MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque

ALDEZABAL I., ARTOLA X., EZEIZA N., URIZAR R.

UNIVERSITY OF THE BASQUE COUNTRY. COMPUTER SCIENCE FACULTY.
DEPARTMENT OF COMPUTER LANGUAGES AND SYSTEMS.
649 P.K. 20.080 DONOSTIA. THE BASQUE COUNTRY
jibecran@si.ehu.es (Nerea Ezeiza)

ADURIZ, I.

UZEI. ALDAPETA 20
20.009 DONOSTIA. THE BASQUE COUNTRY
uzei0005@sarenet.es

Abstract

A lemmatiser-tagger must not only lemmatise word-forms consisting of a single lexical element but it must also be able to detect complex units. In this paper we try to delimit linguistically which complex words deserve to be lemmatised as a unit. Then, we propose a formal description for MultiWord Lexical Units (MWLU) in Basque —resulting of a conscientious analysis of their syntactic and morphological behaviour. Based on that formal description we propose a simple logical formalism to represent those MWLUs so that they can be automatically processed.

1. Introduction

A lemmatiser-tagger is a computational tool used for assigning the correct lemma and grammatical category to each token of a corpus. It is a basic device for corpus analysis, automatic indexation, syntactic and semantic analyses etc. For example, the lemmatiser-tagger for Basque (EUSLEM) (Aduriz *et al.*, 96a) is essential for the second phase of the Systematic Compilation of Modern Basque[1] (EEBS) project (Urkia *et al.*, 91.) being carried out by UZEI[2].

The EUSLEM project is being designed by a group from the Computer Science Faculty of the University of The Basque Country and UZEI. The knowledge in computational linguistics of the former added to the large experience in lexicography of the latter is turning out to be vital for the completion of this project. Some previous outcomes resulted from their collaboration can be considered both the basis and antecedent of this work:

- The Lexical Database for Basque (EDBL) (Agirre *et al.*, 95), which is the basis for many other applications, consists of (i) dictionary entries (the same you can find in any conventional dictionary), (ii) verb forms and (iii) dependent morphemes, all of them with their respective morphological information.
- A tagset system has been also developed. The one we have chosen for Basque is a three level system which the user can parametrise when using the programme. In the first level seventeen general categories are included (noun, adjective, verb, etc). In the second one each category tag is further refined by subcategory tags. The last level includes other interesting morphological information (case, number, etc.).
- With regard to lemmatisation, a morphological analyser was needed, specially for Basque, which is an agglutinative and morphologically complex language. With this purpose in mind, a morphological analyser was developed, based on the Two-Level Morphology (Koskenniemi, 83). Due to the fact that the process of normalisation of Basque is still in progress, the morphological processor had to be able to analyse linguistic variants and to distinguish between standard and non-standar lemmas. For this purpose, the treatment of variant errors was carried out using a two-level subsystem made up of (i) 1,000 items linked to the corresponding correct ones and (ii) twenty rules to cover the most common competence errors.
- If the lexicon-based analyser produced no valid analysis, we needed a lexicon-free lemmatiser that wouldn't let any token unlemmatised or untagged, so that the system might be robust enough. Among the different systems we studied, we chose a two-level mechanism based on the idea used in speech synthesis (Black *et al.*, 1991).

The morphological analyser at that step gives as a result *all* the possible analyses of each *token* in the text. So, on one hand, we must discard as many wrong analyses as possible. For that purpose we will apply a combination of two different strategies: the Constraint Grammar (CG) formalism (Karlsson *et al*., 95; Aduriz *et al.*, 96b) based on linguistic knowledge and a procedure based on statistics and developed within MULTEX project (Amstrong *et al.*, 1995). For the moment, both methods are being tested separately. On the other hand, it is obviously not enough to lemmatise just the word-forms between two blanks. There are several MultiWord Lexical Units that must be lemmatised as a single unit.

Next we will try to delimit linguistically which word combinations deserve to be lemmatised as a unit. Then, we will propose a formal description for Basque MWLUs and a simple logical formalism so that they can be automatically processed. Eventually, the conclusions of this work will be expounded.

---

[1] During the 1987-1992 period UZEI manually compiled and lemmatised a three million word corpus of twentieth century's Basque texts, which is annually renewed.

[2] UZEI is a cultural association created in 1977 the aim of which is to promote Basque lexicon's modernisation within the normalisation process of this language.

