I. Alegria (1), A. Gurrutxaga (2), P. Lizaso (2), X. Saralegi (2), S. Ugartetxea (2), R. Urizar (1)

(1) Ixa taldea –University of the Basque Country
649 Postakutxa. 20080 Donostia
acpalloi@si.ehu.es
(2) Elhuyar Fundazioa
Astesuain Poligonoa, 14 - 20170 Usurbil
agurrutxaga@elhuyar.com

Linguistic and Statistical Approaches to Basque Term Extraction

The development of applications for terminology extraction in Basque demands previous research on linguistic techniques, in order to fulfil the requirements of Basque language processing. Being Basque an agglutinative language, the results of pure statistical methods are not satisfactory and suitable for term extraction. In this work, we have adopted a hybrid approach, based on the selection of term candidates by means of language techniques and the subsequent application of statistical association measures. In this work, we will focus mainly on linguistic technique design, and we will overview the first experimental results. This work is part of Erauzterm, a project for the development of a term extraction tool for Basque. The tool is in its first stage of development, and future improvements are close. Erauzterm is the first attempt to develop such a tool for Basque.

1 Term extraction approaches

As in other areas of automatic language processing, two main approaches have been proposed for terminology extraction from texts: linguistic and statistical approaches.

On the one hand, linguistic techniques rely on the assumption that terms present specific morphosyntactic structures or patterns (Bourigault, 1996). The basic strategy of these techniques is to detect and extract the strings whose structure match some given pattern. Since these patterns are in most cases language-dependent, linguistic techniques demand specific language knowledge processing. On the other hand, statistical approaches take into account that terms have different statistical features from normal words to identify them (for example, the high association grade of multiword constituents). Exactly, in order to estimate the *termhood* of the candidates, we can use statistical models which analyse observed counts of linguistic information related to the candidates. Most of the statistical approaches focus on the extraction of multiword terms, mainly by means of calculating association measures (Chuck & Hanks, 1990; Smadja, 1993; Dias, 1999).

Moreover, some authors adopt hybrid approaches, combining linguistic and statistical techniques. Some of them apply syntactic filters after statistical processing, in order to extract the statistically significant word combinations that match some given morphosyntactic pattern (Samdja, 1993). In other cases, statistical measures are calculated for a list of term candidates previously selected through linguistic techniques (Daille, 1995; Justeson, 1993).

In this project, we have adopted the latter approach. Thus, our first concern was the identification of the main features of terms in Basque.

2 Features of Basque terms

Basque is an agglutinative language with a very rich inflectional system. That is to say, a lemma can appear in a large amount of cases or inflected forms. Besides, word order and phrase structure differ in some aspects from other languages such as English, Spanish or French. Therefore, a language specific approach is needed if we want to use morphosyntactical information for term extraction.

In the last fifteen years, as a result of NLP research for Basque, important developments have been reached in morphological analysis, lemmatisation and syntactic analysis, and nowadays the implementation of linguistic techniques for term extraction is a feasible task.

2.1 Structure of Basque Terms

Prior to this work, a research about terminological structures in Basque was carried out by the IXA group of The University of The Basque Country (Urizar et al., 2000). In the investigation, they worked on three dictionaries from different domains extracting a sample of 150 terms from each of them. The results showed that 42% of the terms were one-word terms, among which, 70% were nouns, 23.6 verbs and 6.4%, adjectives. Noun Phrases (NP) constituted 78.2% of the total, 18.2% were Verb Phrases (VP) and 3.4% adjectives. Regarding multi-word terms, the proportion of NPs increased to 83.2%; VPs amounted to 16.8%.

It was necessary to compare the results of the analysis of terms in dictionaries with the distribution of term structures occurring in real texts. For this purpose, we processed manually a sample of 13,756 words, composed of 28 articles from our corpus. This corpus would also be a reference for a further automatic evaluation of the term extractor.

The methodology for the manual extraction of terminology was defined beforehand for the extraction to be systematic. Firstly, for the selection of the sample, 28 divulgation articles on computer science were chosen randomly. Single words occurring just once in a text (Hapax Legomena) amount to 28.20%; in the case of multiword terms this percentage increases up to 77.47%. The high amount of too low frequency words and the lack of representativeness of the word frequency (small corpus) make statistical inference difficult. A bigger homogeneous corpus would undoubtedly result in a considerable reduction of Hapax, an increase in terms with more representative frequencies, and, therefore, in an improvement in statistical estimation of terms.

