Evaluating different methods for automatically collecting large general corpora for Basque from the web

Igor LETURIA
ELHUYAR FUNDAZIOA, Usurbil, Basque Country (Spain)
i.leturia@elhuyar.com

ABSTRACT

In the last few years, much work has been done to build Basque corpora. But we still lack a large general corpus of a size comparable with those existing in other major languages, and much more so if we take into account the corpora lately built automatically from the web, which nowadays account for billions of word-sized corpora for English, German, Spanish, etc. As Basque is an under-resourced language, it is thus logical that we should also turn to this cheap and fast method of collecting corpora.

In this paper we present the research we have done to build a large general corpus of Basque from the web. We have tried and evaluated which of the two methods mentioned in the literature, that is, by crawling or by using search engines, best suits Basque, in terms of parameters such as speed, cost, size or quality. Our conclusion is that crawling is the one that has the potential for building the largest corpora for Basque. Using this method we have built a good quality corpus of more than 100 million words, and we expect to build a much larger one in the near future.

TITLE AND ABSTRACT IN BASQUE

Webetik euskarazko corpus orokor handiak automatikoki biltzeko metodoen ebaluazioa

Azken urteotan lan handia egin da euskarazko corpusgintzan. Baina oraindik ez dago tamainari dagokionez beste hizkuntza handiagoetakoekin konpara daitekeen corpus orokor handirik; are gehiago kontuan hartzen baditugu azkenaldian automatikoki webetik bildu diren corpusak: milaka milioi hitzetakoak daude ingelesa, alemana, gaztelania eta abarrerako. Euskara baliabide urriko hizkuntza izanik, logikoa da corpusak biltzeko metodo merke eta azkar honetara jotzea.

Artikulu honetan webetik euskarazko corpus orokor handi bat biltzeko egin dugun ikerketa aurkezten dugu. Literatura zientifikoan aipatzen diren bi metodoetatik, hau da, crawling bidez edo bilatzaileak erabiliz, euskararentzat hobekien zeinek funtzionatzen duen aztertu eta ebaluatu dugu, abiadura, kostua, tamaina edo kalitatea moduko parametroei erreparatuz. Gure ondorioa da crawling bidezkoa dela euskarazko corpus handiena eraikitzeko aukera ematen duena. Metodo hori erabiliz 100 milioi hitz baino gehiagoko corpus kalitatezko bat osatu dugu, eta etorkizun hurbilean askoz handiago bat eraikitzea espero dugu.

KEYWORDS: Basque, Corpora, Web as Corpus.

KEYWORDS IN BASQUE: euskara, corpusak, weba corpus gisa.

Proceedings of COLING 2012: Technical Papers, pages 1553–1570,
COLING 2012, Mumbai, December 2012.

1 Summary in Basque

1.1 Motibazioa eta helburuak

Corpusak oso baliabide garrantzitsua dira mundu modernoan biziraun eta komunikabideetan, hezkuntzan eta abar normaltasunez erabilia izan nahi duen hizkuntza batentzat, besteren artean (baina ez bakarrik) beharrezkoak direlako hizkuntza-teknologiak garatzeko. Euskararen kasuan are gehiago, oraindik ere estandarizazio prozesuan baitago eta estatusaren normalizazioan asko baitu egiteko. Baina euskara urria da corpusetan, modu klasikoan eraikitzea garestia delako eta ez dagoelako behar adina baliabide (ekonomikoak zein giza baliabideak) berauek eraikitzeko. Besteak beste, ez du corpus orokor handirik beste hizkuntzen pareko tamainakorik.

Hala ere, azkenaldian weba gero eta gehiago erabili da hizkuntz ikerketetarako edo corpusgintzarako; izan ere, merkeagoa da, corpus handiagoak lor daitezke eta beti eguneratuta dago. Horregatik logikoaz gain ia beharrezkoa da euskararentzat webera jotzea. Azken urteetan egin dira lanak euskarazko corpus mota ezberdinak automatikoki webetik lortzeko, baina oraindik ez da corpus orokor handien gaia landu. Artikulu honetan azaltzen dira helburu horrekin dauden metodo ezberdinak probatuz egin dugun ikerketa-lana eta beronen emaitzak.

1.2 Antzeko lanak

Bi modu nagusi aipatzen dira literaturan webetik corpus orokor handiak lortzeko. Bata crawling metodoa da: URL zerrenda batetik abiatuta, orri horiek jaisten dira eta bertako esteken URLak zerrendara gehitzen dira, eta horrela errekurtsiboki egiten jarraitzen dugu. Metodo honekin egin dira WaCky (Web-as-Corpus kool ynitiative) proiektuko corpusak, milaka milioi hitzetako corpusak alemana, italiera, ingelesa eta frantseserako, eta gehiago daude bidean (Baroni et al., 2009). Bestea bilatzaileak erabiltzean datza. Hitz zerrenda batetik abiatuta, horien konbinazioak bidaltzen zaizkie bilatzaileen APIei eta itzulitako emaitzen orriak jaisten dira. Metodo hau erabiltzen du Sharoff-ek (2006) 100-200 milioi hitz inguruko corpusak osatzeko hainbat hizkuntzatarako.

Bi metodoek emaitza onak lortu badituzte ere, ez dira beraien artean konparatu, beraz galdera asko geratzen dira airean. Zein da azkarragoa? Zeinek lortzen ditu kalitate handiagoko corpusak? Lor daitezke milaka milioietako corpusak bilatzaileen bidez? Gainera, euskararentzat baliteke ezberdin funtzionatzea metodoek: batetik, euskarazko bilaketa lematizatuak egiteko erabili behar diren teknikek emaitzei eragin diezaiekete; bestetik, euskarazko webaren tamaina txikiak eragin lezake crawling-a ez hain eraginkorra izatea. Beraz, biak aztertu eta ebaluatzea beharrezkoa da.