2. Delimiting lexical units

Giving an exact definition for *word* is not an easy task at all. At text level one solution could be to define a *word* as "any string of characters between two blanks" (Fontenelle et al., 94). There are, of course, some lexical units that fit that definition (*etxe* 'house', *zuri* 'white'...). Even a great number of idioms that in non-inflected languages would be multiword, in Basque, which is a highly inflected language, constitute a single typographic unit (*ziurrenik* 'most probably', *aurrerantzean* 'from now on', *aurretiaz* 'in advance',...). But a different definition is obviously necessary for many other idiomatic expressions as well as for several MWLUs (*lan egin* 'to work', *beste bat* 'another', *hutsaren hurrengoa* 'the lowest of the low',...). We have mostly relied on the experience that UZEI gathered in the EEBS project, in which a wider perspective was taken and an extensive range of MWLUs was lemmatised as a unit. We have used those complex units as a basis for our analysis. Next, we will explain what we consider a MWLU.

First of all, we will focus on the treatment of compounds, for they constitute quite a special group within the MultiWord Lexical Units.

2.1. Compounds

Compounds in Basque can appear written mainly in four different ways (Euskaltzaindia, 92):

- attached (*idazmakina* 'writing machine', *plazagizon* 'public man',...)
- hyphenated (*datu-base* 'database', *begi-nini* 'pupil',...)
- separated by a blank and with the first component of the compound phonologically transformed (*Euskal Etxea* 'Basque House', *itsas armada* 'naval forces',... where *euskal* and *itsas* are the respective variants in composition and derivation of *euskara* 'Basque' and *itsaso* 'sea') and
- separated but with the components unaltered (*bake ituna* 'peace treaty',...).

For lemmatising purposes attached compounds are identical to single words and therefore are included in the lexical database as such.

Some of the hyphenated compounds that are considered to be lexicalised are included in our database as if they were single words (*botoi-zulo* 'buttonhole',...). Some others, being freely generated, (*mahai-hanka* 'table leg',...), don't appear in any dictionary and so they are not in our database either. However, the morphological analyser is capable of detecting them. In our database we consider the hyphen a lexical element, as the Two-Level Morphology renders it possible. Therefore, the different hyphens[3], can be followed, like any other lexical element in the database, by a particular set of morphemes and lexicons, and so certain types of compounds can be recognised.

Among compounds in which the first component has been modified there are also lexicalised compounds and freely generated ones. This kind of compounds will only be included in our database as MWLUs when they are lexicalised (*Boli Kosta* 'Ivory Coast'). Nevertheless, all compounds of that type will be detected in a later stage since phonologically transformed components are easy to detect:

* among the words that lose their final *-a*, there are, on one hand, those ending in *-ia* (*filologia* 'philology', *energia* 'energy', *psikologia* 'psychology',...) which are easily recognised by a simple rule. On the other hand, there is a reduced set of exceptions that can also lose their final *-a* in composition (*natura* 'nature', *literatura* 'literature', *kultura* 'culture', *burdina* 'iron', *eliza* 'church' and *hizkuntza* 'language').
* the rest of the cases constitute quite a restricted set (*itsas, erret,...* variants of *itsaso* 'sea', *errege* 'king',...).

---

[3] In the database there is more than one type of hyphen because it has other uses distinct from composition, e.g. declension of foreign words (*Shakespeare-engandik*, 'from Shakespeare').

As for compounds separately written and not hyphenated, the only way of locating them is introducing them into the database. The freely generated compounds, though, cannot be located for the moment. Nonetheless, a computational tool that will retrieve lexical collocations automatically from corpora is being designed. That information will be later used to enrich the MWLU database.

2.2. MultiWord Lexical Units

Sometimes **syntactic patterns** (*zenbat eta ...-ago ... orduan eta ...-ago*, 'the ... -er ... the ... -er') are also lemmatised as MWLUs. Since Basque is an inflected language, frequently the boundary between syntax and morphology is quite vague. We think that those patterns are easier to treat at syntax level so we have left them for a later stage.

To distinguish between **lexical collocation** and **idioms** authors generally use a semantic criterion (Heid, 94). Idioms can hardly be interpreted in terms of the meanings of their constituents (*adarra jo* 'pull s.o.'s leg' ≠ *adarra* 'horn' + *jo* 'to play'), while in collocations their components (or at least one of them) keep their original sense (*zarata atera* 'make noise').

