Analysis and performance of morphological query expansion and language-filtering words on Basque web searching

I. Leturia, A. Gurrutxaga, N. Areta, E. Pociello
Elhuyar Fundazioa, R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country
{igor,agurrutxaga,nereaa,eli}@elhuyar.com

Abstract

Morphological query expansion and language-filtering words have proved to be valid methods when searching the web for content in Basque via APIs of commercial search engines, as the implementation of these methods in recent IR and web-as-corpus tools shows, but no real analysis has been carried out to ascertain the degree of improvement, apart from a comparison of recall and precision using a classical web search engine and measured in terms of hit counts.

This paper deals with a more theoretical study that confirms the validity of the combination of both methods. We have measured the increase in recall obtained by morphological query expansion and the increase in precision and loss in recall produced by language-filtering-words, but not only by searching the web directly and looking at the hit counts –which are not considered to be very reliable at best–, but also using both a Basque web corpus and a classical lemmatised corpus, thus providing more exact quantitative results. Furthermore, we provide various corpora-extracted data to be used in the aforementioned methods, such as lists of the most frequent inflections and declinations (cases, persons, numbers, times, etc.) for each POS –the most interesting word forms for a morphologically expanded query–, or a list of the most used Basque words with their frequencies and document-frequencies –the ones that should be used as language-filtering words–.

1. Introduction

1.1. Basque web searching problems

There are two main reasons why existing web search services are unsuitable for the case of Basque. The first is that Basque is an agglutinative language, and the problems that non-English languages, and agglutinative languages in particular, have with search engines are well known (Bar-Ilan, 2005; Bar-Ilan & Gutman, 2003; Bar-Ilan & Gutman, 2005). In Basque, a given lemma produces many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite) for nouns and adjectives, and the person (me, he, etc.) and the tense (present, past, etc.) for verbs. A brief morphological description of Basque can be found in (Alegria et al., 1996). All this means that looking only for the exact given word or the word plus an "*s*" for the plural is not enough for Basque.

The second reason is that none of the existing search services can discriminate Basque pages in their searches. Searching in any of them for a technical word that also exists in other languages –*anorexia*, *sulfuroso*, *byte* or *allegro*, to cite just a few examples of the many that exist– or a proper noun or a short word, will not only *not* yield results exclusively in Basque, but often not yield any results in Basque at all.

1.2. API based approach

A possible solution could be to set up our own search services, which would only include pages that are in Basque and which would not index the word forms that a page contains, but its lemmas, as proposed in (Langer, 2001). However, instead of taking this approach and implementing and maintaining all the infrastructure that a search engine and its crawling, indexing and serving involves –bandwidth, disk, reliability, etc.–, two recent Basque web search services have been developed that make use of the APIs of classical search engines: Elebila, a search engine (Leturia et al., 2007:1), and CorpEus, a web-as-corpus tool (Leturia et al., 2007:2). They both solve the two problems we have mentioned by means of morphological query expansion and language-filtering words. The approach is a very cost-effective one, and it could be applied to other agglutinative or under-resourced languages as well.

1.3. Looking for conjugations and inflections

When asking the API of a search engine for a word, we need it to return pages that contain its conjugations or inflections, too. The way we have approached this matter is based on morphological query expansion. The importance and use of morphology for various IR tasks has been widely documented (Ambroziak & Woods, 1998; Krovetz, 1993; Woods, 2000; Woods et al., 2000), although it is normally applied by lemmatisation at the indexation stage, which is an unattainable objective for us, as has been stated above. Instead, we apply morphological generation at the querying stage: we use a tool created by the IXA Group of the University of the Basque Country that gives us all the possible inflections or conjugations of the lemma, and we ask the search engine to look for any of them by using an OR operator. For example, if the user asks for *etxe* ("house"), the search engine is asked for "(etxe OR etxea OR etxeak OR etxeari OR etxeek OR etxearen OR…)". This is how a lemma-based search is obtained.

But the APIs of search engines have their limitations with regard to search term count, length of search phrase, etc. These limitations render a proper lemmatised search for Basque impossible, as we cannot search for all the conjugations or inflections. So we send the most frequent ones, which will cover a high enough percentage of all the occurrences. A similar approach is taken in (Kettunen, 2007; Kettunen et al., 2007).

1.4. Language discrimination

We have mentioned already that there is no commercial search engine that can distinguish pages in Basque and return them alone. To achieve this we include, in the search phrase that is sent to the API, the most frequently used words in Basque, in conjunction with an AND operator. We used a corpus to see which these most used words were. But the most frequent words in Basque are short and, as such, the chances of their existing in other languages or being used as abbreviations or acronyms is quite high –in fact, the first two words at least have well-known meanings in other languages–. Therefore, we include a number of these language-filtering words in the queries to obtain a high percentage of Basque results, although this also involves a loss in recall (some Basque pages can be left out because they do not contain one or more of the words).