1.3 Metodologia

1.3.1 Bilatzaileen bidezko metodoa

Bilatzaileen bidezko metodoan, Sharoff-ek (2006) 500 hitz-forma maiz eta orokorren zerrenda bat erabiltzen du. Guk XX. mendeko Euskararen Corpuseko lema maizenak erabili ditugu eta gero sorkuntza morfologiko bidezko kontsultaren hedapena aplikatu (Leturia et al., 2008b). Euskarazko emaitzak soilik lortzeko, hizkuntza iragazteko hitzen teknika baliatu dugu (Leturia et al., 2008b). Sharoff-ek 500 hitz baino gehiago ere erabil daitezkeela iradokitzen du; horren eragina frogatu nahi izan dugu, 500, 1.000, 2.000, 5.000 eta 10.000 hitzeko zerrendak erabiliz corpus ezberdinak jaitsi eta ebaluatuta. Sharoff-ek 4 hitzeko konbinazioak bidaltzen dizkie bilatzaileen APIei eta hizkuntza txikiagoei 3; guk 1, 2, 3, 4 eta 5 hitzezko konbinazioekin egin nahi izan ditugu frogak. Sharoff-ek bilatzaileek itzulitako lehen 10 orriak jaisten ditu, guk badaezpada ere 50.

1.3.2 Crawling metodoa

WaCky proiektuan, crawling egiteko hasierako URLen zerrenda bilatzaileengandik lortzen zuten, hainbat hitzen konbinazioak bidalita; guk DMOZeko "Euskara" ataleko 1.500 URLak erabili ditugu. Haien proiektuan bezala, guk ere ataza anitz paraleloan abiarazita azkartzen dugu orrien deskarga eta webgune aniztasuna lehenesteko estrategia darabilgu.

1.4 Emaitzak

1.4.1 Bilatzaileen bidez

Bilatzaileak erabiliz, corpusik handienak 2.000 edo 5.000 hitzeko hasierako zerrendatik abiatuta lortu dira, 120-130 milioi hitz ingurukoak: lehenak webgune aniztasun handiagoa lortzen du, bigarrenak PDF gehiago (normalean arazoak ematen dituzte bihurtzean eta kalitatea ez da hain ona) eta dokumentu handiagoak (testu jarrai gehiago). Hitzen konbinazioaren luzera egokia 2 dela dirudi, berak lortzen baitu corpusik handiena (130 milioi baino gehiago) eta anitzena, PDF gutxienekin.

1.4.2 Crawling bidez

Bilatzaileen bidezko metodoekin lortu diren corpusek ez dute lortutako tamainak baino askoz gehiago handitzeko gaitasunik, ez behintzat modu produktiboan: bilatzaileei egindako lehen mila galderekin 37 milioi hitz lortzen dira bataz beste, baina egindako azken mila galderekin (12.000raino iritsi gara) 2 milioitik behera lortzen dira. Crawling bidez, aldiz, 100 milioi hitzetik gorako corpusak lortu ditugu (ia PDFrik gabe eta webgune aniztasun handienarekin) eta hazkunde erritmoak hasierakoaren erdia izaten jarraitzen du, beraz handitzen jarraitzeko potentziala du.

1.4.3 Konparazio kualitatiboa

Bi corpus klasikorekin (XX. mendeko Euskararen Corpusa eta Lexikoaren Behatokia) konparatu ditugu crawling bidez eta bilatzaile bidez lortutako web corpus bana. Bilatzaileen bidezkoa besteengandik ezberdintzen duena da testu administratibo gehiago izatea (ziurrenik erakundeen aldizkari ofizialetako PDFengatik), eta crawling bidezkoa webeko berezko generoko testu gehiago izatea. Corpus klasikoen hitz baliagarrien %90etik gora badaude web corpusetan, eta azken hauek %80 inguru hitz baliagarri berri gehiago dituzte lehenek baino.

1.5 Ondorioak

Frogatu dugu weba lehengaitzat hartuta posible dela 100 milioi hitzetik gorako corpusak osatzea bai bilatzaileen bidez bai crawling bidez, eta azken hau erabiliz ziurrenik askoz handiagoak ere osa daitezkeela. Gainera kalitatezkoak dira, corpus klasikoen hitz gehienak barne hartzen dituzte eta berri asko dauzkate. Beraz, webetik corpus orokor handiak biltzeak ekarpen handia egin diezaioke euskarazko corpusgintza eta hizkuntzalaritzari eta baita hizkuntzari orokorrean ere.

2 Motivation and objectives

A language that wants to survive in the modern world and be used normally in the media, in education, etc. needs to have language resources such as dictionaries or corpora. Besides, because language technologies are ever more present in everyday life through the web and our gadgets, it is imperative for any language with perspectives for the future to develop these language technologies; and these in turn need electronic dictionaries and corpora in order to be developed. And dictionaries (lexicographical or terminological) are nowadays produced on the basis of empirical evidence or previous use at least is studied, and these are both provided by corpora (they can even be semi-automatically extracted from corpora using NLP methods). So it is clear that corpora are a very valuable resource for many aspects of the development of a language.

In the case of the Basque language, the need for corpora is even greater, since its standardization did not start until the late 1960s and is still ongoing. But less resourced languages like Basque are not exactly rich in corpora: on the one hand, building a corpus in the classical way, i.e. out of printed texts, is normally a very costly process; on the other, the number of language experts or researchers dealing with these languages is much smaller than that of the major languages. So, the only Basque corpora that are currently available to the public are as follows:

- Orotariko Euskal Hiztegiaren Testu-Corpusa: a 6 million-word non-tagged corpus of classical literary texts produced by Euskaltzaindia, the Royal Academy of the Basque Language.
- XX. mendeko Euskararen Corpusa (http://www.euskaracorpusa.net/XXmendea): a 4.6 million-word balanced corpus produced by Euskaltzaindia; it consists mainly of twentieth century literary texts.
- Ereduzko Prosa Gaur (http://www.ehu.es/euskara-orria/euskara/ereduzkoa/): a 25.1 million-word corpus compiled by the UPV/EHU-University of the Basque Country, composed of literary and press texts regarded as “reference texts” from the years 2000 through 2006.
- Zientzia eta Teknologiaren Corpusa (http://www.ztcorpusa.net): a 8.5 million-word corpus compiled by the Elhuyar Foundation and the IXA Group of the UPV/EHU-University of the Basque Country, consisting of texts on science and technology published between 1990 and 2002 (Areta et al., 2007).
- Klasikoen Gordailua (http://klasikoak.armiarma.com/corpus.htm): a non-tagged 11.9 million-word corpus compiled by the publishing house Susa, consisting of classical texts.
- Lexikoaren Behatokia (http://lexikoarenbehatokia.euskaltzaindia.net): an 18.1 million-word corpus produced by Euskaltzaindia, the Elhuyar Foundation, the IXA Group of the University of the Basque Country and UZEI, made up of 21st century media texts.

