EusBila, a search service designed for the agglutinative nature of Basque

Igor Leturia
Elhuyar R&D
Zelai Haundi 3, Osinalde Industrialdea
20170 Usurbil (Basque Country)
34 - 943 363040
igor@elhuyar.com

Antton Gurrutxaga
Elhuyar R&D
Zelai Haundi 3, Osinalde Industrialdea
20170 Usurbil (Basque Country)
34 - 943 363040
agurrutxaga@elhuyar.com

Nerea Areta
Elhuyar R&D
Zelai Haundi 3, Osinalde Industrialdea
20170 Usurbil (Basque Country)
34 - 943 363040
nereaa@elhuyar.com

Iñaki Alegria
IXA Taldea, University of the Basque Country
649 postakutxa
20080 Donostia (Basque Country)
34 – 943 015076
i.alegria@ehu.es

Aitzol Ezeiza
IXA Taldea, University of the Basque Country
649 postakutxa
20080 Donostia (Basque Country)
34 – 943 018657
aitzol.ezeiza@ehu.es

ABSTRACT
The performance of major search engines for Basque is far from satisfactory, partly due to the agglutinative nature of the language –it is commonly known that search engines do not perform well with such languages– and partly because it is not a language to which search engines restrict their results.

In this paper we present EusBila, a search service for Basque that relies on the APIs of search engines, yet obtains a lemma-based and language-filtered search by means of morphological query expansion and language-filtering words. It is a cost-effective approach, which we think can be used for other agglutinative or minority languages. We also evaluate how well EusBila performs when carrying out a Basque query, and we compare this performance to that of a major search engine in terms of precision and recall, thus demonstrating that EusBila is a very valid solution.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, selection process.*

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language generation, language models.*

General Terms
Performance, Design.

Keywords
Search engine, information retrieval, Basque, agglutinative language, minority language.

SIGIR'07 iNEWS07 workshop, July 27, 2007, Amsterdam, The Netherlands

1. MOTIVATION
The problems that non-English languages, and agglutinative languages in particular, have with search engines are well known [5] [6] [7]. While some search engines do seem to use some sort of additional techniques for languages like German [9], other languages, like Hungarian, have no choice but to implement their own engines in order to have a proper web searching tool available [8].

Basque is also an agglutinative language, so these problems are also applicable, but these are not the only difficulties. Being a minority language, Basque has an additional problem: no search engine offers the possibility of returning pages in Basque alone. Therefore, it is impossible to obtain results for numerous words in Basque, because their forms coincide with words existing in other languages.

So the need for a proper Basque search service is clear. A possible solution could be to set up our own search engine, one that 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 [14] –Basque language detection and lemmatizing were implemented long ago [1]–, but it is beyond our possibilities and objectives to implement and maintain all the infrastructure that a search engine and its crawling, indexing and serving involves –bandwidth, disk, reliability, etc.–. This is why we embarked on a project to develop a proper Basque search service built upon the APIs of existing search engines, so that the solution obtained and the methodology could be applied to other agglutinative or minority languages as well.

2. METHODOLOGY
2.1 Description of the problem
There are two main reasons that make existing search engines unsuitable for the case of Basque. The first is that Basque is an agglutinative language, that is to say, a given lemma makes 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 [3]. For example, the lemma *lan* ("work") forms the inflections *lana* ("the work"), *lanak* ("works" or "the works"), *lanari* ("to the work"), *lanei* ("to the works"), *lanaren* ("of the work"), *lanen* ("of the works"), etc. This means that looking only for the exact word given or the word plus an "*s*" for the plural is not enough for Basque. And the use of wildcards, which some search engines allow, is not an adequate solution, as these can return occurrences of not only conjugations or inflections of the word, but also derivatives, unrelated words, etc. For example, looking for *lan** would also return all the forms of the words *lanabes* ("tool"), *lanbide* ("job"), *lanbro* ("fog"), and many more.

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.

2.2 Looking for conjugations and inflections

When asking a search engine for a word, we need it to return pages that contain its conjugations or inflections, too. Our approach to this matter is based on morphological query expansion. The importance and use of morphology for various IR tasks has been widely documented ([13] [15] [16] [4]), although it is normally applied by lemmatization 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. In order to generate all the possible forms of a given lemma, we use a tool created by the IXA Group of the University of the Basque Country. This tool gives us all the possible inflections or conjugations of the lemma, and we ask the search engine to look for all of them by using an OR operator. For example, if the user asks for *etxe* ("house"), we ask the search engine for "(etxe OR etxea OR etxeak OR etxeari OR etxeek OR etxearen OR…)".

