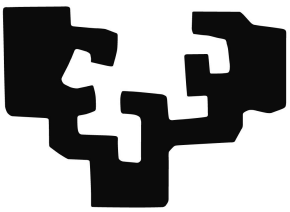
University of the Basque Country

eman ta zabal zazu



Faculty of Computer Science

Department of Computer Languages and Systems

Dr. Xabier Arregi / Dr. Kepa Sarasola

PhD Thesis

The Web as a Corpus of Basque

Igor Leturia

Donostia / San Sebastian, April 2014

To Mum, Dad, and all my family,
and especially to Alaitz, Ximun, and Hur

I don't think there's one word that can describe a man's life

Charles Foster Kane, *Citizen Kane* (1941)

Acknowledgements

There are many people and institutions to thank for their help and collaboration in the writing up of this thesis, without whom it would not have been possible. I will try not to leave anyone out...

In the first place, I would like to thank the Elhuyar Foundation, the organization I work at, for its work to promote the Basque language and, specifically, the resources it devotes to the development of language resources and technologies for Basque. It is within the framework of the department of R&D into Language Technologies that the research described in this Ph.D. thesis has been carried out. With respect to the management of the Foundation, Josu Aztiria deserves a special mention for his belief in and support of our department's work.

All my colleagues from the department of R&D into Language Technologies are also to be thanked. More than just helping me, you have collaborated greatly in the writing up of the thesis, as borne out by the including of your names among the authors of the various papers produced throughout the thesis. Antton, Xabi, Iñaki, Iker, Maddalen, Eli, Nerea... You are all great guys to work with; for me, you are friends more than just colleagues.

Among my colleagues, special thanks go to Antton Gurrutxaga. I started my research career under his guidance, in a department and a project both of which he was the manager, and he was always a great leader and tutor. The idea of initiating the Web-as-Corpus line of research within Elhuyar was his, and he put his trust in me to take responsibility for leading it, which has eventually lead to the production of this thesis. And he has always helped me greatly all along the way.

There are also others who have helped in the writing up of the thesis and whom I would like to thank as well: my friend and co-worker at Elhuyar Dr. Guillermo *Willy* Roa, for his advice and help with my questions, and the translator Sarah J. Turtle, who was my English teacher in my youth, for her corrections and advice in both the thesis itself and in the previously published papers upon which the thesis is based.

I cannot close without mentioning the IXA Group of the University of the Basque Country. They were the pioneers in language technologies for Basque back in the eighties, and it was through them that I discovered this exciting area in which I now work. It is difficult to envisage where the Basque language would be today if they had not set off down the road and done all the work they have done throughout these years.

Among its members, special thanks go to the supervisors of the thesis, Dr. Xabier Arregi and Dr. Kepa Sarasola, for their help and guidance in the conducting and writing up of the thesis.

Thanks also to my friends and, especially, to my relatives and family, for all their interest, support, and blind belief. And Mum and Dad deserve a special place among them, for obvious reasons: I would not be what I am today if it were not for them –in fact, I would not even be!

And finally, last but not least (quite the contrary, as I have left the most important ones for the end), to Alaitz, Ximun, and Hur: thank you for all your love, support, and comprehension, and sorry for the time I have had to spend on the thesis which I could and should have spent with you. You cannot know how much I love you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbr. / Symbol	Meaning
ACL	Association for Computational Linguistics
API	Application Programming Interface
BDST	Basic Dictionary of Science and Technology
BiWeC	Big Web Corpus
BNC	British National Corpus
BootCaT	Bootstrapping Corpora and Terms
BTE	Body Text Extraction
CC	Creative Commons
CLEF	Conference and Labs of the Evaluation Forum
CMS	Content Management System
COW	COrpora from the Web
CRC	CRawling Corpus
DBF	DBase File
deWaC	Web corpus of German collected by the WaCky! initiative
DISC	Disconnected
EDBL	Euskararen Datu-Base Lexikala (Lexical Database of Basque)
EPEC	Euskararen Prozesamendurako Erreferentziazko Corpusa (Reference Corpus for the Processing of Basque)
EPG	Ereduzko Prosa Gaur (Model Prose Today)
ETC	Egungo Testuen Corpusa
EU	European Union
FCG	Frequent Case Generation
frWaC	Web corpus of French collected by the WaCky! initiative
GNU	GNU's Not Unix!
HTML	HyperText Markup Language
ICANN	Internet Corporation for Assigned Names and Numbers
ID	In-Degree
IR	Information Retrieval
itWaC	Web corpus of Italian collected by the WaCky! initiative
KB	KiloByte
KG	Klasikoen Gordailua (Classics Store)
KWiC	KeyWord in Context
LBC	Lexikoaren Behatokiko Corpusa (Corpus of the Observatory of the Lexicon)
LCSR	Longest Common Subsequence Ratio
LLR	Log-Likelihood Ratio
MI	Mutual Information
NLP	Natural Language Processing
OCR	Optical Character Recognition

Abbr. / Symbol	Meaning
OD	Out-Degree
ODP	Open Directory Project
OEHTC	Orotariko Euskal Hiztegiaren Testu-Corpusa (Text Corpus of the General Dictionary of Basque)
OOV	Out Of Vocabulary
NTLD	National Top-Level Domains
PDF	Portable Document Format
POS	Part Of Speech
RDF	Resource Description Framework
RFR	Relative Frequency Ratio
RRR	Relative Rank Ratio
RSS	Really Simple Syndication
RTF	Rich Text Format
SCC	Strongly Connected Component
SEC	Search Engine Corpus
SIG	Special Interest Group
SIGWAC	Special Interest Group on the Web As Corpus
TLD	Top-Level Domain
TM	Text and Markup scoring system
TO	Text Only scoring system
TREC	Text REtrieval Conference
ukWaC	Web corpus of British English collected by the WaCky! initiative
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WaC	Web as Corpus
WaCky!	Web as Corpus kool ynitiative
WHATWG	Web Hypertext Application Technology Working Group
WWW	World Wide Web
X^2	Chi Square
XHTML	EXtensible HyperText Markup Language
XML	EXtensible Markup Language
XXMEC	XX. mendeko Euskararen Corpusa (Corpus of Basque of the 20th Century)
ZTC	Zientzia eta Teknologiaren Corpusa (Corpus of Science and Technology)

1 Introduction

1.1 The ever-growing importance of corpora

Until a couple of decades ago, linguistics-related works such as lexicography, terminology or linguistic research was based mostly on experts' knowledge or intuition. Since there was no fast method to look up usage examples in literature or oral use, it was the brain of these language experts that played the role of storage device for linguistic evidence. Although less trustworthy and smaller in capacity than books or popular knowledge –upon which, of course, these experts relied and which they occasionally queried–, it was the only "device" that could be queried in a reasonable amount of time.

The advent (or, more precisely, the popularization) of computers brought about the possibility of storing and querying large amounts of textual content in a fast and reliable way. However, these same computers have eased the production of text and universalized it. The quantity of written texts is nowadays orders of magnitude larger than in the pre-computer era, and it goes on growing at an ever faster rate. Therefore, it is still not possible for linguistic researchers to have the whole written production of a language at their disposal; we still have to content ourselves with being able to query only a small sample of the whole. Nevertheless, the samples available are much larger, much more reliable, and much more easily and quickly queried than before the computer age.

These samples of written language production are called text **corpora**, and the discipline of doing linguistic work based on the evidence provided by these corpora is called **corpus linguistics**. Linguistic research and other language-related tasks (like lexicography or terminology) are currently done based upon the data obtained from corpora and not upon (or at least not exclusively) experts' intuition, just as in any scientific research. There are various types of corpora, depending on their intended use: general or specialized; monolingual or multilingual; in the latter case, comparable or parallel; etc. More corpora of all kinds are made available every day, and always larger than the previously existing ones.

Due to their size, corpora would be unmanageable if they were not used in conjunction with some corpus querying tool. These usually include many options, for example they can allow one to: query a word (or more than one, even specifying distances

between them) in terms of its surface form, lemma, POS, beginning, with regular expressions, etc.; show the occurrences in a **KWiC** (**Key Word in Context**) form, that is, each example of use in one line with some context around it; sort the results following different criteria (by document, alphabetically, by the context...); show statistics of collocations, n-grams, etc.

The use of corpora is not circumscribed to linguists' work. Dictionaries are very important common-use tools that depend on corpora for being created following modern quality standards because, as we have said, they are produced nowadays on the basis of empirical evidence, or previous use at least is studied, and these are both provided by corpora; and besides, the process of dictionary making can be facilitated and speeded up by the semi-automatic extraction of words, examples, collocations or terminology from corpora using NLP methods. Also language technologies –which are ever more present in everyday life through the web and our gadgets– such as machine translation or cross-lingual information retrieval need electronic dictionaries and corpora in order to be developed. They are also used in sociolinguistic and cultural research (Stubbs, 1996), training of translators (Zanettin et al., 2003), language learning, etc. And finally, they are useful for writers, journalists, translators or anyone who may be writing something and who cannot find a certain word in a dictionary.

Proof of the growing importance of corpora is that the size of general corpora for English periodically grows one order of magnitude, in a similar way to the capacity of computers (processing speed, storage...) doubling every year and a half following Moore's Law (Moore, 1965). The periodicity might not be as exact or small, but the growth is exponential nonetheless: starting at the 1 million-word **Brown Corpus** (Kučera and Francis, 1967) we have gone through the 100 million-word **BNC** or **British National Corpus** (Aston and Burnard, 1998) and arrived to the size of 70 billion-words (Pomikálek et al., 2012).

For all the reasons mentioned above, it is clear that any modern language aiming to be used normally in the media, in education, etc. needs to have corpora at its disposal, because they are a very valuable resource for many aspects of the development of a language.

1.2 Corpus typology

The term **corpus** has been defined in many ways. McEnery and Wilson (2001) stablish four requisites for a collection of texts for it to be considered a corpus: it has to be a representative sample of the object of analysis; it must have a finite and known size; it has to be in machine-readable form; and it has to be a standard reference. Bach et al. (1997) ask for four more conditions: it has to be a large set of real language samples; it has to be collected following some criteria; it has to be stored in electronic format; and it has to be enriched with linguistic information. Biber et al. (1999), on the other hand, require only that it be a text collection compiled following certain predefined criteria. And Kilgarriff and Grefenstette (2003) make an even more open definition, describing it as just a collection of texts and stating that any other requisite is not obligatory.

Although this last definition would also include text collections in paper format, the term is normally used to name only texts in electronic format; as noted above, it is the processing and storage capacity that computers brought that made possible the birth of modern corpus linguistics.

When making a classification of corpora, different criteria can be used. One of these criteria can be the format the corpus is in, and according to this we can distinguish between corpora in paper format and electronic corpora. Another one can be the source of the texts, and taking this criterion into account, corpora can be classified into those scanned from paper, those obtained from writers or publishers, those collected from the web, those transcribed from speech recordings, those involving a combination, etc.

These criteria are somehow related to the way a corpus is constructed. But in practical terms, criteria related to the use that a corpus might have are more interesting, such as domain, genre, representativeness, number of languages, alignment level, etc.

If we consider the domain(s) and/or genre(s) of a corpus, one of the possible types are **general corpora**, that is, those that contain texts in any domain and/or genre. The objective of these corpora is to be of use in the analysis of and be representative of the use of general language. They are also called **reference corpora** (Sinclair, 1996; Leech, 2002). Examples of this kind of corpora are the already mentioned Brown Corpus (Kučera and Francis, 1967) and the BNC (Aston and Burnard, 1998), ukWaC (Ferraresi et al., 2008), etc.

Depending on the domain, a corpus can also be **specialized**. This would be a corpus that contains texts on a specialized domain, *e.g.*, biology, literature... and that has been built to analyse the language used in that domain (Sinclair, 1996). One corpus of this kind is the Medicor (Vihla, 1998), a corpus of medical texts written in American English. When a corpus is general because it contains texts on different domains, but when every text is marked with the domain it belongs to and the corpus allows querying only the texts of a domain, then it can also be considered a specialized corpus –or, more precisely, a collection of specialized corpora. Terminology and automatic terminology extraction are tasks in which a corpus of this type can be useful, as in (Daille, 1995; Smadja, 1993).

If we look at the genre, apart from a general one, a corpus can also be **genre-specific**. For example, there are corpora which consist only of media texts, and others that only contain literary texts. Just as with specialized corpora, if all texts in a general corpus are classified according to their genre, then in practical terms the corpus consists of various genre-specific corpora.

The intended representativeness of a corpus is another parameter for classifying corpora, which divides them into: a) **balanced corpora**, where texts have been chosen at random from the universe to be studied and the corpus obtained has the same features and distribution as the universe; this kind of corpus is the most appropriate for making real objective studies, but are also the most difficult to construct; b) **model corpora**, which consist of texts that the author of the corpus consider as model or exemplary (which is a bit contrary to the objective of a corpus: it does not describe real use but more exactly prescribes how language is to be used); and c) **opportunistic corpora**, where the only building criterion has been to obtain everything that was easily available.

Corpora can also be **monolingual** or **multilingual**. Monolingual ones contain only texts in one language, and they can be used in lexicography to build monolingual dictionaries, in language standardization, etc. Multilingual corpora, on the other hand, contain texts in more than one language, so they can be used in translator training, in multilingual dictionary making, in training machine translation systems, etc.

Among multilingual corpora there are various subtypes, one of which is **parallel corpora**. In these, all the texts that make up the sub-corpora of each language are mutual translations, and they are usually aligned at the sentence-level. That is, a parallel

corpus consists of a collection of sentences, each with its translation into another language or languages. They are usually extracted from software to manage and use translation memories, which are very popular among translators. Examples of this kind are the JRC-Acquis corpus (Steinberger et al., 2006), consisting of 4 million sentences from EU laws in 22 languages, and the Europarl corpus (Koehn, 2005), made up of 28 million words of European Parliament minutes in 11 languages. They can be used for multilingual terminology extraction (Kraif, 2002; Tiedemann, 2003), statistical machine translation (Koehn, 2005), etc.

