CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque

Leturia I.[1], Gurrutxaga A.[1], Alegria I.[2], Ezeiza A.[2]
Elhuyar R&D, IXA Taldea of the University of the Basque Country

Abstract

Having the possibility of querying the web as if it were a corpus is very interesting as well as useful, and even more so in the case of a language like Basque, in view of the fact that the corpora this language has are few in number and limited in size, and that the process to standardize the language started relatively recently. But existing services of this kind are not suitable for Basque, due to its agglutinative nature and the fact that it is not a language to which search engines restrict their results.

In this paper we present CorpEus, a solution that allows the use of the web as a Basque corpus, featuring lemma-based searching, language filtering, variation searching, multiple text formats, parallel downloading, on-the-fly ordering of the KWICs in accordance with different criteria, lemma and POS analysis of the occurrences and occurrence counts and charts.

We also explain the methodology used by CorpEus to obtain a lemma-based and language-filtered search with the option of variant searching, which we think could be useful for building similar tools for other agglutinative and minority languages.

Keywords: corpora, web, web as corpus, Basque, agglutinative language, minority language, lemma-based search, language filtering, variant searching

1. Motivation

There is no doubt whatsoever that a need exists for corpora for linguistic research, for language normalization and for the development of language technologies. Although many corpora are created and used exclusively for this purpose, they are not made publicly available and searchable through the Internet to anyone who wants to consult them, mostly due to a lack of awareness regarding its importance.

However, it is essential for Basque to have corpora that can be queried through a web interface. It must be taken into account that the standardization of Basque did not start until the late sixties, and that many rules, words and spellings have been changing

[1] Elhuyar R&D; Zelai Haundi kalea 3, Osinalde Industrialdea; 20170 Usurbil; Basque Country; {igor,agurrutxaga}@elhuyar.com

[2] IXA Taldea of the University of the Basque Country; 649 Postakutxa; 20080 Donostia; Basque Country; {i.alegria,aitzol.ezeiza}@ehu.es

since. Besides, Basque was not taught in schools until the seventies and in universities until nearly 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 a decision as to the correct word or spelling has not yet been taken. So writers, technical text producers, dictionary makers, translators and even academics in the field of standardization need corpora in order to avail themselves of the data upon which to base their decisions.

And Basque is not exactly a language rich in corpora. The amount and size of corpora is proportional to the number of speakers and the economic resources of the Basque language. These are the only Basque corpora that are currently available to the public:

- XX. mendeko Euskararen Corpusa: a 4.6 million-word balanced corpus owned by the Academy of the Basque Language; it consists mainly of twentieth century literary texts (*XX. mendeko Euskararen Corpusa* [online]).
- Ereduzko prosa gaur: a 23.8 million-word corpus compiled by the University of the Basque Country, composed of literary and press texts regarded as "reference texts" from the years 2000 through 2005 (*Ereduzko prosa gaur* [online]).
- Zientzia eta teknologiaren corpusa: a 7.6 million-word corpus compiled by the Elhuyar Foundation and the IXA Group of the University of the Basque Country, consisting of texts on science and technology published between 1990 and 2002 (Alegria et al. 2006).
- Klasikoen gordailua: a non-tagged 10.7 million-word corpus made by Susa, consisting of classic texts (*Klasikoen gordailua* [online]).

As can be seen, there are very few Basque corpora, and they are small compared to those of major languages. We can also notice that none of them are being updated with recent texts, so they cannot be very helpful in resolving doubts concerning new technical words.

But we do have a huge repository of text that is constantly being updated, and this is the Internet, which contains many more texts in Basque than the other four corpora put together. So, would it not be interesting to be able to consult the Internet as if it were a large Basque corpus? Indeed it would.

However, we are aware that this use poses some major disadvantages, the main ones being the following:

- Such systems will always have some uncertainty due to the Internet's inherent non-linguistically-tagged nature.
- They will never be able to show all the existing information, only what appears in the first 1,000 results returned by the search engines –none that we know of return more–, as they have of necessity to be used to access the web.

