Morphosyntactic structure of terms in Basque for automatic terminology extraction

R. URIZAR, N. EZEIZA, I. ALEGRIA,
Donostia, The Basque Country, Spain

Abstract
This paper describes the morphosyntactic patterns of technical terms in Basque and presents an architecture for a term-extracting tool. As Basque is a highly inflected agglutinative language, part-of-speech information is not enough to define term patterns. The use of morphological and syntactic information is essential to reduce considerably the number of structures. For example, a noun, an adverbial, a postpositive adjectival, the nominal form of a verb and even a determiner in the genitive case may work as a prepositive adjective; however, they all share the same syntactic function. Therefore, for the term-extracting tool to perform properly, the texts must be morphosyntactically analysed and disambiguated. Then a shallow syntactic parser will identify the previously described patterns.

1 Introduction

These last years many tools for extracting terminology from technical texts are being developed for different languages but especially for English. Due to the development in the Natural Language Processing (NLP) and other fields, technology is now ready for creating this kind of applications. However, in the sight of the results, human help is still essential to make the final selection among the possible terms automatically chosen. The candidate terms extracted using these tools can be useful in many aspects of NLP, such as indexing information (text indexing) or constructing term glossaries (either for translation purposes or for dictionary making). Text indexing is a highly topical question since, due to Internet, there is a huge amount of information in the web, which would turn out to be quite useless if it couldn't be efficiently selected. The automatic construction of term glossaries is very useful in the fields of terminology, translation and book publishing. Moreover, in domains in which terminology is being transformed in a highly dynamic way, e.g. in computer science, it would be almost impossible to do any effective terminological work without such a tool.

When trying to develop a similar tool for Basque we encountered extra difficulties. On the one hand, as the unification process of the language is not finished yet, fixing terminology becomes more complicated. On the other hand, there is little research made in this aspect of the language. Besides, being Basque an agglutinative and a highly inflected language, the treatment needed for identifying terminology in texts is much more complex than for the surrounding languages since morphosyntactic information is vital.

2 Terminology extraction

The aim of our project is to extract terms automatically from specialised corpora. However, the first problem arises when defining *term*, or rather when trying to distinguish *terms* from *non-*

terminological units. We will hardly find a formal definition for *term* that satisfies everyone. But intuitively we can say that term is a lexical unit that makes reference to a specific concept in a limited domain and so occurs mostly in specialised (technical) discourses. Sometimes a terminological lexical unit may also be found in general discourses or in different domains, but then its meaning becomes broader or even blurred.

If we have a look at the state of the art we can clearly observe an increasing interest in term extraction in the last few years. This is reflected in several tools such as *LEXTER* [Bourigault 1992], AT&T's *Termight* [Church/Dagan 1994], IBM's *TERMS* [Justeson/Katz 1995], *NPtool* [Arppe 1995] and *ACABIT* [Daille/Jacquemin 1998]. These tools usually combine two different tasks to perform terminology extraction. The first task is usually responsible for the extraction of candidate terms and the second one refines this candidate list selecting those which fit some restrictions. In some projects [Su et al. 1996] the tool for extracting terminological phrases from technical corpora just seeks for combinations of two or three words, without any pattern restriction, in texts that have not been previously analysed neither morphologically nor syntactically. Nevertheless, many other term-extracting tools [Bourigault 1992, Justeson/Katz 1995, Ananiadou 1994] apply shallow analysis to texts before they get searching for terminological phrases.

Here we must distinguish one-word terms from multiword terms. Although, as [Justeson/Katz 1995] state, "judging from data in dictionaries of technical vocabulary, the majority of technical terms do consist of more than one word", in the three dictionaries we analysed one-word terms amounted to 42% of the total, which is by no means negligible.

2.1 Repetition and term recognition

Terminological phrases are more susceptible to repetition than non-terminological phrases, which trend to resort to near synonyms and to use different modifiers in subsequent references to the same entity so as to avoid repetition. Exact repetition of non-terminological phrases causes a monotonous effect and it is only likely to occur if they are separated widely enough in large texts, so some sort of stylistic variation is usual. Besides, unlike in non-terminological phrases adding or changing a modifier in terminological phrases often causes a reference change too. Some statistical methods used in different projects are explained in section 4.