Secondly, the criteria for the manual tagging of terms were defined. Although it is difficult to determine a definition of term that would satisfy everyone, in our case, three terminologists made the selection of terms from the sub-corpus based on the maximal term phrase selection. That is, in the case of compound terms such as *posta elektronikoko mezu* ('e-mail message'), which contains the nested term *posta elektroniko* ('e-mail'), only the longest term was extracted. Sometimes, different terminologists marked different terms from the same terminological phrase. These differences were detected and corrected, to assure that a given text occurrence led to a single extracted term.

Thirdly, all the terms obtained manually from the sample were described by means of morphosyntactic patterns. The collected terms were reviewed and assigned an appropriate pattern. During this phase, the decision on labelling the obtained terms containing numbers, acronyms and foreign terms needed to be considered. For example, the following terms were considered as two and three-word terms, each one attached to its corresponding pattern: *Windows 98* ($N_{nc}N$), *Flip/flop* ($N_{nc}N$), *Pentium II* ($N_{nc}N$), *IBMren*

Aptiva ($A_{prep}N$), *AMDren k6 txip* ($A_{prep}N_{nc}N$). This step is very important since it is essential that the categories used in the patterns match these in the output of the tagger.

Table 1 shows the frequency of the different patterns manually marked in the corpus.

Type	Pattern[1]	Frequency	Overall %	Multiword %	
One-word terms	N	275	39.34		
	V / A / Adv	34	4.86		
Multiword terms	$N_{nc}N$	138	19.74	35.38	83.83
	$N_{nc}A_{pos}$	65	9.30	16.67	
	$A_{prep}N$	55	7.87	14.10	
	$N_{nc}A_{prep}N$	24	3.43	6.15	
	$N_{nc}N_{nc}N$*	9	1.29	2.31	
	$A_{prep}N_{nc}N$	8	1.14	2.05	
	$A_{prep}A_{prep}N$	2	0.29	0.51	
	$A_{prep}N_{nc}A_{pos}$*	5	0.72	1.28	
	$N_{nc}A_{prep}N_{nc}N$*	5	0.72	1.28	
	$N_{nc}N_{nc}A_{pos}$	3	0.43	0.77	
	$N_{nc}A_{pos}A_{pos}$*	3	0.43	0.77	
	$N_{nc}N_{nc}A_{prep}N$*	3	0.43	0.77	
	$N_{abs}V_{gen}N$*	5	0.72	1.28	
	$N_{nc}N_{nc}N_{nc}N$*	2	0.29	0.51	
	Other NPs	7	1.00	1.79	16.15
	Other Patterns	56	8.01	14.36	
		699	100.00	100.00	100.00

Table 1. Frequency of term patterns manually extracted

Noun phrases, either one-word or multiword terms, comprise 87.14% of the total, and 53.22% are multiword units. On the other hand, multiword units are mostly noun phrases (only 7.3% of the multiword units are not noun phrases). As regard to the detected patterns, we have included some patterns not present in the previous work. These new patterns are marked with an asterisk symbol (*).

2.2 Term Variation in Basque

Term variation can be defined generally as the fact that a specialized concept can be expressed by more than one linguistic form. This definition is very loose and includes many different phenomena, from simple graphical variation to synonymy.

From the point of view of information extraction, it is well known that, if the same concept can be formulated in different ways, which are known as variants, an information extraction tool should be able to relate those different linguistic forms or expressions of a concept, in order to avoid missing relevant documents.

In the field of terminology and research in specialized discourse, some authors have pointed out the gap between term representation in terminological glossaries or technical dictionaries, and the linguistic forms used in real texts to express concepts. According to Daille et al. (2000), "describing terms as fixed sequences is obviously an idealised viewpoint". According to Cabré (2001), term variation is one of the central issues of the new theoretical proposal known as Communicative Theory of Terminology.