This is basically how we solve the first problem. It is a straightforward approach, easy to implement, but one which poses, of course, many minor problems and tweaks. The most relevant ones are as follows:

- The API of each search engine has its limitations with regard to search term count, length of search phrase, etc. We found no documentation on this, so we had to discover each limit by trial and error.
- These limitations render a proper lemmatized search for Basque impossible, as we cannot search for all the conjugations or inflections. So we used a corpus to see which the most frequent cases, numbers, tenses, etc. were, and we send their respective forms, in order to make the search results as satisfactory and representative as possible. In those cases in which the search engine is too limited, we make more than one query, each with some of the conjugations or inflections.
- Unfortunately, there is not much documentation about how search engines behave when they are given more than one search term in an OR. Do they start by looking for the first search term and return its results, and only go on to the next term if there are not enough results with the first one? If so, our results would only be better than those of a general search engine if the word in question was very rare. Anyway, we do not think this is what search engines do, as the *snippets* –short extracts of the pages containing the search term(s)– that they return often contain more than one of the search terms. In fact, we have the impression that they try to return pages that have as many different search terms as possible, which is best for our purposes as it improves representativeness. The increase in recall that emerged in the evaluation seems to confirm our previous assumptions.

All in all, we can conclude that this method enables us to obtain a satisfactory lemmatized search for Basque.

2.3 Language discrimination

We have mentioned earlier that there is no commercial search engine that can distinguish pages in Basque and return them alone. This poses a problem when searching for a proper noun or a word that exists in other languages; this often happens with technical words –*anorexia*, *sulfuroso*, *byte*, *allegro*…– and short words. Although there are language detection tools for Basque, a search for such words returns pages in English, Spanish, etc. but rarely any in Basque, so a subsequent filtering of these pages using a language detection tool would be useless.

The approach we have taken to solve this problem is to include, in the search phrase as a filter, the most frequently used words in Basque, in conjunction with an AND operator. Again, we used a corpus to see which these most used words were.

Unfortunately, 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 –the four most used words are *eta* ("and"), *da* ("is"), *ez* ("no") and *ere* ("too"), and the first two at least have well-known meanings used in other languages–. Therefore, we had to include more than one filter word, but how many were needed? The higher the number of these words we included, the higher the precision obtained (fewer non-Basque pages were returned). However, there was also loss in recall (more Basque pages were left out because they did not contain one or more of the words), and vice versa. The logical choice was to opt for precision –showing the user results in other languages would give a poor image of a Basque search and, besides, the user would never know how many results he or she was missing–, so in the default behaviour we include four of these most frequent terms in the search phrase. However, if the number of results is too low, the user is given the option of trying again increasing the recall –that is, with less filtering words.

Nevertheless, this failed to resolve the language-filtering problem completely. Even with the filtering words method, non-Basque pages or bilingual pages in which the search term was in a non-Basque part were returned at times. To filter these results, we use LangId, a free language identifier based on word and trigram frequency developed by the IXA group of the University of the Basque Country. This is applied to the snippet returned by the search engine.

By combining these methods we are able to show results that are exclusively in Basque with a high degree of accuracy.

2.4 Variant searching

Expanding the query using variants of the search term to improve the results was suggested long ago [10]. When performing a

Basque search, having the option of looking not only for the word but also for different variants of a word –archaic spellings, common errors– or even typing errors is very interesting. It must be taken into account that the standardization of Basque only started in the late sixties, and that many rules, words and spellings have changed since. Besides, Basque was not taught in schools until the seventies, nor in universities until nearly into the eighties. All this has led to a scenario in which even written production abounds with misspellings, corrections, uncertainties, different versions of a word, etc. But, above all, the main problem is that there are many areas or words upon which no decision as to the standard word or spelling has yet been taken.

The possibility of looking for variants as well has been added as a user option in our tool. All the linguistic tools made for Basque rely upon EDBL, a lexical database developed by the IXA Group of the University of the Basque Country [2]. This database links each word with its known variants, common errors and archaic spellings. So when sending all the possible inflections or conjugations of a word in an OR to the search engine, it is possible to include these variants, too. If, for example, the user inputs the word *jarduera* ("activity"), the system can ask the search engine to seek , simultaneously, the forms of *iharduera*, a now deprecated spelling widely used until 1998.