- The web is not as balanced as an ideal corpus should be and, for that reason, the data obtained from it might not be representative. But then, what corpus is completely balanced and representative (Kilgarriff A. et al. 2004)?
- There is a lot of redundancy in the web.

Despite the well-known problems of such systems, we thought that the benefits a tool like this for Basque would bring far exceeded those disadvantages. So we embarked on a project to build a web service that would allow the Internet to be queried as a corpus. A word (or a number of them) would be requested from the service, which would return counts, contexts of use and other information on the use of the word in the Internet.

2. Related work

Similar services have already been implemented and are available for public use. Some examples are WebConc (*WebConc* [online]), WebCorp (Kehoe A. et al. 2002) or KWiCFinder (Fletcher W. H. 2006). But all these services rely on search engines, and the problems that non-English languages, and especially agglutinative languages, have with search engines are well known (Bar-Ilan J. et al. 2005 and Bar-Ilan 2005). While some search engines do seem to use some sort of additional techniques for languages like German (Guggenheim E. et al. 2005), other languages, like Hungarian, have no choice but to implement their own engines in order to have proper web searching for their language (Benczúr A. 2003). Basque is also an agglutinative language, so these problems apply to it as well, but these are not the only difficulties. Being a minority language, Basque has an additional problem, of which we have found no mention anywhere: no search engine offers the possibility of returning pages in Basque only. Therefore, it is impossible to obtain results in Basque when looking for many technical words or proper nouns. Those were the main problems we had to contend with in our project, and in this article we will be explaining how we solved them.

3. Problems with existing services

There are two main reasons why services like WebConc or WebCorp are 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, etc.) for words and adjectives, and the person (me, he, etc.) and the time (present, past and future) for verbs. A brief morphological description of Basque can be found in (Alegria I. et al. 1996). 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 given word or the word plus an "*s*" for the plural –which is what the search engines

upon which such systems are based do– is not enough for Basque. And the use of wildcards, which some search engines allow, is not an appropriate solution, as it can return appearances not only of conjugations or inflections of the word, but also of 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 these services can discriminate Basque pages in their searches –which is understandable, since no search engine does, but this still poses a problem–. Searching in any of the aforementioned services for a technical word that exists also in some other language –of which there are many, such as *anorexia*, *sulfuroso*, *byte* or *allegro*–, the results will not be in Basque alone –in fact, there are often no results in Basque at all.

So, taking into account that the existing solutions did not meet our needs, we had no choice but to implement our own.

4. Solving the searching problems

The first job that a 'web as corpus' tool has out of necessity to perform is to make use of search engines. But it is commonly known in the Basque Country that search engines do not provide satisfactory results for Basque, as they do not offer the possibility of looking for pages written only in Basque, and they only return pages containing the exact given word, not its inflections or conjugations. And in our case, as has previously been stated, the search has to be lemma-based and return results in Basque only. In this section we will be explaining how we achieve this.

4.1. 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 generation. 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 that gives us all its possible inflections or conjugations, 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 etxeek etc.)".

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

- The API of each search engine has its limitations as far as search terms, count, length of search phrase, etc. are concerned. We found no documentation on this, so we had to discover each limit by trial and error.

- These limitations make it impossible to conduct a real lemmatized search for Basque, as we cannot search for all the conjugations or inflections. So we used a corpus to see which the most frequent cases, numbers, times, etc. were, and sent 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 made more than one query, each with some of the conjugations or inflections.
- Unfortunately, there is little documentation on how search engines behave when they are given more than one search term in an OR. Do they look for the first search term first and return its results and go for the next term only if there are not enough results with the first? If this were the case, then the results might not be suitable for a corpus-like use. Although we cannot be completely sure about this, we do not think that this is what they do, as the snippets –short extracts of the pages containing the search term(s)– that they return often contain more than one search term. In fact, we have the impression that they try to return pages that have as many different search terms as possible, which is what is best from a corpus point of view.

All in all, we can conclude that, by using this method, we can get a satisfactory lemmatized search for Basque.