But in recent years the web has been used increasingly for linguistic research, both via tools like WebCorp (Renouf et al., 2006) or KWiCFinder (Fletcher, 2006) that query search engines directly and show concordances, or via tools that use the Internet as a source of texts for building corpora to be used the classical way, after linguistic tagging and indexation (Baroni and Kilgarriff, 2006). As Kilgarriff and Grefenstette (2003) put it, although the use of the web as a source for building linguistic corpora has its detractors (who basically object to its lack of representativeness and reproducibility and to the quality of its texts), this approach offers undeniable advantages, mainly that the corpora that can be obtained are much larger, that the cost of the automatic building processes is much lower and that the web is constantly up to date.

So it is not only logical but also almost unavoidable that a less resourced language like Basque should proceed as other languages have and turn to the Internet, and this is what has been done in recent years. The following are the tools or resources built by using this web-as-corpus approach:

- CorpEus (http://www.corpeus.org/): a web service to query the web live as if it were a Basque corpus, by showing KWiCs of pages in Basque (Leturia et al., 2007).
- AutoCorpEx: a tool to automatically collect specialized corpora from the web (Leturia et al., 2008a)
- Co3: a tool to obtain domain comparable corpora from the web (Leturia et al., 2009).
- PaCo2: a tool to build parallel corpora from the web (San Vicente and Manterola, 2012).

Using the last three tools, various corpora (monolingual specialized, comparable and parallel) have been built. But there is still no large general corpus in Basque with a state-of-the-art size; as we have seen, the largest Basque corpus has 25 million words, whereas, taking English as an example, the BNC was finished in 1994 and runs to 100 million words (Aston and Burnard, 1998), and the web-derived corpus ukWaC contains more than 2 billion words (Ferraresi et al., 2008). So, the main objective of the research described in this paper was to build a general Basque corpus that was as large as possible (comparable to the sizes of the corpora we have mentioned, if possible). But in order to achieve this, we have had to test the different methods mentioned in the literature for collecting large general corpora from the web to see which performed best for Basque, because the features of the language might affect the results; so the results of the evaluation of the different methods are also shown.

3 Related work

There are roughly two methods that are mentioned in the literature when it comes to building large corpora out of the web. One of them is the crawling method: starting from a list of seed URLs, the pages they point to are downloaded, and the links found in them are added to the list of URLs to do likewise with them; we apply this recursively until the list is finished or we reach a predefined endpoint. As the web is a collection of interconnected pages, starting from almost any seed list sufficiently large and applying this method, the whole public web can be downloaded. This is the method used in the WaCky project (Web-as-Corpus kool ynitiative), an initiative to build gigantic web corpora for many languages (Baroni et al., 2009), with which they have already built four corpora for four languages of around or more than 2 billion words each (and more are on the way): deWaC for German (Baroni and Kilgarriff, 2006), itWaC for Italian (Baroni and Ueyama, 2006), ukWaC for English (Ferraresi et al., 2008) and frWaC for French.

The other method relies on the use of search engines. A list of seed words is used, combinations of them are sent to the APIs of search engines and the resulting pages are downloaded, until the goal size is reached, no more combinations are left or no new pages are returned. This is the method used by Sharoff (2006) to build BNC-sized corpora (around 100-200 million words) for various languages.

Both methods report success stories. They are able to obtain corpora of the desired size and the word frequencies are comparable with those in classical corpora such as BNC. However, the two methods or the corpora obtained with them have not been compared with each other, so many questions remain in the air. Which is the fastest method? Which obtains the best quality corpora? Is it possible to obtain corpora of billions of words with the search engines method?

And even if it was clear which of the methods is the best, it does not necessarily have to be so for obtaining a corpus in Basque, due to the singularities of the language. For example, no search engine offers the possibility of restricting its results to pages that are in Basque or to perform a search taking the rich morphology of Basque into account, so some hacks have to be used when querying search engines for content in that language, which might affect the results of the corpora obtained. Or, due to the smaller size of the Basque web, the crawling method might not obtain a sufficient size because it might leave out a significant part of the Basque web if the seed URLs list is not good or large enough. So, testing and evaluating both methods (and with different parameters) for collecting a large general corpus of Basque is a necessary task.

4 Methodology

4.1 Search engine method

Sharoff (2006) uses the search engine method to build 100-200 million-word corpora for various languages. He uses a seed list of 500 words, which have to meet certain requirements: they must be frequent, they have to be general (i.e., they should not indicate a specific topic) and they must not be function words (prepositions, articles, pronouns, conjunctions, etc.).

In our case, we acted likewise. We took the list of frequent words from XX. mendeko Euskararen Corpusa (see above), and we removed the non-desired ones. We took out the pronouns and conjunctions, but there was no need to remove articles or prepositions (Basque is an agglutinative language and these are appended to the words). Topic-specific words were not removed, because there were not many of them among the first 500 (the most frequent words tend to be general).

However, we take a different approach to Sharoff's regarding lemmatization. As he points out, general search engines do not perform lemmatization, so his seed words list is formed by word forms. We use a list of lemmas and apply morphological query expansion when calling the API of the search engine. Basically, this consists of obtaining the inflections of a word by morphological generation and sending the most frequent ones within an OR operator. For example, if the lemma of a word is *etxe* ("house"), the search engine is asked for "etxe OR etxea OR etxeak OR etxeari OR etxeek OR etxearen OR…". This method has proven to be effective for obtaining from search engines a lemma-based search for Basque, and is the method usually used in services or projects that need to search for content in Basque (Leturia et al., 2008b; Leturia et al., 2009).

In order to obtain from search engines only the pages in the language of the corpus to be collected, some studies consider the need to select words that are unique to the language (Kilgarriff and Grefenstette, 2003; Ghani et al., 2003), rejecting words like "restaurant" that exist in several different languages. The above-mentioned work by Sharoff uses the language filter of search engines, except for Ukrainian (which is not covered by search engines), for which the query is complemented with a couple of very frequent function words that are not used in cognate languages. We do not reject words existing in other languages but we are not in a position, either, to use the language filters of search engines because none of the existing search services can limit the results to pages in Basque. We apply the technique of the language-filtering words, like Sharoff for Ukrainian, by appending the most frequent words in Basque to the query ("… AND eta AND da AND ez AND ere"). As proven by Leturia et al. (2008b), this is the most effective method for obtaining results in Basque alone from search engines, although it means a loss in recall.

