Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques

Antton Gurrutxaga
Elhuyar Foundation
a.gurrutxaga@elhuyar.com

Iñaki Alegria
IXA group/Univ. of the Basque Country
i.alegria@ehu.es

Abstract

Taking as a starting-point the development on cooccurrence techniques for several languages, we focus on the aspects that should be considered in a NV extraction task for Basque. In Basque, NV expressions are considered those combinations in which a noun, inflected or not, is co-occurring with a verb, as *erabakia hartu* ('to make a decision'), *kontuan hartu* ('to take into account') and *buruz jakin* ('to know by heart'). A basic extraction system has been developed and evaluated against two references: a) a reference which includes NV entries from several lexicographic works; and b) a manual evaluation by three experts of a random sample from the n-best lists.

1 Introduction

The last decade has witnessed great advances in the automatic identification and processing of MWEs. In the case of Basque, advances are limited to terminology extraction and the tagging in corpora of the MWEs represented in lexical databases.

Furthermore, the work on both theoretical and practical phraseology in Basque has been mainly focused on idiomatic expressions, leaving aside collocations (Pérez Gaztelu et al., 2004). As a consequence, Basque NLP and lexicography have not benefited from the approach that emphasized the importance of such units, and very important areas are underdeveloped.

With the aim of taking steps to turn this situation, we undertake the task of extracting NV combinations from corpora. As a preliminary step, we must face the morphosyntactic aspects of Basque that might influence the efficiency of the process.

2 MWE: basic definition and extraction techniques

As a basis for our work, we take idiomaticity as the key feature for the definition and classification of MWE. Idiomaticity could be described as a non-discrete magnitude, whose "value", according to recent investigations (Baldwin and Kim, 2010; Fazly and Stevenson, 2007; Granger and Paquot, 2008), has turned to depend on a complex combination of features such as institutionalization, non-compositionality and lexico-syntactic fixedness.

The idiomaticity of MWEs appears rather as a continuum than as a scale of discrete values (Sinclair, 1996; Wulff, 2010). Thus, the classification of MWEs into discrete categories is a difficult task. Taking Cowie's classification as an initial basis (Cowie, 1998), our work is focused on phrase-like units, aiming, at this stage, to differentiate MWEs (idioms and collocations) from free combinations. Specifically, NV combinations with the following characteristics are considered as MWEs:

- Idioms: non-compositional combinations, as opaque idioms (*adarra jo*: 'to pull somebody's leg'; lit: 'to play the horn') and figurative idioms (*burua hautsi*: 'to rack one's brain'; lit: 'to break one's head').
- Collocations:
  - Semicompositional combinations, in which the noun keeps its literal meaning,

whereas the verb acts as a support verb (*lan egin*: 'to work'; lit. 'to do work'), or has a meaning which is specific to that combination (*atentzioa eman*: 'to catch someone's eye'; lit. 'to give attention' (sth to sb)); *legea urratu*: 'to break the law'; lit. 'to tear the law').

– Compositional combinations with lexical restriction, in which it is not possible to substitute the verb with its synonyms, or that present a clear statistical idiosyncrasy in favor of a given synonym choice (*elkartasuna adierazi*: 'to express solidarity'; *konpromisoa berretsi*: 'to confirm a commitment').

Among the different techniques that have been proposed to extract and characterize MWEs, the cooccurrence of the components is the most used heuristic of institutionalization, and the use of association measures (AM) goes back to early research on this field (Church and Hanks, 1990; Smadja, 1993). In recent years, the comparative analysis of AMs has aroused considerable interest, as well as the possibility of obtaining better results by combining them (Pearce, 2002; Pecina, 2005). Cooccurrence techniques are usually used in combination with linguistic techniques, which allow the use of lemmatized and POS-tagged corpora, or even syntactic dependencies (Seretan, 2008).

3 Special features of Basque NV combinations

These are some characteristics of the NV combinations in Basque to be considered in order to design the extraction process efficiently:

- Basque being an agglutinative language, MWE extraction must work on tagged texts, in order to identify different surface forms with their corresponding lemma. Thus, pure statistical methods working with raw text are not expected to yield acceptable results.

- Some combinations with a noun as first lemma do not correspond to NV combinations in the sense that is usually understood in English. For example, the expression *kontuan hartu* can be translated as *take into account*, where *kontu* is a noun in the inessive case. We are interested in all types of combinations that a noun can form with verbs.

