Universidad del País Vasco · Euskal Herriko Unibertsitatea · eman ta zabal zazu

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

WG1, WG2, WG3

Uxoa Iñurrieta*, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka and Kepa Sarasola
IXA NLP group, University of the Basque Country
uxoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu, a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

ixa

Aim: to carry out **linguistic investigations into improving the treatment of word combinations in a rule-based MT system**, *Matxin* (Mayor et al., 2011), which translates **Spanish into Basque**, two languages of very different typology.

Double challenge

- **Detection** in the source language (ES, Spanish)
- **Transfer** into and generation in the target language (EU, Basque)

Done so far

- **Morphological analysis** of verb+noun combinations in a bilingual dictionary: Spanish into Basque and Basque into Spanish
- **Syntactic analysis** of the most frequent Spanish verb+noun combinations
- **Detection experiment** of Spanish verb+noun combinations
- Creation of a **public database** which collects the information achieved from the linguistic analysis

1 Morphological analysis

Analysed combinations

- **Source:** Elhuyar bilingual dictionary for Spanish-Basque and Basque-Spanish
- 2,650 Spanish combinations along with 6,392 Basque equivalents
- 2,945 Basque combinations along with 6,587 Spanish equivalents

Spanish into Basque analysis

- Spanish combinations gathered**: verb + (prep) + (det) + noun**

Basque equivalents	%
noun (abs) + verb	**35.24%**
verb	23.53%
noun (cas/pos) + verb	**13.30%**
other	27.93%

Table 1: Morphological structures of the Basque equivalents

Basque into Spanish analysis

- Basque combinations gathered**: noun + verb**
- Many different cases and postpositional marks attached to the nouns

Spanish equivalents	%
verb	58.07%
verb + (prep) + (det) + noun	**30.02%**
other	11.90%

Table 2: Morphological structures of the Spanish equivalents

Non-word-for-word translations

- Only **48.54%** of the Spanish verb+noun combinations are translated by noun+verb combinations into Basque
  - Out of those, only **21.79%** are translated regularly, that is, by substituting the noun and the verb with their usual equivalents (ex. 1)
- **58.07%** of the Basque noun+verb combinations are translated by a verb only into Spanish (ex. 2), and only **30.85%** are translated by verb+noun combinations
  - Out of the ones translated by verb+noun combinations, only **28.01%** are translated regularly (ex. 3)

(1) '(to) pay attention'
EU: **jaramon egin**
attention.ABS do.INF
ES: **hacer caso**
do.INF attention

(2) '(to) work'
EU: **lan egin**
work.ABS do.INF
ES: **trabajar**
work.INF

(3) '(to) laugh heartedly'
EU: **barrez ito**
laughter.INS suffocate.INF
ES: **morirse de risa**
die.INF PREP laughter

2 Syntactic analysis

Analysed combinations

- The **150 most frequent combinations** out of the ones previously analysed morphologically
- **Frequency information gathered from a parallel corpus** consisting of 491,853 sentences from many different sources

Analysed features

- **Definiteness or indefiniteness** of the noun phrase. Always consistent? (ex. 4)
- **Number** of the noun phrase. Always consistent? (ex. 5)
- **Possibility to add a modifier** to the noun phrase (ex. 6)
- **Possibility to separate** the noun phrase and the verb (ex. 7)
- **Possibility to change the order** of the elements (ex. 8)

(4) **fijar un plazo**
fix.INF IND.DET.S deadline.S
'(to) fix a deadline'

(5) **fijar el plazo**
fix.INF DEF.DET.S deadline.S
'(to) fix the deadline'

(6) **fijar los plazos**
fix.INF DEF.DET.P deadline.P
'(to) fix the deadlines'

(7) **fijarán un nuevo plazo**
fix.3P.FUT IND.DET.S new.S deadline.S
'(to) fix a new deadline'

(8) **fijarán mañana el plazo**
fix.3P.FUT tomorrow DEF.DET.S deadline.S
'(to) fix a deadline tomorrow'