In the aforementioned paper, Sharoff also suggests that more than 500 words can be used. It would indeed be interesting to test the effect of the length of the seed word list in the corpus collection process and the obtained corpora; so, we tried with seed words lists of 500, 1,000, 2,000, 5,000 and 10,000 words.

In that same work, 4-word combinations are sent to the APIs, in order to get pages that contain relatively long pieces of connected text and a smaller number of noisy pages, i.e. tables or lists of links. However, he states that it is possible to relax the condition for four words in a query for languages which do not have sufficient number of Internet pages, and in fact he used 3-word combinations for Romanian. Because the Basque web may be orders of magnitude smaller than that in other languages, there is justification in seeing if there is in fact improvement with a shorter combination length; and there is no reason why the effect of longer ones should not be checked as well. That is why we also tried and evaluated 1, 2, 3, 4 and 5-word combinations.

Regarding the search engine, we used Google's API, just like Sharoff. From the results returned by the API, Sharoff downloads the first 10 pages. We decided to download the first 50, for one reason: because of the smaller size of the Basque web, many searches return no results (especially in the longer seed words lists and the longer combinations); so in order to build larger corpora while making the least possible number of queries, we downloaded more results from the productive queries.

4.2 Crawling method

The crawling method needs a list of seed URLs as a starting point. The corpora collected by the WaCky initiative (Baroni et al., 2009) obtain these seed URLs by making random queries of 2-word combinations to search engines (they make about 1,000 queries for getting around 10,000 seed URLs). In our case, we took the 1,500 URLs of the *Euskara* (meaning Basque language) section of DMOZ, the Open Directory Project. Although it is not an exhaustive list of all the websites in Basque and it is not as active and updated as it used to be, all the most important sites are undoubtedly there, and by following the links present in them recursively, we believe that it is almost certain that ultimately no site would be left out (except for island sites that have no inbound links from anywhere, but neither would these be indexed by search engines). However, this is one of the points that we wanted to test in our experiments.

The crawling is done in a multi-threaded parallel way with a breadth first strategy (prioritizing website variety above website completeness), just as in the WaCky initiative.

4.3 Common filters in both methods

The pages that are downloaded, whether by using search engines or by crawling, need to go through some cleaning and filtering if we want to build a quality corpus. Here we will describe how we implemented these steps.

Language filtering. When building a corpus, one is usually looking for texts in a language. When using search engines, the language filter is done by telling the search engine to return only results in that specific language. But when using the crawling method or, in some cases, also the search engine method (if search engines do not offer filtering by the language we want), it is up to us to do the language filtering after downloading. For this task we make use of LangId, a language identifier based on character and word trigram frequencies specialized in Basque, applied at paragraph level so that we can also extract content from bilingual documents.

Length filtering. Fletcher (2004) proved that filtering web documents by their size improved the quality of the web corpora. Those that do not reach a minimum are usually error messages from web servers or tend to have little textual content once page headers, menus, etc. are removed. On the other hand, those that are too large are not good for linguistic corpora, since they are often not representative of real language and tend to be lists, catalogues, spam and such things. We do in fact apply a length filter but, unlike most projects, it is not based on the size of the downloaded file. We reject documents the length of which after conversion to plain text is under 1,000 characters or over 100,000 characters.

Spam and porn filtering. The web is full of spam, porn and other kinds of noise. When we build a corpus out of web documents it is essential to get rid of these elements, but it is not always easy. The size filter proposed by Fletcher (2004) decreases this kind of noise but does not eliminate it completely. If we use search engines, we will most probably get less spam and porn, since they already do this filtering. But it is always desirable, and in the case of crawling methods necessary, to implement the detection of spam and porn. We do not apply any specific filter for spam and porn, because there is hardly any in Basque. People with commercial intentions target larger audiences and do not bother about minority languages spoken by communities that speak some other major language. Therefore, the language filter does the job perfectly.

Boilerplate removal. Web pages are full of "*boilerplate*", which is the linguistically uninteresting material that web server software automatically creates and which is repeated throughout every page in a website: headers, navigation menus, copyright notices, ads, etc. It is advisable to remove this boilerplate for various reasons: it makes ugly KWiCs, it distorts word frequencies and it makes the work of other filters (near-duplicate filtering, for example) more difficult. For boilerplate removal, we use Kimatu (Saralegi and Leturia, 2007), a language-independent system based on heuristics and features like tag density, punctuation signs, function words, etc. that scored second (74.3%) in the Cleaneval competition (Baroni et al., 2008).

Near-duplicate detection. The detection of exact duplicates is a straightforward task easily accomplished by hashing techniques. But much content is repeated across different websites (news from agencies in media sites, CC licensed articles in many blogs...) which are not exact duplicates, and these cannot be detected by hashing methods. The method most used for this job is Broder's algorithm (2000). We included a near-duplicate detection module based on Broder's shingling and fingerprinting algorithm.

Containment detection. It is very common for a web page containing an article with its own URL to be included in its entirety in the main page of its home newspaper or blog. Broder also implemented an algorithm for detection of already contained documents (1997). It is not as optimized as near-duplicate detection, but it is possible to use it for small- and medium-sized corpora building. In our downloading process, we included a containment detection method also based on Broder's works.

5 Evaluation

5.1 Corpora obtained by the search engine method

5.1.1 Effect of length of seed word list

As we have already stated, as a first experiment we tested and downloaded 5 different corpora using 5 different lengths of seed word list: 500, 1,000, 2,000, 5,000 and 10,000 words. In all of them, we used combinations of 3 words (as Sharoff suggested for languages with a smaller presence on the web and applied to Romanian), made 12,000 queries and downloaded the first 50 results of each query. The sizes obtained can be seen in table 1. We will now analyse various aspects of the corpora obtained.