- Representing NV combinations as lemma-lemma pairs is by no means satisfactory; we would not be able to differentiate the aforementioned *kontuan hartu* from *kontu hartu* ("to ask for an explanation"). So it is necessary to deal with the form or type of every noun.

- In order to propose canonical forms for NV combinations, we need case and number annotations for nouns in bigram data. The next examples are different forms of the canonical *erabakia hartu* ('to make a decision'): *ez zuen erabakirik hartu* ('he did not make any decision'), *zenbait erabaki hartu behar ditugu* ('we have to make some decisions'). Canonical forms can be formulated by bigram normalization (see section 4.5 for details).

4 Experimental setup

4.1 Corpora resources

In our experiments, we use a journalistic corpus from two sources: (1) Issues published between 2001-2002 by the newspaper *Euskaldunon Egunkaria* (28 Mw); and (2) Issues published between 2006-2010 by the newspaper *Berria* (47 Mw). So, the overall size of the corpus is 75 Mw.

4.2 Corpus-processing

For linguistic tagging, we use EUSTAGGER by the IXA group of the University of the Basque Country (Aduriz et al., 1996). After linguistic processing, we obtain information about the lemma, part-of-speech, subcategory, case, number and other morphosyntactic features.

We used EUSTAGGER without the module to detect and annotate MWEs in order to evaluate the automatic extraction, regardless of wheter the candidates are in the lexical database.

4.3 Preparing tagged corpora for bigram generation

For bigram generation, we use the Ngram Statistics Package-NSP (Banerjee and Pedersen, 2010). In

order to retain in the text sent to NSP the linguistic information needed according to section 3, we add different types of linguistic information to the tokens, depending on the POS of the components of the combination we are dealing with. In the case of NV combinations, the nouns are represented in the following form:

```
token_lemma_POS_subcategory_case_number
```

In the case of verbs, only lemma and POS are used, as verb inflection has no influence on the canonical form of the expression. In future work, verb inflection will be one of the parameters to measure syntactical flexibility. All other types of tokens are discarded and considered as 'non-token' for NSP processing.

Before this step, some surface-grammar rules are defined to detect and filter the participle forms that are not part of a NV combination, but must be analyzed as adjectives or nouns (eg. *herrialde aurreratuak* 'developed countries', and *gobernuaren aliatuak*, 'government's allies').

4.4 Bigram generation

We generated bigram sets for two different window spans: ± 1 and ± 5. In both sets, the frequency criterion for a bigram to be generated is $f > 30$. Also, the following punctuation marks are interpreted as a boundary for bigram generation: period, colon, semicolon, and question and exclamation marks. Then, all counts of bigrams in NV and VN order are combined using NSP, and reordered in NV order.

Additionally, a heuristic is used to filter some combinations. The first member of many "compound verbs" like *nahi izan* ('to want'), is a noun, and some of them combine usually with a verb, in VN order: *ikusi nahi (zuen)* ('he wanted to see'). In order to reduce this noise, the combinations occurring mostly in VN order are removed. The combinations generated from passive constructions (*hartutako erabakien ondorioak*, 'the consequences of the decisions made') are not affected by this filtering.

4.5 Bigram normalization

In order to get more representative statistics, and to get information that would enable us to propose a canonical form for each MWE candidate, different inflection forms of the same case in nouns are normalized to the most frequent form, and bigram counts are recalculated. I.e. [*erabakia / erabakiak / erabakiok / erabakirik / erabaki*] *hartu* are collapsed to *erabakia hartu* ('to make a decision'), because all the mentioned forms of the lemma *erabaki* appear in the absolutive case. In contrast, the combinations *kontu hartu* ("to ask for an explanation") and *kontuan hartu* ("take into account") are not normalized, as their noun forms correspond to different cases, namely, absolutive (*kontu*) and inessive (*kontuan*). A Perl script detects in the dataset the bigrams to be normalized, using the combined key noun_lemma/noun_case+verb_lemma, creates a single bigram with the most frequent form, and sums the frequencies of bigrams and those of the noun unigrams.

