morphosyntactic interpretations for the MWE.

HABIL takes as input the analyses of simple words given by the morphosyntactic analyser for Basque *Morfeus* (Aduriz *et al*., 1998).

[1] http://xxmendea.euskaltzaindia.eus/Corpus

If a MWE has been described as unambiguous, HABIL adds the interpretations corresponding to the expression and eliminates those belonging to the components.
When a MWE is ambiguous in a given surface realization pattern, HABIL adds the interpretations corresponding to the MWE without erasing those belonging to the components.

```
"<egin>"<1-2>"
   "egin" ADI SIN ADOIN NOTDEK @-JADNAG
   "egin" ADI SIN PART BURU NOTDEK @-JADNAG
   "egin" IZE ARR ABS MG @OBJ @PRED @SUBJ
   "lan egin" ADI ADK ADOIN NOTDEK mw1 @-JADNAG
   "lan egin" ADI ADK PART BURU NOTDEK mw1 @-JADNAG
   [...]
"<zuen>"
   "*edun" ADL B1 NOR NORK NR HURA NK HARK @+JADLAG
   "ukan" ADT PNT B1 NOR NORK NR HURA NK HARK @+JADNAG
   "zuek" IOR PERARR ZUEK GEN NUMP MUGM ZERO @ < IZLG @IZLG>
   [...]
"<lan>"<2-2>"
   "landu" ADI SIN ADOIN NOTDEK @-JADNAG
   "lan" IZE ARR ABS MG @OBJ @PRED @SUBJ
   "lan egin" ADI ADK ADOIN mw1 NOTDEK @-JADNAG
   "lan egin" ADI ADK PART BURU NOTDEK mw1 @-JADNAG
   [...]
```

For instance, in the sentence *egin zuen lan* '(s)he worked' the components of the MWE *lan egin* —'to work', lit. 'to do/make work'— are split and do not occur in the canonical order. In the figure above, each indented line represents a possible interpretation of the word forms in the sentence. We can see that the analyses corresponding to the MWE have been added (marked mw1) without eliminating the analysis of the components, since this verbal expression is considered ambiguous in this surface realization.

Disambiguation of ambiguous MWEs

At that stage, the ambiguous word combination has only been identified as a MWE candidate. Therefore, the next step is to determine in which contexts the candidate is actually a MWE and in which ones it is not.
For the identification of ambiguous MWEs, some authors (Baldwin and Kim, 2010) apply WSD techniques in which the literal and the MWE interpretation of a given word combination are considered to have different meanings and therefore a different context of usage.
But another approach (Li *et al*., 2003; Hashimoto *et al*., 2006) takes into account the morphosyntactic fixedness of MWEs. This approach is based on the premise that several MWEs only realize in some specific morphosyntactic configurations, and they undergo a more restricted variation than literal usages.
At present, we are developing a grammar based on the Constraint Grammar (CG) formalism (Karlsson et al., 1995; Tapanainen, 1996) for MWE ambiguity resolution. CG rules only make use of morphosyntactic information in the sentence context so as to remove or select some readings, add or replace tags... For our research, we have chosen the 20 most frequent MWEs described in the lexical database having at least one ambiguous realization pattern. For the development of the grammar, we built a sub-corpus of 21,125 sentences from the *Statistical Corpus of 20th Century Basque* which contained occurrences corresponding to both MWEs and literal occurrences.
The grammar we have built consists of 111 rules, many of which can be reused to disambiguate MWEs of the same type. The grammar has proved to have 98.90 % coverage and 98.88 % accuracy. This shows that, for many Basque MWEs, morphosyntactic context can be enough to resolve ambiguity.

References

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G. and Lersundi M. (2001). EDBL: a general lexical basis for the automatic processing of Basque. *IRCS Workshop on Linguistic Databases*, Philadelphia, USA.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A. and Urizar R. (1998). A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages*, Granada, Spain.

Baldwin T. and Kim S.N. (2010). Multiword expressions. In Indurkhya N. and Damerau F. (Eds.), *Handbook of Natural Language Processing*, 267-292. CRC Press, 2nd edition.

Copestake A., Lambeau F., Villavicencio A., Bond F.B., Baldwin T., Sag I. and Flickinger D. (2002). Multiword Expressions: linguistic precision and reusability. *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, 1941-1947, Las Palmas, Spain.

Euskaltzaindia (2000). Hiztegi Batua. *Euskera*, XVL(2):475-757, Bilbao, Spain.

Hashimoto C., Sato S. and Utsuro T. (2006). Japanese idiom recognitions: Drawing a line between literal and idiomatic meanings. *Proceedings of the COLING/ACL 2006 Interactive Poster Session*, 353-360, Sydney, Australia.

Karlsson F., Voutilainen A., Heikkila J. and Anttila A. (1995). *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Prentice-Hall, Berlin, Germany.

Li W., Zhang X., Niu C., Jiang Y. and Srihari R. (2003). An expert lexicon approach to identifying English phrasal verbs. *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 513-520, Sapporo, Japan.

Tapanainen P. (1996). *The Constraint Grammar parser CG-2*. Publications of the University of Helsinki, 27, Helsinki, Finland.