3. EUSBILA

EusBila is the solution we have developed for a Basque search service, making use of the APIs of major search engines and applying the methods mentioned above –lemma-based searching, language-filtering words and variant searching option–. In this section we will explain in more detail how EusBila works, and what its features are.

3.1 System architecture

The general architecture of the system is as follows:

- The user enters a search term.
- If the user has selected the corresponding option, EusBila uses EDBL to obtain the variants of the search term.
- The morphological generator is called to obtain the inflections and conjugations of the search term.
- A search phrase is built by combining the conjugations and inflections of the search term within an OR operator, and the filtering words with an AND operator.
- The APIs of the search engines are queried with the search phrase.
- The snippets returned by the engines are subjected to a final language test using LangId.
- The results are returned to the user.

3.2 Features

Some of the features of EusBila are as follows:

- Lemma-based and language-filtered search: EusBila performs an internet search for Basque by making use of the APIs of search engines, but simultaneously using morphological generation to obtain a lemma-based search and filtering words to obtain a language-filtered search.
- Variant searching: The user can also choose to look for known variants –common errors, archaic forms…– of the word.
- More than one search term: The user can enter more than one search term, and the lemma-based search is performed for all of them.
- Exact phrase searching: Search engines usually offer the possibility of performing an exact phrase search by enclosing the search terms in double quotes. EusBila offers this possibility too, but it applies the morphological generation to the last word of the phrase, thus performing a proper lemma-based search for whole noun phrases or terms –in Basque only the last component of the noun phrase is inflected.

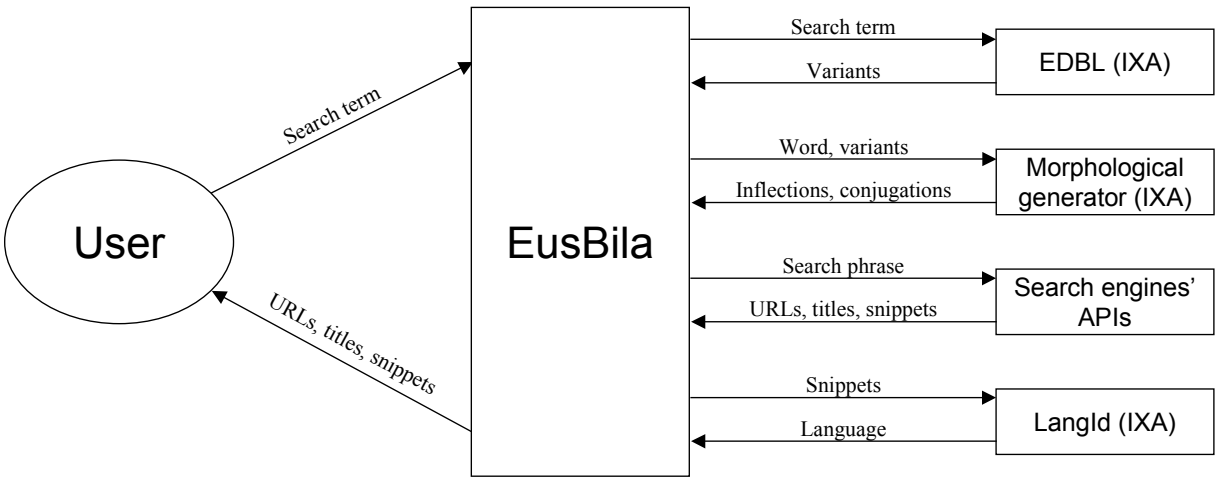

Figure 1. Diagram showing EusBila's architecture.