4.2. Language discrimination

We have mentioned earlier that there is no commercial search engine –at least none that we know of– that can distinguish pages in Basque and return them only. This poses a problem when searching for a word that exists in other languages, which often happens with technical words: *anorexia*, *sulfuroso*, *politekniko*, *byte*, *allegro*… And the search for technical words is, we think, a very usual and useful application for a Basque corpus, because technical terminology is the least normalized area of the Basque language.

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, joined with an AND. 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, their chance of existing in other languages or being used as abbreviations or acronyms is quite high. Therefore, there is no single "magic" word that exists only in Basque and in all Basque texts that we could use as a filter. The most frequent word in Basque is *eta* ("and"), but it is also the name of an armed group widely mentioned in the media in any language; the next most frequent is *da* ("is"), which is also "yes" in many Slavic languages; and the next ones are two- or three-character words, too.

Therefore, we had a problem deciding how many of these most frequent words we should include in the query. Once again, we came up against the omnipresent dichotomy in language technologies: precision vs. recall. The higher the number of

these words that we included, the more we gained in precision (fewer non-Basque pages were returned) but we also lost in recall (more Basque pages were left out because they did not contain one or some 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 behavior we included four of the frequent terms in the search phrase. In any case, if the number of results is insufficient, the user is given the option of trying again while increasing the recall, though this does not always improve the result, as many non-Basque pages are often returned.

Nevertheless, this does not completely resolve the language-filtering problem. The multilingual nature of the web gave us another difficulty. Experience has shown us that there are many bilingual (or even trilingual) pages in the Internet. With the filter words method explained above, we obtain pages that have Basque in them, but they are not necessarily monolingual, because the searched-for word is sometimes in a piece of text that is in another language. To filter out these occurrences, 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, which we apply to some context around each occurrence of the search term. Choosing the right length of the context caused a slight dilemma too: if it was too short, the language identifier would not have enough data to decide the right language correctly; if too long, bits of text in other languages could be included. Empirically we found that the best result was obtained if we first tried with quite a broad context. Then, if LangId said that the text was not Basque, which would normally be due to parts in other languages being included, a second attempt was made by reducing its length progressively until a minimal length was reached. The occurrence would be included in the result if some of the attempts said that the language was Basque.

By combining these methods we are able to show, with great accuracy, only those results that are in Basque.

4.3. Variant searching

One of the disadvantages of using the web for linguistic purposes is that the web is not linguistically tagged. This has two main consequences:

- There will always be some uncertainty with ambiguous words, that is to say, words that can have more than one lemma. When trying to search for one of the lemmas, we cannot be sure if the lemma of the returned words is the one we searched for or another one. The dative form of the lemma *pilota* ("ball") is *pilotari*, which also means "pelota player". Logically, a search for the lemma *pilota* will have to look for *pilotari* too, and it will return not only those occurrences referring to the first meaning (dative form of *pilota*), but also those

meaning *pelota player*. In a corpus that has been tagged and manually disambiguated we would not have this problem.

- In linguistically tagged and manually disambiguated corpora, different variants of a word –old spellings, common errors– or even typing errors have their correct lemma assigned, so searching for a certain lemma would also return these forms.

At the moment there is nothing we can do about the first problem, but we have somehow managed to deal with the second. All the linguistic tools made for Basque rely upon the EDBL, a lexical database developed by the IXA Group of the University of the Basque Country (Aduriz I. et al. 1998). This database links each word to its known variants, common errors and old spellings. So when sending the search engine all the possible inflections or conjugations of a word in an OR, it is possible to send these variants too. This possibility has been added as a user option in our tool. If, for example, we are interested in the collocations or terms in which the noun *jarduera* ("activity") is the head, we could ask our system to simultaneously retrieve the occurrences of *iharduera*, a now deprecated spelling widely used until 1998.

4.4. Evaluation of the methodology

The methodology used in CorpEus to obtain a lemma-based search and for getting results in Basque only has been tested in an evaluation performed on EusBila, a search service for Basque that we have developed and which is based on the same techniques that CorpEus uses.