[1] N: noun; N_{nc}: non-case noun; A: adjective; A_{prep}: prepositive adjective; A_{pos}: postpositive adjective; V: verb; V_{gen}: verb plus genitive; Adv: adverb

This is important not only from the practical point of view of indexing for information retrieval, but also from the perspective of terminologists' needs, whose main goals are to describe term usage and to provide writers, translators, etc. with efficient and useful resources that must be usable in real contexts. Moreover, even from the restricting and prescriptive perspective of term standardization, it is necessary to know which are the terms actually used in technical texts since that information is essential in making decisions that can help term normalization.

Most research on term variation, and on the ways to manage it for appropriate information extraction, takes as a starting point 'controlled terms', also named 'original terms' or 'base terms'. Some linguistic approaches (Daille, 1995; Jacquemin, 2001) define rules that associate controlled terms of length 2 with a set of possible variations. Pure statistical approaches, mostly based on association measures, extract binary associations by means of these measures, and, afterwards, apply enticement techniques to acquire longer terms, which include variations of terms of length 2.

2.2.1 Classification of term variants

Different kinds of term variants are distinguished in the bibliography: orthographic variants, inflectional variants, morphosyntactic variants, syntactic variants and semantic variants.

a. Orthographic variants

This type of variation is habitual mainly due to capitalization (induced by punctuation or not). Table 2 shows different kinds of orthographic variants and examples.

Variant	Examples in Basque	English translation
Capitalization	*Informatika Sail / informatika sail* *Internet / internet*	*Department of Computer Science*
Inner hyphen insertion	*programazio lengoaia / programazio-lengoaia*	*programming language*
Dropped final *a* (in some old words and words ending *–ia*)	*hizkuntza-tresna / hizkuntz tresna* *telefonia-sare / telefoni sare /*	*language tool* *telephone network*

Table 2. Different orthographic variants

These variations are managed by the lemmatiser-tagger *Euslem* (Ezeiza et al., 1997), which relates them to only one lemma or canonical form.

b. Inflectional variants

When defining these kinds of variants, Jacquemin et al. (2000) include the singular or plural forms of words, and the infinitive, past participle and gerund forms of verbs. Similar criteria are applied in other works (Nenadic et al., 2002). However, in Basque inflectional variation is a central issue in language processing, as Basque is an agglutinative language, with a very rich inflectional system. The amount of inflectional forms in which a given lemma or the 'canonical form' of a given term can appear in texts is extensive. For instance, the following are some examples of *sistema eragile* ('operating system'): *sistema eragilearen* (gen. sing.), *sistema eragileari* (dat. sing.), *sistema eragileetan* (loc. pl.), *sistema eragiletan* (loc. indef.). These types of inflections is not relevant for term analysis, but some inflections of words 'within' the term ought to be accounted for, for instance, *Interneteko konexio* / *Internerako konexio* ('Internet conexion'). Even though inflections are involved in this type of inner variations, our opinion is that they would be more appropriately classified as morphosyntactic variants in the next section.

Language technology for Basque is at present ready for the task of morphological analysis and lemmatisation, due to the work carried out by the IXA group, and inflections of canonical forms can be processed in a straightforward way by means of lemmatisation tools (*Euslem*). Inner inflectional variants can be normalized by means of the extraction of lemma-lemma bigrams, and, at the same time, information about inner variation suffixes can be retained in the canonical forms (or form-lemma bigrams).

c. Morphosyntactic variants

In this type of variations, related words as derivatives are involved. The main components are nouns (N), prepositive adjectives (A_{prep}) [2], postpositive adjectives (A_{pos}) and nominal form of the verb (V_{nom}). The most important equivalent cases with examples are shown in table 3.

Variant model	Examples in Basque	English translation
$N_1N_2 \Leftrightarrow A_{prep1}N_2$ [3]	*sare-kudeaketa ⇔ sare(ar)en kudeaketa*	network management
$N_1N_2 \Leftrightarrow N_1V_{nom2}$	*sare-kudeaketa ⇔ sarea(k) kudeatzea*	network management
$N_1N_2 \Leftrightarrow N_2A_{pos1}$	*informatika-ekipo ⇔ ekipo informatiko*	computing equipment

Table 3. Different morphosyntactic variants

In the third case, the order of the words is inverted, but morphological changes occur as well; thus, this type of permutation is different from permutation in English or languages with similar word order. In those languages, permutation variants are syntactic variants, in which words remain the same, not being substituted by derivatives: (*birth date ⇔date of birth*). From the point of view of meaning equivalence, this type of Basque variants is similar to morphological English variants as *cell density ⇔cellular density*.

d. Syntactic variants

Insertion, juxtaposition and coordination are habitual variants.[4] Insertion and juxtaposition can be associated with inflectional variations. In table 4, some examples are shown.