2. Motivation

The combination of morphological query expansion and language-filtering words has been successfully applied in the search service Elebila and the web-as-corpus tool CorpEus. But some details of their implementation that could be of importance in their performance have not been tested sufficiently thoroughly. For example, the choice of the most frequent word forms for the morphological query expansion has been made quite intuitively, without analysing a corpus for the most frequent inflections of each POS; choosing the number of language-filtering words has been done empirically by observing the results of some searches, instead of properly analysing their effect on precision and recall in order to find a compromise between them; and the only evaluation the methodology has been subjected to is based on hit counts of real web searches.

So the aim of this work is to make a more detailed analysis of the effects of the methodology: to perform corpora-based precision and recall measurements on both morphological query expansion and language-filtering words, and to compare these with the precision and recall observed by applying them in web searches. We also wanted to produce corpora-based lists of the most frequent cases for each POS.

3. Design of the study

As stated above, the study described in this paper consists of various corpus-based measurements. One of the corpora used for carrying it out is the ZT Corpus (Areta et al., 2007), a lemmatised Basque corpus on science and technology made up of 7.6 million words. But since the typology of the documents that form a classical corpus and those that form the WWW might differ (Sharoff, 2006), we have considered it advisable to use not only a classical corpus, but also a web corpus. So we have compiled a web corpus by crawling the Basque branch of the Google Directory (http://www.google.com/Top/World/Euskara/). We downloaded the 3,000 plus page present there and recursively followed all the links found in pages that LangId, a language identifier developed by the IXA Group of the University of the Basque Country, identified to be in the Basque language. The downloading process has been designed to ensure as much website variety as possible, by queuing the links found, prioritising different domains in each parallel downloading stage, etc. The web corpora obtained is made up of over 44,000 documents and approximately 20 million words.

The various measurements using these corpora had to be done by employing many different words. Instead of choosing random words, we used the search logs corresponding to the four months that the Elebila search engine has been in existence. These accounted for over 400,000 searches involving over 800,000 words, which after lemmatisation made over 70,000 different words. We ordered them by decreasing frequency and used the topmost ones for our work. This way, by basing our study on these most searched-for words and afterwards optimising the aforementioned IR tools with the results of the study, we will be maximising their performance for real-life searches.

4. Morphological query expansion

4.1. Most frequent cases

Morphological query expansion consists of sending the different conjugations or inflections of a lemma with an OR operator to the API, but since the number of words that the APIs allow in a query is limited, it is important, in order to improve recall as much as possible, that these be the most frequent ones. In the morphological query expansion implemented in CorpEus and Elebila, the most frequent cases had been chosen intuitively, without making a corpora-based analysis.

So in order to base the expansion on more empirical data, we have looked for these most frequent cases in both of the aforementioned corpora. For each of the morphologically productive POSs in Basque –nouns, proper nouns, place names, adjectives and verbs–, we took the most searched-for words of the Elebila logs. Because of the non-tagged nature of the web corpus, the words chosen had to be non-ambiguous. Then each different surface form of the words was assigned its case. By grouping them by case and ordering them by decreasing frequency, we produced a list of the most frequent cases for each POS, both in the classical corpus and the web corpus. The lists of each corpus, although similar, reveal some differences between them, so we have preferred to choose the web corpus lists. These were the most frequent cases of each POS:

□ Verb:

1. Participle / perfective aspect (*sortu*)
2. Imperfective aspect (*sortzen*)
3. Verbal noun + *-ko* (*sortzeko*)
4. Unrealized aspect (*sortuko*)
5. Short stem (*sor*)
6. Verbal noun + Nominative singular (*sortzea*)
7. Adjectival participle (*sortutako*)
8. Participle + Nominative singular (*sortua*)
9. Dynamic adverbial participle (*sortuz*)
10. *-ta/-da* stative adverbial participle (*sortuta*)
11. Participle + Nominative plural / Ergative singular (*sortuak*)
12. Verbal noun + Inessive singular (*sortzean*)
13. *-(r)ik* stative adverbial participle (*sorturik*)
14. Verbal noun + Allative singular (*sortzera*)
15. Adjectival participle + Nominative plural / Ergative singular (*sortutakoak*)
16. Verbal noun (*sortze*)

□ Adjective:

1. Nominative singular (*berria*)
2. Nominative plural / Ergative singular (*berriak*)
3. Nominative indefinite (*berri*)
4. Genitive plural (*berrien*)
5. Inessive singular (*berrian*)
6. Genitive singular (*berriaren*)
7. Associative singular (*berriarekin*)
8. Ergative indefinite (*berrik*)
9. Dative singular (*berriari*)
10. Instrumental indefinite (*berriz*)
11. Inessive indefinite (*berritan*)
12. Sociative plural (*berriekin*)
13. Inessive plural (*berrietan*)
14. Genitive locative singular (*berriko*)
15. Partitive (*berririk*)

