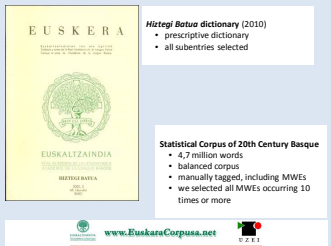
# Computational Treatment of Basque Multiword Expressions

(WG1, WG2, WG3)

Ruben Urizar

IXA research group – University of the Basque Country (UPV / EHU)

## Sources




***Hiztegi Batua* dictionary** (2010)
- prescriptive dictionary
- all subentries selected

**Statistical Corpus of 20th Century Basque**
- 4,7 million words
- balanced corpus
- manually tagged, including MWEs
- we selected all MWEs occurring 10 times or more


| POS | |
|---|---|
| Verbs | 837 |
| Nouns | 695 |
| Adverbs | 343 |
| Quantifiers | 113 |
| Conjunctions | 93 |
| Adjectives | 53 |
| Interjections | 33 |
| Pronouns | 20 |
| Others | 20 |
| **TOTAL** | **2,207** |

## Representation
## Lexical Database for Basque (**EDBL**)

- The purpose of the description is to formally encode all the possible surface realizations of each MWE.
- We worked out a single representation covering all types of MWEs ranging from fixed expressions to those of highest morphosyntactic flexibility.

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. their **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.
   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 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**.

## Raw text

```
Erlijio  alorrean  Zaitegik   berak     egin      zuen      lan   batez ere.
   |        |         |         |         |         |        |        |
religion  area.INE  Zaitegi.ERG himself.ERG do-PART  AUX.3S.PST work  above all

                It was specially Zaitegi himself who worked on the religion area.
```

## Morphosyntactical analysis:
## **EUSTAGGER**

```
"<Erlijio>"<HAS MAI>"
  "erlijio" IZE ARR ZERO HAS MAI @KM>
  "erlijio" IZE ARR ABS MG HAS MAI @OBJ @PRED @SUBJ
"<alorrean>"
  "alor" IZE ARR INE NUMS MUGM @ADLG
"<Zaitegik>"<HAS MAI>"
  "Zaitegi" IZE LIB ERG NUMS MUGM HAS MAI @SUBJ
"<berak>"
  "bera" ADJ ARR ERG MG AORG @SUBJ
  "bera" ADJ ARR ABS NUMP MUGM AORG @OBJ @PRED @SUBJ
  "bera" ADJ ARR ERG NUMS MUGM AORG @SUBJ
  "berak" DET ERKIND NMGP ABS NUMP MUGM @OBJ @PRED @SUBJ
  "bera" DET ERKIND NMGS ERG NUMS MUGM AORG @SUBJ
"<egin>"
  "egin" ADI SIN ADOIN NOTDEK @-JADNAG
  "egin" ADI SIN PART BURU NOTDEK @-JADNAG
  "egin" IZE ARR ABS MG @OBJ @PRED @SUBJ
  [...]
"<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>"
  "landu" ADI SIN ADOIN NOTDEK @-JADNAG
  "lan" IZE ARR ABS MG @OBJ @PRED @SUBJ
  [...]
"<batez>"
  "bate" IZE ARR INS MUG @ADLG
  "bat" DET DZG INS MUG @ADLG
  "bat" IZE ZKI INS MUG @ADLG
"<ere>"
  "ere" LOT LOK EMEN @LOK
"<$.>"<PUNT PUNT>"
  PUNT PUNT
```

## MWEs detected:
## **HABIL**

```
"<Erlijio>"<HAS MAI>"
  "erlijio" IZE ARR ZERO HAS MAI @KM>
  "erlijio" IZE ARR ABS MG HAS MAI @OBJ @PRED @SUBJ
"<alorrean>"
  "alor" IZE ARR INE NUMS MUGM @ADLG
"<Zaitegi-k>"<HAS MAI>"
  "Zaitegi" IZE LIB ERG NUMS MUGM HAS MAI @SUBJ
"<berak>"
  "bera" ADJ ARR ERG MG AORG @SUBJ
  "bera" ADJ ARR ABS NUMP MUGM AORG @OBJ @PRED @SUBJ
  "bera" ADJ ARR ERG NUMS MUGM AORG @SUBJ
  "berak" DET ERKIND NMGP ABS NUMP MUGM @OBJ @PRED @SUBJ
  "bera" DET ERKIND NMGS ERG NUMS MUGM AORG @SUBJ
"<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
  [...]
"<batez_ere>"
  "batez_ere" LOT LOK EMEN mw2 @LOK
"<$.>"<PUNT PUNT>"
  PUNT PUNT
```