4.4.1. What is EusBila

As the method CorpEus uses to perform a lemmatized search for pages in Basque alone proved to be successful, and as the performance of major search engines for a Basque query is far from satisfactory, we started to think about the possibility of launching a search engine for Basque, based on the principles of CorpEus. The main problem of such a service lay in the limit on the number of calls per day of the APIs of search engines. While CorpEus is a tool aimed mainly at language professionals and its use is unlikely to exceed the limit of these APIs –we know the use a Basque corpus tool normally has, as we have access to the logs of our previous corpus tool, Zientzia eta Teknologiaren Corpusa–, these limits would clearly be insufficient for a general-purpose search service.

The recent change in the license of the API of Microsoft's Live Search has solved this problem and made possible the development of EusBila, a Basque search service based on Microsoft's API that uses CorpEus' methodology. Microsoft has augmented the number of calls per day it admits in its free version to 25,000, which might be enough for our service. Furthermore, should this number be exceeded, they have also added the possibility of a commercial license that charges on a per-different-user-over-one-million basis, which could never be too expensive in our case, as there are fewer

than one million Basque speakers. EusBila will be launched by the company Eleka Ingeniaritza Linguistikoa in September 2007, under the commercial name Elebila.

4.4.2. Results of the evaluation

We presented a paper on EusBila in the iNEWS 07 workshop –Improving Non-English Web Searching– that was held during the SIGIR 07 conference. For this paper we designed and carried out an evaluation to measure the effects of each improvement of EusBila: morphological query expansion and language-filtering words. These improvements in the searching are the same as those used by CorpEus, so the evaluation can be applied to CorpEus too.

For the evaluation, we compared the results of each improvement of EusBila with those of Microsoft's search engine. These results were compared in terms of precision and recall. The indicator we used for precision was the percentage of results that were actually in Basque, and the one for recall was the estimated hit counts returned. The words we chose for the evaluation were taken from the search logs spanning a whole year from a very popular science portal in Basque, Zientzia.net (*Zientzia.net* [online]).

Here follows a summary of the results of the evaluation:

Evaluation	**Condition**			**Measured variable**	**Result**
	Language-filtering words	**Morphological query expansion**	**Words**		
Gain in recall due to morphological query expansion	Not applied	-	Only Basque	Hit counts	89.43% increase
Gain in precision due to language-filtering words	-	Not applied	Any kind	% of results in Basque	70.55 points increase, from 27.19% to 97.74%
Loss in recall due to language-filtering words	-	Not applied	Only Basque	Hit counts	Decrease from 6.48% to 57.69%, depending on the number of language-filtering words*
Gain in recall due to morphological query expansion	Applied	-	Any kind	Hit counts	40.19% increase

Table 1. Summary of the results of the evaluation

More detailed data of the evaluation can be found in the aforementioned paper (Leturia et al. 2007).

5. CorpEus

CorpEus is the result of our efforts to devise a tool for consulting the Internet as if it were a Basque corpus. It can be queried in the address http://www.corpeus.org. In this section we will be explaining in further detail how CorpEus works and what its features are.

5.1. System architecture

The general architecture of the system is as follows:

- First it makes use of the APIs of search engines to obtain the pages in which the search term appears. We use the methods explained above to obtain a proper Basque search.
- Then it downloads the pages returned by the search engines.
- Finally, it shows the results in a "corpus way" –occurrence counts, KWICs of every occurrence, etc–, but only of the occurrences that are in a phrase in Basque, as we have mentioned earlier.