Besides, to identify one-word terms, the frequency of the words/lemmas that occur in a specialised corpus must be compared with their frequency in a general corpus. Terms, by definition, occur in specialised discourses. When these terms appear in general types of discourse or in a variety of domains it often has broader or more diverse meanings [Justeson/Katz 1995] and so a lower frequency.

3 Morphosyntactic structure of Basque terms

Unlike surrounding languages Basque is an agglutinative tongue with quite a complex morphology. For example, the morphological analyser for Basque is theoretically capable of recognising about 460.000 inflected forms for each noun taking into account only two levels of recursion [Agirre et al. 1992], while in English it is just enough to make a number distinction.

Consequently, it is necessary to perform morphosyntactic analysis, as will be described in section 4, before we face terminology extraction task. Once the text is morphosyntactically analysed, we need to modelise the structure of the terms, that is, we must specify the morphosyntactic structure that terms may adopt, in order to be able to detect candidate terms. These candidate terms will then be filtered using statistical techniques and eventually a linguist will make the final selection.

In order to describe the structure of terms in Basque, we analysed three technical dictionaries from different domains and at random we extracted a sample of 150 terms from each. We chose Computer Science [Euskalterm 1993], Civil Service [HAEE/IVAP 1995], and Football [Uzei 1985] dictionaries, since it is easier for us to get corpora on these subjects for further evaluation and application.

Among all terms extracted at random from the dictionaries above mentioned NPs constitute 78.2% of the total, 18.2% are VPs and 3.4% are adjectives. 41.9% are one-word units, 70.0% of which correspond to nouns, 23.6% are lexical verbs and 6.4% adjectives.

Dictionary Type	Football	Computer Science	Civil Service	Overall
Nouns	68.9	62.3	79.0	70.0
Verbs	26.9	30.2	13.8	23.6
Adjectives	4.5	7.5	7.2	6.4

Table 1: Frequencies (%) of one-word terms.

Noun phrases amount to 83.2% of the multiword terms and the rest (16.8%) are verb phrases although in the sample from computer science dictionary we found a non-significant amount of them.

Overall, the average length of terms in the samples subject to investigation is 1.71 with no significant variations among the different dictionaries studied; football terms seem to be slightly shorter with 1.61 of average length. This is lower than the 1.91 average length for English terms provided by [Justeson/Katz 1995]. Such a difference is easily explained by the fact that Basque is an agglutinative language.

Only 2 of the terminological NPs studied have a conjunction (*eta* 'and') and just 5 have a determiner (ordinal and cardinal numerals) other than the article[1]. Among the multiword NPs 37.8% are compounds made up of only nouns, usually two, since Basque rarely accepts compounds of more than two nouns. Postpositive adjectives[2], which always go after the noun, occur in up to 22.1% of the terminological noun-phrases subject to investigation. 15.4% of the postpositive adjectives are actually past participles of lexical verbs e.g. *kopia egiaztatu* 'certified copy'. Prepositive adjectivals occur in 38.1% of the multiword terminological NPs examined.

Most of the prepositive adjectives in Basque, which mainly go before the noun in NPs[3], are of the kind of *barruko* 'internal' (lit. 'of inside'); in fact the morphological analyser also gives the "noun + locative genitive case" (*barru+ko*) interpretation along with the lexicalised adjective *barruko*. Therefore, any word, regardless of its grammatical category, taking any of both

genitives in Basque (locative and possessive) may function as a *prepositive adjectival* and this is reflected in the morphosyntactic analysis by the syntactic function[4] @IZLG> (*prepositive adjectival*). Thus, the head of the prepositive adjectival may be either a noun as in *gobernuaren ordezkari* 'delegate of the government', or an adverbial as in *txikizkako salmenta* retail sale (lit. 'sale of by retail'), or a postpositive adjective as in *zuzeneko esleipen* 'direct designation', the nominal form of the verb as in *irekitzeko lizentzia* 'opening licence' (lit. 'licence for/of opening') or even a determiner as in *hiruko torneo* 'triangular tournament' (lit. 'tournament of three'). Although most prepositive adjectivals are actually built directly over a noun or an adverbial and the ones built over determiners, for example, are quite residual, they all share the same syntactic function. So we used this information instead of the one about part of speech to make the term patterns, reducing considerably their number.