As an example, this is normalization data for *kalean ibili* ('to walk on the street'):

```
kalean_kale_IZE_ARR_INE_NUMS<>ibili_ADI<>223 3354 10880
kaleetan_kale_IZE_ARR_INE_NUMP<>ibili_ADI<>119 243 10880
→
kalean_kale_IZE_ARR_INE_NUMS<>ibili_ADI<>342 3597 10880
```

Besides, ergative-singular → absolutive-plural normalization is carried out when the ratio is greater than 1:5. This heuristic is used in order to repair some mistakes from the tagger. Finally, partitive case (PAR) is assimilated to absolutive (ABS) for bigram normalization; partitive is a case used in negative, interrogative and conditional sentences with subjects of intransitive verbs and objects of transitive verbs. I.e. *ez zuen erabakirik hartu* ('he did not make any decision').

Thus, this is the normalization of *erabakia hartu*:

```
erabakia_erabaki_IZE_ARR_ABS_NUMS<>hartu_ADI<>2658 6329 88447
erabakiak_erabaki_IZE_ARR_ABS_NUMP<>hartu_ADI<>1632 2397 88447
erabakiak_erabaki_IZE_ARR_ERG_NUMP<>hartu_ADI<>88 141 88447
erabakirik_erabaki_IZE_ARR_PAR_MG<>hartu_ADI<>211 211 88447
→
erabakia_erabaki_IZE_ARR_ABS_NUMS<>hartu_ADI<>4589 9361 88447
```

4.6 AM calculation

The statistical analysis of cooccurrence data is carried out using Stefan Evert's UCS toolkit (Evert, 2005). The most common association measures are calculated for each bigram: f, t-score (also t-test), log-likelihood ratio, MI, MI^3, and chi-square (χ^2).

4.7 Evaluation

In order to evaluate the results of the bigram extraction process, we use as a reference a collection of

NV expressions published in five Basque resources: a) *The Unified Basque Dictionary*, b) *Euskal Hiztegia* (Sarasola, 1996); c) *Elhuyar Hiztegia*; d) *Intza* project; and e) EDBL (Aldezabal et al., 2001).

The total number for NV expressions is 3,742. Despite the small size of the reference, we believe that it may be valid for a comparison of the performance of different AMs. Nevertheless, even a superficial analysis reveals that the reference is mostly made up of two kinds of combinations, idioms and typical "compound verbs"[1].

Every evaluation against a dictionary depends largely on its recall and quality, and we envisage, as recommended by Krenn (1999), to build a handmade gold standard. To this end, we extract an evaluation sample merging the 2,000-best candidates of each AM ranking from the w = ± 1 extraction set. There are 4,334 different bigrams in this set. This manual evaluation is an ongoing work by a group of three experts (one of them is an author of this paper). Annotators were provided with an evaluation manual, with explanatory information about the evaluation task and the guidelines that must be followed to differentiate MWEs from free combinations, based on the criteria mentioned in section 2. Illustrative examples are included.

At present, a random sample of 600 has been evaluated (13.8%), with a Fleiss kappa of 0.46. Even though some authors have reported lower agreements on this task (Street et al., 2010), this level of agreement is comparatively low (Fazly and Stevenson, 2007; Krenn et al., 2004), and by no means satisfactory. It is necessary to make further efforts to improve the discriminatory criteria, and achieve a better "tuning" between the annotators.

5 Results

Figure 1 shows the precision curves obtained for each AM in the automatic evaluation. Frequency yields the best precision, followed by t-score, log-likelihood and MI^3. MI and χ^2 have a very low performance, even below the baseline[2]. These results are consistent with those reported by Krenn and Evert (2001) for support-verbs (FVG). Accordingly, this is the type of combination which is very much present in our dictionary reference.

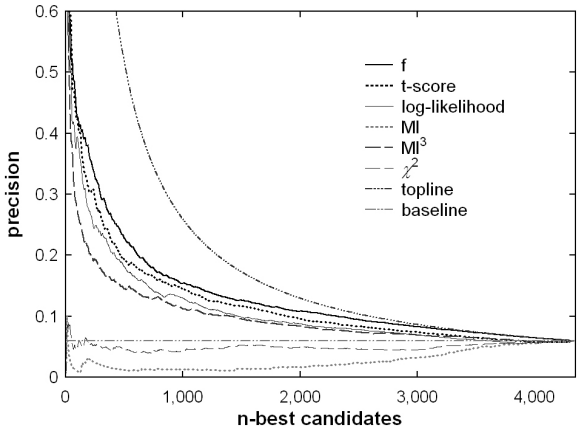

Figure 1: Precision results for the extraction set with w = ± 1 and $f > 30$.