However, it is very difficult to outline the boundary between idioms and lexical collocations since from totally opaque idioms to open lexical collocation there is a wide range of word-combination arranged along a scale or continuum (Cowie, 90):

- *Pure idioms*. These are totally opaque idioms, that is, idioms *strictu sensu*. Historically, opaque idioms are "the end-point of a process in which word combinations first establish themselves through constant re-use, then undergo figurative extension and finally petrify or congeal" (Cowie, 90). Among opaque idioms we can find 'pure lexical idioms' (*ahuntzaren gauerdiko eztula* 'trifle') as well as those we call 'grammatical' (*harik eta* 'until', *hala eta guztiz ere* 'nevertheless',...).
- *Figurative idioms*. These word-combinations are idiomatic in the sense that variations are seldom found but for the speakers the primitive 'literal' sense is not as distant as those in the opaque idioms (*hutsaren hurrengoa* 'the lowest of the low').
- *Restricted collocations*. In these combinations, also called *semi-idioms*, one word has a figurative sense not found outside that limited context. The other elements are used in their usual literal sense (*bideak urratu* 'to make a way').
- *Open collocations*. Each element of these word-combination is used in its ordinary literal sense (*hego haizea* 'south wind').

As we mentioned above, we have relied on UZEI's criterion and taken a wider perspective when deciding what to consider a MWLU. We deal with both types of idioms, pure and figurative, as well as with restricted collocations. Open collocations, though, have only been taken into account if they express a particular concept and so they deserve an entry in Basque modern dictionaries (*Euskal Herria* 'The Basque Country'*, Amerikako Estatu Batuak* 'United States of America',...).

We also deal with **foreign words** such as *in situ, in fraganti, a priori,...* those MWLUs can be somehow considered idiomatic as well since the meaning of their constituents is totally opaque in Basque.

However, we do not consider **proverbs**, although among them there are also idiomatic (*txakur zaunkaria ez da horzkaria* 'his bark is worse than his bite') and non-idiomatic ones (*San Bizente hotza, neguaren bihotza* 'The cold St. Vincent 's day is the middle of the winter'. Locating **catch-phrases** (*Hauxe behar genuen!* 'that's all we needed') and **similes** (*berakatz-atala baino finagoa* '(as) busy as a clove of garlic') is not our purpose either.

3. Representation of MWLUs.

After delimiting linguistically what we consider MWLUs, we needed a formal description so that they could be automatically processed. For that, we made a conscientious analysis of the syntactic and morphological behaviour of those lexical units. From that analysis we deduced the relevant formal

features that had to be considered. In order to describe MWLUs in Basque, we have functionally established the following features:

- Sure/ambiguous: We say a MWLU is *sure* when its components can only be analysed as a whole lexical unit and therefore no other interpretation is possible (*behin eta berriro* 'over and over again'). In that case, the analyses that don't belong to the MWLU interpretation are eliminated (see Fig. 1).

```
/<behin>/
     ("behin eta berriro"   ADB ADO HAT 1/3)
/<eta>/
     ("behin eta berriro"   ADB ADO HAT 2/3)
/<berriro>/
     ("behin eta berriro"   ADB ADO HAT 3/3)
/<begiratzen>/
     ("begira"  ADI SIN + ASP EZBU)
     ("begira"  ADI SIN + LOT MEN KONP @-JADNAG_MP @-JADLAG_MP)
/<nuen>/
     ("*edun"  ADL B1 NR_HU NK_NI + LOT MEN @+JADNAG_MP @+JADLAG_MP)
     ("*edun"  ADL B1 NR_HU NK_NI + LOT MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>)
     ("*edun"  ADL B1 NR_HU NK_NI)
     ("*edun"  ADT B1 NR_HU NK_NI + LOT MEN @+JADNAG_MP @+JADLAG_MP)
     ("*edun"  ADT B1 NR_HU NK_NI + LOT MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>)
     ("*edun"  ADT B1 NR_HU NK_NI)
```

Fig. 1 Morphological analysis of the sentence ***behin eta berriro*** *begiratzen nuen...* 'I was looking **over and over again**...'.

- contiguous/dispersed: We say a MWLU is dispersed when its components do not necessarily occur one after another. In that case, the processing gets more complicated since we have to seek the components in subsequent words (see Fig. 2). If a MWLU has more than two components, some of them may be contiguous and some others may not.

```
/<gure>/
     ("gu"  IOR PER NUMP + DEK GEN @IZLG> @<IZLG + DEK ABS MG @OBJ @SUBJ)
     ("gu"  IOR PER NUMP + DEK GEN @IZLG> @<IZLG)
/<sukalderaino>/
     ("sukalde"  IZE ARR + DEK NUMS MUGM + DEK ALA @ADLG + DEK ABU @ADLG)
/<heldu>/
     ("hel"  ADI SIN + ASP PART + DEK ABS MG @OBJ @SUBJ)
     ("hel"  ADI SIN + ASP PART)
/<nahi>/
     ("nahi izan"   ADI ADP HAT 1/2)
     ("nahi"  ADI ADP)
     ("nahi"  ADI SIN)
     ("nahi"  IZE ARR + DEK ABS MG @OBJ @SUBJ)
     ("nahi"  IZE ARR)
/<baldin>/
     ("baldin"  PRT)
/<bada>/
     ("nahi izan"   ADI ADP HAT 2/2)
     ("izan"  ADB ADO + ADL A1 NR_HU)
     ("izan"  ADB ADO + ADT A1 NR_HU)
     ("izan"  AUR BALD + ADL A1 NR_HU)
     ("izan"  AUR BALD + ADT A1 NR_HU)
     ("bada"  LOT LOK @LOK)
```