[2] The genitive case may work as a prepositive adjective

[3] Such variants are not always equivalents or synonymous, just in the same sense of English morphological variants like *corn kernel* and *kernel of corn* have not necessary the same meaning (Jacquemin, 2001). In other cases, meanings are not equivalents at all: *sistema eragile ≠ sistemaren eragile* ('operating system', 'promoter of the system').

[4] In Jacquemin (2001), this type is not considered as variation (it does not comply the fourth condition: the variant should not contain the original term). Nevertheless, according to Daille (1995), juxtaposition is one of the "operations that lead to a term of length 3 from a term of length 1 or 2". Moreover, Daille et al. (2000), classify juxtaposition under syntactic variations.

Variants	Model examples	Examples in Basque	English translation
Insertion of determiner	$N_1A_{prep}N_2 \Leftrightarrow$ $N_1A_{prep}DetN_2$	*posta elektronikoko mezu* ⇔ *posta elektronikoko zenbait mezu*	*e-mail message / some e-mail messages*
Other lexical insertions[5]	$N_1N_2 \Leftrightarrow N_1A_{pos}N_2$	*telefonia-sare* ⇔ *telefonia mugikorreko sare*	*telephone network / mobile telephone network*
Juxtaposition[4]	$N_1N_2A_{pos1} \Leftrightarrow$ $N_1N_2A_{pos1}A_{pos2}$ $N_1N_2 \Leftrightarrow AprepN_1N_2$	*harpidedun-linea digital* ⇔ *harpidedun-linea digital asimetriko* *Internet-zerbitzu* ⇔ *Interneteko zerbitzu-hornitzaile*	*digital subscriber line / asymmetrical digital subscriber line* *Internet service / Internet service provider*
Head Coordination	N_1N_2 Conj $N_3 \Leftrightarrow$ N_1N_2 Conj N_1N_3	*programazio-lengoaiak eta -metodoak*	*programming languages and methods*
Argument Coordination	N_1 Conj $N_2N_3 \Leftrightarrow$ N_1N_3 Conj N_2N_3	*irrati- eta telebista-emanaldiak*	*radio and television broadcastings'*

Table 4. Different syntactic variants

e. Semantic variants

This type of variation involves the semantic relationship between constituents. For example, *idazkera bitar* eta *notazio bitar* ('binary notation') or *hizkuntz atlas* and *atlas linguistiko* ('language atlas', 'linguistic atlas') are equivalent terms.

3 Experimental work

The term extraction process is performed in two major steps: a) the selection of term candidates by means of linguistic techniques, and b) ranking and filtering of candidates by means of statistical techniques. Prior to the selection of candidates, a module to corpus-building module was provided. After the statistical processing, also an evaluation module was provided for human term validation. This was also used to assess recall and precision, in which case a reference list of terms extracted manually was required. As regard to variants, only orthographic, inflectional and syntactic variants of the type $N_1N_2 \Leftrightarrow A_{prep}N_2$ have been treated for the moment. A tool was designed and implemented for the integration of the different tasks. The tool is composed of the following main elements: the corpora builder, the terminology tagger, and the corpora navigator. The application has been designed to accept various document formats. The context of the extracted terms will be also available for the user. The physical design lies on a web browser, a web server (Apache+mod_perl), and a native XML database (Berkeley DB XML).

3.1 Linguistic process

The system for the automatic extraction of term consisted not only in the accomplishment of a grammar but also in the acquisition of the maximal balance between recall and precision. For this purpose, we took the following steps.

3.1.1 The grammar

For the automatic detection of terms a grammar was written using an xfst syntax based on the morphosyntactic patterns derived from the study of the manual term tagging. The

[5] In the majority cases, insertion and juxtaposition variants modify the meaning of the base term, and they are usually new concept denominators (but not always: *telefono-sare kommutatu* ⇔ *telefonia-sare publiko kommutatu* ('switched telephone network', 'public switched telephone network').

method applied for terminology extraction was therefore based on the selection of morphosyntactic patterns. As we said before, previous to this grammar, terms were manually extracted in order to be compared to the automatically extracted ones. All the terms obtained manually from the sample were described by means of morphosyntactic patterns. However, only the morphosyntactic patterns with a higher frequency in the corpus were described in the grammar of the transducer.