□ Noun:

1. Nominative indefinite (*hiztegi*)
2. Nominative singular (*hiztegia*)
3. Nominative plural / Ergative singular (*hiztegiak*)
4. Genitive locative singular (*hiztegiko*)
5. Genitive singular (*hiztegiaren*)
6. Dative singular (*hiztegiari*)
7. Inessive singular (*hiztegian*)
8. Partitive (*hiztegirik*)
9. Instrumental indefinite (*hiztegiz*)
10. Instrumental singular (*hiztegiaz*)
11. Genitive singular + Nominative singular (*hiztegiarena*)
12. Genitive plural (*hiztegien*)
13. Sociative singular (*hiztegiarekin*)
14. Ablative singular (*hiztegitik*)
15. Allative singular (*hiztegira*)
16. Inessive plural (*hiztegietan*)
17. Allative singular + Genitive locative (*hiztegirako*)

□ Proper noun:

1. Nominative (*Mikel*)
2. Ergative (*Mikelek*)
3. Genitive (*Mikelen*)
4. Dative (*Mikeli*)
5. Associative (*Mikelekin*)
6. Genitive + Nominative singular (*Mikelena*)
7. Partitive (*Mikelik*)
8. Genitive + Nominative Plural / Ergative singular (*Mikelenak*)
9. Instrumental (*Mikelez*)
10. Inessive (*Mikelengan*)

□ Place name:

1. Nominative (*Egipto*)
2. Genitive locative (*Egiptoko*)
3. Inessive (*Egipton*)
4. Allative (*Egiptora*)
5. Ablative (*Egiptotik*)
6. Genitive (*Egiptoren*)
7. Dative (*Egiptori*)
8. Genitive locative + Nominative singular (*Egiptokoa*)
9. Allative + Genitive locative (*Egiptorako*)
10. Associative (*Egiptorekin*)
11. Genitive locative + Nominative plural / Ergative singular (*Egiptokoak*)
12. Destinative (*Egiptorentzat*)
13. Instrumental (*Egiptoz*)
14. Terminal allative (*Egiptoraino*)
15. Genitive locative + Inessive singular (*Egiptokoan*)

These lists will be used to improve the morphological query expansion of subsequent versions of Elebila and CorpEus.

4.2. Gain in recall

Once the most frequent cases of each POS were known, we tried to measure the increase in recall we would obtain for each POS by including 1, 2, 3… of the cases of the same words as before in an OR. We have performed this using both corpora and also by looking at the increase in hit counts returned by Microsoft's Live Search API.

For the overall measure of the POSs, we made a weighted average of them, taking into account the frequency of use of each POS. To calculate these frequencies, we classified approximately the first 900 words (all that have a query frequency over 100) out of the more than 70,000 of the Elebila logs into one of the categories. This may not seem very much, but they do in fact account for more than 44% of the queries.

The global increase in recall for each corpus is shown in Figure 1. A conclusion we can draw from the graph is that with as few as 5 cases, we can obtain an increase in recall of 50%, thus proving the validity of the morphological query expansion method.

The gain shown in the chart is an average of the gains obtained by each POS; the individual gains for the web corpus are shown in Figure 2. The differences between them are obvious: some POSs, namely verbs, adjectives and place names, really benefit from the query expansion while the others do so to a lesser extent.