Seed word list length	Documents	Words	Words per document	Different websites
500	49,387	81,508,628	1,650.41	4,452
1,000	83,941	105,374,227	1,255.34	4,849
2,000	83,147	119,474,991	1,436.91	4,675
5,000	52,913	129,342,982	2,444.45	4,398
10,000	25,350	85,271,975	3,363.79	3,021

TABLE 1 – Sizes of the collected corpora for each length of seed word list.

In the table we observe that the smaller the seed words list we use, the smaller the resulting corpora are, although the APIs return many more results. The reason for this is that the words in the smaller lists are more common and many pages contain them, but search engines will always be returning the same ones (the ones rated highest in their page rank) and the duplicate filters will remove them; if the words are more rare, fewer pages will contain them and they will not be repeated as much. Nevertheless, for the 10,000 list, the words are so rare that very few pages contain 3 of them and so the corpus obtained is smaller, and will probably be likewise for seed words lists above that.

For all these reasons, it can be concluded that, unlike for English or other languages, a 500-seed word list is not optimal for a language with a moderate presence on the web like Basque. Looking at the sizes, the optimal seed word list length seems to be 5,000, because that is the one that obtains the largest corpus. However, the type of documents from which the corpus has been built is something to take into account, which is shown in fig. 1. The 5,000 seed word corpus is the one containing more words from PDF documents, and PDFs are problematic: it is a visual format instead of an edition one (it does not contain the original continuous text, but rather the coordinates in the page of each line of text, word or even character), so original text extraction from them is never perfect and often very bad. PDF to text converters commit many errors when trying to rearrange the original paragraphs: two-column documents' lines are all messed up, header and footer text are repeated for every page and inserted into other paragraphs... As Fletcher (2007) points out, “PDF does not encode the logical formatting of the text (headings, paragraphs, captions etc.)” and “one problem that plagues all PDF to text converters persists: spaces are occasionally dropped or inserted between or within words”. Because of all of these, the 2,000 seed word corpus may be more appropriate (it is the one that has more words from HTMLs and second in total size), depending on our preference for size or quality.

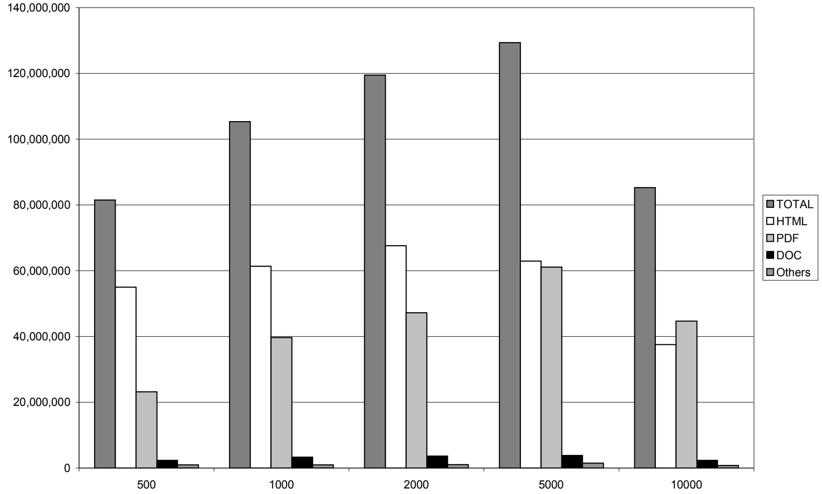

FIGURE 1 – Size in words of the corpora obtained for each seed word list length, total and by type of document.

Another clear difference in the corpora is the average size of the documents (table 1), which grows with the size of the seed word list (logically: if the words are rarer, they are more likely to be found in larger documents). Because corpora are used for linguistic research, the interesting documents for corpora are those that contain a reasonable amount of connected text (Sharoff, 2006). Although we apply the length filter in the collection process and all texts in the corpus have a minimum text, if we are interested in obtaining texts that are as long as possible, then we should opt for corpora obtained with longer seed word lists.

One more thing we have studied is the website variety of the corpora. It is usually interesting for a corpus to be from as many different sources as possible, to be able to analyse more diversity in the use of language; otherwise, style books of media, internal glossaries, etc. can lead to corpora that are too homogeneous. The number of different websites of each corpus is also shown in table 1. As can be seen there, in the last one the number of different websites falls drastically (again, it is logical if that corpus is composed of bigger documents and is smaller), but there is no significant difference among the rest.

Finally, there is one more point worth mentioning: using the search engines method, corpora do not grow continuously at a constant rate. Due to the page ranking these engines use, the same pages tend to appear over and again and are discarded by the duplicates detection, so the bigger the corpus is, the lower its growth rate gets, as the graph in fig. 2 shows. The growth rate is the number of new words obtained for each call to the search engine API, and is represented by the inclination of the curve in the graph.

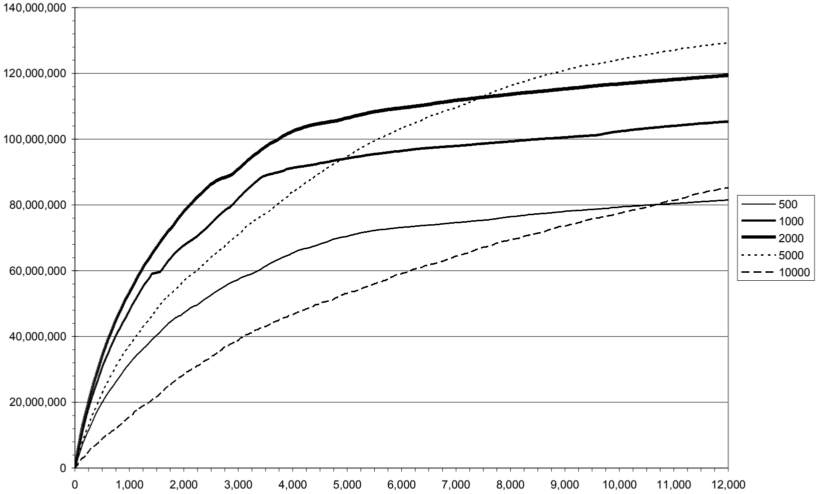

FIGURE 2 – Growth rate of the corpora obtained for each seed word list length.

So, it is not clear whether by using the search engines method we can build corpora as large as we would like to; and even if we could, it would be in a very unproductive way: while with the first 1,000 queries we obtain 37 million words on average (with a maximum 53 million words), in the last 1,000 queries we obtain less than 2 million words on average. And queries to the search engines are not an infinite resource: either they are paid services or have a maximum of calls per month.