- Lemma and POS of the search term: The user can enter a search term that is not a plain lemma but a form of a lemma –conjugation or inflection–. The search term is analyzed to get its lemma and POS, and the morphological generation is made according to them. If the form is ambiguous, the most probable lemma and POS are taken for the morphological generation, but when the results are returned, the user is given the option of trying with the other analysis.
- Calls for showing proper snippets: Snippets are the short extracts of the pages that search engines return. As EusBila includes some language-filtering words in the search phrase, the snippets sometimes show these language-filtering words, rather than the word the user was looking for. In these cases EusBila shows no snippet, as the information it contains is irrelevant to the user. But snippets are very useful to help the user decide which link may contain the information he or she is looking for, so EusBila offers the possibility of trying to show as many snippets as possible. This is done by making another call to the APIs of the search engines' for each result without a proper snippet, but restricted to the site and without the filtering words. Naturally, activating this option makes the search slower.
- Various search engines: EusBila can choose among different search engines (Google, Yahoo, Microsoft, Alexa…). But each of these APIs have their own limit in terms of the number of queries per day. So when opening the service to the public, these limits have been taken into account, and we have chosen to offer EusBila's Basque search service through Microsoft's API. The other choices will either be insufficient for the use a Basque search service might have, or else a fee must be paid to use them. We are of the opinion that the number of queries per day offered by Microsoft's API will be enough for EusBila; if not, the commercial license is possible too. In any case, for other minority languages, the other choices might possibly be suitable. The following table shows the limits and licensing possibilities of the APIs we have implemented.

Table 1. Limits and licensing possibilities of the APIs

API	Free access		Commercial license
	Queries / day	Results / Query	
Google	1,000	10	No
Yahoo	5,000	100	No
Microsoft	25,000	50	Yes
Alexa	-	-	Yes

4. EVALUATION

The overall impression of any EusBila user is positive. It is clear that it outperforms the major search engines for a Basque search, as it solves the two problems mentioned above. But in order to translate these impressions into objective figures, we have designed and carried out a quantitative evaluation, comparing the results of EusBila with those of a major search engine.

4.1 Design of the evaluation

To carry out the evaluation, we decided to assess the two improvements of EusBila –morphological generation and language-filtering words– separately, and see the effect they had on precision and recall.

In order to do this, we ran searches for a sample of Basque words both through a commercial search engine and through EusBila (using the API of that same engine), in which only the improvement method being evaluated was activated, and then we compared the first 100 results. We thought it was best to use only one API throughout the whole evaluation, and we chose Microsoft, as it is the one that offers the highest number of queries per day –the intensive use of the API needed for the evaluation would easily surpass the daily limit of the others and would many days just to retrieve the results.

For evaluating the effects of the improvements in recall –either loss or gain–, we measured two variables: the difference in the estimated hit counts returned by the API and the number of different results in the improved query. We are aware that hit counts returned by search engines do not constitute an exact or reliable measure [12], but they are used by many researchers as an acceptable approximation [11]. For our case, we think that hit counts are a clearer indicator of recall than the other measure. Nevertheless, we show the results of the two variables. Both of them were measured and compared automatically, without human intervention.

For evaluating the gain in precision, we measured the difference in the percentage of Basque pages. This was done by language experts, who recorded the language each page returned was in.

With respect to the words, we thought it would be better to carry out the evaluation using real, ordinary Basque search terms, rather than choosing random words. For this purpose, we obtained the search logs spanning a whole year from a very popular science portal in Basque, Zientzia.net (http://www.zientzia.net), which meant that we had more than 500,000 searches that made up a total of more than 50,000 different words. We lemmatized these words and ordered them according to decreasing frequency, and took the topmost ones.

We mentioned above that EusBila's language-filtered search 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 when possible, the evaluation variables were measured 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 this category (and consequently the ones used for our evaluation) were: *ur* ("water"), *herri* ("people", "town"), *lur* ("earth", "ground"), *zuri* ("white", "to you"), *baso* ("wood"), *euri* ("rain"), *HIES* ("AIDS"), *berri* ("new"), *hartz* ("bear"), *nola* ("how").
- Proper nouns: Proper nouns are usually the same in other languages. The words for this category were *Egipto* ("Egypt"), *Galileo*, *Edison*, *Newton*, *Pluton* ("Pluto"), *Darwin*, *Galilei*, *Thomas*, *Franklin*, *Einstein*.

- International words: Words that we know definitely exist in another language (usually English, Spanish or French). These were the most searched for words in this category: *energia* ("energy"), *historia* ("history"), *mota* ("kind"), *sistema* ("system"), *ozono* ("ozone"), *planeta* ("planet"), *mineral* ("mineral"), *droga* ("drug"), *biografia* ("biography"), *natural* ("natural").
- Words that are probably 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. These were the words used: *animalia* ("animal"), *petrolio* ("petrol"), *zelula* ("cell"), *nuklear* ("nuclear"), *zentral* ("central"), *klima* ("climate"), *efektu* ("effect"), *zientzia* ("science"), *elektriko* ("electric"), *aparatu* ("system", "device").
- Basque words: Words that we are almost sure do not exist in any other language. The most searched for words in this category were *kutsadura* ("pollution"), *berriztagarri* ("renewable"), *elikadura* ("feeding"), *gaixotasun* ("illness"), *ugalketa* ("reproduction"), *berotegi* ("greenhouse"), *gizaki* ("human"), *basamortu* ("desert"), *elikagai* ("food"), *minbizi* ("cancer").