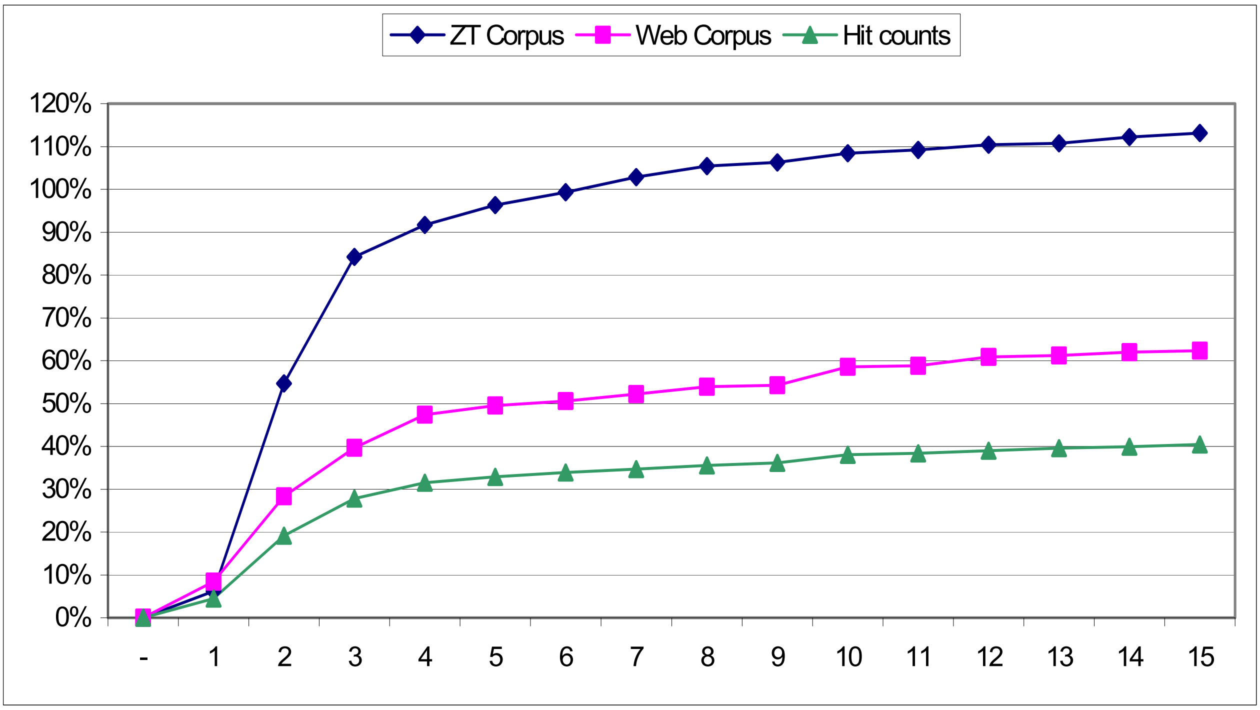

Figure 1: Gain in recall produced by including more cases in the queries

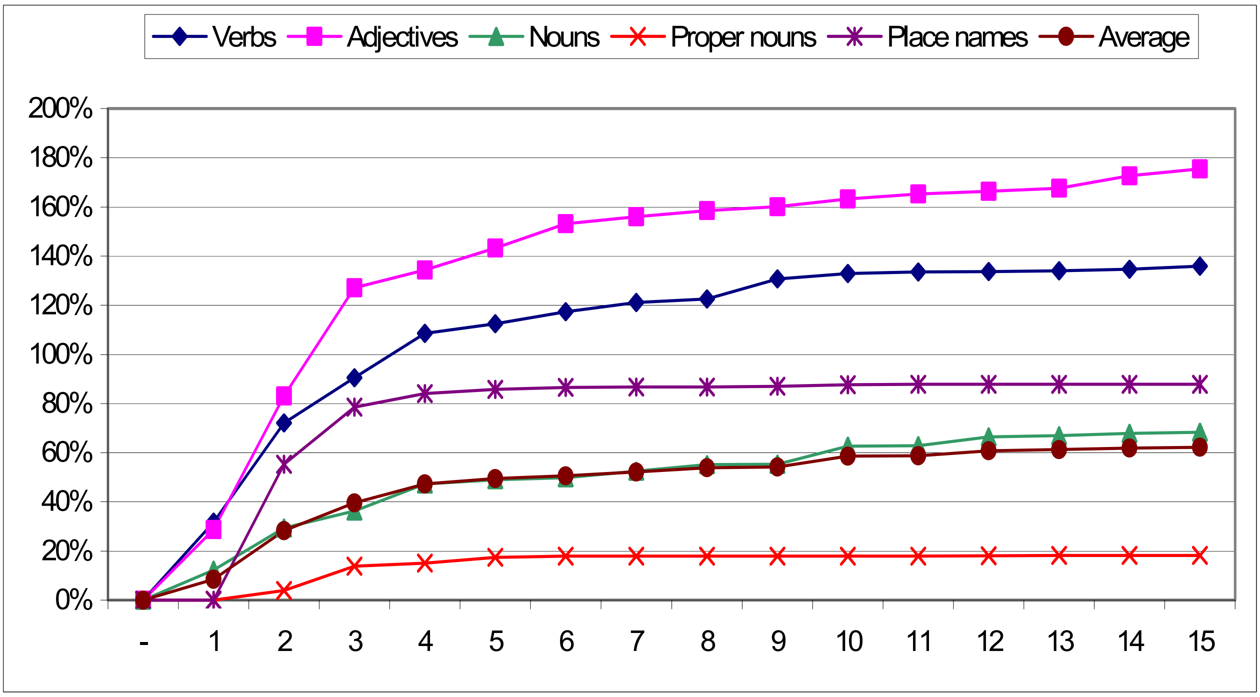

Figure 2: Gain in recall produced by including more cases in the queries, for each POS and in the Web Corpus

5. Language-filtering words

5.1. Choosing the words

To obtain pages only in Basque from the API of a search engine, we use the language-filtering words method, which consists of adding the most frequent Basque words to the search terms. But the selection of these most frequent words had been done using a classical corpus. In this piece of work we have carried out the same study on the two aforementioned corpora.