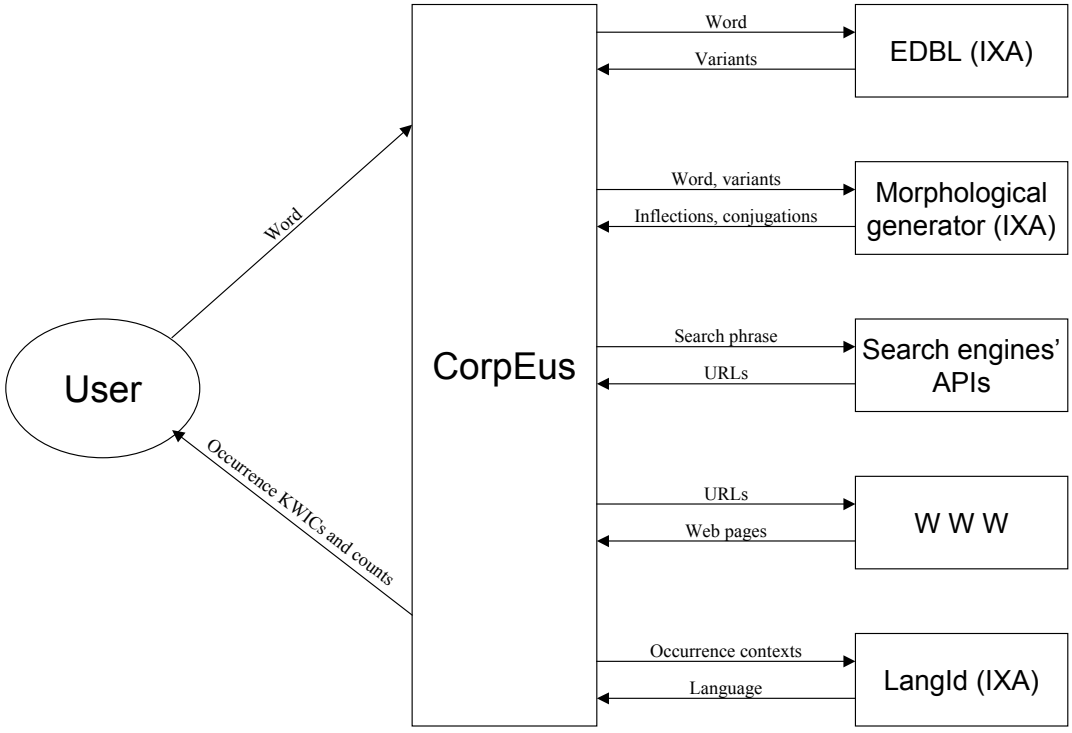

Figure 1. Diagram showing the architecture of CorpEus

5.2. Features

These are some of the features of CorpEus:

- Lemma-based and language-filtered search: The search of the Internet that CorpEus performs makes use of the techniques we have already described, in order to obtain pages in Basque alone but at the same time containing inflections and conjugations of the searched word. It also offers the possibility of looking for variants 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. Then all the appearances of any of the search terms are shown.
- Exact phrase searching: CorpEus offers the possibility, just as search engines do, of performing an exact phrase search by enclosing the search terms in double quotes. But EusBila applies the morphological generation to the last word in 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.
- Parallel downloading: Once the search engine has returned its results, each of the returned pages is downloaded. For the downloading, different processes are launched concurrently and the contexts are served in the order the pages arrive, so that a slow or blocked page does not stop the complete process.