Another type of multilingual corpora are **comparable corpora**. In these, the texts that comprise the sub-corpora of each language share some features: they can be of the same domain, genre, timespan, etc. Although their alignment level is smaller than in parallel corpora and thus they have less explicit information to extract knowledge (*e.g.*, bilingual terminology) from them, they are easier to obtain in large quantities than parallel corpora and can be used in similar tasks with similar results if larger corpora are used, as research in fields such as machine translation (Munteanu and Marcu, 2005), bilingual terminology extraction (Fung and Yee, 1998; Rapp, 1999) or cross-language information retrieval (Talvensaari et al., 2007) has shown. That is why comparable corpora are becoming increasingly popular. They are also interesting, for example, for examining how things are said in different languages without the translation bias; Manca (2008) analysed bilingual phraseology in the tourism domain using comparable corpora instead of parallel ones to avoid this bias.

1.3 Basque corpora

The Basque language, if it aims to survive in the future and be used in everyday life, also needs corpora for its development, just like any other language. Probably even more so than many others, for various reasons.

It must be taken into account that the standardization of Basque did not start until the late sixties of the last century and is still ongoing, and that many rules, words, and spellings have been changing since. Furthermore, Basque was not taught in schools until the seventies and did not become a medium of instruction at universities nearly until 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. There are also many areas or words upon which a decision as to which form or spelling should be used has not yet been taken. And Basque still lacks terminological

dictionaries for many domains and its dictionaries are not as many or as large as would be desirable. 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. Finally, not having as many research institutions or as much economical resources devoted to their development, language technologies for Basque are not as advanced as they might be for many other languages (Hernáez et al., 2012). And corpora are needed if all those tasks (standardization, dictionary making, and language-technology development) are to be accomplished.

Yet the corpora available in Basque are not as abundant, varied or large as would be necessary, due again to the lack of all kinds of resources (economic, human, etc.) smaller languages suffer from, and building a corpus in the classical way, *i.e.*, out of printed texts, is normally a very costly process. That is why the number and size of Basque corpora is proportional to the number of speakers and the economic resources of the Basque language.

These are the only general corpora in Basque available for public use:

- Orotariko Euskal Hiztegiaren Testu-Corpusa (Text Corpus of the General Dictionary of Basque) or OEHTC (Euskaltzaindia, 1984): 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 (Corpus of Basque of the 20th Century) or XXMEC (Euskaltzaindia, 2002): a 4.6 million-word balanced corpus produced by Euskaltzaindia, which consists mainly of twentieth century literary texts.
- Ereduzko Prosa Gaur (Model Prose Today) or EPG (University of the Basque Country, 2006): 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.
- Klasikoen Gordailua (Classics Store) or KG (Susa, 2005): a non-tagged 11.9 million-word corpus compiled by the publishing house Susa, consisting of classical texts.

- Euskararen Prozesamendurako Erreferentziazko Corpusa (Reference Corpus for the Processing of Basque) or EPEC (Aduriz et al., 2006): a 300,000-word corpus tagged at the morphological, syntactic, and semantic levels and manually disambiguated.
- Lexikoaren Behatokiko Corpusa (Corpus of the Observatory of the Lexicon) or LBC (Euskaltzaindia, 2009): a 26.5 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.

Regarding specialized corpora of Basque, there is only one that we are aware of, namely:

- Zientzia eta Teknologiaren Corpusa (Corpus of Science and Technology) or ZTC (Areta et al., 2007): an 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.

There are also a few parallel corpora, coming from the translation memories that diverse organizations, institutions or public bodies have made available for public use:

- EIZIE's translation memories (EIZIE, 2002): a translation memory made of 87,000 Spanish-Basque sentences from the Association of Translators, Correctors, and Interpreters of Basque Language.
- Translation memories from the Provincial Council of Gipuzkoa (Provincial Council of Gipuzkoa, 2011): 520,000 sentences (mostly Spanish-Basque) from the Provincial Council of Gipuzkoa.
- Consumer Corpus (Eroski Foundation, 2010): a parallel corpus of the retail domain in four languages (Spanish, Basque, Galician, and Catalan), with 263,000 sentences in the Spanish-Basque pair, made available by the Eroski Foundation.
- Translation memories from the Provincial Council of Biscay (Provincial Council of Biscay, 2011): Spanish-Basque translation memory from the Provincial Council of Biscay.

- Translation memories of the Basque Government's Official Translation Service (Basque Government, 2010): Spanish-Basque translation memory from the Basque Government.

This information has been taken from (Areta et al., 2008) and afterwards updated. As we can see, there are no more than a dozen corpora in Basque, and they are generally small in comparison with those in other major languages, not specialized, and not updated or enlarged; thus, their usefulness for detecting the most recently incorporated words, terms, and neologisms is severely limited.

1.4 The Web-as-Corpus approach

The enormous difference in size between the corpora in Basque and in, say, English is not explained solely by the difference in resources (human, economic...) devoted to each. If English and other languages have attained corpora in the order of a billion-words, it is thanks to a new approach that emerged a few years ago: the **Web-as-Corpus** approach, based on using the web as a source of linguistic evidence. The term was probably coined by Kilgarriff in his 2001 paper entitled *Web as corpus* (Kilgarriff, 2001), in which he made one of the first apologies for using the web for linguistic purposes and sparked off a whole new discipline.

This approach offers many advantages. One of them is that immense quantities of text have been put on the web in recent years, resulting in an unaccountably huge amount of them, and very large corpora can be built from the web, much larger than by traditional methods. And it is widely accepted that, regarding corpora and NLP, "more data is better data": Banko and Brill (2001) proved that a simple disambiguation algorithm trained over a very large corpus obtained better results than a more sophisticated algorithm trained over a cleaner but smaller corpus. And Keller and Lapata (2003) found evidences of bigrams on the web that could not be found in any corpus.

Another advantage is the format of the texts of the web. Building corpora has traditionally been a very laborious and costly process. Not so long ago almost every corpus was built by means of digitalizing paper books. This involved the tiring process of typing them or, more recently, scanning, OCRing, and correcting them. Later on, texts were usually obtained directly in electronic format, but the process was not made much easier: authors and/or publishers still had to be contacted, texts still had to be converted from the proprietary and secret (Word, Quark...) or presentation- and not

edition-oriented (PDF...) formats they were made available in to a standard and single format... The World Wide Web or WWW, on the other hand, provides a huge number of texts, publicly and openly available to the public, in a standard and easy-to-handle format (HTML).

That the web is constantly being updated and enlarged is one more point in its favour. Due to the work needed to build a corpus the traditional way, so far most of the corpora have been static, *i.e.*, they were built and finished once and were left that way forever. And even in the few cases where a corpus was continuously updated, a significant period of time usually elapses between the production of a text to its availability through a corpus (Baroni and Ueyama, 2006). This renders a considerable number of corpora unsuitable for many types of tasks, like analysing recent linguistic phenomena, building a terminological dictionary on a new topic, etc. The web, on the contrary, is constantly being updated (Fetterly et al., 2004) and is therefore appropriate for such tasks, or for quickly building a corpus upon which to develop such tasks.

Finally, practically any language, way of speaking or domain is present on the web. So it is probably the cheapest and, therefore, most appropriate way of building corpora for many less-resourced languages (maybe even the only one for some of them). Scannell (2007), for example, built corpora for 416 languages from the web, and Ghani, Jones, and Mladenić (2003) also used the web for building corpora for minority languages. Besides, many specialized domains have hardly any presence in general corpora, so the web might be the only source of evidence for analysing the language, terms, etc. of those domains.

The Web-as-Corpus approach has its detractors with their objections, too. There are many that say the web cannot be considered a corpus because it does not comply with its definition –at least not with its most restrictive definitions mentioned above, that is, McEnery and Wilson's (2001) or Bach's (1997) and others'. Sinclair (2005), for example, says that its dimensions are unknown and constantly changing, and that it has not been designed from a linguistic perspective. Others add that it is not representative of real language –Thelwall (2005), for example, claimed that the language of young people with above average computer skills was overrepresented and normal written and spoken language under-represented–, or that it is not linguistically tagged, etc.

However, these objections apply to the direct using of the web as a corpus; a classical corpus that would not have any of these disadvantages can be built using web pages as a source, as Ferraresi (2007) noted.

Yet another major objection to the web is the lack of quality of its texts (Thelwall et al., 2003). Many of them (blogs, wikis, fora, etc.) are non-revised texts written by non-experts or non-professionals of language, as opposed to texts from books or the media which have been revised and/or written by professional writers or journalists. But the web is not only that: books and texts from the media that have been edited are also present on the web. Besides, the spontaneous production of texts from normal people on the web is currently an undeniable linguistic phenomenon, and the only source to study it is the web itself. For this reason, some people –*e.g.*, Schäfer and Bildhauer (2013)– complain about the exact opposite: to them, traditional corpora are "sometimes too close to the respective standard language […] for certain types of research questions."

Kilgarriff and Grefenstette (2003) object to all these objections and defend the Web-as-Corpus approach. They categorically affirm that the web is a corpus and that it is as representative (or as unrepresentative) as any other corpus, total representativeness being impossible. Baroni and Ueyama (2006) add that written language is always overrepresented in traditional corpora regarding real language use –not so many people write, whereas everyone speaks– and that the web is getting more people to write –making web corpora more representative of real use of language.

Although not in the form of objections, some authors aptly note some disadvantages of the approach:

- Schäfer and Bildhauer (2013) and Ferraresi (2007) warn about the huge amount of noise in the web (that is, linguistically non-interesting content such as spam, boilerplate or duplicate content, to which we will refer in detail later); since the web texts are collected or treated automatically by software tools, even if they try to filter the unwanted texts, these tools are never perfect, and some of this noise will make it to the corpus.
- Schäfer and Bildhauer (*ibid.*) and Fletcher (2011) point out the copyright issues: whereas the terms of use and distribution of classical corpora are pre-agreed with the text providers (media, publishers...), among texts downloaded from the web, although freely accessible, many are copy-

righted, and what one can do with them is unclear since it depends on the law of the country of origin and other complications; it is not easy to know even the country of origin of a web text or if it is under copyright!

- Fletcher (*ibid.*) mentions the difficulty of knowing the sources or other information about a web page: indeed, in classical corpora one can know the authors, creation date, and other things about a text, which is important for linguistic purposes in order to know how much credit to give to a piece of evidence; but with texts from the web only the website from which it was downloaded can be known for certain.

However, the three drawbacks we have just mentioned might cease to exist in the future, because the World Wide Web Consortium or W3C (the organization that rules and defines the standards of the web), in its new version of the HTML standard, called HTML 5 (W3C, 2013), has included semantic tags and attributes that will enable real content to be distinguished from boilerplate and the authors and licenses of web pages to be known.

Advantages, disadvantages, objections, and defences aside, the fact that the web is being more and more used for linguistic research or as a source for texts for building corpora is an undeniable reality. As proof of this, the ACL (Association for Computational Linguistics) has a SIG or Special Interest Group on the Web as Corpus, SIGWAC, which has been organizing an annual international workshop on the subject of the Web as Corpus since 2005, where many works using this approach have been presented throughout these years.

1.5 The web as a corpus of Basque

Throughout this introduction we have pointed out some facts, which we will sum up here:

- There are many types of corpora, and the more of each kind there are and the bigger they are, so much the better. They are necessary for linguistic research, language learning, translations, dictionary making, development of language technologies, etc.
- The Basque language, due to its situation (minority language, still in standardization process, language technologies not as developed...) needs corpora at least as much as any other language, if not more.

- The corpora in Basque are generally few, small, and not up to date compared with those of other major languages.
- Many major languages currently have corpora of the size of billions of words, or corpora on many specialized areas, big multilingual corpora, etc.
- This has been possible thanks to the Web-as-Corpus approach. Using the web as a source is a fast and cheap way to build corpora.
- The Basque language does not have many resources, neither human nor economic, to build corpora.

The conclusion that can be drawn from here is evident: if other languages have been able to build varied and huge corpora thanks to the Web-as-Corpus approach because of its low cost, and if Basque is a language with few resources that is in great need of corpora, then we must make use of the Web-as-Corpus approach to build corpora for the Basque language.

As obvious as it may seem, it is not so clear that this approach would be a successful one for many reasons. For one, the Basque web is clearly not nearly as big as that of the major languages we have mentioned. In those other languages it is enough to collect a small portion of what there is online to build corpora on any domain, of any size, etc. But it remains to be seen whether obtaining a portion of the Basque web (or even obtaining all of it) will be enough to satisfy the current needs of Basque concerning corpora.

Another possible problem might arise from the morphology of the language. Basque is an agglutinative language with a very rich morphology that search engines (the only access door for many Web-as-Corpus techniques) do not take into account. In fact, it is not only the morphology: search engines do not even offer the possibility of returning results in Basque alone (just as they do not for all but 40 languages, in fact). These problems might render the techniques unsuitable in our case, and it is not clear how we might circumvent them.

However, our hypothesis is that the Web-as-Corpus approach can be valid to make a significant change in the situation of corpora for Basque. This thesis is aimed at trying to confirm this hypothesis and, by doing so, improve the state of the Basque corpora.

1.6 Specific objectives

To confirm the validity of our hypothesis (*i.e.*, that the Web-as-Corpus approach can improve the situation of Basque corpora), we set ourselves some specific objectives, which were the following:

- To build a tool that would enable the web to be queried as if it were a Basque corpus.
- To develop tools that would automatically collect from the web a general corpus of Basque that would outdo existing corpora by an order of magnitude and reach a size of at least 100 million words, and which would be of a quality comparable with the other ones.
- To build a tool for automatically collecting domain-specialized Basque corpora from the web, of a sufficient size and quality for terminological uses (evaluated with an automatic terminology extraction task), and to collect some domain-corpora using it.
- To develop a tool for automatically collecting Basque-English domain-comparable corpora good enough and large enough to be used for automatic bilingual terminology extraction, and use it to collect some comparable corpora.
- To make these tools and corpora publicly available to the greatest possible extent.

We would regard the hypothesis as having been confirmed if we were able to obtain good results for all, or at least most, of them.