Ambiguous MWE

Unambiguous MWE

## MWE processor
## **HABIL**

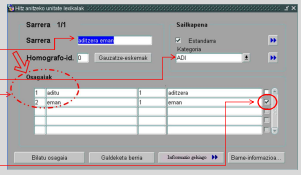
- It deals with both **contiguous and split MWEs**
- It takes into account **all the possible orders** of the components
- It checks that **inflectional restrictions** are complied with
- It generates morphosyntactic **interpretations** for the MWE

## Example (EDBL)

Lexical Database for Basque (EDBL)

**Composition of the MWE**

**COMPOSITION**
- Lemma (sarrera): ***aditzera eman*** 'to announce' lit. 'to give to understand'
- Part of Speech (kategoria): verb (ADI)
- Components (osagaiak): VB *aditu* (understand) + VB *eman* (give)
- Component conveying morpho-syntactical information to the whole MWE: *eman*

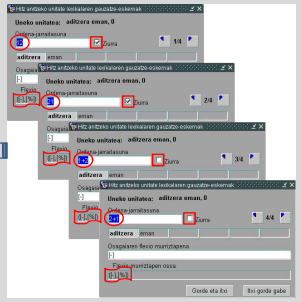



**Surface Realization Pattern of MWE**

**SURFACE REALIZATION**

For MWE ***aditzera eman*** 'to announce' there are 4 realization patterns corresponding to 4 possible orders of components:

- Order of components (ordena-jarraitasuna):
  - ✓ contiguous: 12 and 21
  - ✓ split: 1+2 and 2+1
- Inflectional restrictions (flexio-murriztapena):
  1. first component (*aditzera*, 'to understand') is fixed [-]
  2. second component (*eman*, 'to give') may take any inflection [%]
- Unambiguousness (ziurra):
  - ✓ orders 12 and 21 are unambigu-ous
  - ✓ 1+2 and 2+1 are ambiguous.




## Disambiguation grammar
## Constraint Grammar (CG)

**PREMISE**: Many MWEs may undergo more restricted variations than literal uses.

- We 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*
- The sub-corpus which contained occurrences of word combinations corresponding to both MWE and literal interpretations.
- 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.

## Example (CG grammar rule)

CG disambiguation grammar

**RULE**

```
ADD (%MWE)                                   TARGET: MWEs in list LAN-EGIN
TARGET LAN-EGIN
IF (0 EGIN)                                  CONDITIONS (IF):
   (1 EDUN/EZAN) (NOT 1 NOR-HAIEK)           0 position: 2nd component is a verb in list EGIN
   (2 LAN AND LAN-EGIN)                      1 position: TR AUX (edun or ezan) ; OBJ NOT 3P
   (NOT 3 IZENONDO OR POSDET);               2 position: 1st component is a verb in list LAN and
                                                marked as MWE in list LAN-EGIN
                                             3 position: NOT post.ADJ or post.DET
```

**LISTS**

```
LIST LAN-EGIN = "lan_egin" "hitz_egin" "ihes_egin" "parte_hartu" ;
LIST LAN = "lan" "hitz" "ihes" "parte" ;
LIST EGIN = "egin" "hartu" ;

LIST EDUN/EZAN = "*edun" "*ezan" ;
LIST NOR-HAIEK = NR HAIEK ;
LIST IZENONDO = (ADJ IZAUR-) ;
LIST POSDET = (DET ERKARR) (DET BAN)
   "berbera" "bat" "batzuk" "bi" "anitz" "aski" "asko" "dena"
   "franko" "guzti" "gutxi " "gehiago" "gehiegi" ... ;
```

## MWEs disambiguated:
## **CG grammar**

```
"<Erlijio>"<HAS MAI>"
  "erlijio" IZE ARR ZERO HAS MAI @KM>
  "erlijio" IZE ARR ABS MG HAS MAI @OBJ @PRED @SUBJ
"<alorrean>"
  "alor" IZE ARR INE NUMS MUGM @ADLG
"<Zaitegi-k>"<HAS MAI>"
  "Zaitegi" IZE LIB ERG NUMS MUGM HAS MAI @SUBJ
"<berak>"
  "bera" ADJ ARR ERG MG AORG @SUBJ
  "bera" ADJ ARR ABS NUMP MUGM AORG @OBJ @PRED @SUBJ
  "bera" ADJ ARR ERG NUMS MUGM AORG @SUBJ
  "berak" DET ERKIND NMGP ABS NUMP MUGM @OBJ @PRED @SUBJ
  "bera" DET ERKIND NMGS ERG NUMS MUGM AORG @SUBJ
"<egin>"<1-2>"
  "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>"
  "lan_egin" ADI ADK ADOIN mw1 NOTDEK @-JADNAG
  "lan_egin" ADI ADK PART BURU NOTDEK mw1 @-JADNAG
  [...]
"<batez_ere>"
  "batez_ere" LOT LOK EMEN mw2 @LOK
"<$.>"<PUNT PUNT>"
  PUNT PUNT
```

Ambiguity resolved

**CONCLUSION:** 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.