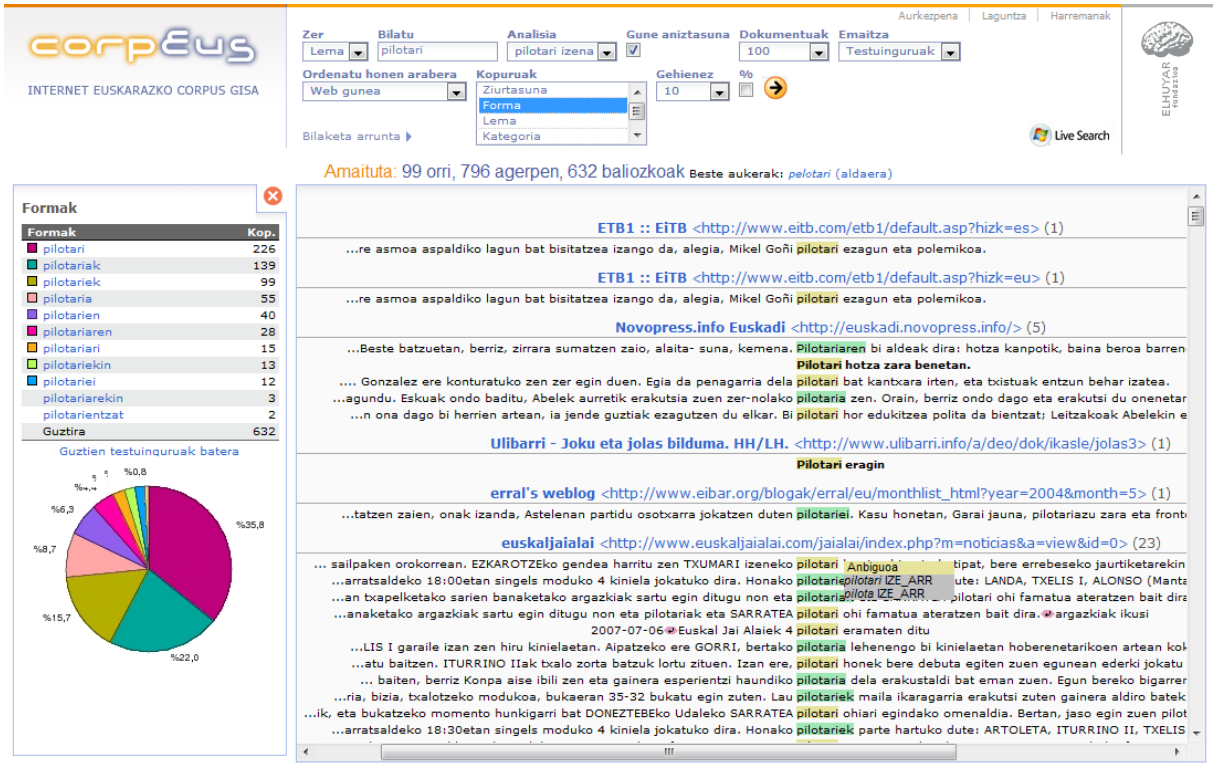

Figure 2. Screen capture of CorpEus with results for "anorexia"; as can be seen, the results are lemma-based and in Basque only

- Different page types: The web is made up not only of HTML files, but also of many other formats and kinds of files too (PDF, video files, sound files…). Corpus tools are interested in textual content, so in CorpEus we try to show the occurrences of the word in as many types of text content pages as possible. So far we have been able to access the content from HTML, XML, RSS, RDF, TXT, DBF, PDF, DOC, RTF, PPT, PPS and XLS files, using various free software tools to convert them or to extract their content.
- Ordering criteria: The KWICs can be ordered following different criteria. The default is the order in which the pages arrive, but the user can choose to order them by form of the searched word, context after, context before… And they are ordered on the fly as they come in, without having to wait until all the results have arrived.
- Searched word analysis: In the KWICs, each form of the searched word shows its possible lemma and POS analysis in a floating box that appears if the mouse is moved over it. The words that have only one possible analysis are shown in green, whereas ambiguous words are shown in yellow, and words that the analyzer does not recognize are shown in red.
- Count charts: CorpEus can show different charts with counts of word forms, possible lemma or POS, word before, word after…

6. Conclusions

The web is a huge and constantly-up-to-date text repository, so for minority languages the possibility of using it as a corpus is very interesting. However, tools that offer this kind of service do not work well for agglutinative or minority languages. With CorpEus we have developed a 'web as corpus' tool designed to solve the main problems of searching the web for Basque, performing a lemma-based search and offering results in Basque only. It also offers the choice of variant searching, features multiple text formats, parallel downloading, on-the-fly ordering of the KWICs according to different criteria, lemma and POS analysis of the occurrences and occurrence counts and charts. Furthermore, we have developed a methodology that could be useful for creating similar tools for other minority or agglutinative languages that have the same or similar problems as Basque.

References

ADURIZ I., ALDEZABAL I., ANSA O., ARTOLA X. and DIAZ DE ILARRAZA A. *EDBL: a Multi-Purpose Lexical Support for the Treatment of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 1998, vol. II p. 821-826.