In the following table we show the 16 most frequent words of each corpora with the document-frequency of each of them:

Web corpus		ZT Corpus	
eta (“and”)	91.94%	*eta* (“and”)	98.44%
da (“is”)	74.37%	*da* (“is”)	92.67%
ez (“no”)	64.51%	*ez* (“no”)	79.05%
du (“has”)	64.11%	*dira* (“are”)	78.65%
bat (“a”)	62.81%	*ere* (“too”)	78.27%
ere (“too”)	55.65%	*du* (“has”)	75.49%
dira (“are”)	55.45%	*izan* (“be”)	73.45%
izan (“be”)	54.24%	*dute* (“have”)	72.14%
egin (“do”)	52.77%	*bat* (“a”)	67.66%
beste (“other”)	47.74%	*baina* (“but”)	64.41%
edo (“or”)	42.94%	*den* (“that is”)	64.04%
dute (“have”)	41.72%	*egin* (“do”)	62.56%
den (“that is”)	39.19%	*beste* (“other”)	57.21%
egiten (“doing”)	38.98%	*baino* (“than”)	56.77%
baina (“but”)	36.94%	*egiten* (“doing”)	55.78%
baino (“than”)	27.29%	*edo* (“or”)	55.59%

Table 1: Most frequent word forms in both corpora.

The 16 most frequent words in both corpora are the same, but not their order. In view of this, we preferred to choose the candidates to act as language-filtering words from the first list, as this corpus is supposedly more similar to the one to which we will apply our tools, that is, the Internet. So the candidates will be the topmost six words from the web corpus list: *eta*, *da*, *ez*, *du*, *bat* and *ere*.

Also, previous to this work, choosing how many language-filtering words should be added to the queries had been done by observing their behaviour in a number of searches. Now we have performed precision and recall studies on different combinations of the six candidates to be language-filtering words.

Looking at the document-frequencies of the candidate words, it is clear which should be the words to choose for one-word and two-words language-filters, since there are significant gaps between the frequencies of the first three words in both corpora. Choosing which should be the third and fourth words is more difficult, because the next words have quite similar document-frequencies. For these ones we can even consider OR combinations. So the combinations for which we will analyse the precision and recall in the following subsections are listed below:

❑ 0 words:

1. -

❑ 1 word:

2. *eta*

❑ 2 words:

3. *eta* AND *da*

❑ 3 words:

4. *eta* AND *da* AND (*ez* OR *du* OR *bat* OR *ere*)
5. *eta* AND *da* AND (*ez* OR *du* OR *bat*)
6. *eta* AND *da* AND (*ez* OR *du* OR *ere*)
7. *eta* AND *da* AND (*ez* OR *bat* OR *ere*)
8. *eta* AND *da* AND (*du* OR *bat* OR *ere*)
9. *eta* AND *da* AND (*ez* OR *du*)
10. *eta* AND *da* AND (*ez* OR *bat*)
11. *eta* AND *da* AND (*ez* OR *ere*)
12. *eta* AND *da* AND (*du* OR *bat*)
13. *eta* AND *da* AND (*du* OR *ere*)
14. *eta* AND *da* AND (*bat* OR *ere*)
15. *eta* AND *da* AND *ez*
16. *eta* AND *da* AND *du*
17. *eta* AND *da* AND *bat*
18. *eta* AND *da* AND *ere*

❑ 4 words:

19. *eta* AND *da* AND *ez* AND (*du* OR *bat* OR *ere*)
20. *eta* AND *da* AND *du* AND (*ez* OR *bat* OR *ere*)
21. *eta* AND *da* AND *bat* AND (*ez* OR *du* OR *ere*)
22. *eta* AND *da* AND *ere* AND (*ez* OR *du* OR *bat*)
23. *eta* AND *da* AND *ez* AND *du*
24. *eta* AND *da* AND *ez* AND *bat*
25. *eta* AND *da* AND *ez* AND *ere*
26. *eta* AND *da* AND *du* AND *bat*
27. *eta* AND *da* AND *du* AND *ere*
28. *eta* AND *da* AND *bat* AND *ere*

5.2. Loss in recall

To measure the loss in recall produced by the language-filtering words, we measured their document-frequency in the classical corpus and the web corpus. We also measured the decrease in hit counts obtained by searching the web for Basque words alone, using Microsoft Live Search’s API. The results are shown in Figure 3.

By taking a look at the graph, we can see the remarkable similarity between the web corpus and hit counts series, proving that the web corpus we have crawled from the web is a good sample for predicting the behaviour of the web. Furthermore, we can observe that the recall in the ZT Corpus is significantly greater, most likely due to the fact that the type of documents of which this corpus is composed of –books and articles on science and technology– is, on average, greater in size than most web pages, which confirms our previous supposition that it was better to base our study on a corpus collected from the web.

5.3. Gain in precision

The addition of more of the language-filtering words to the query leads to a gain in language precision. For quantifying this gain the ideal thing would be, as before, to measure it on the corpora, but this is not possible, since we would need a multilingual corpus that would have the same proportion of each language as the web does, which is very difficult, if not impossible, to obtain. So we had no other option but to measure the gain in precision by searching the web through Microsoft’s API and looking at the percentage of results in the Basque language. For classifying the results into Basque or non-Basque we used LangId again, applying it to the snippets returned.