On the other hand, 70% of the terminological VPs include a noun in the absolutive case functioning as a direct object (*errekurtsoa aurkeztu* 'to appeal a case') and the remaining 30% contain an adverbial (*lanetik bota* 'to make redundant'). As with prepositive adjectivals, an adverbial in VPs can be either an adverb or a word with the adverbial syntactic function (@ADB). This word is usually a noun with an adverbial case mark, e.g. *argitara eman* 'to make public', *indarrean sartu* 'to come into force', *pilatik atera* 'to pop' (lit. 'to extract from the stack'), *oinez urrundu* 'to clear (a ball) with the foot' but sometimes, it can also be an adjective as in *laburrean pasatu* (lit. 'to pass in short') or even the past participle of a lexical verb as in *ibilian jaurti* 'to shoot on the run'.

The regular expression below covers 97.5% of the multiword terms thoroughly studied, ranging from the highest 98.7% in the football dictionary to the lowest 95.9% in the Civil Service dictionary.

$$((N_{nc}|A_{prep})^{+}(N|A^{+}_{pos})|((ADV|(N^{*}_{nc}\ N_{abs}))\ V)$$

Besides, the most frequent patterns shown in table 2 cover up to 92.2% of the samples.

Type	**Dictionary Patterns**[5]	Football	Computer Science	Civil Service	Overall
Noun	N_{nc} N	21.6	36.2	33.9	30.6
	$A_{prep}\ N_{nc}$? N	30.7	25.5	14.9	23.7
Phrases	$N_{nc}\ N_{nc}$? A_{pos}	14.8	23.4	13.8	17.3
	$A_{prep}\ A_{prep}$ N	3.4	2.1	3.4	3.0
	$N_{nc}\ A_{prep}$ N	0.0	5.3	1.7	2.3
Verb Phrases	N_{nc}? N_{abs} V	18.2	0.0	14.9	11.0
	ADV V	4.5	2.1	6.3	4.3

Table 2: Frequencies (%) of the main morphosyntactic patterns.

4 Automatic term recognition in Basque

As we pointed out in section 3, it is essential to have the text previously analysed (or at least lemmatised and tagged). [Koskenniemi 1996] shows that for Finnish, a language with very rich

morphology, the use of an inflectional analyser monotonically improves recall. Also for Dutch [Kraaij/Pohlmann 1996] and French [Jacquemin/Tzoukermann 1999] inflectional and derivational analysis has proved to be efficient to improve the quality in textual IR.

4.1 Architecture of the term-extracting tool

The IXA research group (http://ixa.si.ehu.es) has already developed a set of tools for Basque, which will be used in this project for the basic analysis. The term-extracting tool is based on the results of lemmatisation.

We use EUSLEM, a lemmatiser/tagger for Basque [Ezeiza et al. 1998]. It has two main modules (see figure 1):

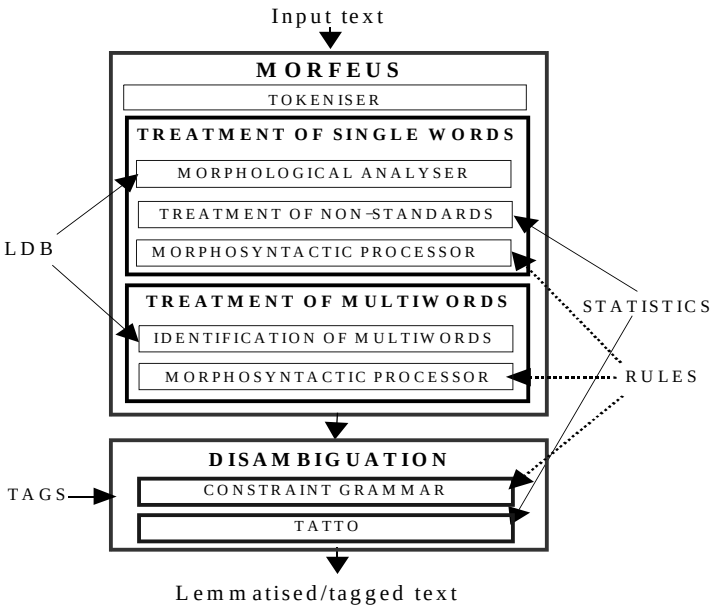

Figure 1: Architecture of EUSLEM

- MORFEUS, a robust morphosyntactic analyser. MORFEUS is performed in two main phases:

  a) The morphosyntactic treatment of single word units:

     1. Morphological analysis, which assigns to each input word all the corresponding interpretations [Alegria et al. 1996] in three incremental phases. In the first phase the analysis (both inflectional and derivational) of standard words is made, in the second one dialectal variants and competence errors are analysed, and finally the analyser deals with the words that don't have an entry in the lexical database. This last phase is very important in terminology extraction, because it is quite probable that a term or one of its components be a loan word or a recently coined item that is not included in the lexicon.

2. Disambiguation of lemmas of non-standard words using both statistical and linguistic information. This treatment is necessary because the average number of interpretations in non-standard words is significantly higher than in standard ones.
3. Morphosyntactic processing of the result of the analysis to elaborate the morphological information [Aduriz et al. 2000]

b) The morphosyntactic treatment of multiword units in two steps: first, it identifies widely used multiword lexical units [Aduriz et al. 1996] and named entities, and then, morphosyntactically processes the results.

- The morphosyntactic disambiguation module, which is achieved in two steps: first we apply the constraint grammar for Basque [Aduriz et al. 1997] and, then a HMM-based disambiguator [Armstrong et al. 1995].

Based on this tool, we have designed the architecture of the term extraction tool shown in figure 2. First, the text is morphologically analysed and disambiguated, assigning to each token its corresponding lemma, POS tag, and morphosyntactical information. Then, we use the shallow syntactic parser to identify the NP and VP patterns described in section 3. Eventually, we select terms from the candidate list using statistical measures.

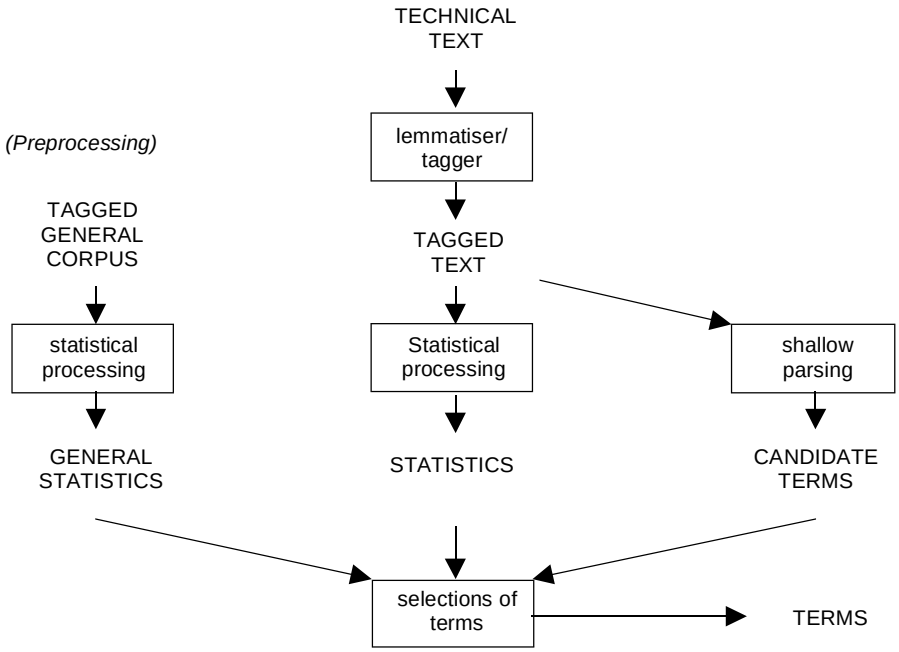

Figure 2: Architecture of the term extraction tool

4.2 Statistical methods for term recognition

The statistical methods applied vary considerably depending on the project. The simplest one would consist on getting a minimal absolute frequency [Justeson/Katz 1995], but in

most of them many probabilistic formulae are combined, among which *mutual information* [Church/Hanks 1990] stands out. This formula is the basis of many systems but sometimes it is combined with more sophisticated ones. For instance, [Su et al. 1996] compare the probability of co-occurrence of a word combination in a given text with its co-occurrence probability in a big and balanced corpus, so as to obtain the relative frequencies, since these can be vital in order to discriminate terminology.