For the overall measure, we made a weighted average of them, taking into account the frequency of use of each category. To calculate these frequencies, we classified approximately the first 400 words out of the more than 50,000 into one of the categories. This may not seem very much, but they do in fact account for more than 40% of the queries.

Table 2. Frequency and query percentage of each category of word

Category of word	Word		Query	
	Count	%	Count	%
Short words	72	18.65%	44,214	18.64%
Proper nouns	46	11.92%	17,491	7.37%
International words	63	16.32%	46,853	19.76%
Words probably in other languages	100	25.91%	63,266	26.68%
Basque words	105	27.20%	65,345	27.55%
Total categorized	**386**	**0.73%**	**237,169**	**40.27%**
Total	**52,701**		**588,996**	

4.2 Results

4.2.1 Gain in recall due to morphological query expansion

As we decided to evaluate each improvement of EusBila separately, in order to evaluate the effects of morphological generation without using the language-filtering words, it was necessary that it should be done only with the Basque words. We searched for them in Microsoft's search API, and then we repeated the operation, but using morphological generation. These were the results obtained:

Table 3. Gain in recall due to morphological query expansion for Basque words alone

Word	Hit counts		Increase	New results among the first 100	
	without morphological query expansion	with morphological query expansion		Count	%
kutsadura	2,778	3,373	21.42%	37	37.00%
berriztagarri	65	2,729	4,098.46%	88	135.38%
elikadura	10,804	11,818	9.39%	41	41.00%
gaixotasun	4,113	7,617	85.19%	75	75.00%
ugalketa	1,474	1,467	-0.47%	34	34.00%
berotegi	226	247	9.29%	34	34.00%
gizaki	4,897	12,853	162.47%	85	85.00%
basamortu	210	845	302.38%	69	69.00%
elikagai	2,579	8,957	247.31%	84	84.00%
minbizi	147	1,795	1,121.09%	84	84.00%
Total	**27,293**	**51,701**	**89.43%**	**631**	**65.39%**

4.2.2 Gain in precision due to language-filtering words

We then evaluated the effect of language-filtering words without applying morphological query expansion. We first made a normal search and then an additional one with language-filtering words. We measured the increase in the percentage of Basque results for each category of word, and obtained the following results:

Table 4. Gain in precision obtained by language-filtering words for each category of word, and weighted average

Category of word	Weight	% of Basque pages		Increase
		without filtering words	with filtering words	
Short words	18.64%	9.82%	97.38%	87.56
Proper nouns	7.37%	0.20%	76.41%	76.21
International words	19.76%	0.00%	97.18%	97.18
Words probably in other languages	26.68%	18.40%	100.00%	81.6
Basque words	27.55%	77.80%	99.57%	21.77
Weighted average		**27.19%**	**97.74%**	**70.55**

4.2.3 Loss in recall due to language-filtering words

In order to measure the loss in recall that language-filtering words could cause, we needed to have some Basque results before applying them, so it was essential that the chosen words should be exclusively Basque words. Thus we searched for such words in Microsoft's search API, and then carried out the same search, but using language-filtering words. Again, we measured the difference in the hit counts returned by the API and the number of

results that did not appear in the first 100 results of the non-language-filtered-search.

We have pointed out above that EusBila gives the option of choosing between precision and recall, and accordingly includes more or fewer language-filtering words. We have made searches with all the different options, from 1 filtering word to 4, so the result of this evaluation is a range of percentages, as shown in the following tables.