5.1.2 Effect of length of combination sent to search engine

In the other experiment, we collected 5 corpora using 5 different search engine word combination lengths: 1, 2, 3, 4 and 5 words. For all of them, the rest of the parameters were the same: a seed word list of 2,000 words was used, 12,000 queries were made and the first 50 results of each query were downloaded. The details of the collected corpora are shown in table 2. Again, we will take a look at some features of these.

Combination length	Documents	Words	Words per document	Different websites
1	36,093	44,692,614	1,238.26	4,089
2	85,562	131,738,927	1,539.69	6,095
3	83,147	119,474,991	1,436.91	4,675
4	41,568	116,371,032	2,799.53	3,824
5	23,108	89,139,248	3,857.51	2,547

TABLE 2 – Sizes of the collected corpora for each combination length.

We can see that there is not a direct correlation between combination length and corpus size. If we send 1-word combinations of a 2,000 word seed list, there are only 2,000 different combinations, as the rest are repeated and do not return new results; therefore, we get the smallest corpus by far. With 2-word combinations we get the largest corpus, but from then on it gradually decreases again, because there will be fewer pages that have all the words. And in this case it is also the 2-word combination corpus that has the most words coming from HTML documents.

Regarding the document length, the same phenomenon as with the seed list length happens: for longer combinations, the size of the documents grows. But the website variety in this case falls for combinations longer than 2. And the dramatic fall in the growth rate also occurs in all these cases.

5.2 Corpus obtained by the crawling method

With the crawling method, and starting with the 1,500 seed URLs from the *Euskara* section of DMOZ, we have so far queued 39,163,290 links, tried to download 6,520,000 of them, successfully downloaded 3,472,166 pages (the rest were not available at the time, or had been discontinued, or gave errors) and included 168,991 out of them in the corpus. The rest were discarded because they were not in Basque (a high percentage of pages in Basque point to pages in other languages, mainly Spanish and English), or were in a format that could not be converted into text, or did not get through the filters (length, duplicate, etc.). The size in words of the downloaded corpus is 115 million. Its features can be seen in table 3.

Documents	Words	Words per document	PDFs	Different websites
168,991	114,565,240	677.94	1,276	5,060

TABLE 3 – Size of the corpora collected by crawling.

Only 1,276 documents come from PDFs, that is, only 0.75%. But the average document length is small, 678 words.

The website variety that could be obtained by the crawling method was one of our concerns. Starting from a set of seed URLs, there is a risk that they may not be enough or good enough, and that many websites are left out because they are not linked to in the initial pages or in the ones recursively linked by these. However, we can see that we have got a number of different websites larger than all but one of the search engines corpora, and compared to that single larger one it is proportionally not much smaller.

It is also interesting to take a look at the growth rate of the corpus (fig. 3). There is certainly a decrease in it, but not that pronounced: in the first million links followed, 23.3 million words were collected, whereas with the last million links we obtained 11.3 million words. It has gone down to 48.5% of the initial rate, while in the search engine corpora, this number is 5.4% on average. This proves that this method has the potential to collect a still much larger corpus.

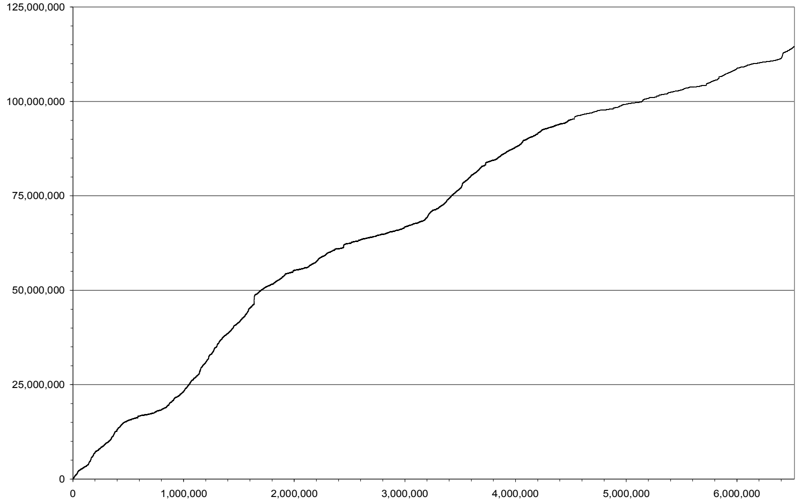

FIGURE 3 – Growth rate of the corpus obtained by the crawling method.

5.3 Qualitative analysis

In the previous subsections, the different corpora obtained were evaluated quantitatively (size, cost, etc.), but a more qualitative evaluation is necessary when corpora are involved, that is, an analysis according to linguistic criteria, because that is what corpora are used for.

5.3.1 Most characteristic words by LLR

There is no absolute method or measure to evaluate the linguistic quality of a corpus. Instead, what is usually done is to compare it with another by using the log-likelihood ratio or LLR association measure (Dunning, 1993) to identify the words that are more characteristic of one with regard to the other (Rayson and Garside, 2000); this is the method used both by Sharoff (2006) and Ferraresi et al. (2008) for evaluating the search engine method corpora and the ukWaC, respectively.

To carry out the evaluation, we chose the largest of the search engine method corpora, i.e. the one obtained with 2000 seed words and 2 combinations (we will call this corpus SE henceforth) and the crawling method corpus (CR henceforth). The number of URLs coinciding in both is only 6,815 (SE is made of 85,562 URLs and CR of 168,991).

Apart from comparing these two corpora with each other, we compared them with two reference corpora: XX. mendeko Euskararen Corpusa (a 4.6 million-word balanced corpus of twentieth century literary texts), henceforth XX, and Lexikoaren Behatokia (an 18.1 million-word corpus of 21st century media texts), henceforth LB.

Ferraresi et al. (2008) used nouns, adjectives and verbs for their analysis, but we also took adverbs and pronouns. We used the lemmas of words. Because in the case of XX we did not have access to the corpus but only to a list of lemma-frequencies and because it was lemmatized with a tagger different to the one we use, we had to discard proper nouns and numbers (the XX frequency list did not contain them) and make some adjustments (there were some deprecated lemmas that are now written in another way).