1.7 Outline

So, the following chapters of this thesis will follow a structure according to the objectives mentioned above.

First, in Chapter 2, we explain the state of the art of the Web-as-Corpus approach in the different modalities we have addressed: querying the web live as if it were a corpus and using the web as a source of texts for building large general corpora, specialized corpora, and domain-comparable corpora. In addition, we also detail the state of the art of the various cleaning and filtering stages involved in any web corpus collection method.

Chapter 3 will detail the work we have carried out in order to build a tool for querying the web as a Basque corpus. We explain the problems we have found for building such a tool and the solutions we have devised. Specifically, we describe the techniques of morphological query expansion and language-filtering words, which we will be using in the rest of the tools developed throughout the thesis. And we detail the experiments we have carried out to evaluate and measure the performance of these techniques.

In Chapter 4, we briefly describe the techniques and tools we have used in each of the filtering and cleaning stages (boilerplate removal, duplicate detection, etc.) that have to be applied to pages collected from the web before including them in a corpus.

Then, Chapter 5 deals with the experiments we have performed to collect a large general corpus of Basque using two different methods (the search engine method and the crawling method) with different parameters, and also with the qualitative evaluation to which we have subjected the different corpora obtained and its results (with regard to other reference corpora).

Chapter 6 deals with the experiments carried out to collect domain-specialized corpora in Basque, their results, and the performance they obtain when used in an automatic terminology extraction task, compared with manually built corpora.

Chapter 7, similarly, refers to the experiments for obtaining Basque-English comparable corpora and the evaluation of the corpora obtained in an automatic bilingual terminology extraction task, also compared with manually built corpora.

Finally, Chapter 8 sums up the thesis by explaining the results obtained and the resources, tools, and publications produced throughout its development. Chapter 9 concludes the thesis by listing the bibliography used and cited in it.

2 State of the art of the Web-as-Corpus approach

The term **Web-as-Corpus** is a generic name used to describe diverse techniques and methods, all of which share the use of the web as a corpus, that is, as a source of evidence for linguistic research. But the techniques for using the web as a corpus are different for each type of corpus: it is not the same to try to employ the web as a general corpus, or as a specialized one, or as a parallel one, etc.

In this chapter we will describe the state of the art of the Web-as-Corpus approach in the four areas this thesis aims to address: direct querying of the web with linguistic intentions and the use of the web for collecting specialized, comparable, and general corpora.

2.1 Direct queries of the web as if it were a corpus

The Web-as-Corpus approach is divided into two main sub-approaches: one of them makes use of the web to easily obtain texts with which to build a classical corpus (linguistically tagged, indexed, etc.); the other (the one we will refer to in this section) directly queries the whole web in search of linguistic evidence.

This sub-approach has received different names. De Schryver (2002) calls it **web as corpus** as opposed to the other one which he refers to as **web for corpus** (although later the generic name Web-as-Corpus was adopted for both). Others (Bernardini et al., 2006) call it **web as a corpus surrogate**.

In the introduction, when we listed the objections some people raised against the Web-as-Corpus approach, we said that most of them (having an unknown size, not being linguistically tagged, etc.) did not apply to the cases when we use the web as a source to obtain texts and build a classical corpus with them. But the use we are dealing with in this section, *i.e.*, the live consultation of the web for linguistic evidence, cannot claim to be one of those cases, and the objections are applicable without any excuses.

Besides, the live querying of the web must necessarily be done with the intermediation of search engines. And since search engines were not designed to be used for linguistic purposes, more factors emerge to justify the inappropriateness of this use.

The problems of the use of search engines for linguistic purposes were very well enumerated by Kilgarriff (2006) and are as follows:

- Search engines do not linguistically tag the pages they index, so when asking for all occurrences of a lemma, many queries with each of the possible inflections have to be made. Lately, search engines have been introducing more "linguistic" intelligence –not based on real linguistic knowledge but in analyses of query logs (Guo et al., 2008)– and return inflections, variants, corrections or even synonyms of the words asked for (Cucerzan and Brill, 2004); but how they do it or to what extent is obscure, so one cannot rely on it.
- The search syntax is limited. We cannot use distances between words or wild cards, for example.
- The numbers returned by search engines refer to pages containing the search terms, not to actual occurrences of them.
- The counts returned are very arbitrary. They change not only between different search engines, but also in the same search engine for different users (because of the personalization of results they try to provide), days (because of updates in the index), etc. This makes any result non-reproducible (Lüdeling et al., 2007), a must for any valid research.
- Hits are returned ordered by search-user-satisfaction criteria –an unknown combination of PageRank (Brin and Page, 1998), query-result language model similarity and others–, not linguistic ones, and only some results can be retrieved; thus, we can never know what linguistically interesting content we might be missing.

Volk (2002) also noted the drawbacks of using search engines for linguistic research and claimed the need for a "Corpus Query Tool for the Web."

Disadvantages notwithstanding, there is still room for some legitimate uses for Googleology, as Kilgarriff (*ibid.*) called it. Looking for use examples of very rare words or constructions that cannot be found in any other corpus might be one of them. The same can be said about very recent neologisms, for example technology-related terms. Also word counts returned by search engines, although they are not to be trusted or taken as an exact or absolute measure, can serve as relative word-frequency indicators in certain cases (monosemic words, looked up at the same time from the same computer...). And it may also be the only option for languages which have no corpora, or only few and small ones (as is the case of Basque).

The simplest form of using the web for linguistic evidence is to enter a word in the interface of a search engine and to look at the hit counts as a rough estimate of their frequency or at examples of use in the pages returned (Bergh et al., 1998; Grefenstette, 1999; Turney, 2001; Chklovski and Pantel, 2004). But this method does not provide the same results as a typical corpus querying tool would. We do obtain hit counts (to be handled with care, as we have said), but not KWiCs with occurrences of the word: search engines only return a list of pages where the search term appears. They do include a snippet for each result (a piece of text from the page) that does contain the word we were looking for, but we can only see one occurrence of it and with a very limited context; besides, if we looked for more than a word, they might be a long way from each other on the page, and the snippet does not give us an idea of the distance. So if we want to see the exact sentences the word is used in, we have to click and go inside each of the pages and look for it. A most tiresome process, as we can see.

To avoid this, tools and services have been built on top of search engines that spare us these steps. We enter the word in them, they ask some search engine for it, and they download the pages that the search engine has returned, look for the desired word, and show it in a KWiC form. This concordancing feature was something that users of the above method were demanding when it did not exist, like Bergh, Seppänen and Trotta (*ibid.*).

These tools still have some more drawbacks to add to the ones we have already mentioned, according to Kilgarriff (*ibid.*):

- The APIs that search engines offer and that are the way by which these tools query the search engines are limited regarding number of uses per time period, and have been getting more so over time. Some have become paid services, and others have even just given up the service.
- Once in a while, the constraints, syntax or any other factor of APIs change, making the tool we have built upon it useless until we adapt it to the new conditions.

For all these reasons, he advocates the other approach, that is, to use the web as a source for texts with which to build corpora, emphasizing the need for these to be really large in order to compete with search engine usage.

However, services and tools of this kind have been built. Some of them are web services that anyone can use through a web browser, such as WebCONC (Hüning, 2001), The Linguist's Search Engine (Resnik et al., 2005), WebCorp (Kehoe and Renouf, 2002; Renouf et al., 2007) or Web Concordancer (Fletcher, 2007b). Others are programs that have to be installed on one's computer, like KWiCFinder (Fletcher, 2006), or CGI scripts to build a web service, like NetKwic (van Noord, 1997a).

Due to the changes in the APIs we have mentioned, most of these tools do not work any more. Web Concordancer has a sign announcing that it dropped its service in August 2012 due to the retirement of the Bing API. The Linguist's Search Engine also announced its going out of service in April 2010. WebCONC does not seem to be working. We have not tried NetKwic or KWiCFinder because they require downloading and installation, but their websites have not been updated for many years (their last updates date back to 2000 and 2007, respectively), so it is highly improbable that they have miraculously survived the many changes or disappearances search engine APIs have suffered since. WebCorp is the only one we can confirm is still working.

However, none of these services work well for Basque, due partly to some characteristics of the language itself and partly to the treatment (or, better, non-treatment) that search engines give to the Basque language (and many other languages, for that matter). Therefore, one of the objectives of this thesis has been to build a tool that would allow the web to be queried live as if it were a corpus of Basque. The tasks we have carried out in pursuit of this objective and the results we have obtained are described in Chapter 3.

2.2 Using the web as a source of texts for building corpora

The other main strategy of the Web-as-Corpus is to make use of the web to collect texts for building a classical corpus, that is, one that we will tag and index properly. This can be applied to the collection of corpora of any type: general, specialized, comparable, parallel...

As we have already mentioned, De Schryver (2002) called this strategy **web for corpus**, although later the name he initially gave to the other strategy (web as corpus) has been adopted to name both in general. Bernardini, Baroni, and Evert (2006) used the term **web as a corpus shop** to name this sub-approach.

This modality of the Web-as-Corpus is nowadays much more popular and well-regarded than the direct consultation one. But it has not always been like this. A few years ago, there were many tools for querying the web live for linguistic purposes and APIs that allowed it. However, as we have pointed out in the previous section, most of them have ceased to exist, whereas there are really plenty of corpora of many kinds built with documents obtained from the web, as we will confirm in the next subsections.

Kilgarriff has been one of the major defenders of using the web as a source for building huge classic corpora instead of using search engines for linguistic purposes, with two well known papers that probably set the milestones for the mentioned shift in trend to happen (Kilgarriff, 2003; Kilgarriff, 2006). And he is one of the leading actors in the web-as-a-corpus-source strategy, taking part in projects for collecting large general corpora (Pomikálek et al., 2009; Baroni and Kilgarriff, 2006; Kilgarriff et al., 2010) or specialized corpora (Baroni et al., 2006), apart from being behind the impressive Sketch Engine (Kilgarriff et al., 2004), a commercial infrastructure allowing the fast treatment of very large monolingual or parallel corpora in any language, this treatment including linguistic processing, concordance search, and word sketch calculation (one page summaries of a word's grammatical, collocational, and translational behaviour).

2.2.1 Obtaining large general corpora using the web as source

There are roughly two methods that are mentioned in the literature when it comes to building large corpora out of the web: the crawling method and the search engine method.

2.2.1.1 Crawling method

This is what the **crawling method** consists of: 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 pre-specified endpoint. As the web is a collection of interconnected pages, starting from an appropriate seed list and applying this method, most of the whole public web (we call the public web that which is neither behind a form, like dictionar-

ies for example, nor is private and needs a password) can be reached and thus, provided we have enough processing power and storage capacity, potentially downloaded.

Choosing the seed URL list is the first step for a crawl, and a very important one too. To understand why, it is necessary to explain the **graph structure of the web** introduced by Broder et al. (2000). The web is a collection of **hypertext** (Berners-Lee, 1989), that is to say, a collection of documents that link to each other by means of unidirectional links. Broder et al. (*ibid.*) defined the following sets to classify pages according to their link interconnection degree:

- **SCC** (Strongly Connected Component): the set of pages that for any p_1 and p_2 inside it a link path can be found that leads from p_1 to p_2.
- **IN**: pages that have no link path arriving to them from the pages on SCC, but from which a link path can be found to SCC.
- **OUT**: pages that can be reached starting from the SCC but that do not have any outgoing link path to it.
- **TUBE**: the pages with which a link path can be created from IN to OUT but without links to or from the SCC.
- **TENDRILS**: the set of pages that have incoming link paths from IN but cannot be connected to the SCC plus the set of pages that have link paths to OUT but cannot be reached from the SCC.
- **DISC** (Disconnected): conformed by the pages that cannot be reached from or have no way to access any page of the other sets.

Figure 2.1 illustrates the graph structure of the web they defined with these sets and interconnections. The graph is now known as the “bow-tie” graph, due to the shape the authors gave to the graph in the original paper.

Broder et al. (2000) calculated empirically the approximate sizes of each of the sets and found that, with the exception of DISC and TUBE, the others were, surprisingly, roughly the same size. The explanation for the size of IN can be that many new pages are created all the time and at the beginning they do not have any inbound links. Typical OUT pages can be company websites that do not link to the outside, which are numerous. However, later studies (Serrano et al., 2007) report an ample predominance of SCC with respect to the rest.

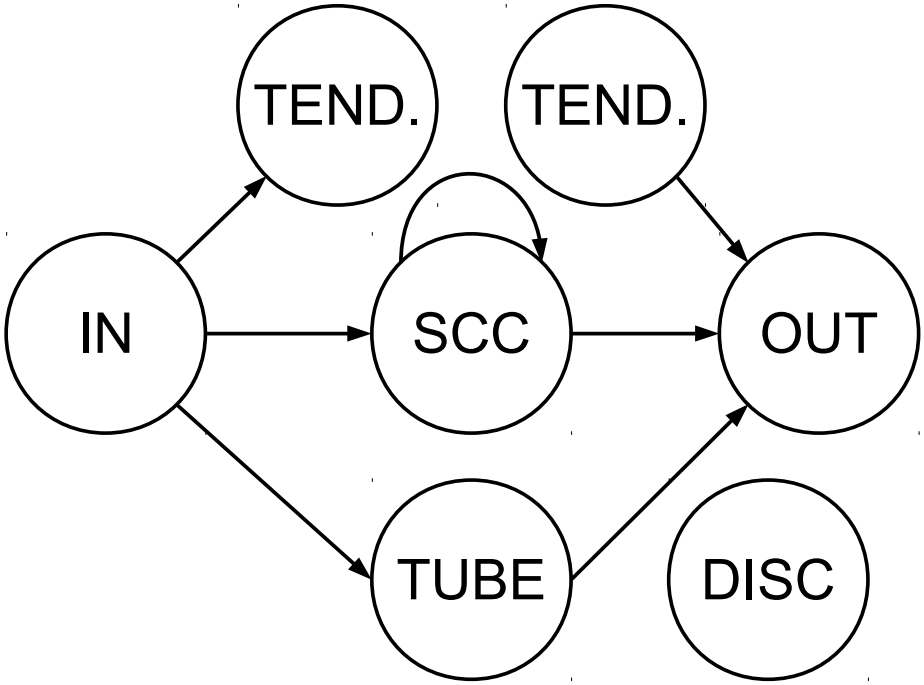


Figure 2.1: Graph structure of the web
Source: Own creation, adapted from the bow-tie graph from (Broder et al. 2000)

The concept of "appropriate" seed URLs list we talked about earlier, starting from which most of the public web could be reached, can be understood in terms of this graph. For example, if all the initial URLs are from OUT, we will only be able to download some pages from OUT and would be missing most of the web. The same can be said about DISC. On the other hand, if the pages of the seed are from IN, apart from having some pages from IN, we can eventually reach all SCC and OUT plus a part from TUBE and TENDRILS. And starting from SCC we can also have all SCC and OUT. However, it is not easy to know beforehand which of the sets an URL belongs to, making this criterion of choice for the seed URLs very difficult to apply.