Table 5. Loss in recall due to language-filtering words for Basque words alone, measured in hit count decrease

Word	Decrease in hit counts with			
	1	2	3	4
	language-filtering words			
kutsadura	4.72%	19.26%	35.39%	42.84%
berriztagarri	-44.62%	-38.46%	-13.85%	-4.62%
elikadura	4.69%	45.82%	69.40%	73.85%
gaixotasun	1.56%	10.60%	24.48%	35.52%
ugalketa	60.65%	86.30%	83.45%	84.74%
berotegi	3.10%	13.72%	17.26%	21.68%
gizaki	2.37%	8.35%	14.03%	45.62%
basamortu	22.38%	7.62%	26.67%	28.10%
elikagai	0.58%	28.15%	44.44%	54.91%
minbizi	11.56%	13.61%	19.05%	76.19%
Total	**6.48%**	**30.67%**	**46.40%**	**57.69%**

Table 6. Loss in recall due to language-filtering words for Basque words alone, measured in pages no longer among the first 100

Word	% of pages no longer among the first 100 with			
	1	2	3	4
	language-filtering words			
kutsadura	31.43%	34.29%	37.14%	42.86%
berriztagarri	28.07%	35.09%	50.88%	47.37%
elikadura	41.79%	44.78%	67.16%	74.63%
gaixotasun	38.75%	40.00%	50.00%	58.75%
ugalketa	61.54%	58.97%	61.45%	65.38%
berotegi	34.09%	40.91%	46.59%	52.27%
Gizaki	46.91%	43.21%	49.38%	59.26%
basamortu	37.68%	34.78%	43.48%	56.52%
elikagai	30.77%	33.33%	46.15%	55.13%
minbizi	25.61%	24.39%	34.15%	75.61%
Total	**37.87%**	**39.07%**	**48.40%**	**59.07%**

Although the loss in recall is not negligible quantitatively speaking, it is not so important in terms of real user experience. The results that are left out because they do not have one or more of the filter words do not usually have very much content. Any text in Basque that is sufficiently long normally contains the filter words. Therefore, even if some results are left out, the ones that remain are usually longer and, therefore, more relevant. This is an impression we have; it has not been evaluated. And in any case, if there are not enough results or if the user does not find the desired result, the system gives the option of trying again with increased recall –that is, with fewer filter words.

4.2.4 Gain in recall due to morphological query expansion with language-filtering words applied

After measuring the two improvements separately, we thought it would be interesting to evaluate both of them together. The application of language-filtering words would let us measure the effect of morphological generation in words that do not exist exclusively in Basque.

This time we used the most searched for words for each category of word once again. Firstly, we tried a search with the language-filtering words and then with both language-filtering words and morphological generation. Again we measured the difference in the approximate hit counts returned by the API and the number of new results that did not appear in the first 100 results of the non-morphological-query-expansion search.

The results of each category of word and the weighted average can be seen in the following table:

Table 7. Gain in recall obtained by morphological generation for each category of word and weighted average

Category of word	Weight	Gain in hit counts	% of new results
Short words	18.64%	43.75%	71.30%
Proper nouns	7.37%	11.83%	37.85%
International words	19.76%	16.51%	53.47%
Words probably in other languages	26.68%	64.37%	61.05%
Basque words	27.55%	57.36%	59.50%
Weighted average		**40.19%**	**59.94%**

4.3 Summary

This is a summary of the results obtained in the evaluation:

- Gain in precision due to language-filtering-words: increase of 70.55 points –from 27.19% to 97.74%– in the percentage of Basque pages.
- Loss in recall due to language-filtering words: a decrease ranging between 6.48% and 57.69% in hit counts, depending on the number of words
- Gain in recall due to morphological generation:
    - With words that exist only in Basque and without language-filtering words: an 89.43% increase in hit counts
    - With any word and applying language-filtering words: a 40.19% increase in hit counts

The evaluation shows that the benefits obtained with our methodology for a Basque search are considerable, so we can conclude that EusBila is a valid service for searching in Basque. Although the loss in recall due to language-filtering words is significant in quantitative terms, we have the impression that those fewer results are qualitatively better, and in any case, the user can reduce the amount of filter words if necessary.

5. CONCLUSIONS

Using search engines for making a query in a minority and agglutinative language like Basque is often a frustrating experience, as they do not perform lemma-based searching or return results in Basque alone.

With EusBila we have built a Basque search service that doesn't need to crawl or index anything, as it makes use of the APIs of the main search engines. To obtain a lemma-based search it uses morphological query expansion, and to obtain pages in Basque alone it uses language-filtering words.

The evaluation has shown that the methodology used is valid, as the increase in performance –gain in precision due to language-filtering words and gain in recall due to morphological generation– is significant. Even if there is a loss in recall due to the language-filtering words, the reduced result set seems to be qualitatively better; moreover, it can be avoided as the inclusion of filter words –and the number of them– is optional.