We mentioned above that the performance of the language-filtering words method is most noticeable when the search term exists in other languages, or when it is short, or when it is a proper noun. If the word only exists in Basque, the language-filtering words might bring little benefit or even none at all. So we have measured the gain in precision separately for different categories of words:

- ❑ Short words: Words with 5 characters or less. The probability of their existing in other languages is high. The most searched for words in Elebila from this category (and consequently the ones used for our evaluation) were words like *herri* (“people”, “town”), *berri* (“new”), *haur* (“child”), *ipuin* (“tale”), *gabon* (“Christmas”) or *mapa* (“map”).
- ❑ Proper nouns: Proper nouns are usually the same in other languages. Some of the words for this category were *Wikipedia*, *Google*, *Elhuyar*, *Egipto*, *Euskadi* (“Basque Country”), etc.
- ❑ International words: Words that we know definitely exist in another language (usually English, Spanish or French). These were some of the most searched for words in this category: *biografia* (“biography”), *historia* (“history”), *energia* (“energy”), *animalia* (“animal”), *mitologia* (“mitology”), *arte* (“art”)…
- ❑ Words that are likely to be found in other languages: Technical words which, despite not being exactly the same in the three languages mentioned above, have quite similar spellings in all of them, so the probability of their existing in some other language is high. Some examples of these words are *musika* (“music”), *informazio* (“information”), *eskola* (“school”), *definizio* (“definition”) and *didaktiko* (“didactic”)
- ❑ Basque words: Words that we are almost sure do not exist in any other language. The most searched for words in this category were *euskal* (“Basque” as adjective), *euskara* (“Basque language”), *hiztegi* (“dictionary”), *hezkuntza* (“education”), *hizkuntza* (“language”), *ariketa* (“exercise”) and various others.

For the overall measure, we made a weighted average of them, taking into account the frequency of use of each category, again calculated by classifying the first 900 most searched words in the Elebila logs.

Category of word	**Word**		**Query**	
Short words	191	21.75%	98,867	30.40%
Proper nouns	287	32.69%	70,611	21.71%
International words	98	11.16%	40,562	12.47%
Words likely in other languages	94	10.71%	31,856	9.80%
Basque words	208	23.69%	83,297	25.61%
Total categorized	**878**	**1.22%**	**325,193**	**40.42%**

Table 2: Frequency and query percentage of each category

The gain in precision produced by the language-filtering words for each category of word and overall is shown in Figure 4.

The peaks and valleys of the graph provide us with hints as to the filtering properties of the last four words (*ez*, *du*, *bat* and *ere*). All the valleys are combinations containing *du* and the highest peaks contain the word *ere*, so these two are, respectively, the worst and best words of the four for filtering. Between *ez* and *bat* there is no significant difference, although *ez* seems to behave a little better. These conclusions are logical: *du* is a word that is present in almost any text in a big language like French; *bat* is a word that, although not very frequent, exists in the language with the highest presence on the web, that is, English; and, as far as we know, *ez* and *ere* are not widely used words in at least three major languages, such as English, Spanish and French, but *ere* is longer and hence yields better results.

5.4. Choosing the number of language-filtering words

In Figure 5 we put together the precision, recall and F-measure of the different language-filtering word combinations.

The conclusions we can draw from it are that by using 4-word combinations we can achieve very good precision (even high above 90%), but with fairly bad recall (near or below 50%). So maybe it is a better idea to use 3-word combinations that do not include the word *du* –like *eta* AND *da* AND (*ez* OR *bat* OR *ere*), *eta* AND *da* AND (*ez* OR *ere*) or *eta* AND *da* AND (*ez* OR *bat*)–, with which we can achieve a precision of 86-87% and a recall of 68-65%. In fact, these are the combinations with the highest F-measure. But we must take into account that for proper nouns or international words the precision would fall to around 70%.

The best thing might be to keep a list of the most searched proper nouns and international words, and when someone wants to search for one of them, use 4-word combinations, and otherwise use 3-word ones. Or we could also prioritise precision and normally use 4 words, and if the user is not happy with the results, then he or she can be given the option of searching again by increasing the recall (using 3 words).