However, there are more factors to consider. It is important that the pages we are interested in are reached as soon as possible, so that we do not waste time, storage or bandwidth downloading and following several uninteresting pages and links (we are talking of millions or billions, in some cases). Therefore, we should design the seed list with this in mind. This might be quite easy when we are looking for specialized

corpora (we have to locate sites with many pages that deal with a particular domain or subject), but not so easy when we want to collect a general corpus: we need linguistically interesting pages, but how can we design a seed list that, in a crawl, will reach these as early as possible?

Using search engines is the method used in many web corpus crawling projects for obtaining the seeds, for example in (Baroni et al., 2009; Schäfer and Bildhauer, 2012). Word combinations are sent to a search engine and the links of the pages they return make up the seed list. Regarding the number of pages in the seed for this case, including many of them will make the initial performance of the crawl very good (because the seeds are the pages downloaded first and when asked for word combinations, search engines return content-rich and long pages), but afterwards there is not much difference with other seeds, as Schäfer and Bildhauer (2013) proved. Another method suggested by those same authors is to use the links in the language categories from the **Open Directory Project** (ODP, 1998) or the links in **Wikipedia** dumps (Wikimedia Foundation, 2001). Liu and Curran (2006), for example, use the Open Directory Project to gather the seeds for a 10 billion-word corpus of English, and also Ravichandran, Pantel, and Hovy (2005) for collecting a corpus of 31 million web pages.

Apart from the seed list, the crawling process also has to be designed so that the interesting pages are reached early and as few as possible useless links are followed. This thing we have just mentioned is very important because of the great differences between the probabilities of reaching different pages. Let's explain this. The **in-degree** or *ID(p)* of a web page *p* is defined as the number of pages linking to it and its **out-degree** or *OD(p)* as the number of pages it links to. Many studies (Kumar et al., 1999; Barabási and Albert, 1999; Broder et al., 2000; Manning et al., 2009) have empirically proved that the in-degrees of web pages are distributed according to a power law with an α value of 2.1, which means that the probability of a page having a certain in-degree can be calculated using the following formula (where *i* is the in-degree and *P(i)* the probability):

$$P(i) = i^{-2.1} \tag{2.1}$$

And the graph representing this formula can be seen in Figure 2.2.

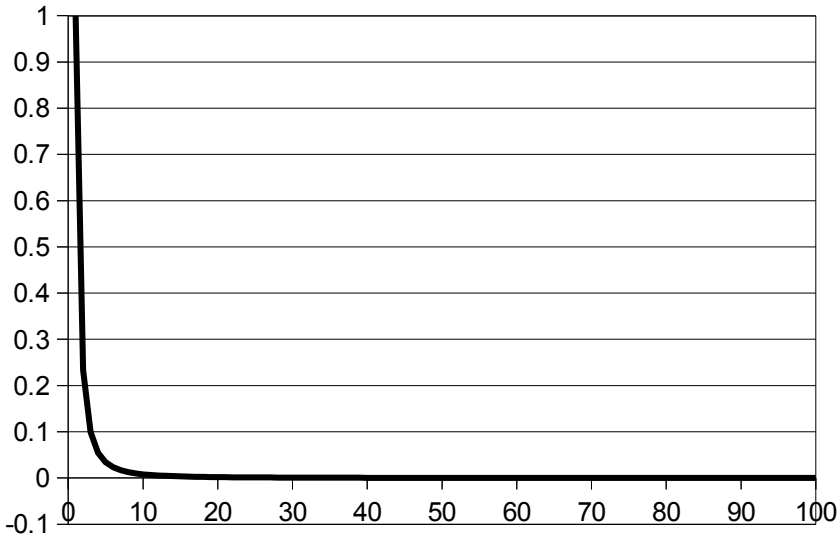


Figure 2.2: Graph representing the probability of a page to have a certain in-degree

This means that the number of pages with a high in-degree is relatively small and the number of pages with a low in-degree large. This fact is widely accepted even though other studies (Serrano et al., 2007) cast doubt on the α value of 2.1 and even that the in-degree shows a power law distribution.

Therefore, in a crawl we will always be bumping into the same few pages with a high in-degree, which might not be the best suited for a corpus (and even if they were, we only need each of them once). This is an effect that should be minimized in the name of effectiveness, and the crawl should be designed to find the low in-degree pages. But again, it is not so easy to decide *a priori* which links will be interesting when we are collecting a general corpus. One technique that is possible and easy to implement (and that, therefore, is implemented almost always) is to follow only the links of pages that are in the target language.

Another decision to take before starting a crawl is whether we are going to follow a breadth-first or a depth-first strategy. With a **depth-first strategy**, whenever we discover a new website or domain, we follow all the internal links until we have downloaded it all; a **breadth-first strategy** means that we prioritize website diversity and that we first download pages in new domains, or else that we download the pages in the order the links have been queued. The most widely used strategy is the breadth-first one, according to (Schäfer and Bildhauer, 2013), most likely because a depth-first

corpus would not be as varied as a breadth-first one of the same size: all the texts would be from fewer sources, and this is not generally good for a corpus (unless it is done intentionally, *e.g.*, in the case of a prescriptive "model" corpus where we just want to include certain sites).

In some cases, the crawl is restricted to **national top-level domains** (hereinafter **NTLD**) for obtaining monolingual corpora, as in (Baroni et al., 2009; Schäfer and Bildhauer, 2012; Pomikálek et al., 2009; Halácsy et al., 2004). But it depends greatly on the NTLD, because the case of each is different. Some NTLDs are successful (`.uk`) and others are not (`.us`), some belong to countries with a single dominant language (`.de`) and others do not (`.be`)... However, limiting the crawl to a NTLD is not enough to ensure that all documents will be in the desired language –as is obvious–, so some sort of additional language filtering is necessary.

Finally, there are many details to take into account when crawling in order to improve the performance. The parallelization of the downloads by multi-threading can achieve a great gain in time, because there are many waiting times in this part. It is also interesting to implement a system to detect that a page has already been downloaded even if no page with exactly the same URL has not (for example, when parameters are in a different order). Downloading only text pages that are in a format we can handle, convert, and use without videos, images or others is also important. Politeness delays should be respected to avoid being banned by sites if we ask them for too many pages too continuously. And there are more aspects to fine tune a crawl; we can find a more detailed reference of them in (Schäfer and Bildhauer, 2013), for example.

The crawling method is the usual choice for collecting large general corpora from the web. It is the one used, for example, 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 in the pipeline): 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. Kehoe and Gee (2007) also used this method to build the WebCorp Linguist's Search Engine (not to be confused with WebCorp, the live querying service from the same authors), a corpus of English containing 340 million words. Schäfer and Bildhauer (2012) built the 9 billion-word COW corpora with the crawling method too. And ClueWeb09 (Callan et al.,

2009), the 1.2 billion web pages dataset from which Pomikálek, Jakubícek and Rychlỳ (2012) extracted their 70 billion-word corpus was also collected this way. Finally, BiWeC (Pomikálek et al., 2009), a 5.5 billion-word corpus, was also obtained by crawling.

There are many tools that one can download and use to perform a crawl, and some of them are free software too. They include:

- **Heritrix** (Mohr et al., 2004): This is the web crawler that the Internet Archive project (Internet Archive, 1996) developed and uses. The corpora of the WaCky! initiative (Baroni et al., 2009) and BiWeC (Pomikálek et al., 2009) have been compiled using this crawler.
- **GNU wget**: A command line tool included in most Linux distributions, developed by the GNU project (Free Software Foundation, 1983). It can download a single page or a whole site. It is used, for example, by Zhang et al. (2006) to compile parallel corpora from the web.
- **Nutch** (Khare et al., 2004): It is the crawler of the Apache project (Apache Software Foundation, 1999). ClueWeb09 (Callan et al., 2009) was collected using Nutch.
- **HTTrack** (Roche, 2004): Software for downloading copies of websites, that can also be used for crawling.

These tools are configurable to allow many of the requisites we have mentioned. If we have some special need, such as the implementation of page language detection we mentioned earlier, the free software nature of the tools allow them to be modified: ClueWeb09 (Callan et al., 2009) used a modified version of Nutch, for example. Another option, if our needs are quite special, is to develop our own crawler.

2.2.1.2 Search engine method

The other method relies on the use of **search engines**. Although this method was initially used for collecting specialized corpora (see BootCaT below), it is also used to build large general corpora. 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 either the goal size is reached, no more combinations are left, or no new pages are returned. According to (Schäfer and Bildhauer, 2013), there are known problems with the use of search engines for general corpora: the bias introduced by the engine towards documents interesting for search purposes but not linguistic ones; the bias towards long

pages induced by the use of various search terms (although this can be interesting for linguistic purposes, it sometimes results in word lists or dictionaries being returned); and, finally, the restrictions in the use of APIs that search engines have been including lately. For all this reasons, the crawling method is the most used nowadays.

However, the search engine method was successfully employed by Sharoff (2006) to build BNC-sized corpora (around 100-200 million words) for various languages. (Chinese, English, German, Romanian, Ukrainian, and Russian). Ueyama and Baroni (Baroni and Ueyama, 2004; Ueyama and Baroni, 2005; Ueyama, 2006) used it for building and evaluating different fairly small Japanese general corpora. And Kilgarriff et al. (2010) employed it to build around 100 million-word corpora for six languages (Dutch, Indonesian, Norwegian, Swedish, Thai, and Vietnamese) and smaller ones for another two (Hindi and Telugu).

The approach for obtaining the list of seed words is different in each of the above mentioned works, although the words for the list generally share some characteristics in most of them: they are high-frequency words, of general use (*i.e.*, not specialized) and they are not function words (prepositions, articles, pronouns, conjunctions, etc.). Since a corpus collection process is usually aimed at obtaining texts in a certain language, some works (Kilgarriff and Grefenstette, 2003; Ghani et al., 2003) use words that are unique to the language and reject words like *hotel*. Others, like, like the above-mentioned work by Sharoff, use the language filter of search engines, except for languages for which this feature is not available in search engines, in which case they complement the query with a couple of very frequent function words that are not used in cognate languages. Some authors, for example Kilgarriff et al. (*ibid.*), only use words longer than 5 characters to avoid the above-mentioned possibility.

The seed words are usually obtained from corpora (BNC, Wikipedia...). The length of the seed words list also differs: for example Sharoff (*ibid.*) uses 500 words, whereas Kilgarriff et al. (*ibid.*) used those words whose frequency ranking was between 1,000 and 6,000. The words they use in those works are surface forms, not lemmas.

Regarding the length of the query sent to the APIs of search engines, that is, the length of the combinations randomly made out of the seed words, this must be long enough to avoid documents in other languages and short enough to get enough results from the search engines. Depending on the work and the language, 2-, 3- or 4-word combinations are used.

Regarding the search engine, Sharoff's work used Google's API and Kilgarriff et al. used Yahoo's and Bing's. From the results returned by the API, Sharoff and Kilgarriff et al. download the first 10 pages. Finally, regarding the number of queries, Sharoff made 5,000 and Kilgarriff et al. 30,000.

2.2.1.3 Choice of method

As we can observe, both methods, the crawling one and the search engine one, 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? Is it possible to obtain corpora of billions of words with the search engine method? (Some authors, namely Schäfer and Bildhauer (2013) and Baroni and Ueyama (2006) suggest that it is not.) Which obtains the best quality corpora?

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 again 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. That is why testing and evaluating both methods (and with different parameters) for collecting a large general corpus of Basque is a necessary task, which has been performed in this thesis and whose development and results are described in Chapter 5.

2.2.2 Using the web to build specialized corpora

Regarding specialized corpora, the web is a good source of specialized texts, and tools can be (and have in fact been) built to automatically collect texts on a specialized subject from it.

Before BootCaT (Baroni and Bernardini, 2004) came onto the scene, collecting corpora on a certain topic from the web was mainly done by **focused crawling**. This concept, introduced in (Chakrabarti et al., 1999), is a special type of crawling that im-

plements some bias towards the kind of pages we need. For example, we could decide to queue the links found in a page only if the page belongs to the domain and language in which we are building the corpus. Or we could implement a language detection based only on the URL and queue a link only if we are sure it is in the desired language, as in (Baykan et al., 2008). The same can be done with the domain, trying to guess it by the URL as they do in (Baykan et al., 2009).

However, a focused crawling normally needs a subsequent filtering of the downloaded pages using some sort of topic classifier, which is usually done by training *ad hoc* machine learning filters, as in (Chakrabarti et al., 1999). The features used for training domain-filters are normally content words or terms from the domain (Sharoff, 2007).

The seed URLs that are used in a crawl for a domain-specialized corpus are usually websites related to the target domain. And if we are sure that most of the documents in the seed websites are in the target domain, it might be interesting to use a depth-first strategy not to get out of those initial sites, or at least to download the entirety of those websites first. This simpler way has been used in the WebCorp Linguist's Search Engine (Kehoe and Gee, 2007) to obtain 125-million-word corpora in some specialized domains, by just downloading the pages that the Open Directory Project (ODP, 1998) indicated as belonging to the domains.