The structures defined in the grammar involve only NPs, both one-word and multiword terms. The main reason why only NPs have been taken into account for the first phase of the investigation is that NP terms constitute the vast majority of the terms extracted manually (87.14%). Besides, VPs are by far much more difficult to detect than NPs, mainly because the order of sentence elements in Basque is quite free, and VP constituents occur very often in non-contiguous positions.

The reason for rejecting some multiword patterns in the grammar was mainly their frequency. A frequency higher than one was necessary for a term structure to be included in the grammar. However, term structures of low frequency but thought to be productive for Basque were accepted to be part of the grammar, for example, $A_{prep}A_{prep}N$, as in *telekomunikazioen munduko enpresa* ('telecommunication company'). In the same way, patterns containing appositions, proper and common nouns that determine or add additional information about the head noun are only marked with entities in the terms. Only the most frequent entities were chosen in $N_{nc}N_{nc}N$, and the rest together with the $N_{nc}N_{nc}N_{nc}N$ pattern were left aside to reduce noise. With the same aim, $N_{abs}V_{gen}N$ and $N_{nc}N_{nc}A_{prep}N$ were excluded. Adverbs, except for those which are part of complex postpositions, were also discarded because they originated too much noise and few terms of the same structure were found in the patterns that include an adverb. In regard to one-word terms, nouns and acronyms are inserted in the grammar.

After applying the grammar, the expected recall should be the same as the one obtained in the manual term extraction, which reached the 84.69% of the total of patterns, taking into account the structures defined in the grammar.

The information of the patterns was described through a system of morphosyntactic rules, which were matched against the tagged corpus. The morphosyntactic information of the patterns defined in the grammar was associated to the tagged corpus using the mappings in table 5.

Pattern	Grammar	Part-of-speech	Subcategory	Case
N	N_o | N_{nc} | N_{prep}	IZE | SIG	ARR | IZB | LIB	DEK |ABS | GEN | GEL
A_{prep}	A_{prep}	N | ADJ | ADI	IZE | IZO | IZL	GEN | GEL | BAN | DESK
A_{pos}	A_{pos}	ADJ | ADI	IZO	

Table 5. Mapping of the items in the patterns and the tagger's output items

An Xfst transducer (Beesley & Karttunen, 2003) used the word combinations derived from the patterns to extract by means of the mentioned grammar. The grammar developed had been previously elaborated by the IXA group from the University of the Basque Country. Afterwards, this grammar was expanded, in order to adapt it to the new morphosyntactic patterns.

3.1.2 First results and improvements

Initially, the recall in the automatic term extraction was lower than expected. The automatically extracted terms summed up 67% of the morphosyntactic patterns selected to form the grammar. From those automatically extracted terms, one-word terms amounted to 49% of the total and multiword terms were 51%.

Due to a series of facts regarding the transducer as well as the automatic linguistic resources which are explained below, some terms were not obtained. Results depend very much on the quality of the linguistic tools. As we mentioned before, the corpus was morphosyntactically analysed by the lemmatiser-tagger for Basque *Euslem*.

Apart from this morphological analysis, a shallow syntactic analysis was necessary for an efficient detection of morphosyntactic patterns. However, due to errors on the disambiguation process, the recognition of foreign words or other syntactic constituents such as postpositions, not all the analysed tags were given an appropriate analysis. Therefore, some terms were not detected and even parts of terms included in longer term candidates were obtained.

For these reasons, an additional dictionary was created in order to increase the number of lemmas recognized by the lemmatiser-tagger *Euslem* in technical and scientific corpus. Needless to say, the identification of the new lemmas upgraded the results in the analysis and disambiguation of the tagger. Consequently, term extraction was more efficient. In the sample created for this research, 3.5% of the word forms in the texts were not detected. The task of creating a personal dictionary was assisted by the spell checker *Euspell* developed by the IXA group. This way we got a list of unknown words from the spelling checker, which was listed in order of frequency. The first 100 lemmas were chosen to enlarge the vocabulary of *Euslem*.