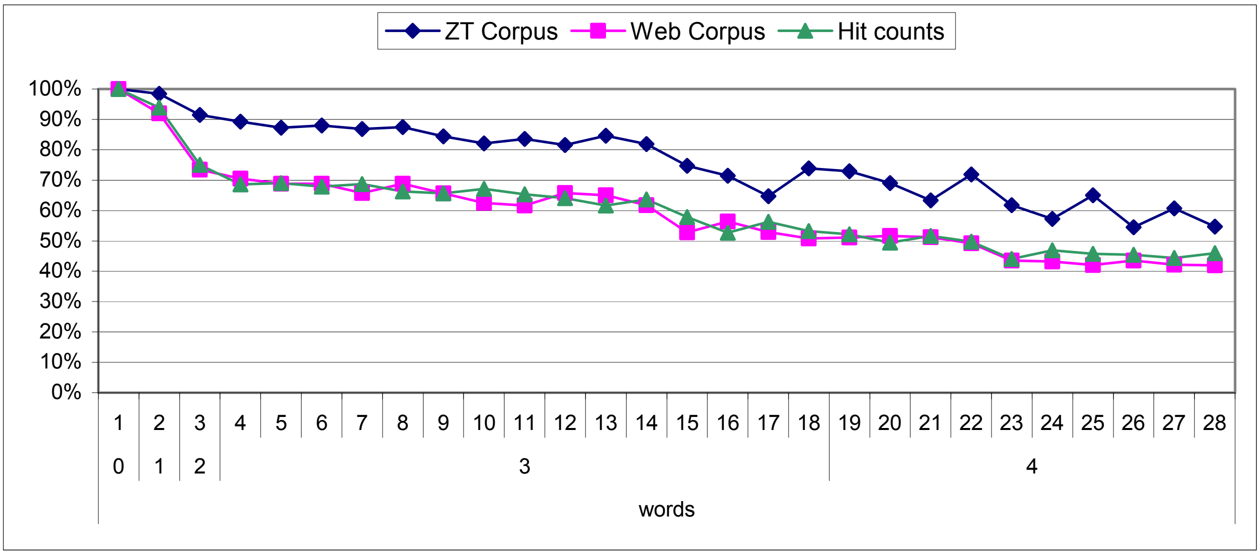

Figure 3: Loss in recall produced by the different language-filtering word combinations.

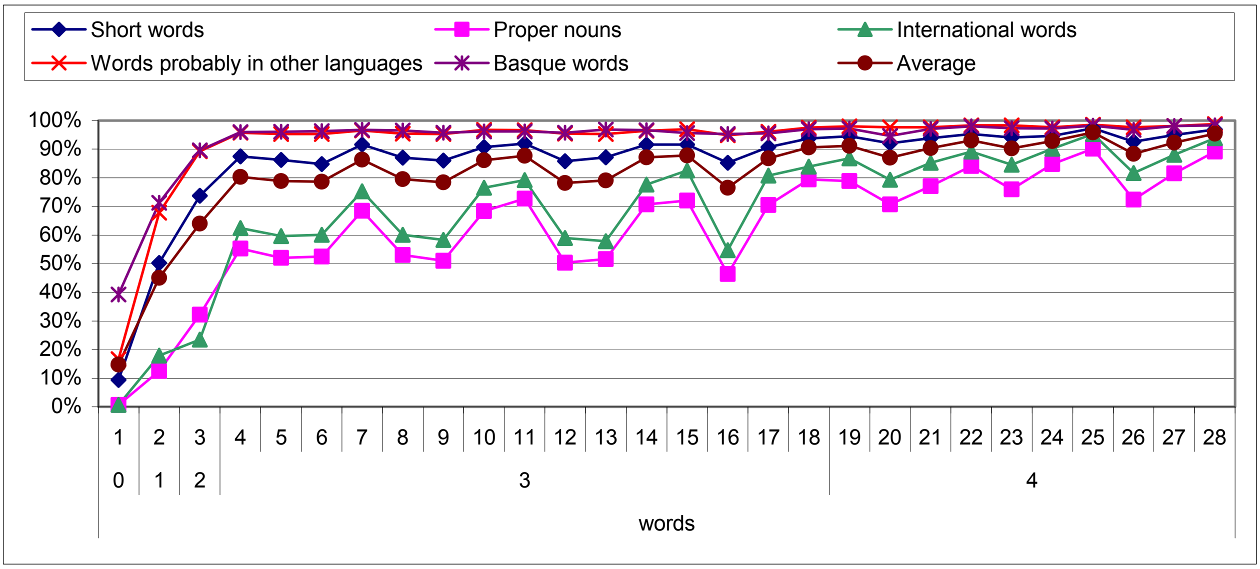

Figure 4: Gain in precision produced by the different language-filtering word combinations.