BootCaT (Baroni and Bernardini, 2004) introduced a new methodology: a list of seed words is given as input (which are words on the topic), the APIs of search engines are queried for combinations of these seed words, and the pages are downloaded. This methodology has in some cases been used to build big general corpora (Sharoff, 2006; Ueyama, 2006; Kilgarriff et al., 2010), but for collecting smaller specialized corpora, it has become the *de facto* standard. There is also a web version of BootCaT, **WebBootCaT** (Baroni et al., 2006), which allows those interested in the tool to make use of it without having to install it.

Since the advent of BootCaT, the topic-filtering stage that focused crawling once used has been abandoned, as it has been assumed that the search for words on a topic suffices for obtaining the corresponding texts on it alone. Yet there are not many studies on the precision obtained by the word-list method of BootCaT, and the results of the few that have been done give us reason to believe that a topic-filtering stage *is* necessary: in the original BootCaT paper, an evaluation was performed on a small

sample of 30 texts of each of the two corpora collected, and a third of them proved to be uninformative or unrelated to the topic. Depending on the application, this amount of noise in the corpora may be considered to be unacceptable.

Besides, using this methodology to collect Basque specialized corpora has been found to obtain an even worse domain-precision, due again to some features inherent in the language and to the neglect by search engines with respect to all but some major languages. And this phenomenon is likely to happen not only with Basque but also with other languages with similar features (inflectional complexity, minority language...). That is why this thesis will also be exploring ways to obtain specialized corpora in Basque from the web; Chapter 6 will describe the work performed to achieve this objective.

2.2.3 Collecting domain-comparable corpora from the web

There is not much literature, at least that we are aware of, about the process of collecting comparable corpora. Most of the literature concerning comparable corpora deal with the exploitation of such resources, and simply briefly mention how they got the corpora.

Comparable corpora have traditionally been obtained in a supervised or directed way. A very common source are news agencies (Barzilay and Lee, 2003; Munteanu and Marcu, 2005), who provide news collections in different languages but from the same period; since a great proportion of the news are often the same in all countries, the corpora rendered are considered comparable. Established research corpora (*e.g.*, TREC or CLEF collections) are some other usual comparable corpora used by many researchers. When the web has been used as a source, downloading some pre-chosen web sites for texts on the domain has been the preferred method, or starting a crawling from them and subsequently applying some machine-learning domain filter, as in (Talvensaari et al., 2008).

There are some works that deal with converting comparable corpora from ***light*** to ***hard*** (Sheridan and Ballerini, 1996; Braschler and Schäuble, 1998; Bekavac et al., 2004; Huang et al., 2013). The light and hard comparability levels for corpora were first introduced by Bekavac et al. (*ibid.*). A light comparable corpus would be composed of corpora from two (or more) languages composed according to the same principles (*i.e.*, corpora parameters) which are defined by features such as domain, size, time-span, genre, gender, and/or age of the authors, etc. The hard type comparability is

dependent on already collected and established light comparable corpora. It is derived from them by applying certain language technology tools/techniques and/or document meta-descriptors to find out which documents in lightly comparable corpora really deal with the same or similar topic. A subset of lightly comparable corpora which has been selected by these tools/techniques, whether document-level aligned or not, can be regarded as hard comparable corpora. But again, these studies deal with the obtaining of the light corpora, which is done by one of the methods mentioned above, very superficially.

As we can see, methods that make use of search engines, like BootCaT or similar, are not mentioned in the literature for compiling domain-comparable corpora, although they have become the main method for domain-specialized monolingual corpora and the difficulties in obtaining both types of corpora can be considered similar.

But the crawling approach poses some problems. First, there is a human choice of the sources (websites on the target domain), which is normally limited and small, and this makes the corpora at least biased and often not very diverse. Besides, for small languages (like Basque, for example), in many domains it would not be easy to identify good sources that would contain a significant quantity of documents on the domain. Finally, crawling for domain-specialized corpora, as we said, often requires machine-learning domain filtering, which needs to have some training corpora available beforehand, and this is not always the case.

So this thesis sets out to experiment with search-engine based ways of collecting comparable corpora which include Basque as one of the languages. These experiments and their results are detailed in Chapter 7.

2.3 Corpus cleaning

When we download pages from the web to form a corpus, no matter whether we are building a general, specialized or comparable corpus, whether we are using search engines or crawling, or whether we are interested in one language or another, these pages need to go through some cleaning and filtering if we want to build a quality corpus. The following stages at least should be implemented in any web-corpus building process: **length filtering**, **language filtering**, **spam filtering**, **porn filtering**, **boilerplate removal**, **near-duplicate detection** and **containment detection**. All the web corpus building projects have most of them implemented in one way or another. We will mention the most significant ones in the following subsections.

This corpus cleaning process is usually done by **post-processing**, *i.e.*, first all the documents are downloaded and then the different cleaning and filtering stages are applied to the pages (Schäfer and Bildhauer, 2013). According to these authors, more than 90% –in one of their experiments, 94%– of the downloaded content is discarded in the cleaning and filtering process.

2.3.1 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.

The most well-known method (and probably the most widely used) for language detection is the one used by **TextCat** (van Noord, 1997b), based on n-gram frequencies (Cavnar and Trenkle, 1994). It obtains a very good precision if the text is long enough (200 characters and upwards is enough). Another technique is to use a list of the most frequent function words and allow only documents with a minimum proportion of them, because according to Baayen (2001) real language (the one of interest for a corpus) fulfils this requirement. This last method is the one used in the corpora of the aforementioned WaCky! initiative (Baroni et al., 2009) and also by Ravichandran, Pantel, and Hovy (2005) for collecting a corpus of 31 million web pages.

But there are cases when the texts whose language we must identify are short, for example when mining text from microblogging sites like Twitter (Twitter, Inc., 2006), which allows 140 characters at most –and where messages typically include links or the name of the person to whom an answer is being addressed. Or when we want to implement a language filter at the paragraph or sentence level. In this last case, it is best to decide whether to include or reject of a short paragraph or sentence by looking at its vicinity: if the paragraphs surrounding it are not in the language, it is unlikely that it will be either.

2.3.2 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 (Fletcher put the threshold at 5 KB) are usually error messages from web servers or tend to have little textual con-

tent once page headers, menus, etc. are removed. On the other hand, those that are too large (according to Fletcher, larger than 200 KB) 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. This method has also been used in the corpora collected by the WaCky! initiative (Baroni et al., 2009).

2.3.3 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) and mentioned above decreases this kind of noise but does not eliminate it completely. If we use search engines, we will most likely 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.

The WaCky! initiative (Baroni et al., 2009), for example, removes spam through the function words list used for language detection, since spam pages are often made up of nonsense word lists or links. For porn, they have a black list of words usual in porn pages and remove pages that have more than a certain number or proportion of them.

2.3.4 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 (most of the boilerplate is just single words or very short sentences), it distorts word frequencies (words like *contact*, *home*, *copyright*, and others will be enormously overrepresented) and it makes the work of other filters –near-duplicate filtering, for example– more difficult (two pages with the same content but in different servers might be identified as not duplicate because of the different boilerplate; likewise, two different but short pages from the same site might be considered duplicates if the boilerplate, which is always very similar for a website, is proportionally long). The task of removing these unwanted sections of web pages and keeping just the real content is called boilerplate removal.

When crawling a few websites, it is possible to build filters adapted to the structure of those sites with a performance of 100%, but this is not valid for a web-wide corpus collection. There are also site-level methods for the detection of boilerplate (Pomikálek, 2011), which try to detect the structure of the boilerplate of a site by looking at the similarities in different pages from the same site. But they cannot always be applied: they need a minimum number of pages for each site, sometimes the boilerplate is different for different sections of a site, etc.

The reference in many Web-as-Corpus projects for the detection of unwanted sections in web pages –for example, in the WaCky! corpora (Baroni et al., 2009) or in BiWeC (Pomikálek et al., 2009)– is the **BTE** (**Body Text Extraction**) algorithm developed by Finn et al. (Finn et al., 2001) or some adaptation of it. This algorithm uses the HTML tag density to detect real content, since this contains proportionally fewer tags than boilerplate.

The other main reference in this area is the **CleanEval competition** organized in 2007 in the 3rd Web as Corpus workshop (Baroni et al., 2008). 10 systems took part in it, some based on heuristics and some based on machine learning, but they all took into account more features apart from tag density: markup related features (length of the markup, tag density, link count...), linguistic features (stop words, punctuation signs, function words, sentence length...), document features (length...), etc.

Ferraresi et al. (2008) and Pomikálek et al. (2009) reported that the BTE algorithm performed better than any CleanEval system; but Pomikálek recognized in his Ph.D. thesis that such a claim was made based on his own work and that further research he himself conducted "revealed though that the performance of the BTE algorithm is problematic and the good results here are mostly due to the characteristics of the CleanEval collection and scoring method," concluding that the "BTE algorithm –widely used for cleaning Web corpus data– achieves much lower precision scores than other algorithms, which makes it a rather inappropriate choice for this application" (Pomikálek, 2011).

One of the participants in that shared task (Gao and Abou-Assaleh, 2007) used a visual approach, that is, they tried to detect the rectangular sections in the rendered version of the page and choose the central and largest one.

Pomikálek, in his Ph.D. thesis, developed the jusText system, which is a heuristics-based system that uses link density, function words proportion, block size, and proximity to other already identified boilerplate blocks (Pomikálek, 2011). This system is used in the 70 billion-word corpus described in (Pomikálek et al., 2012).

Some other approaches we can mention are (Kohlschütter et al., 2010) or (Pasternack and Roth, 2009).

The boilerplate removal task is a very difficult one. We are still far from obtaining a system that will attain a 100% effectiveness for any website, and we probably never will. However, in a near or medium-term future this task might become unnecessary. The new version of the standard for the format of web pages, HTML 5 (W3C, 2013), includes semantic tags to distinguish these sections in a web page: `<header>`, `<footer>`, `<nav>` (for the navigational menus), `<article>` and others. Work on HTML 5 was begun in 2004 by the Web Hypertext Application Technology Working Group or WHATWG, their work was taken up by the W3C in 2007, started gaining popularity around 2010 and received the status of candidate recommendation in 2012. Although HTML 5 is being used more and more every day, the adoption of its different parts –both from Internet browsers and web developers– is very unequal. In the case of the semantic tags we have mentioned, all of the browsers have been supporting them in their latest versions (Deveria, 2013), but its use in websites is not still widespread. It might take some time until CMS (Content Management System) developers include it and websites update their CMSs to the latest versions, but boilerplate removal will hopefully become obsolete eventually.

2.3.5 Near-duplicate detection

The detection of exact duplicates is a straightforward task easily accomplished by **hashing techniques**. These consist of representing each document by a fixed length number, where the probability of two documents that differ even very slightly producing the same hash is extremely low, and two equal documents produce the same hash. Looking for equal hashes suffices for detecting exact duplicates.

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** for the detection of duplicates (Broder, 1997). It takes all the **shingles** or n-grams of a document, and if two documents share many of them, it means they are very similar. This simple yet efficient technique owes its popularity to the level of optimization later achieved by the author (Broder, 2000), which makes it usable even on a web scale. Keeping all the shingles of many web documents and comparing all of them for each document pair is computationally very expensive. But Broder (*ibid.*) made his algorithm much more efficient by taking a fingerprint of each shingle, ordering them, and grouping them again into **supershingles** which are again fingerprinted. Just a few numbers must be kept for each document (the fingerprints of these supershingles), and the coincidence in only one of them is enough to ensure very high similarity.

The corpora of the WaCky! initiative (Baroni et al., 2009) have employed a variant of Broder's method: from each document, they randomly take 25 shingles of length 5, but made only of content words –they do not take into account function words–; if two documents share two of them, they are considered duplicates, since "the chances that, after boilerplate stripping, two unrelated documents will share two sequences of five content words are very low" (Baroni and Ueyama, 2006).

Charikar (2002) used a hashing technique that produces similar hashes for similar objects, and detects near-duplicate documents by looking for hashes with a small Hamming distance; this method can also achieve quite high efficiency levels. In (Ravichandran et al., 2005), they use a method described in (Kołcz et al., 2004) that creates equal signatures for duplicate or near duplicate documents, and is reported as being remarkably fast and with a high accuracy.

There are many other methods (Pugh and Henzinger, 2008; Shivakumar and Garcia-Molina, 1999; Manber, 1994; Heintze, 1996; Schleimer et al., 2003), but most of them are some sort of variant of Broder's algorithm (Pomikálek, 2011).

However, the methods mentioned above have been developed in the context of search engines and are thus aimed at detecting almost near-duplicate documents (search engines do not want to return duplicate results). But in the context of corpus building, detecting smaller similarities is also interesting; two long documents might have a similarity of 50% and including them in the corpus would mean that the half in

which they coincide, which might be very long, is repeated twice. And the same can be said for any other smaller percentage too: detecting and avoiding duplicities of 10% –or even of a single paragraph– can be interesting in corpus building.

Another problem to be addressed when looking for duplicate content is what to do once it has been found. If we have only looked for near-duplicate documents, when we find two almost duplicate documents, it suffices to keep one of them and discard the other. But if we were looking for a resemblance of 50% and if we were to find two documents with that similarity level, what should we do with them? Including both would mean including duplicate content in the corpus –considerable duplicate content if the documents are long. On the other hand, if we just keep one and discard the other, half of the discarded one that was not in the other document is left out for no reason. The obvious solution is to include one of the documents and the part of the other that is not repeated, but this poses two main problems: firstly, it adds complexity to the system in terms of storage and processing, and might render the system non-scalable for very large corpora; and secondly, if the resemblance level we were looking for is small or if we were applying it at the paragraph level, removing the small duplicate parts can render a document fragmented.