On the other hand, some of the terms in the manually extracted list that were missing in the automatically produced one (also called silence) were listed, and we tried to find a solution for them. Firstly, typographical elements such as slash or hyphen were included in the grammar. For example, terms having a slash such as *flip/flop* or *TCP/IP protokoloa* ('TCP/IP protocol') were found in the corpus. Moreover, the use of the hyphen is very regular and productive in Basque and the tagger analysed this type of compounds as a single word (*software-teknologia* 'software technology'). However, they were considered two-word patterns in the manual term extraction and that is why the hyphen had to be treated in the grammar. Secondly, multiword foreign terms were added too. Terms like *HyperText Markup Language* or *Netscape Navigator 2.0 browser* were also extracted. As for numbers, they were only considered in final position of terms (*Windows 98*, *Word 6.0*), and therefore, terms including numbers in any other position were left out.

3.1.3 Nested terms

Not only the grammar and linguistic tool were improved but also the list of the automatically extracted candidate terms was reviewed in order to refine it and increase the occurrence of some terms. In the manually tagged corpus, only the maximal patterns, that is, the longest possible word combinations, were marked. This list was later used to asses the results of the automatic term extractor, which turned out to be very poor; only 67% of the terms extracted in the first phase were detected in the automatic process. In order to improve these results maximal patterns were decomposed into sub-structures, so that, all the nested terms —that is, terms included in bigger candidate terms— could be recovered.

In order to do so, we proceeded in the following way. On the one hand, to add the terms included in longer combinations, maximal NPs were decomposed into sub-structures. The syntactic constituents that follow the maximal probability of including a term were kept and the least probable syntactic constituents were discriminated among the considered patterns. Needless to say, nested terms must be composed of more than two constituents in multiword terms. Trigrams and tetragrams were divided into head and modifier sub-structures. To obtain these syntactic constituents (head and modifier) a very simple grammar was built based on the probability of the components of each pattern. For example, from the $N_{nc}N_{nc}$Apos pattern *RAM memoria handi* ('big RAM memory'), *RAM memoria* ('RAM memory') is the only interesting term.

On the other hand, an additional analysis of the morphosyntactic structures that produce non-terminological units was done to reduce noise. This analysis showed that, for instance, sometimes the element A_{prep} (prepositive adjective) deserves to be considered as the constituent of a term while some others it does not. For example, postpositions produce a non-negligible amount of noise, and to avoid it the insertion of a new grammar (the module *Zatiak*, developed by the IXA group) has been foreseen. Many complex postpositions in Basque have a noun constituent (*ordenagailuen artean* 'between computers'). For example, in the maximal NP *ordenagailuen arteko komunikazioa* ('communication between computers'), the strings *ordenagailuen arte* and *arteko komunikazio* are not nested terms. The way to exclude those strings from the list of candidate terms is to tag *ordenagailuen arte* as a postposition.

As for inflected forms, only the case marks with a high relevance in the terminological language were included in the grammar. Some case marks caused significant noise and were therefore excluded from the grammar. The excluded forms belonged to the A_{prep}N pattern and the prepositive adjective belonged to the type of V_{gen}, as in *egiteko leku* ('place to do') or *garatutako tresna* ('developed tool'). The omission of inappropriate term-structures because of noise is not significant for term extraction has showed favourable results for our system, although, there will be some terms we cannot extract.

The treatment of nested terms, together with the improvements in the treatment of typographical elements, numbers and foreign words, resulted in a better recall, which at this stage reaches 87% of the terms corresponding to the patterns of the grammar.

3.2 Statistical process

The statistical methods applied in this kind of applications vary considerably depending on the system. In our approach, we decided to apply two different strategies for multiword and for one-word terms. *Unithood* is used by means of word association measures for the treatment of multi-word candidates, and *termhood* measures (Kageura, 1996) for one-word term candidates. In our experiments, the association measures were empirically modified trying to introduce a simple *termhood* paradigm. In this case, the changes improve the ranks.