Figure 2 offers an evaluation of the influence of window span and bigram normalization. The best results are obtained by the f ranking with a narrow window and without bigram normalization. Regarding bigram normalization, it could be concluded, at first sight, that the canonical forms included in the dictionary are not the most frequent forms of their corresponding MWEs. Thus, the frequency criteria used to normalize different forms of the same case and assign canonical forms must be reviewed. As for window span, the hypothesis that, since Basque is largely a free-word-order language, a wider window would yield more significant cooccurrence statistics, is not confirmed at the moment. Further analysis is needed to interpret these results from a deeper linguistic point of view.

Even though the manually evaluated random sample is small (600 combinations), some provisional conclusions can be drawn from the results. The amount of candidates validated by at least two of the three evaluators is 153, whereas only 29 of them are included in the dictionary reference. Even though MWE classification has not yet been undertaken by the annotator's team, a first analysis by the authors shows that most of the manually validated combina-

[1] Support verbs with syntactic idiosyncrasy (anomalous use of the indefinite noun), as *lan egin* ('to work') and *min hartu* ('to get hurt').

[2] Following Evert (2005), our baseline corresponds to the precision yielded by a random ranking of the n candidates from thedata set"; and our topline is "the precision achieved by an "ideal" measure that ranks all TPs at the top of the list".

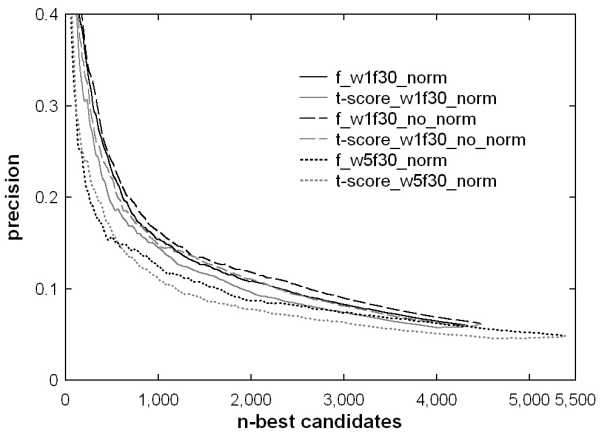

Figure 2: Precision results of f and t-score for three different extraction sets ($f > 30$): a) w = ± 1 with bigram normalization; b) w = ± 1 without bigram normalization; and c) w = ± 5 with bigram normalization.

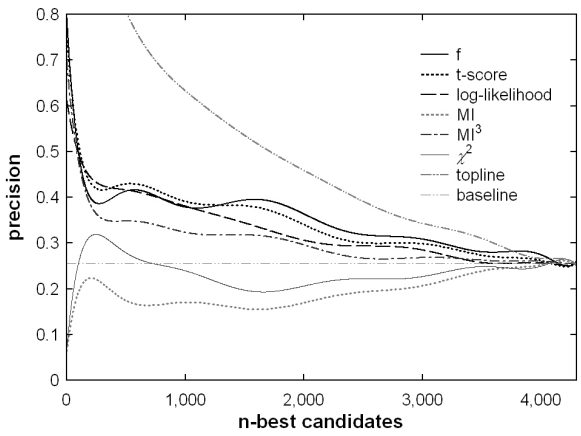

Figure 3: Precision results estimated from a 13.8% randon sample manually evaluated (600 conbinations).

tions not included in the dictionary (108 out of 124) are restricted collocations (mainly support-verb constructions that are not "compound verbs") or statistically idiosyncratic units. This is the first clue that confirms our suspicions about the limited coverage and representativeness of the reference. At the same time, it could be one of the possible explanations for the low inter-annotator agreement achieved, as far as those types of MWEs are the most difficult to differentiate from free combinations.

Figure 3 presents the precision curves for the complete evaluation set estimated from the manually evaluated random sample using the technique proposed by Evert and Krenn (2005). As expected, precision results increase compared with the evaluation against the dictionary. Frequency and t-score outperform the other AMs, but frequency is not the best measure in the whole range, as it is overtaken by t-score in the first 1,200 candidates.