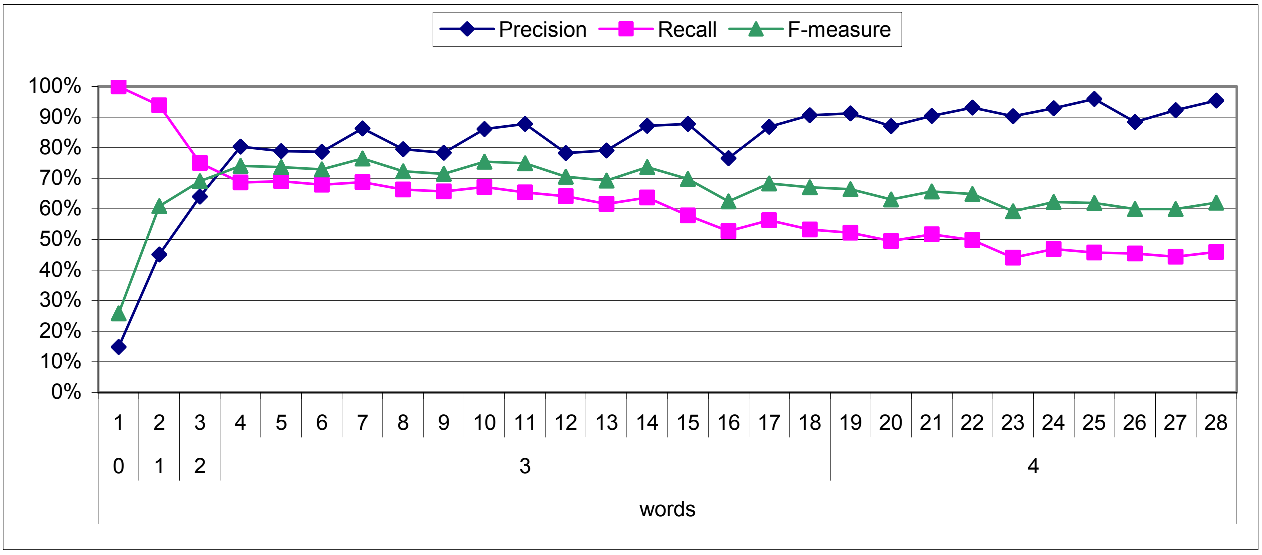

Figure 5: Precision, recall and F-measure produced by the different language-filtering word combinations.

6. Conclusions

The studies performed in this piece of work provide more detailed and corpora-based precision and recall data that confirm the validity of morphological query expansion and language-filtering words as a means for obtaining satisfactory Basque web searches from the APIs of classical search engines. Moreover, the precision and recall data and frequency lists obtained in this work will be very helpful in further improving existing tools that use this methodology, such as the Basque search engine Elebila and the web-as-corpus tool CorpEus, and they will also constitute very valuable documentation for future IR projects for Basque. Besides, the methodology of the study could point the way ahead for building IR tools for other agglutinative or minority languages.

7. References

Alegria, I., Artola, X., Sarasola, K. (1996). Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, 4(2), pp. 193--203.

Ambroziak, J., Woods, W.A. (1998). Natural Language Technology in Precision Content Retrieval. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. Moncton, New Brunswick, Canada: University of Moncton.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. (2007). ZT Corpus - Annotation and tools for Basque corpora. In *Proceedings of Corpus Linguistics conference*. Birmingham, UK: University of Birmingham.

Bar-Ilan, J. (2005). Expectations versus reality - Search engine features needed for Web research at mid 2005. *Cybermetrics, International Journal of Scientometrics, Informetrics and Bibliometrics*, 9(1), paper 2.

Bar-Ilan, J., Gutman, T. (2003). How do search engines handle non-English queries? - A case study. In *Proceedings of the 12th international World Wide Web Conference*. Budapest, Hungary, pp. 415--424.

Bar-Ilan, J., Gutman, T. (2005). How do search engines respond to some non-English queries?. *Journal of Information Science*, 31, pp. 13--28.

Kettunen, K. (2007). Managing keyword variation with frequency based generation of word forms in IR. In *Proceedings of NODALIDA Conference*. Tartu, Estonia: University f Tartu, pp. 318--323.

Kettunen, K., Airio, E., Järvelin, K. (2007). Restricted inflectional form generation in management of morphological keyword variation. *Information Retrieval*, 10(4-5), pp. 415--444.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. Pittsburgh, Pennsylvania: ACM, pp. 191--202.

Langer, S. (2001). Natural languages and the world wide web. *Bulletin de linguistique appliquée et générale*, 26, pp. 89--100.

Leturia, I., Gurrutxaga, A., Alegria, I., Ezeiza, A. (2007). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, pp. 69--81.

Leturia, I., Gurrutxaga, A., Areta, A., Alegria, I., Ezeiza, A. (2007). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of iNEWS'07 workshop in SIGIR*. Amsterdam, The Netherlands: ACM, pp. 47--54.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S Bernardini (Eds.), *WaCky! Working papers on the Web as Corpus*. Bologna, Italy: Gedit Edizioni, pp. 63--98.

Woods, W.A. (2000). Aggressive morphology for robust lexical coverage. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle, Washington: ACM, pp. 218--223.

Woods, W.A., Bookman, L.A., Houston, A., Kuhns, R.J., Martin, P., Green, S. (2000). Linguistic knowledge can improve information retrieval. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle, Washington: ACM, pp. 262--267.