3.2.1 Treatment of multiword terms

Word association measures are used in order to rank multiword units according to the association grade among their components. Most of the association measures proposed in the literature are intended to rank bigrams and rely on different concepts. For example, Mutual Information (MI), introduced in this field by Church and Hanks (1989), was taken from Information Theory. Other measures such as the log-likelihood ratio (LL) introduced by Dunning (1994), t-score and Chi-square are based on hypothesis testing. In order to rank MWUs composed of two or more words, Dias et al. (1999) introduced Mutual Expectation (ME), a measure based on Normal Expectation, which is a generalization of Dice coefficient for n-grams. Blaheta and Jonhson (2001) use measures based on parameters of certain Log-linear models to rank verbs composed of two or more words.

As for the bigrams, the input is the list of candidates extracted in the linguistic process. We have carried out experiments with two lists: with and without processing nested terms, in order to find out the best starting point to get maximum precision and recall. The results for bigrams using different association measures are shown in Table 6 (number of terms, precision, recall and F-score).

Type	# of terms	# of extr. terms	# of correct terms	P (%)	R (%)	F (%)
MI	255	1156	210	18.17	82.35	29.77
MI^3	255	1156	210	18.17	82.35	29.77
LL	255	612	135	22.06	52.94	31.14
t-score	255	681	143	20.99	56.08	30.55
Chi-square	255	1156	210	18.17	82.35	29.77

Table 6. First results for bigrams (nested included)

We tried to improve empirically association measures. In order to improve the representativeness of word frequency, we use for the frequency (marginal frequency) of the components their normal frequency, instead of the observed frequency in the corpus. This normal frequency is calculated from a global character corpus. Table 7 shows the results for bigrams, including nested bigrams.

Type	P (%)	R (%)	F (%)
MI	30.91	66.67	42.24
MI^3	33.26	58.04	42.29
LL	30.55	65.88	41.74
t-score	31.11	60.39	41.07
Chi-square	28.76	69.02	40.60

Table 7: Precision, recall and F-score (nested included)

For the treatment of terms of length higher than 2, we have followed two strategies to rank trigram and tetragram candidates. In those strategies, candidates are ranked by their *unithood*, but this *unithood* is estimated between different groups of constituents. In the first strategy, it is calculated between the head and modifier of the candidate. In the second strategy, it is calculated among all the components.

We have observed that, as in the case of the bigrams, the classification improves when normal frequencies are taken into account[6]. The results are shown in table 8. The two last rows (ME and measures based on Log-linear models) are calculated using the second strategy. LL, t-score and ME measures show the best performance.

Type	P (%)	R (%)	F (%)
MI	20.69	40.00	27.27
MI^3	20.93	45.00	28.57
LL	21.71	46.67	29.63
t-score	21.71	46.67	29.63
Chi-square	20.69	40.00	27.27
ME	27.03	33.33	29.85
Log-Linear models	18.14	61.67	28.03

Table 8. Precision, recall and F-score for n-grams ($2< n \leq 4$)

[6] This has been applied to all measures except for ME, where frequencies of monograms are not used.

3.2.2 Treatment of one-word terms

There are several methods to obtain the *termhood* of a one-word candidate. In our case, we considered that the relation between the relative frequency of the nouns and the relative frequency of a general corpus (*normal frequency*) might be a good measure to classify individual word candidates. This way, the *termhood* of the candidates is obtained dividing the observed frequency in the corpus by the normal frequency. Damerau (1993) defines this as relative frequency ratio (RFR). When RFR is applied, the best F-score occurs when the nested candidates are included (see results in Table 9).

Type	# of terms	# of extracted terms	# of correct terms
Monograms RFR	245	407	153
	P (%)	**R (%)** (marginal frequency)	**F (%)**
	37.59	62.45	46.93

Table 9: Precision, recall and F-score (nested included)

4 Conclusions and future work

In this paper, we have presented the application of a hybrid approach for the extraction of terminology in Basque, combining the detection of term candidates through linguistic techniques with the subsequent ranking of candidates according to different statistical measures. Both one-word and multiword terms were extracted. The results show that precision and recall improve when normal frequencies are used in the calculation of association measures and relative frequencies. The main sources of errors are problems identifying foreign words and postpositions, especially in the treatment of nested terms. The tagger is actually being improved in order to manage foreign words more efficiently. Besides, most postpositions in Basque are analysed as nouns and, therefore, they produce a non-negligible amount of noise. In order to avoid it, new rules to treat postpositions are being developed. Moreover, the treatment of morphosyntactic and syntactic term variation will be taken into account in future developments of the tool. Finally, we are compiling a bigger test-corpus to gain credibility in our experiments

Acknowledgements

This project is part of the Hizking21 project (www.hizking21.org), which is being developed to support and achieve technical and scientific aims in the field of language technologies. This research has been partially funded by the *Etortek* and *Saiotek* programmes of the Department of Industry, Trade and Tourism of the Government of the Basque Country.