Fig. 2 Morphological analysis of the sentence *gure sukalderaino heldu* ***nahi*** *baldin* ***bada***... 'If you **want** to reach our kitchen...'.

- ordered/order-free: Regardless of it being continuous or dispersed the elements of a MWLU may not necessarily keep an order[4] (see Fig. 3). A clear example of this are verb periphrasis such as *lo egin* 'to sleep, *min eman* 'to hurt' etc. since their constituents usually shift their positions, e.g. in imperative clauses (we say *lo egin dut* 'I have slept', but *egizu lo*, 'sleep').

```
/<Ez>/
      ("ez"  ADB ADO)
      ("ez"  IZE ARR + DEK ABS MG @OBJ @SUBJ)
      ("ez"  IZE ARR)
/<zitzaion>/
      ("falta izan"   ADI ADP HAT 2/2)
      ("izan"  ADL B1 NR_HU NI_HU + LOT MEN @+JADNAG_MP @+JADLAG_MP)
      ("izan"  ADL B1 NR_HU NI_HU + LOT MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>)
      ("izan"  ADL B1 NR_HU NI_HU)
/<korbata>/
      ("korbata"  IZE ARR + DEK ABS MG @OBJ @SUBJ)
      ("korbata"  IZE ARR + DEK ABS NUMS MUGM @OBJ @SUBJ)
      ("korbata"  IZE ARR)
/<eramatea>/
      ("eraman"  ADI SIN + LOT MEN KONP @-JADNAG_MP @-JADLAG_MP)
      ("eraman+te"  ADI SIN + ASP IZE + DEK ABS NUMS MUGM @OBJ @SUBJ)
/<besterik>/
      ('beste'  DET DZG + DEK PAR MG @OBJ @SUBJ)
/<falta>/
      ("falta izan"   ADI ADP HAT 1/2)
      ("falta"  ADI ADP)
      ("falta"  IZE ARR + DEK ABS MG @OBJ @SUBJ)
      ("falta"  IZE ARR + DEK ABS NUMS MUGM @OBJ @SUBJ)
      ("falta"  IZE ARR)
```

Fig. 3 Morphological analysis of the sentence *Ez **zitzazion** korbata eramatea besterik **falta**...,* 'All he **needed** was to wear a tie...'

- Inflected/invariable: The components of a MWLU may either appear in an invariable form (*hurrenez hurren* 'respectively') or inflect. For instance, both constituents of the verb periphrasis *bizi izan* 'to live' can inflect and therefore the possible combinations are countless (*bizi naiz*, 'I live', *biziko banintz* 'if I lived', *bizi izanik* 'living'...). In the case that the components of the MWLU can inflect, they can accept either any inflection or just a restricted set of inflected forms. Thus, restrictions are needed for components accepting just a few inflected forms. Moreover, lack of restrictions, especially in "dispersed + order-free" lexical units, may increase the ambiguity rate considerably. The more restrictions we make the less ambiguity we get.

  Summing up, when defining a MWLU we must give the following information:
  - We must declare whether each component is invariable or can inflect and specify the restrictions of inflection when necessary. We use a logical formalism to represent these restrictions.

Examples:	(1)	adarra jo +SEG ADI ADK	adar jo	((ABS NUMS MUGM) or (PAR MG)) /
	(2)	hala ere +SEG LOT LOK	- -	

[4] However, the "order-free + contiguous" combination is not necessary. The analysis of the syntactic and morphological behaviour of MWLUs —previous to their formal description— led us to the conclusion that in Basque whenever a MWLU is order-free, it is also dispersed.

(3) lan -
egin egin
-SEG
ADI ADK

(4) beste beste *
bat bat
-SEG
DET DZG

In the example (1) above —*adarra jo* 'to pull s.o.'s leg'— the lemma *adar* can only take the singular absolutive ("ABS NUMS MUGM") or the partitive ("PAR MG"). The lemma *jo*, on the contrary, accepts any possible inflection (*jo, jotzen, joko, jotzeko, jotzera,...*).
In the example (2) —*hala ere* 'nevertheless'— both components are invariable and in (3) —*lan egin* 'to work'— the first component cannot vary but the second one can inflect freely.