(9) **el plazo fue fijado**
DEF.DET.S deadline.S be.3S.PST fix.PRT
'the deadline was fixed'

3 Detection experiment

Experiment

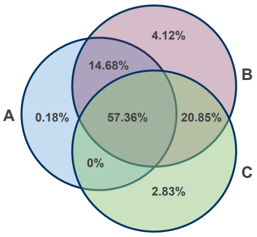
- **Corpus used:** 15,182,385 Spanish sentences, taken from the parallel English-Spanish corpus made public for the shared task in the ACL 2013 workshop in staistical MT
- **MWEs searched:** the 150 combinations previously syntactically analysed (see section 2)
- **Compared methods:**
  A. The old one, based on the words-with-spaces strategy
  B. A second one, based on our linguistic data and automatic chunking information
  C. A third one, based on our linguistic data and automatic syntactic dependencies

Results

- MWEs detected in all: **433,092**
  - **27.05%** not detected by method A
- **Evaluation** carried out by linguists: 500-sentence set per system

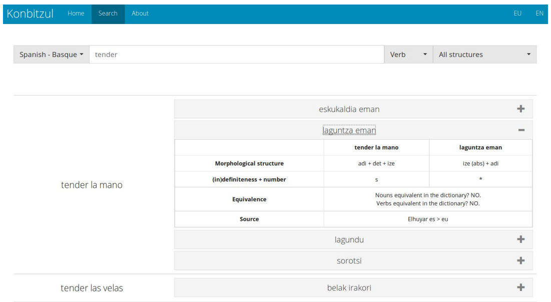
Method(s)	% detected	Precision
B and C only (not A)	20.85%	96.60%
B only (not C and A)	4.12%	92.60%
C only (not B and A)	2.83%	83.20%

Table 3: Results of the detection experiment



Picture 1: Results of the detection experiment

4 The *Konbitzul* database



tender la mano

eskukaldia eman

laguntza eman

	tender la mano	laguntza eman
Morphological structure	adi + det + ize	ize (abs) + adi
(in)definiteness + number	s	*
Equivalence	Nouns equivalent in the dictionary? NO. Verbs equivalent in the dictionary? NO.	
Source	Elhuyar es > eu	

lagundu

sorotsi

tender las velas

belak irakori

Features

- Publicly available at http://ixa2.si.ehu.es/konbitzul
- Linguistic information from our analysis
- Various search criteria:
  - Language direction: Spanish-Basque or Basque-Spanish
  - Verb, noun or whole combination
  - Morphological structure

5 Conclusions and future work

Conclusions

- Very few verb+noun combinations are translated word-for-word between Spanish and Basque, so **MT systems need a sophisticated treatment of such MWEs**
- Linguistic information specific to MWEs is helpful for their detection; the number of **identified combinations increased by 27.05%** when combining it with chunking information and dependency parsing

Future work

- Analyse what linguistic information is needed for **MWE transfer into Basque**
- **Integrate the results** in *Matxin*
- Analyse what **semantic information** could help for MWE translation

References and further information

Elhuyar gaztelania-euskara hiztegia. Elhuyar Fundazioa. http://hiztegiak.elhuyar.eus/

Iñurrieta, U. (2015). Translation of Spanish Multiword Expressions into Basque: linguistic analysis and detection experiment. In *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.*

Mayor, A., Alegria, I., De Ilarraza, A. D., Labaka, G., Lersundi, M., & Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, *25*(1), 53-82.

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

* Uxoa Iñurieta's work is funded by a PhD scholarship from the Ministry of Economy and Competiivenes (BES-2013-066372).

Abbreviations

ABS	absolutive
ADI	verb
DEF	definite
DET	determiner
EU	Basque
ES	Spanish
FUT	future tense
IND	indefinite
INF	infinitive
INS	instrumental
IZE	noun
S	singular
PREP	preposition
PRT	participle
PST	past tense
3S	3rd person singular
3P	3rd person plural