In order to normalise the mentioned terms, i.e. to remove the terms included in longer combinations, a project in Manchester University [Frantzi/Ananiadou 1996] proposes a formula named *C-value*. Applying this formula they are capable of distinguishing, for example, *soft contact lenses, hard contact lenses*, and *contact lenses* as different terms, excluding, however, *soft contact*.

Nevertheless, researchers using these formulae are conscious of the limits of statistics. For instance, [Jacquemin/Tzoukermann 1999] propose a further treatment to expand multiword terms using derivational morphology and syntax. They apply some transformational rules so as to detect not only morphosyntactic variants such as *variation de climat* 'climate variation' and *variation climatique* 'climatic variation' but also syntactical ones e.g. *fruits et agrumes tropicaux* 'tropical fruits and citrus' *agrumes tropicaux* 'tropical citrus' and *fruits tropicaux* 'tropical fruits'. However, they don't deal with semantic variations containing synonyms such as *kidney function* and *renal function*. Thus, there seems to be a lot to do in this aspect if we want the results to be accurate enough.

We have adopted the approach of [Su et al. 1996], but, in the first version of the tool, we are using lemmas instead of words to calculate statistical measures. However, we are also planning to identify n-grams combining different sources of information – e.g. word, lemma, POS tag, syntactic tag... due to the morphosyntactic features of the language.

At the moment we are implementing the statistical engine and combining the general shallow parser [Aduriz et al. 1997] and another finite state parser as proposed in [Gojenola/Oronoz 2000] to filter out only the phrases that fit the constraints described in section 3. Finally, we are also compiling corpora for each selected subject and, even if they are still quite small, we think they will soon be big enough for the tool to obtain preliminary results.

5 Conclusions and future work

We have described a terminology extraction tool for Basque, a highly inflected agglutinative language. This feature of the language makes it essential to process the input texts morphosyntactically.

In the near future we intend to extend the term-extracting tool so that it identifies morphological (both inflectional and derivational) variables too. It would be also interesting to use semantic information for the normalisation task so as to detect semantic variants of the terms such as synonyms, but for the moment, we don't intend to treat this kind of variables.

We want to point out that the limited availability of technical texts in Basque makes this work harder. It may seem that the statistical measures will not be significant enough to make conclusions about the results. However, as this work is based not only on statistical measures but

also on the morphosyntactic description of terms, we hope the results will be an acceptable first approach to automatic recognition of Basque terms.

Notes

[1]We don't take into account the article since in Basque it goes before the case mark and so attached to the lemma, e.g. *etxearen* 'of the house' (lit. 'house-the-of').

[2]In Basque we can distinguish two main different types of adjectives, postpositive and prepositive.

[3]Prepositive adjectives in Basque may sometimes occur after the noun in NPs but we didn't find a single example of this use since it is a stylistic variant that is avoided in technical terminology.

[4]We use Constraint Grammar [Karlsson et al.1995] notation to express syntactic tags.

[5]We have used a POS-based notation for simplicity, but we also use syntactic information to define patterns.

[6]N stands for noun, A_{pos} is a postpositive adjective, A_{prep} is a prepositive adjectival, V is a lexical verb, and ADV is an adverbial. Nabs means the noun must be in the absolutive case and N_{nc} is a noun with no case mark on it. Note that in Basque only the last word in a NP can take a case mark other than any of both genitives.

6 Acknowledgements

This research has been partially supported by the Basque Government (EX1998-30 and pre-doctoral grant BFI97.202) and by the University of the Basque Country (UPV-19/1999).

References

[Aduriz et al. 2000] Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. A Word-Level Morphosyntactic Analyzer for Basque *Second Int. Conf. on Language Resources and Evaluation LREC2000*. Athens (Greece). 2000.

[Aduriz et al. 1997] Aduriz, I., Arriola, J.M., Artola, X., Díaz de Illarraza, A., Gojenola, K., Maritxalar, M. Morphosyntactic Disambiguation for Basque based on the Constraint Grammar Formalism. *Proceedings of RANLP'97*, Bulgaria. 1997.