- When two components of the MWLU may appear dispersed but are not order-free we mark with the symbol * between them (see example (4)).
- When two components can be both dispersed and order-free we mark with the symbol / between them (see example (1)).
- When a MWLU is sure we mark it +SEG and when it is ambiguous we mark it -SEG (see examples above).
- Finally the category of each MWLU is specified (see examples above).

4. Conclusions

In this paper we have tried to show that it is not sufficient to lemmatise just the word-forms between two blanks. There are several Multiword Lexical Units such as idioms, restricted collocations, foreign words etc. that must be lemmatised as a single unit. Nevertheless, processing those complex lexical units involves some additional difficulties. For instance, the constituents of the MWLUs do not always occur one after another or in the same order and, although some components have an invariable form, some others can inflect. Thus, it is essential to develop a formal representation so as to be able to describe in detail the MWLU.

Currently we are testing a prototype for the treatment of MWLUs. A few significative examples of its output are also shown in this paper (see Fig. 1, 2 and 3).

5. Bibliography

Aduriz, I.; Aldezabal, I.; Alegria, I; Artola, X; Ezeiza, N.; Urizar, R. (1996a): "EUSLEM: A lemmatiser/tagger for Basque", Computational Lexicology and Lexicography. *Euralex'96*. Göteborg.

Aduriz, I.; Arriola, J.M.; Artola, X.; Díaz De Ilarraza, A.; Gojenola, K.; Maritxalar, M.; Urkia M. (1996b): *Euskararako murriztapen-gramatika: lehen urratsa.* Department of Languages and Computing Systems, Computer Science Faculty, University of The Basque Country.

Agirre, E.; Arregi, X.; Arriola, J.M.; Artola, X.; Díaz De Ilarraza, A.; Insausti, J.M.; Sarasola, K. (1995): "Different issues in the design of a general-purpose Lexical Database for Basque". *First Workshop on application of Natural Language to Data Bases, NLDB'95.*

Alegria, I. (1995): *Euskal morfologiaren tratamendu automatikorako tresnak.* Ph. D. thesis. University of The Basque Country.

Amstrong, S, Russel, G., Petitpierre, D., Robert, G. (1995): "An open Architecture for Multilingual Text Processing", in: Proceedings of the 5th Conference of the EACL, Volume 1, 101-106.

Black, A., van de Plassche, J., Williams, B. (1991): "Analysis of Unknown words through Morphological Decomposition", in: *Proceedings of the 5th Conference of the EACL*, Volume 1, 101-106.

Clausen, U.; Lyly, E. (1994): "Criteria for identifying and Representing Idioms in a Phraseological Dictionary", The way words work / combinatorics, *Euralex'94*. Amsterdam. 258-262.

Cowie, A.P.; Mackin R.; McCaig I.R. (1990): *Oxford Dictionary of Current Idiomatic English.* v2.

Euskaltzaindia (1992): *Hitz elkartuen osaera eta idazkera*. Hitz-elkarketa / 4. LEF batzordea.

Fontenelle, T.; Adriaens, G.; De Braekeleer, G. (1994): "The Lexical Unit in the Metal® MT System", *MT*.The Netherlands. v9. 1-19.

Heid, U. (1994): "On Ways Words Work Together - Topics In Lexical Combinatorics", The way words work together / combinatorics, *Euralex'94*. Amsterdam. 226-257.

Izagirre, K. (1981): *Euskal lokuzioak. Espainolezko eta frantsesezko gida-zerrendarekin.*

Karlsson, F.; Voutilainen, A.; Heikkila, J.; Anttila, A. (1995): *Constraint Grammar: Language-independent System for Parsing Unrestricted Text.* Mouton de Gruyter.

Koskenniemi, K. (1983): *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production.* Ph D. thesis, University of Helsinki.

Longman (1989): *Dictionary of English Idioms.*

Segond, F.; Tapainen, P. (1995): "Using a Finite-State Formalism to Identify and Generate Multiword Expressions", *MLTT*. Grenoble.

Urkia, M. (Forthcoming): *Euskal morfologiaren analisi automatikorantz*. Ph. D. thesis. University of The Basque Country.

Urkia, M.; Sagarna, A. (1991): "Terminología y lexicografía asistida por ordenador. La experiencia de UZEI". *Actas del VII Congreso de SEPLN*. v9. 193-202.