Pomikálek developed in his Ph.D. thesis (Pomikálek, 2011) a system that can find duplicate paragraphs in a very large corpus in a scalable way. It uses fingerprinted n-grams, but for the sake of scalability only retains those that occur more than once in the corpus. For calculating which n-grams happen twice or more, the system uses an iterative algorithm that first calculates frequencies of 1-grams and only stores those that appear more than once, then calculates occurring-more-than-once 2-grams (2-grams will only occur twice or more if each of its 1-gram components does) and only stores those, etc. In the end, for each document only the n-grams occurring more than once are kept (about 10%). This way, duplicated paragraphs can easily be detected. Then we can remove only those paragraphs if we are not concerned about fragmentation, or we can keep them if there are not many of them and are concerned about fragmentation, or we can remove the whole document if many different paragraphs are duplicated and we are not concerned about losing some recall. This method was used in (Pomikálek et al., 2012) for building a 70-billion-word corpus and in (Pomikálek et

al., 2009) to build 5.5-billion corpora. The problem with this method is that it can only be applied for post-processing a raw corpus already downloaded, not for on-the-fly detection of duplicates.

2.3.6 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 to detect already contained documents (Broder, 1997). It is not as optimized as near-duplicate detection, but it is possible to use it for small- and medium-sized corpora building.

To our knowledge, no web corpus collection project has implemented a specific containment detection technique. However, a duplicate detection method implemented at the paragraph level, such as Pomikálek's (2011), detects containment also, obviously.

3 Querying the web directly as if it were a corpus of Basque

In the previous chapter we have seen that the live querying of the web for linguistic evidence, either by using search engines or services that make use of them, has in recent years been abandoned in favour of making use of large web-derived corpora. The reasons for having given up this approach are its many disadvantages, which were very well expressed in (Kilgarriff, 2006) and which we have summed up in the state-of-the-art chapter.

But even admitting all those drawbacks, we think there are cases in which this kind of use is still legitimate and acceptable. We pointed out that one of them is the case when a language has no corpora or just few and small ones, as happens with many minority and under-resourced languages. And this is the case of Basque.

The problem is that existing services of this kind –currently, as far as we know, only WebCorp (Renouf et al., 2007) is still working– or search engines in general, for that matter, do not produce good results when used for studies on the Basque language or other minority languages, due to the very limited support that search engines give to them. That is why we embarked on the work of building a web service that would allow the web to be queried as a corpus of Basque. 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 web. This would be done by making use of the APIs of search engines, just as other similar services do, but implementing techniques to improve its performance for Basque. This chapter describes the work performed and the results obtained in this pursuit.

3.1 Problems of search engines with Basque

When using search engines to look for information in Basque, some problems arise. One of the most noticeable is that Basque is an agglutinative language. The problems that non-English languages, and agglutinative languages in particular, have with search engines have been widely addressed (Bar-Ilan and Gutman, 2005; Lazarinis, 2007; Lazarinis et al., 2007; Efthimiadis et al., 2009).

In the case of Basque, one of the problems comes from its rich morphology. A given lemma produces many different surface forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, etc.) for nouns and adjectives, and the

person (me, he, etc.) and the time (present, past, future, etc.) for verbs. A brief morphological description of Basque can be found in (Alegria et al., 1996). Quoting Wikipedia in its article on inflection (Wikimedia Foundation, 2013b), Basque "is a highly inflected language, heavily inflecting both nouns and verbs. [...] Verb forms are extremely complex, agreeing with the subject, direct object and indirect object; and include forms that agree with a "dative of interest" for intransitive verbs as well as allocutive forms where the verb form is altered if one is speaking to a close acquaintance. These allocutive forms also have different forms depending on whether the addressee is male or female." And its article on the Basque language (Wikimedia Foundation, 2013a) says that "a Basque noun phrase is inflected in 17 different ways for case, multiplied by 4 ways for its definiteness and number. These first 68 forms are further modified based on other parts of the sentence, which in turn are inflected for the noun again. It is estimated that at two levels of recursion, a Basque noun may have 458,683 inflected forms."

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 for the exact given word alone or applying some simple stemming rules of other languages (such as appending an *s* for the plural, which is what major search engines do) is not sufficient for Basque. Neither is the use of wild cards –which some search engines used to allow– 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. Even the progressively increasing "linguistic" intelligence that search engines have been introducing lately, returning inflections, variants or corrections of the search terms (Cucerzan and Brill, 2004) often harms more than helps because, as Guo et al. (2008) noted, this intelligence is not based on real linguistic knowledge but on analyses of query logs, and since more queries are made in languages other than Basque, if the original word exists in other languages (see next paragraph), the supposedly inflections that are returned are often not Basque words at all.

The other major obstacle when web searching in Basque is that none of the existing search services offers the possibility of restricting the results to pages that are in Basque. Searching in any of them for a technical word that also exists in other lan-

guages (*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. And local (Spanish) versions of search engines do not perform better: at best, a few results in Basque might appear lost among the results in Spanish, even when using the Basque UI.

3.2 Proposed solution

Even if search engines have these limitations when being used for Basque searching, we have no choice other than to make use of them to make a live query of the web with linguistic purposes. Implementing a continuously up-to-date search engine for Basque by crawling, indexing, and ranking on a daily –or even weekly, or monthly– basis is out of the question. It is both too complicated and too costly (bandwidth, disk, reliability, etc.). Therefore, we will make use of the APIs of search engines but we will apply two NLP techniques for trying to solve the problems mentioned above and to improve performance significantly: **morphological query expansion** and **language-filtering words**.

3.2.1 Morphological query expansion

We described the morphology problem above. In order to obtain a real lemma-based search for Basque, when the API of a search engine is requested for a word, we need it to return pages that contain the word's conjugations or inflections, too; but no search engine does that. The way we propose for approaching this matter is based on **morphological query expansion**.

The importance and use of morphology for various IR tasks has been widely documented (Ambroziak and Woods, 1998; Krovetz, 1993; Woods, 2000; Woods et al., 2000; Langer, 2001). But morphological variation processing is usually approached by lemmatization or stemming at the indexing stage (this is the case of the above papers), since it has been proved to be very effective. This is also the method used so far in Basque IR to deal with the agglutinative nature of Basque language (it is the preferred method in the search boxes of Basque websites).

Instead, since our intention is to use major search engines that do not apply Basque lemmatization or stemming at the indexing stage, morphological generation at the querying stage is applied in our approach.

Some works do propose using query expansion for dealing with morphology (Xu and Croft, 1998; Moreau et al., 2007). However, they rely on corpora and statistical co-occurrence methods or machine learning to find the morphological derivations of the words in the query. These techniques are mostly language independent, but they can expand the query not only with inflections or conjugations of the words but also compounds, other kinds of derivatives and sometimes even unrelated words. Besides, they have not been evaluated on a highly inflectional language: the more derivations we try to get using these methods, the bigger the probability of getting wrong words. By using a morphological generator based on lexica and rules, we always get correct inflections and conjugations. Keller and Lapata (2003), when looking for examples of rare verb-object pairs in the web, created 36 combinations that covered all the inflections of both verb and noun and inserted definite and indefinite articles between them; but they created these combinations manually and made a query for each combination. Stanković (2008) does use a rule-based morphological generator for expanding queries in a highly inflectional language, but focuses on correctly inflecting compounds and phrases.

The approach most similar to ours is that used by Kettunen, Airio, and Järvelin (2007) and Kettunen (2007) for Finnish and other morphologically rich languages, which they call FCG or Frequent Case Generation; they also use corpus-based studies to obtain the most frequent cases of each kind of word and then use morphological generators to produce the forms of these cases for the searched words.

Specifically, we use a tool created by the IXA Group of the UPV/EHU-University of the Basque Country which gives us all the possible inflections or conjugations of the lemma, and the search engine is asked to look for any of them by using an OR operator. For example, if the user asks for *etxe* (*house*), the search engine is asked for *etxe OR etxea OR etxeak OR etxeari OR etxeek OR etxearen OR*…. But the APIs of search engines have their limitations with regard to search term count, length of search phrase, etc. These limitations render a proper lemmatized search for Basque impossible, as searching for all the conjugations or inflections is not feasible. So the most frequent ones are sent, and this will cover a high enough percentage of all the occurrences, as Kettunen et al. (*ibid.*) have proved.

Many previous works dealing with the query expansion problem have shown the importance of weighting the expanded words and the query original terms differently. Although this has been usually applied for attaching different degrees of importance to the original search term and its synonyms, it might also be interesting to weigh the original search term and its inflections or inflections with different frequencies differently. By using the OR operator, we are giving equal weight to all query terms and thus losing this potential benefit, but we have no other choice: we are using APIs of web search engines, and they do not provide the possibility of weighting the search terms.

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. This impression is shared by Schäfer and Bildhauer (2013), who observed that pages coming "from search engine queries are biased toward long and content-rich documents […] because of the conjunct tuple query method used to obtain them."

3.2.2 Language-filtering words

The search engine result filtering in a given language is a well-known problem in IR. There are many tools and techniques for language classifying of texts: N-gram based, trigram frequencies based, Markov models based, etc. (Padró and Padró, 2004). The best known among these tools is probably TextCat (Cavnar and Trenkle, 1994; van Noord, 1997b). However, the one that offers the best results for Basque is **LangId**, a free language identifier based on word and trigram frequencies developed by the IXA group of the UPV/EHU-University of the Basque Country, and which specializes in recognizing Basque and its surrounding languages (Spanish, French, and English).

Having such a tool available, the most obvious and straightforward approach for showing only results that are in Basque would be to filter the results returned by the API by applying LangId to the snippets, since this is the method most used in the liter-

ature (Osinski et al., 2004; Ghani et al., 2003). But this method works when we have enough data in both the desired language and in others; in our case, as we have already stated, searching for technical words that also exist in other languages, proper nouns or short words will often yield very few results in Basque if any at all, so this subsequent filtering would leave almost no results.

In order to obtain from the APIs results in Basque alone, we propose an approach that we call **language-filtering words**; this consists of adding to the search phrase, in conjunction with an AND operator, some Basque words to act as language filters. The features these words need to share are as follows: 1) they should be very frequent, so that practically any document in the Basque language will contain them, and 2) they should be specifically Basque, so that no documents in other languages will contain them. Unfortunately, the most frequent words in Basque are short and, as such, the chances of their presence in other languages or being used as abbreviations or acronyms is quite high. In fact, at least the two most frequent words have well-known meanings in other languages. 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, several of these language-filtering words need to be included in the queries in order to obtain a high percentage of Basque results, although this also involves a loss in recall (some Basque pages may not be returned because they do not contain one or more of the words).

Scannell (2007) also adds very frequent words to the queries in order to obtain pages in a language in his Crúbadán project to collect corpora for many minority languages from the web. He uses one word or two at most. For choosing the words, he uses native speakers whenever possible. Otherwise he employs similar criteria to ours to choose the words. He takes the highest frequency words that are not high frequency words in another language. He also emphasizes the importance of the words not to be short to avoid collisions with words in other languages, but points out that "for 121 of the 416 Crúbadán languages (29%), none of the top 10 most frequent words have four or more letters." Kilgarriff et al. (2010) also pointed out that "it tends to be short words which are words in multiple languages."

The words to be used as language-filtering words are what can be considered as **stopwords** (very frequent words present in almost any page that are not representative of the textual content and which are therefore discarded by search engines when indexing). However, as we have already stated, there are no Basque-aware search engines, so these words are not included in their stopwords list.

3.3 Implementation details and quantitative evaluation

In order to obtain optimum performance, it is important to fine-tune certain details of the morphological query expansion and language-filtering words methodology as much as possible. The choice of how many and which language-filtering words to use, and expanding the query with the most frequent inflections of the words, are crucial for the effectiveness of our approach. These choices had to be made on the basis of precision and recall studies over different corpora. Incidentally, these studies have also produced quantitative measurements of the level of improvement offered by these services. We describe these studies and results in this section. They have been previously published in (Leturia et al., 2008a; Leturia et al., 2013).

3.3.1 Design of the study

As stated above, the study described in this section consists of various corpus-based measurements. One of the corpora used for carrying it out is the Zientzia eta Teknologiaren Corpusa or ZTC (Areta et al., 2007), a lemmatized Basque corpus on science and technology made up of 8.5 million words. Since, as Sharoff (2006) observed, the typology of the documents that form a classical corpus and those that form the web might differ, we considered it advisable to use not only a classical corpus, but also a web corpus. So a web corpus was compiled by crawling the Basque branch of the Google Directory (Google, Inc., 2004), a service based on the Open Directory Project (ODP, 1998) to which PageRank-based ordering (Brin and Page, 1998) was applied and which was shut down in July 2011. We downloaded the 3,000 plus pages present there and recursively followed all the links found in pages that LangId identified to be in the Basque language. The downloading process was designed to ensure as much website variety as possible and used a breadth-first approach, by queuing the links found, prioritizing different domains in each parallel downloading stage, etc. The web corpus obtained is made up of over 44,000 documents and approximately 20 million words.

The various measurements using these corpora had to be done by employing many different queries. We are aware that quasi-standard query collections to evaluate IR systems, such as the TREC test questions, exist, but we opted to use queries that people doing Basque searches really use, so we took the search logs of Elebila (Leturia et al., 2007b), a search engine for Basque that we launched in 2007 using a naive implementation of our morphological query expansion and language-filtering words methodology (cases for the morphological expansion and how many and which language-filtering words were chosen quite intuitively and without making any measurements of the improvement obtained). That way, since our study was based on these most frequent queries, by optimizing the tools with the results of the study, we would be maximizing their performance for real-life searches. The Elebila logs we used accounted for over 400,000 searches involving over 800,000 words, which after lemmatization made over 70,000 different words. The lemmatized queries were subsequently ordered according to decreasing frequency, and the topmost ones were used for our work. All these most frequent queries are one word long, which suits our experiments well. Examples of these queries are *berri* (*new*), *didaktiko* (*didactic*), *eoliko* (*eolic*), *hiztegi* (*dictionary*), *musika* (*music*), *energia* (*energy*), *ikasi* (*learn*), *Galileo* (*Galileo*), *Mozart* (*Mozart*), *Egipto* (*Egypt*) and *Bilbo* (*Bilbao*).

3.3.2 Language-filtering words

When choosing how many and which language-filtering words to include, we find ourselves again facing 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. So in order to choose the words, we had to measure exactly those parameters: the gain in precision and the loss in recall obtained by each filter word or combination of them. We did this with the corpora and words mentioned above.