Bibliography

1 BEESLEY, K.R. & KARTTUNEN, L. (2003). *Finite State Morphology*. CSLI. Standford University.

2 BLAHETA, D. & JOHNSON, M. (2001). "Unsupervised learning of multi-word verbs." In: *Proceedings of the 39th Annual Meeting of the ACL* (pp. 54-60). Toulouse.

3 BOURIGAULT, D. (1994). *LEXTER, un Logiciel d'Extraction de Terminologie. Application a l'acquisition des connaissances a partir de textes*. Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

4 BOURIGAULT, D. (1996). "Lexter, a Natural Language Processing Tool for Terminology Extraction." In: *Proceedings of 7th EURALEX International Congress*.

5 CABRÉ, T. (2001). "Consecuencias metodológicas de la propuesta teórica (I)." In: *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*. Barcelona: IULA-UPF (pp. 27-36).

6 CHURCH, K.W. & HANKS, P.P. (1989). "Word association norms, mutual information and lexicography." In: *Proceedings of the 27th Annual Meeting of the ACL* (pp. 76-83). Vancouver.

7 DAILLE, B. (1995). "Combined approach for terminology extraction: lexical statistics and linguistic filtering". In: *UCREL Technical Papers*, 5, University of Lancaster.

8 DAILLE, B., HABERT, B., JACQUEMIN, C. & ROYAUTÉ, J. (2000) "Empirical Observation of Terms Variations and Principles for their Description." In: *Terminology*, 3(2), 197-258. Amsterdan: John Benjamins

9 DAMERAU, F.J. (1993). "Generating and evaluating domain-oriented multi-word terms from texts." In: *Information Processing & Management*, 29, 433–447. Elsevier

10 DIAS, G., GUILLORÉ, S., BASSANO, J.C. & LOPES, J.G.P. (2000). "Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?" In: *Proceedings of Recherche d'Informations Assistée par Ordinateur* (pp. 1-20). Paris.

11 DIAS, G., GUILLORÉ, S., LOPES, J.G.P. (1999). "Mutual Expectation: a Measure for Multiword Lexical Unit Extraction." In: *Proceedings of VExTAL Venezia per il Trattamento Automatico delle Lingue*. Universitá Cá Foscari. Venezia, Italy.

12 DUNNING, T. (1994) "Accurate Methods for the Statistics of Surprise and Coincidence." In: *Computational Linguistics* 19(1): 61-74. Cambrigde, Mass: The MIT Press.

13 EZEIZA N., ADURIZ I., ALEGRIA I., ARRIOLA J.M. & URIZAR R. (1998). "Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages." In: COLING-ACL'98, Montreal.

14 JACQUEMIN, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Mass.: The MIT Press.

15 JUSTESON, J. (1993). "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text." In: *IBM Research Report, RC 18906 (82591)*.

16 KAGEURA, K. & UMINO, B. (1996). "Methods of Automatic Term Recognition." In: *Terminology*. 3(2), 259-289. Amsterdan: John Benjamins.

17 MAYNARD, D. & ANANIADOU, S. (2000). "Trucks: A Model for Automatic Multi-Word Term Recognition." In: *Journal of Natural Language Processing*, 8(1), 101-126.

18 NENADIC, G., SPASIC, I., ANANIADOU, S. (2002). "Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts." In: *Proceedings of 3rd International Conference on Language, Resources and Evaluation, LREC-3*, Las Palmas, Spain.

19 SMADJA, F. (1993). "Retrieving Collocations from Text: XTRACT." In: *Computational Linguistics*, 19(1) 143-177.

20 Urizar R., Ezeiza N. & Alegria I. (2000). "Morphosyntactic structure of terms in Basque for automatic terminology extraction." In: Proceedings of the 9th EURALEX International Congress. (pp. 373-382). Stuttgart.