Furthermore, it seems to us that the methodology used in EusBila could be used by other minority and agglutinative languages to build a search service suited to them, even more so if we take into account that the requirements of the system are very low, as it makes use of the APIs of the search engines.

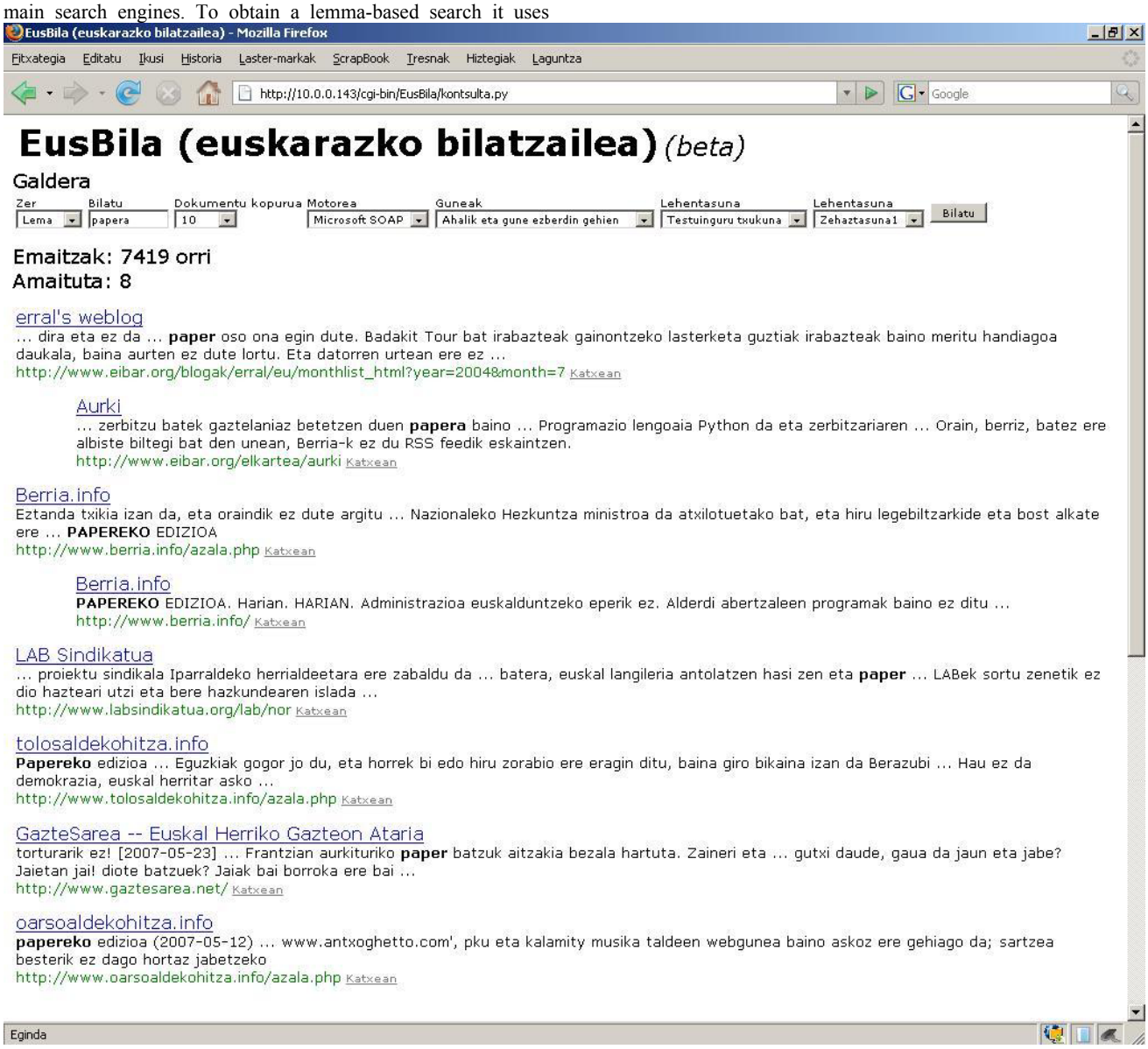

Figure 2. Screen capture of EusBila with results for *paper*. As can be seen, the results are lemma-based and in Basque alone

6. REFERENCES

[1] Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N., and Urizar, R. *EUSLEM: A lemmatiser / Tagger for Basque*. In *Proceedings of Euralex Conference* (Göteborg, Sweden, 1996), vol. I 17-26.
Also [online] [date: 2007-05-20]: <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911639/publikoak/96EUSLEM.ps>