[Aduriz et al. 1996] Aduriz I., Aldezabal J.M., Artola X., Ezeiza N.,Urizar R. MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque . *Papers in Computational Lexicography, Complex'96.* pp. 1-8. Linguistics Institute, Hungarian Academy of Sciences. Budapest. 1996.

[Agirre et al. 1992] Agirre E., Alegria I., Arregi, X., Artola X., Díaz de Ilarraza A., Maritxalar, M., Sarasola K., Urkia M. Xuxen: a Spelling Checker/Corrector for Basque based in Two-Level Morphology. *Proceedings of ANLP'92*, 119-125. Povo Trento. 1992.

[Alegria et al. 1996] Alegria I., Artola X., Sarasola K., Urkia M. Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing* 11 (4): 193-203. Oxford University Press. 1996.

[Ananiadou 1994] Ananiadou S. A Methodology for Automatic Term Recognition. *Proc. of the Conference on Computational Linguistics (Coling-94)*, 1034-1038, Kyoto, Japan. 1994.

[Armstrong et al. 1995] Armstrong S., Russel G., Petitpierre D., Robert G. An open architecture for Multilingual Text Processing. *Proceedings of EACL'95*. vol 1, 101-106. 1995.

[Arppe 1995] Arppe A. Term Extraction from Unrestricted Text. *http://www.lingsoft.fi/doc/nptool/term-extraction.html*.

[Bourigault 1992] Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases. *Proc. of the Conference on Computational Linguistics (Coling-92),* 977-981, Nantes, France. 1992.

[Church/Hanks 1990] Church K., Hanks P. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics***16**:22-29. 1990.

[Church/Dagan 1994] Church K., Dagan I. Termight: Identifying and Translating Technical Terminology. *Proc. of the 4th Conference on Applied Language Processing,* Stuttgart, Germany. 1994.

[Daille/Jacquemin 1998] Daille B., Jacquemin C. Lexical Database and Information Access: A Fruitful Association?. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),669-673*. Granada. 1998.

[Euskalterm 1993] Euskalterm, Informatika Fakultatea. *Informatika Hiztegia,* Ed. Elkar. 1993.

[Ezeiza et al. 1998] Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. Combining stochastic and rule-based methods for desanbiguation in agglutinative languages. *Proc. of the Conference on Computational Linguistics (Coling-ACL'98)*. Montreal 1998.

[Frantzi/Ananiadou 1996] Frantzi K.T., Ananiadou S. Extracting Nested Collocations. *Proc. of the Conference on Computational Linguistics (Coling-96),* 41-46. 1996.

[Gojenola/Oronoz 2000] Gojenola, K. Oronoz, M. Corpus-Based Syntactic Error Detection Using Syntactic Patterns, *NAACL-ANLP00, Student Research Workshop,* Seattle. 2000

[HAEE/IVAP 1995] HAEE/IVAP. *Administrazioko mila hitz,* Ed. HAEE/IVAP, Gasteiz. 1995.

[Jacquemin/Tzoukermann 1999] Jacquemin, C., Tzoukermann, E., "NLP for term variant extraction: synergy between morphology, lexicon and syntax", in: Tomek Strzalkowski (Ed.): *Natural Language Information Retrieval*, (Dordrecht: Kluwer Academy Publishers) 1999: 25-74.

[Justeson/Katz 1995] Justeson J.S., Katz S.M. Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering***1** (1): 9-27. Cambridge University Press. 1995.

[Karlsson et al. 1995] Karlsson F., Voutilainen A., Heikkila J., Anttila A. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Mounton de Gruyter. 1995.

[Koskenniemi 1996] Koskenniemi, K., Finite-state morphology and information retrieval. *Porceedings of the ECAI-96 Workshop on Extended Finite State Models of Language*, pp. 42-45, ECAI, Budapest, Hungary. 1996.

[Kraaij/Pohlmann 1996] Kraaij, W., Pohlmann, R. Viewing stemming as recall enhancement. *Proceedings, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'96), pp. 40-48, Zurich.

[Su et al. 1996] Su K., Wu M., Chang J. A Corpus-based Approach to Automatic Compound Extraction. *Proceedings of the Conference on Computational Linguistics (Coling-96),* 243-247. 1996.

[Uzei 1985] Uzei. *Kirola.1. Futbola,* Ed. Elkar, Donostia. 1985.