The most outstanding words of XX compared with any of the other three corpora can be put into three groups: religious words (*jaungoiko* and *jainko* “God”, *eliza* “church”, *apaiz* “priest”, *santu* “saint”, *otoitz* “prayer”, etc.), pronouns (*hura* “he”, *neu* and *ni* “me”, *zu* and *hi* “you”, *gu* “we”, etc.), and words that are scarcely used any more, either because they are now usually said another way, or because they were dialectal or incorrect forms of the times before the standardization, or because they are objects that do not exist or are no longer used (*gizaldi* “century”, *eroan* “to take”, *ipini* “to put”, *ezkero* “if”, *pezeta* old currency of Spain, etc.). The prominence of words from the first and third groups is easily understood in view of the difference in temporal deixis across XX and the other three corpora. The greater presence of words from the second group is a normal phenomenon in fiction and narrative texts compared with media and web texts, as Sharoff (2006) also confirmed.

The words characteristic of LB in comparison with any of the others can be divided into two groups: adverbs of time (*atzo* “yesterday”, *gaur* “today”, *herenegun* “the day before yesterday”, *iaz* “last year”, *bihar* “tomorrow”, etc.) and words from typical media sections such as sports (*talde* “team”, *partida* “match”, *jokatu* “to play”, etc.), politics (*presidente* “president”, *gobernu* “government”, *nazioarte* “international”, etc.), society (*atxilotu* “to arrest”, *auzitegi* “court”, *epaile* “judge”, etc.), culture (*film* “film”, *disko* “record”, *kontzertu* “concert”, etc.) or economy (*euro* “euro”, *krisi* “crisis”, *lan* “to work”, etc.). Both word groups are typical of media texts.

The web corpus we collected using search engines, SE, differs from the other three in words from the administrative domain (*prozedura* “procedure”, *lege* “law”, *artikulu* “article”, *administrazio* “administration”, *eranskin* “appendix”, *dekretu* “decree”, etc.) or the educational domain (*hezkuntza* “education”, *ikasle* “pupil”, *ikastetxe* “school”, *irakaskuntza* “teaching”, *irakasle* “teacher”, etc.). The prominence of administrative words is greater when compared with the CR corpus. The cause of this might lie in the fact that regional, provincial and local governments publish their official gazettes in PDF format; and, as we saw before, the SE corpus has a large proportion of PDFs, so these might be mostly of an administrative nature.

Finally, the words characteristic of the corpus obtained by crawling, CR, are words typical of web pages (*iruzkin* “comment”, *orri* “page”, *sare* “net”, *erabiltzaile* “user”, *web* “web”, *blog* “blog”, *erantzun* “to comment”, *internet* “Internet”, *lizentzia* “license”, *software* “software”, etc.) or of media websites (*albiste* “news”, *argazki* “photo”, *bideo* “video”, *emisio* “broadcast”, *kanal* “channel”, *telebista* “TV”, etc.), month and weekday names (*azaro* “November”, *urri* “October”, *igande* “Sunday”, *astearte* “Tuesday”, etc.) or words from the cultural domain (*dantza* “dance”, *euskara* “Basque language”, *kultura* “culture”, *ikastaro* “course”, *antzoki* “theatre”, etc.). Except for the last, all the groups of words are common in web pages, so we can say that the main feature of this corpus is that it is mostly composed of genuine web pages.

5.3.2 Number of distinct and 'useful' words

Baroni et al. (2009) compared ukWaC and itWaC with reference corpora in each of those languages (the BNC and la Repubblica corpus) looking at three parameters: the number of distinct words in a corpus, the coverage of a corpus within another and the enrichment a corpus gives to another. We have done the same with the four corpora analysed in the previous subsection. We counted the lemmas of all types of words, except proper nouns and numbers (because of the reasons already explained).

Just as in the aforementioned work by Baroni et al., we show the number of distinct words in terms of absolute numbers and of words that occur at least 20 times. This frequency threshold was chosen by them as a rough way of estimating the number of 'useful' words in a corpus, following Sinclair's (2005) claim that at least 20 occurrences of a word are usually needed for an experienced lexicographer to describe its behaviour, and taking into account that low frequency words will not be of any use in NLP applications either. Although admittedly arbitrary, we also used the 'Sinclair cutoff'. The number of distinct words that each corpus has is shown in table 4.

Corpus	Total words	Words f ≥ 20
XX	53,993	9,147
LB	36,311	12,922
SE	74,132	33,056
CR	64,424	27,238

TABLE 4 – Number of distinct words in each corpus.

As we can see, the number of total and 'useful' words is much greater in the web corpora; this is logical due to their much greater size. However, the high number of total words of the XX corpus is striking: it has almost as many words as the web corpora (which are more than 20 times larger) and much more than the LB corpus (which is almost 4 times bigger). This is due to the fact that a considerable part of the XX corpus is made up of texts from before the standardization of the Basque language and it contains many obsolete, outdated, out-of-use or non-standard words that were tagged manually but which Basque taggers do not usually recognize.

5.3.3 Coverage and enrichment

In order to prove that those 'useful' words attested in the web corpora are the sort of words linguists and lexicographers would be typically interested in, rather than, say, web-related terms of limited general interest, Baroni et al. looked at two measures of overlap, namely coverage and enrichment. The coverage of a corpus in another one is the proportion of words that are above the Sinclair cutoff in both over the total words above this threshold in the first corpus; it can be considered as a rough measure of the extent to which the first corpus is “substitutable” by the second, because it gives an idea of how many of its useful words are also present in the other. The enrichment of a corpus in another one is defined as the proportion of words that are above the Sinclair cutoff in the second corpus but below it in the first, over the total words below the threshold in the first one (to avoid noise in the form of typos or loanwords, only words with at least 10 occurrences are considered); this gives a rough idea of the number of words for which the first does not have enough information, but the second does. We have also calculated these measures, and the statistics obtained are reported in table 5.

Corpora type	Corpora	Coverage	Enrichment	Corpora	Coverage	Enrichment
Classical	XX / LB	57.14%	14.36%	LB / XX	80.71%	38.36%
Classical / Web	XX / SE	26.13%	0.48%	SE / XX	94.44%	83.67%
	XX / CR	31.43%	1.01%	CR / XX	93.59%	77.43%
	LB / SE	36.74%	0.81%	SE / LB	93.99%	83.85%
	LB / CR	44.21%	1.48%	CR / LB	93.19%	79.11%
Web	SE / CR	95.80%	56.24%	CR / SE	78.94%	9.52%

TABLE 5 – Coverage and enrichment of each corpus with regard to each of the others.