6 Conclusions and Future work

The first results for the extraction of NV expressions in Basque are similar to the figures in Krenn and Evert (2001). Frequency and t-score are good measures and it seems difficult to improve upon them. Nevertheless, in light of the results, it is essential to complete the manual evaluation and build a representative gold standard in order to have a more precise idea of the coverage of the reference, and get a more accurate view of the behaviour of AMs in function of several factors such as the type of combination, corpus size, frequency range, window span, etc. Bigram normalization is, in principle, a reasonable procedure to formulate representative canonical forms, but requires a deeper analysis of the silence that it seems to generate in the results. Finally, the first evaluation using a small gold-standard is encouraging, because it suggests that using AMs it is possible to find new expressions that are not published in Basque dictionaries.

In the near future, we want to carry out a more comprehensive evaluation of the AMs, and study how to combine them in order to improve the results (Pecina and Schlesinger, 2006). In addition of this, we want to detect lexical, syntactic and semantic features of the expressions, and use this information to characterize them (Fazly et al., 2009).

Acknowledgments

This research was supported in part by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03-01) and by the Basque Government (Berbatek: Tools and Technologies to promote Language Industry. Etortek - IE09-262). Our colleagues Ainara Estarrona and Larraitz Uria are kindly acknowledged for providing their expertise as linguists in the manual evaluation process.

References

Aduriz, I., I. Aldezabal, I. Alegria, X. Artola, N. Ezeiza, and R. Urizar (1996). EUSLEM: A lemmatiser/tagger for Basque. *Proc. of EURALEX'96*, 17–26.

Aldezabal, I., O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández, and M. Lersundi (2001). EDBL: A general lexical basis for the automatic processing of Basque. In *IRCS Workshop on linguistic databases*, pp. 1–10.

Baldwin, T. and S. Kim (2010). Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*.

Banerjee, S. and T. Pedersen (2010). The design, implementation, and use of the Ngram Statistics Package. *Computational Linguistics and Intelligent Text Processing*, 370–381.

Church, K. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational linguistics 16*(1), 22–29.

Cowie, A. (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press, USA.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Ph. D. thesis, University of Stuttgart.

Evert, S. and B. Krenn (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language 19*(4), 450–466.

Fazly, A., P. Cook, and S. Stevenson (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics 35*(1), 61–103.

Fazly, A. and S. Stevenson (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pp. 9–16. Association for Computational Linguistics.

Granger, S. and M. Paquot (2008). Disentangling the phraseological web. *Phraseology. An interdisciplinary perspective*, 27–50.

Krenn, B. (1999). *The usual suspects: Data-oriented models for identification and representation of lexical collocations*. German Research Center for Artificial Intelligence.

Krenn, B. and S. Evert (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pp. 39–46.

Krenn, B., S. Evert, and H. Zinsmeister (2004). Determining intercoder agreement for a collocation identification task. In *Proceedings of KONVENS*, pp. 89–96.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Proc. of LREC 2002*, pp. 1530–1536.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pp. 13–18. Association for Computational Linguistics.

Pecina, P. and P. Schlesinger (2006). Combining association measures for collocation extraction. pp. 651–658.

Pérez Gaztelu, E., I. Zabala, and L. Grácia (2004). *Las fronteras de la composición en lenguas románicas y en vasco*. San Sebastián: Universidad de Deusto.

Sarasola, I. (1996). *Euskal Hiztegia*. Kutxa Fundazioa / Fundación Kutxa.

Seretan, V. (2008). *Collocation extraction based on syntactic parsing*. Ph. D. thesis, University of Geneva.

Sinclair, J. (1996). The search for units of meaning. *Textus 9*(1), 75–106.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics 19*(1), 143–177.

Street, L., N. Michalov, R. Silverstein, M. Reynolds, L. Ruela, F. Flowers, A. Talucci, P. Pereira, G. Morgon, S. Siegel, M. Barousse, A. Anderson, T. Carroll, and A. Feldman (2010). Like finding a needle in a haystack: Annotating the american national corpus for idiomatic expressions. In *Proc. of LREC 2010*, Valletta, Malta.

Wulff, S. (2010). *Rethinking Idiomaticity*. Corpus and Discourse. New York: Continuum International Publishing Group Ltd.