3.3.2.1 Choosing the words

For choosing the language-filtering words, the first step was to see which the most frequent words in Basque were. In Table 3.1 the 16 most frequent words of each corpora with the document-frequency of each of them are shown.

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

Table 3.1: Most frequent word forms in both corpora

The 16 most frequent words in both corpora are the same, but their order is different. In view of this, we chose the candidates to act as language-filtering words from the first list, as this corpus is supposedly more similar to the one to which we will apply our tools, that is, the Internet. So the candidates will be the topmost six words from the web corpus list: *eta*, *da*, *ez*, *du*, *bat*, and *ere*. Next, precision and recall studies were performed on different combinations of these six candidates.

If one looks at the document-frequencies of the candidate words, it is clear which words would have been chosen if the filter had consisted of one or two words, since there are significant gaps between the frequencies of the first three words in both corpora. Choosing which should be the third and fourth words is more difficult, because the next words have quite similar document-frequencies. For these ones we can even consider OR combinations. So the combinations for which the precision and recall will be analysed in the following subsections are shown in Table 3.2.

Number of words	Combination
0 words	1. -
1 word	2. *eta*
2 words	3. *eta* AND *da*
3 words	4. *eta* AND *da* AND (*ez* OR *du* OR *bat* OR *ere*)
	5. *eta* AND *da* AND (*ez* OR *du* OR *bat*)
	6. *eta* AND *da* AND (*ez* OR *du* OR *ere*)
	7. *eta* AND *da* AND (*ez* OR *bat* OR *ere*)
	8. *eta* AND *da* AND (*du* OR *bat* OR *ere*)
	9. *eta* AND *da* AND (*ez* OR *du*)
	10. *eta* AND *da* AND (*ez* OR *bat*)
	11. *eta* AND *da* AND (*ez* OR *ere*)
	12. *eta* AND *da* AND (*du* OR *bat*)
	13. *eta* AND *da* AND (*du* OR *ere*)
	14. *eta* AND *da* AND (*bat* OR *ere*)
	15. *eta* AND *da* AND *ez*
	16. *eta* AND *da* AND *du*
	17. *eta* AND *da* AND *bat*
	18. *eta* AND *da* AND *ere*
4 words	19. *eta* AND *da* AND *ez* AND (*du* OR *bat* OR *ere*)
	20. *eta* AND *da* AND *du* AND (*ez* OR *bat* OR *ere*)
	21. *eta* AND *da* AND *bat* AND (*ez* OR *du* OR *ere*)
	22. *eta* AND *da* AND *ere* AND (*ez* OR *du* OR *bat*)
	23. *eta* AND *da* AND *ez* AND *du*
	24. *eta* AND *da* AND *ez* AND *bat*
	25. *eta* AND *da* AND *ez* AND *ere*
	26. *eta* AND *da* AND *du* AND *bat*
	27. *eta* AND *da* AND *du* AND *ere*
	28. *eta* AND *da* AND *bat* AND *ere*

Table 3.2: Candidate combinations for different numbers of language-filtering words

3.3.2.2 Loss in recall

To measure the loss in recall produced by the language-filtering words, their document-frequency in the classical corpus and the web corpus were measured. The decrease in hit counts obtained by searching the web using the API of Microsoft Live

Search –now Bing– (Microsoft Corporation, 2006), for words that only exist in Basque (otherwise, occurrences of the words in other languages could have distorted the results), was also measured.

We are aware that hit counts are known to be an unreliable source of information (Uyar, 2009) and that it would be better to at least average hit counts from all major search engines. But the studies performed in this paper involved making many thousands of queries to the APIs, and using APIs other than Microsoft's, due to the limitations they impose on number of queries per day, would have meant several weeks or even months for doing them.

The results are shown in Figure 3.1.

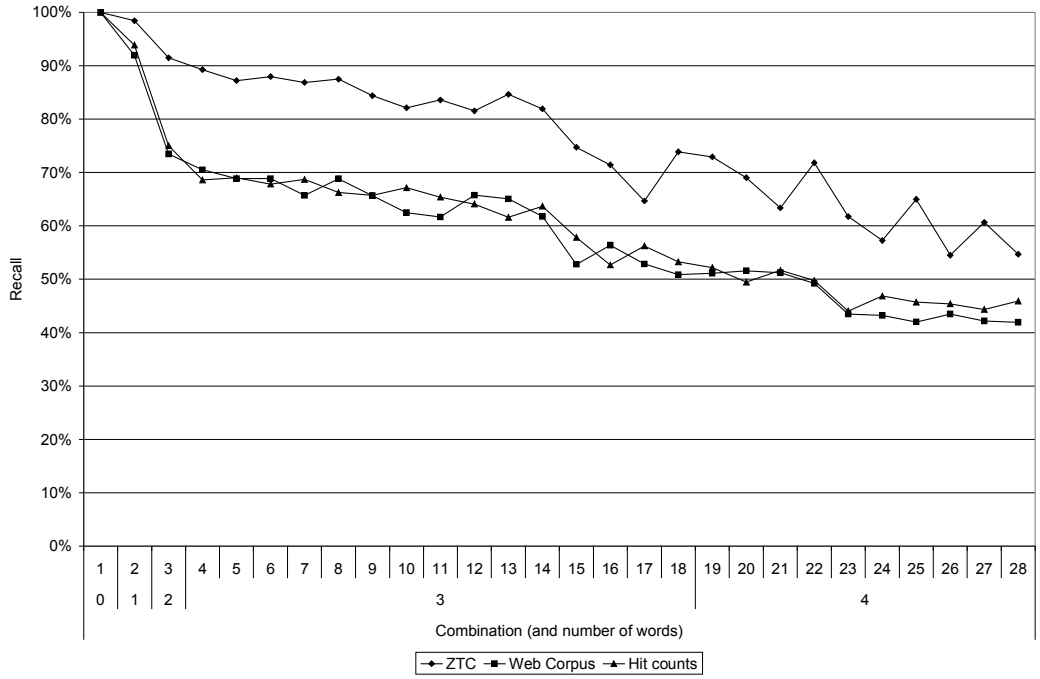


Figure 3.1: Loss in recall produced by the different language-filtering word combinations

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

The loss in recall from one to two filter words is significant. Also, in the groups of three or four filter words, there is a gap when passing from the combinations with an OR to those without it. The graph shows that including three or four filter words without an OR reduces recall to half, which is a significant reduction, so one or two filter words would be best if sufficiently large language-precision is achieved.

3.3.2.3 Gain in precision

The addition of more of the language-filtering words to the query leads to a gain in language precision. To quantify this gain, the ideal procedure would be, as before, to measure it over the corpora, but this is not possible, since we would need a multilingual corpus that would have the same proportion of each language as the web does, which is very difficult, if not impossible, to obtain. So we had no other option but to measure the gain in precision by searching the web through Microsoft's API and looking at the percentage of results in the Basque language. To classify the results into Basque or non-Basque we used LangId again, by applying it to the snippets returned. LangId is specialized in Basque detection and obtains an accuracy of practically 100%, so it works very well even with such short texts.

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

- Short words: Words with 5 characters or less. The probability of their existing in other languages is high. The most searched for words in Elebila from this category (and consequently the ones used for our evaluation) were words like *herri* (*people*, *town*), *berri* (*new*), *haur* (*child*), *ipuin* (*tale*), *gabon* (*Christmas*) or *mapa* (*map*).
- Proper nouns: Proper nouns are usually the same in other languages. Some of the words for this category were *Wikipedia*, *Google*, *Elhuyar*, *Egipto*, *Euskadi* (*Basque Country*), etc.

- International words: Words that we know definitely exist in another language (usually English, Spanish or French). These were some of the most searched for words in this category: *biografia* (*biography*), *historia* (*history*), *energia* (*energy*), *mitologia* (*mythology*) and *arte* (*art*).
- Words that are likely to be found in other languages: Technical words which, despite not being exactly the same in the three languages mentioned above, have fairly similar spellings in all of them, so the probability of their existing in some other language is high. Some examples of these words are *musika* (*music*), *informazio* (*information*), *eskola* (*school*), *definizio* (*definition*) and *didaktiko* (*didactic*).
- Basque words: Words that we are almost sure do not exist in any other language. The most searched for words in this category were *euskal* (*Basque* as adjective), *euskara* (*Basque language*), *hiztegi* (*dictionary*), *hezkuntza* (*education*), *hizkuntza* (*language*), *ariketa* (*exercise*) and various others.

For the overall measuring of the categories, a weighted average of them was made by taking into account the frequency of use of each category. To calculate these frequencies, we classified approximately the first 900 words (all that have a query frequency of over 100) out of the more than 70,000 words of the Elebila logs into one of the categories. This may not seem very much, but they do in fact account for more than 40% of the queries. The percentage of words and queries of each category is shown in Table 3.3.

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

Table 3.3: Frequency and query percentage of each category

The gain in precision produced by the language-filtering words for each category of word and overall is shown in Figure 3.2.

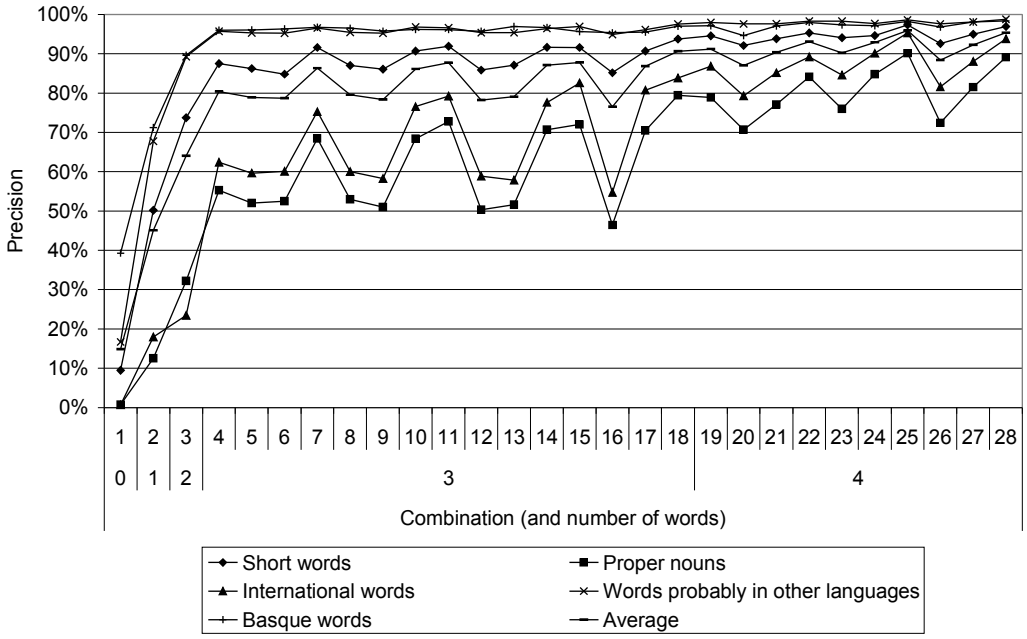


Figure 3.2: Gain in precision produced by the different language-filtering word combinations

The peaks and valleys of the graph provide us with hints as to the filtering properties of the last four words (*ez*, *du*, *bat*, and *ere*). All the valleys are combinations containing *du* and the highest peaks contain the word *ere*, so these two are, respectively, the worst and best words of the four for filtering. Between *ez* and *bat* there is not a big difference, although *ez* seems to behave a little better. These conclusions are logical: *du* is a word that is present in almost any text in a big language like French; *bat* is a word that, although not very frequent, exists in the language with the highest presence on the web, that is, English; and, as far as we know, *ez* and *ere* are not widely used words in at least three major languages, such as English, Spanish, and French, but *ere* is longer and hence yields better results.

The graph also shows that the average language-precision obtained without any language-filtering words is around 15%. This means that if we did not use language-filtering words and then filtered the results with a language classifier, we would get far fewer results.

3.3.2.4 Choosing the number of language-filtering words

In Figure 3.3 we put together the precision, recall, and F-measure of the different language-filtering word combinations.

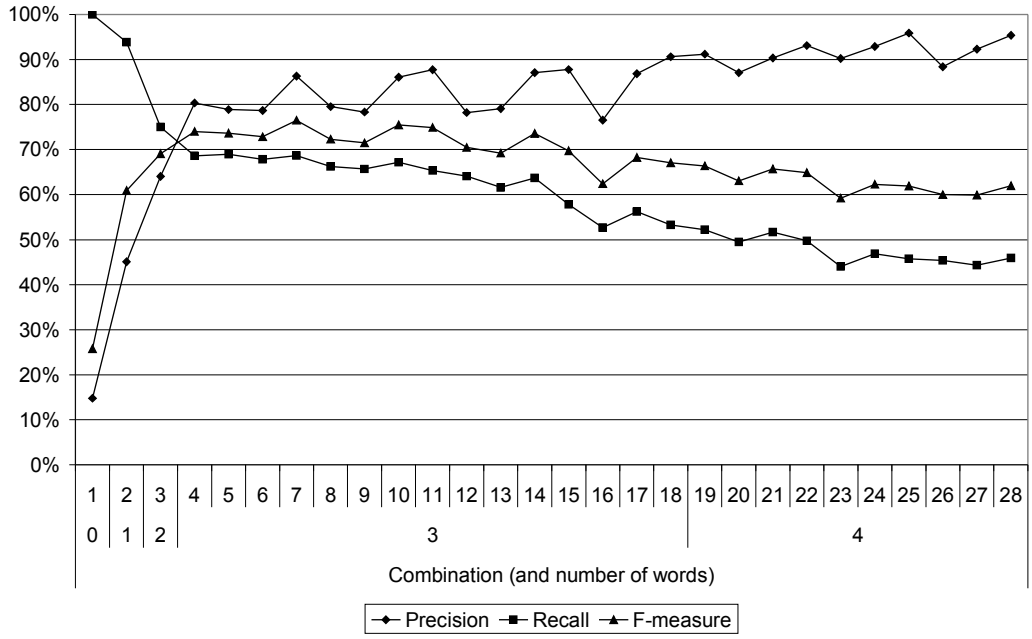


Figure 3.3: Precision, recall, and F-measure produced by the different language-filtering word combinations

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

The most appropriate step might be to keep a list of the most searched proper nouns and international words, and when someone wishes to search for one of them, use 4-word combinations, and otherwise use 3-word ones. Or we could also prioritize precision (showing the user results in other languages would give a poor image of the tool) instead of recall (the user would never know how many results he or she was missing) and normally use 4 words, and if the user is not happy with the results, or if there are too few of them, then he or she can be given the option of searching again by increasing the recall (using 3 words). This last option is the one chosen to implement our system.