It shows that the web corpora cover high above 90% of the classical corpora with an enrichment over them of around 80%, whereas the coverage of the classical corpora over the web ones is normally below 40% and their enrichment is always below 2%; these data are similar to the ones obtained in the aforementioned research by Baroni et al. with ukWaC/BNC and itWaC/La Repubblica.

However, the comparison between the two web corpora, SE and CR, offers surprising results. Although they are of almost equal size, we have seen in the previous subsection that SE contains many more distinct and 'useful' words than CR, and the coverage and enrichment are not symmetrical: CR is almost completely covered by SE (95.80%), but in the other direction this number is only 78.94%; and SE enriches CR by 56.24%, whereas CR only contributes to SE with 9.52% of new words. It looks as if, for equal sizes, the search engine method obtains more linguistically varied corpora than the crawling method. Nevertheless, we have shown that the crawling method can collect much larger corpora, so this deficiency will supposedly be corrected if we continue crawling and enlarging the corpus.

Conclusions

We have proven that both crawling and using search engines are valid methods for obtaining BNC-sized corpora for Basque. With the search engines method, using 2,000 or 5,000 seed words we obtained the largest corpora: the former obtains greater website variety, the latter obtains more PDFs (usually problematic) and larger documents (more connected text). The optimal word-combination length to send to the APIs seems to be 2, because it obtains the largest and most varied corpus with the least number of PDFs. However, if more than 100-150 million words are needed, crawling is the way to go: we have collected a corpus of a size and website variety comparable with those obtained via search engines, with much fewer PDFs and the potential to get much bigger. This corpus is now 115 million words big, but we expect to make it much larger in the near future.

When compared with classical corpora, these web corpora differ in that the search engine ones contain more administrative texts (most probably due to the PDFs of official gazettes) and the crawling one more web-domain texts. Since almost all of the words in the classical corpora are present in the web ones, whilst they provide many new words, we can conclude that collecting large corpora from the web can make a great contribution to Basque corpus building, linguistics and the language in general.

Acknowledgments

This research has been partially funded by the Industry Department of the Regional Government of the Basque Autonomous Community in its Saiotek 2011 call (ref. SA-2011/00216).

References

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N. and Sologaistoa A. (2007). ZT corpus: Annotation and Tools for Basque Corpora. In *Proceedings of Corpus Linguistics 2007*, Birmingham, U.K.

Aston, G. and Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh, U.K.

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. In *Proceedings of the 11th Conferenceof the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 87-90), Trento, Italy.

Baroni, M. and Ueyama, M. (2006). Building general- and special purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium* (pp. 31-40), Tokyo, Japan.

Baroni, M., Chantree, F., Kilgarriff, A. and Sharoff, S. (2008). Cleaneval: a competition for cleaning web pages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation Journal*, 43(3): 209-226.

Broder, A. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997* (pp. 21-29), Salerno, Italy.

Broder, A. (2000). Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium* (pp. 1-10), Montreal, Canada.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61-74.

Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (2008) Introducing and evaluating ukWaC, a very large Web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC4)* (pp. 47-54), Marrakech, Morocco.

Finn, A., Kushmerick, N. and Smyth, B. (2001). Fact or fiction: Content classification for digital libraries. In *Proceedings of Personalisation and Recommender Systems in Digital Libraries Workshop*, Dublin, Ireland.

Fletcher, W. H. (2006). Concordancing the Web: Promise and Problems, Tools and Techniques. In Hundt, M., Nesselhauf, N. and Biewer, C. (Eds.), *Corpus Linguistics and the Web* (pp. 25–46). Amsterdam, The Netherlands: Rodopi.

Fletcher, W. H. (2007). Implementing a BNC-compare-able web corpus. In Fairon, C., Naets, H., Kilgarriff, A. and De Schryver G.-M. (Eds.), *Building and exploring web corpora* (pp. 43–56). Louvain-la-Neuve, Belgium: Cahiers du Cental.

Ghani, R., Jones, R. and Mladenić, D. (2003). Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems*, 7(1): 56-83.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. *Computational Linguistics*, 29(3): 333-347.

Leturia, I., Gurrutxaga, A., Alegria, I. and Ezeiza, A. (2007). Corpeus, a 'web as corpus' tool designed for the agglutinative nature of basque. In *Proceedings of the 3rd Web as Corpus Workshop (WAC3)* (pp. 69–81), Louvain-la-Neuve, Belgium.

Leturia, I., San Vicente, I., Saralegi, X. and Lopez de Lacalle, M. (2008). Collecting basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th Web as Corpus Workshop (WAC4)* (pp. 40–46), Marrakech, Morocco.

Leturia, I., Gurrutxaga, A., Areta, N., Pociello, E. (2008). Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Leturia, I., San Vicente, I. and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the internet. In *Proceedings of the 5th International Web as Corpus Workshop (WAC5)* (pp. 53–61), Donostia/San Sebastian, Spain.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of Workshop on Comparing Corpora of ACL 2000* (pp. 1–6), Hong Kong, China.

Renouf, A., Kehoe, A. and Banerjee, J. (2006). WebCorp: an Integrated System for WebText Search. In Hundt, M., Nesselhauf, N. and Biewer, C. (Eds.), *Corpus Linguistics and the Web* (pp. 47–67). Amsterdam, The Netherlands: Rodopi.

San Vicente, I. and Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Saralegi, X. and Leturia, I. (2007). Kimatu, a tool for cleaning non-content text parts from html docs. In *Proceedings of the 3rd Web as Corpus Workshop (WAC3)* (pp. 163–167), Louvain-la-Neuve, Belgium.

Sharoff, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Queries. In Baroni, M. and Bernardini, S. (Eds.), *WaCky! Working Papers on the Web as Corpus* (pp. 63-98). Bologna, Italy: Gedit Edizioni.

Sinclair, J. McH. (2005). Corpus and text – Basic principles. In: Wynne , M. (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford, U.K.: Oxbow Books.