[2] Aduriz, I., Aldezabal, I., Ansa, O., Artola, X., and Diaz de Ilarraza, A. *EDBL: a Multi-Purpose Lexical Support for the Treatment of Basque*. In *Proceedings of the First International Conference on Language Resources and Evaluation* (Granada, Spain, 1998), vol. II 821-826.
Also [online] [date: 2007-05-20]: <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911709/publikoak/98LREC.ps>

[3] Alegria, I., Artola, X., and Sarasola, K. *Automatic morphological analysis of Basque*. In *Literary & Linguistic Computing* (Oxford University Press, Oxford, 1996), vol. II nº 4 193-203.
Also [online] [date: 2007-05-20]: <http://hal.ccsd.cnrs.fr/docs/00/08/13/51/PDF/96LITER_M.pdf>

[4] Ambroziak, J., and Woods, W.A. *Natural Language Technology in Precision Content Retrieval*. In *Proceedings of the International Conference of Natural Language Processing and Industrial Applications* (Moncton, Canada, 1998).
Also [online] [date: 2007-05-20]: <http://www.sun.com/research/techrep/1998/smli_tr-98-69.pdf>

[5] Bar-Ilan, J. *Expectations versus reality – Search engine features needed for Web research at mid 2005*. In *Cybermetrics, International Journal of Scientometrics, Informetrics and Bibliometrics* (vol. 9, 2005), nº 1 paper 2.
Also [online] [date: 2007-05-20]: <http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html>

[6] Bar-Ilan, J., and Gutman, T. *How do search engines handle non-English queries? – A case study*. In *Proceedings of the 12th international World Wide Web Conference* (Budapest, Hungary, 2003), 415-424.
Also [online] [date: 2007-05-20]: <http://www2003.org/cdrom/papers/alternate/P415/415.pdf>

[7] Bar-Ilan, J., and Gutman, T. *How do search engines respond to some non-English queries?*. In *Journal of Information Science* (vol. 31, 2005), nº 1 13-28.

[8] Benczúr, A. A., Csalogány, K., Fogaras, D., Friedman, E., Sarlós, T., Uher, M., and Windhager, E. *Searching a small national domain - a preliminary report*. In *Poster of the 12th international World Wide Web Conference*, (Budapest, Hungary, 2003), 184-.
Also [online] [date: 2007-05-20]: <http://www2003.org/cdrom/papers/poster/p184/p184-benczur.html>

[9] Guggenheim, E., and Bar-Ilan, J. *Tauglichkeit von Suchmaschinen für deutschesprachige Abfragen*. In *Information, Wissenschaft und Praxis* (vol. 56, 2005), nº 1 35-40.

[10] Jones, K.S., and Tait, J.I. *Automatic search term variant generation*. In *Journal of Documentation* (vol. 40, 1984), nº 1 50-66.

[11] Keller, F., and Lapata, M. *Using the web to obtain frequencies for unseen bigrams*. In *Computational Linguistics* (vol. 29, 2003), nº 3 459-484.
Also [online] [date: 2007-05-20]: <http://acl.ldc.upenn.edu/J/J03/J03-3005.pdf>

[12] Kilgarriff, A. *Googleology is bad science*. In *Computational Linguistics* (vol. 33, 2007), nº 1 147-151.

[13] Krovetz, R. *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, 1993), 191-202.

[14] Langer, S. *Natural languages and the world wide web*. In *Bulletin de linguistique appliquée et générale* (vol. 26, 2001), 89-100.
Also [online] [date: 2007-05-20]: <http://www.cis.uni-muenchen.de/people/langer/veroeffentlichungen/bulag.pdf>

[15] Woods, W.A. *Aggressive morphology for robust lexical coverage*. In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (Seattle, Washington, 2000), 218-223.
Also [online] [date: 2007-05-20]: <http://acl.ldc.upenn.edu/A/A00/A00-1030.pdf>

[16] Woods, W.A., Bookman, L.A., Houston, A., Kuhns, R.J., Martin, P., and Green, S. *Linguistic knowledge can improve information retrieval*. In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (Seattle, Washington, 2000), 262-267.
Also [online] [date: 2007-05-20]: <http://acl.ldc.upenn.edu/A/A00/A00-1036.pdf>