Also [online] [date: 2007-04-25].
<http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911709/publikoak/98LREC.ps>

ALEGRIA I., ARETA A., ARTOLA X., DIAZ DE ILARRAZA A., EZEIZA N., GURRUTXAGA A., LETURIA I., POLIN Z., SAIZ R., SOLOGAISTOA A., SOROA A. and VALVERDE A. *Structure, Annotation and Tools in the Basque ZT Corpus*. Proceedings of LREC 2006 Conference, Genova, Italy, 2006, p. 1406-1411.

Also [online] [date: 2007-04-25]. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1141404023/publikoak/pdf>

ALEGRIA I., ARTOLA X. and SARASOLA K. *Automatic morphological analysis of Basque*. Literary & Linguistic Computing, Oxford University Press, Oxford, 1996, vol. II nº 4 p. 193-203.

BAR-ILAN J. *Expectations versus reality – Search engine features needed for Web research at mid 2005*. Cybermetrics, International Journal of Scientometrics, Informetrics and Bibliometrics vol. 9, 2005, nº 1 paper 2.

Also [online] [date: 2007-04-25]. <http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html>

BAR-ILAN J. and GUTMAN T. *How the search engines respond to some non-English queries?*. Journal of Information Science vol. 31, 2005, nº 1 p. 13-28.

BENCZÚR A. A., CSALOGÁNY K., FOGARAS D., FRIEDMAN E., SARLÓS T., UHER M. and WINDHAGER E. *Searching a small national domain - a preliminary report*. Poster Proceedings of Conference on World Wide Web, Budapest, Hungary, 2003.

Also [online] [date: 2007-04-25]. <http://www2003.org/cdrom/papers/poster/p184/p184-benczur.html>

Ereduzko prosa gaur [online] [date: 2007-04-25]. <http://www.ehu.es/euskara-orria/euskara/ereduzkoa/araka.html>

FLETCHER, W. H. *Concordancing the Web: Promise and Problems, Tools and Techniques*. Corpus Linguistics and the web, Editions Rodopi BV, 2006, p. 25-46.

Also [online] [date: 2007-04-25]. <http://www.kwicfinder.com/FletcherConcordancingWeb2005.pdf>

GUGGENHEIM E. and BAR-ILAN J. *Tauglichkeit von Suchmaschinen für deutschesprachige Abfragen*. Information, Wissenschaft und Praxis vol. 56, 2005, p. 35-40.

HÜNNING M. *WebConc* [online] [date: 2007-04-25]. <http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi>

KEHOE A. and RENOUF A. *WebCorp: Applying the Web to Linguistics and Linguistics to the Web*. WWW2002 Conference, Honolulu, Hawaii, 2002.

Also [online] [date: 2007-04-25]. <http://www2002.org/CDROM/poster/67/>

KILGARRIFF A. and GREFENSTETTE G. *Introduction to the special issue on the Web as corpus*. Computational Linguistics vol. 29, 2004, p. 333-348.

Also [online] [date: 2007-04-25]. <http://mitpress.mit.edu/journals/pdf/coli_29_3_333_0.pdf>

Klasikoen gordailua [online] [date: 2007-04-25]. <http://klasikoak.armiarma.com/corpus.htm>

LETURIA I., GURRUTXAGA A., ARETA A., ALEGRIA I. and EZEIZA A. *EusBila, a search service designed for the agglutinative nature of Basque*. Proceedings of iNEWS'07 workshop, SIGIR, Amsterdam, Holland, 2007, p. 47-54.

Also [online] [date: 2007-07-30]. <http://rea.teimes.gr/~lazarinf/ir7w/3DiNEWS07_Proceedings_N.pdf>

WebCorp [online] [date: 2007-04-25]. <http://www.webcorp.org.uk/>

XX. mendeko Euskararen Corpusa [online] [date: 2007-04-25]. <http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html>

Zientzia eta Teknologiaren Corpusa [online] [date: 2007-04-25]. <http://www.ztcorpusa.net>

Zientzia.net [online] [date: 2007-04-25]. <http://www.zientzia.net>