3.3.3 Morphological query expansion

3.3.3.1 Most frequent inflections

In order to maximize the performance of morphological query expansion, it is important for the inflections used to be the most frequent ones. We must take into account that search engines allow, in the worst case, up to only 18 words in the queries; to this limitation we have to subtract three or four for the language-filtering words; so in some cases we can only send 14 morphologically generated words; and if the user has requested more than one word, we have to divide the inflections by the number of words requested.

So we have looked for these most frequent inflections in both of the aforementioned corpora. We took the most searched-for words of the Elebila logs and classified them into the five morphologically productive POS in Basque: nouns, proper nouns, place names, adjectives, and verbs (strictly speaking, place names are not a POS, but they are inflected differently from other proper nouns). Because of the non-tagged nature of the web corpus, the words chosen had to be non-ambiguous regarding their POS. Then we looked at the document-frequency that every different surface form of the words had in both corpora, and we assigned its inflection name to each of them. By grouping them by inflection name and ordering them by decreasing frequency, we produced a list of the most frequent inflections for each POS, both in the classical corpus and the web corpus. The lists of each corpus, although similar, reveal some differences between them. Since they were to be applied in a web search application, we chose the web corpus lists. The most frequent inflections of each POS are shown in Table 3.4.

3.3.3.2 Gain in recall

Once the most frequent inflections of each POS were known, we measured the increase in recall we would obtain for each POS by including 1, 2, 3… of the inflections in an OR. We performed this measurement using the same words as before. Again both of the aforementioned corpora were used, and we also looked at the increase in hit counts returned by Microsoft's Live Search API.

	Verb	Adjective	Noun	Proper noun	Place name
1	Participle / perfective aspect (*sortu*)	Nominative singular (*berria*)	Nominative indefinite (*hiztegi*)	Nominative (*Mikel*)	Nominative (*Egipto*)
2	Imperfective aspect (*sortzen*)	Nominative plural / Ergative singular (*berriak*)	Nominative singular (*hiztegia*)	Ergative (*Mikelek*)	Genitive locative (*Egiptoko*)
3	Verbal noun + *-ko* (*sortzeko*)	Nominative indefinite (*berri*)	Nominative plural / Ergative singular (*hiztegiak*)	Genitive (*Mikelen*)	Inessive (*Egipton*)
4	Unrealized aspect (*sortuko*)	Genitive plural (*berrien*)	Genitive locative singular (*hiztegiko*)	Dative (*Mikeli*)	Allative (*Egiptora*)
5	Short stem (*sor*)	Inessive singular (*berrian*)	Genitive singular (*hiztegiaren*)	Associative (*Mikelekin*)	Ablative (*Egiptotik*)
6	Verbal noun + Nominative singular (*sortzea*)	Genitive singular (*berriaren*)	Dative singular (*hiztegiari*)	Genitive + Nominative singular (*Mikelena*)	Genitive (*Egiptoren*)
7	Adjectival participle (*sortutako*)	Associative singular (*berriarekin*)	Inessive singular (*hiztegian*)	Partitive (*Mikelik*)	Dative (*Egiptori*)
8	Participle + Nominative singular (*sortua*)	Ergative indefinite (*berrik*)	Partitive (*hiztegirik*)	Genitive + Nominative Plural / Ergative singular (*Mikelenak*)	Genitive locative + Nominative singular (*Egiptokoa*)
9	Dynamic adverbial participle (*sortuz*)	Dative singular (*berriari*)	Instrumental indefinite (*hiztegiz*)	Instrumental (*Mikelez*)	Allative + Genitive locative (*Egiptorako*)
10	*-ta/-da* stative adverbial participle (*sortuta*)	Instrumental indefinite (*berriz*)	Instrumental singular (*hiztegiaz*)	Inessive (*Mikelengan*)	Associative (*Egiptorekin*)
11	Participle + Nominative plural / Ergative singular (*sortuak*)	Inessive indefinite (*berritan*)	Genitive singular + Nominative singular (*hiztegiarena*)		Genitive locative + Nominative plural / Ergative singular (*Egiptokoak*)
12	Verbal noun + Inessive singular (*sortzean*)	Sociative plural (*berriekin*)	Genitive plural (*hiztegien*)		Destinative (*Egiptorentzat*)
13	*-(r)ik* stative adverbial participle (*sorturik*)	Inessive plural (*berrietan*)	Sociative singular (*hiztegiarekin*)		Instrumental (*Egiptoz*)
14	Verbal noun + Allative singular (*sortzera*)	Genitive locative singular (*berriko*)	Ablative singular (*hiztegitik*)		Terminal allative (*Egiptoraino*)
15	Adjectival participle + Nominative plural / Ergative singular (*sortutakoak*)	Partitive (*berririk*)	Allative singular (*hiztegira*)		Genitive locative + Inessive singular (*Egiptokoan*)
16	Verbal noun (*sortze*)		Inessive plural (*hiztegietan*)		
17			Allative singular + Genitive locative (*hiztegirako*)		

Table 3.4: Most frequent inflections for each POS

In a couple of cases, there were inflections of a word that formed a word that also had another completely different sense. When this happened, the recall would go up abruptly and form peaks. These exceptional cases were removed and not taken into account for the measurement.

For the overall measure, we made a weighted average according to the frequency of use of each POS, calculated again by classifying the first 900 most searched words in the Elebila logs. The percentage of words and queries of each POS is shown in Table 3.5.

POS	Words		Queries	
Verb	12	1.66%	3,915	1.52%
Adjective	26	3.59%	16,708	6.49%
Noun	406	56.00%	169,244	65.78%
Proper noun	193	26.62%	39,618	15.40%
Place name	88	12.14%	27,819	10.81%
Total categorized	**725**	**1.01%**	**257,304**	**31.98%**

Table 3.5: Frequency and query percentage of each POS

The global recall obtained for each corpus is shown in Figure 3.4.

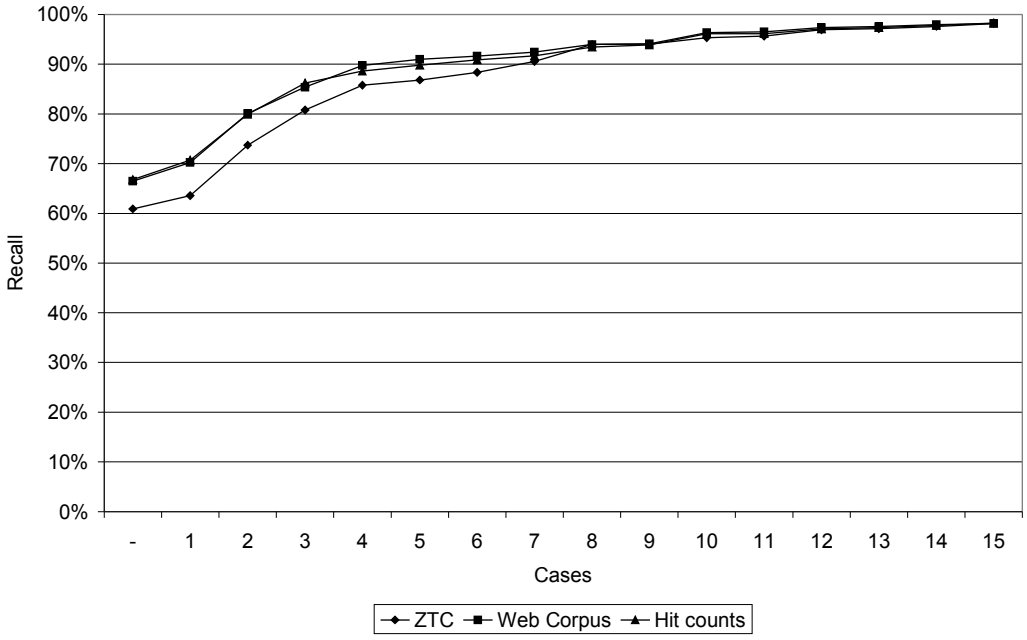


Figure 3.4: Evolution of recall produced by including more inflections in the queries

And the increase in recall obtained over the baseline can be seen in Figure 3.5.

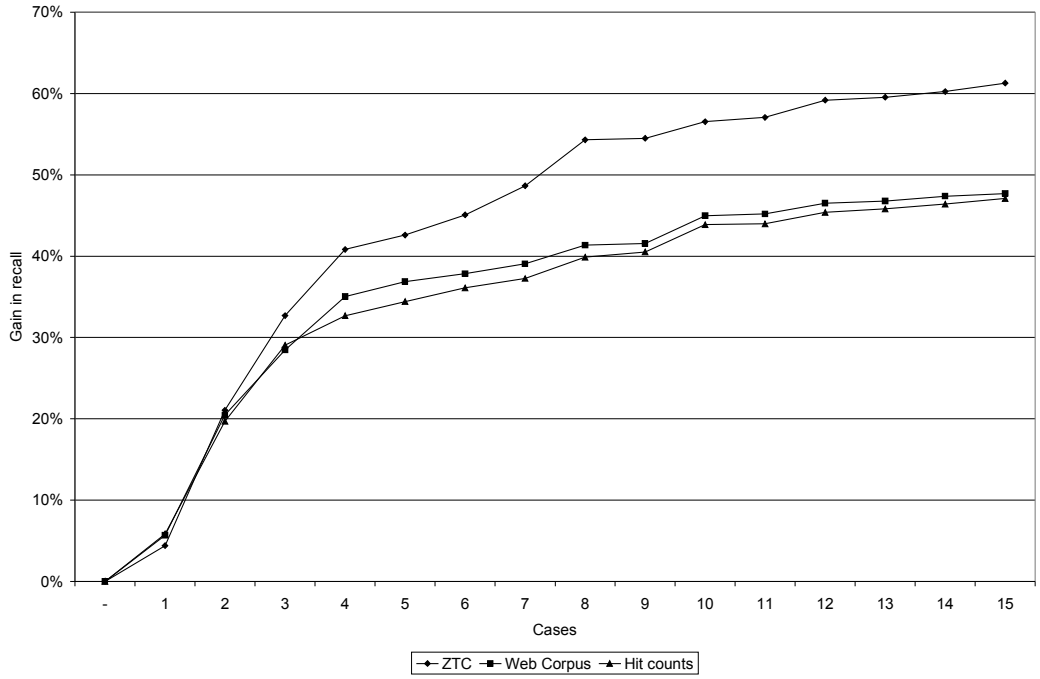


Figure 3.5: Gain in recall obtained by including more inflections in the queries

The high level of coincidence between the web corpus and hit counts series must be noted once again.

With as few as 4 inflections, an increase in recall of about 35% can be obtained, and with more inflections we can even achieve an increase of 47%. The recall obtained without applying morphological query expansion is only two thirds of what can be achieved by applying it. Thus the validity of the morphological query expansion method can be considered proven.

There is no decision to be taken as to the number of inflections that will be sent in an OR; as many as possible will be included, since there is no drawback in doing so. In the query, the word form entered by the user is sent first, and then the inflections sorted by decreasing case-frequency; in any case, the order does not seem to affect the results.

The gain shown in the chart is the weighted average of the gains obtained by each POS; the individual gains for the web corpus are shown in Figure 3.6.

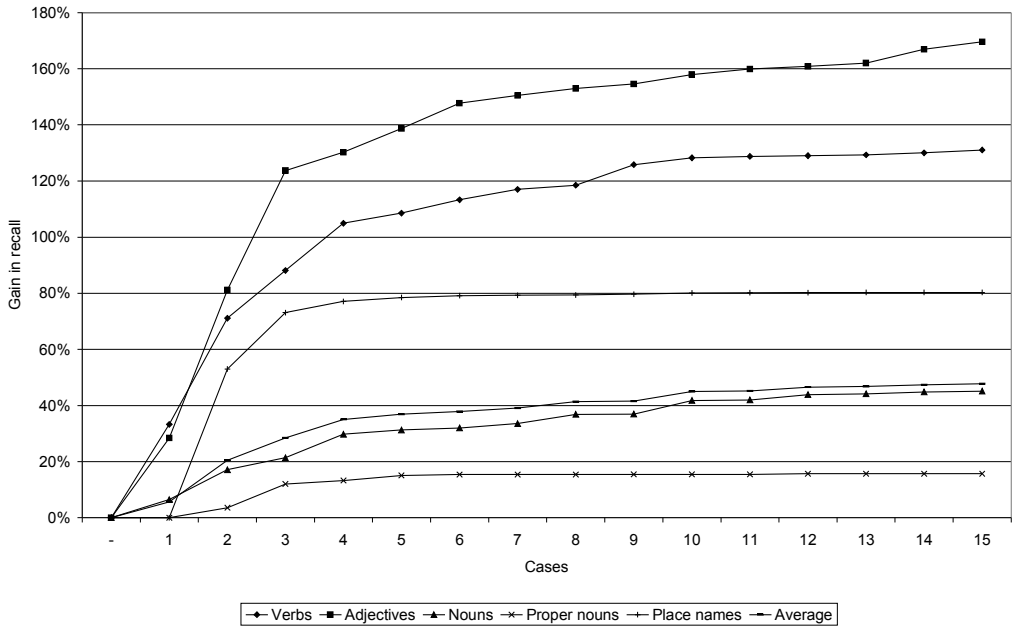


Figure 3.6: Gain in recall obtained in the Web corpus by including more inflections in the queries, for each POS

The differences between the various POS are obvious: some of them, namely verbs, adjectives, and place names, really benefit from the query expansion while the others (nouns and proper nouns) do so to a lesser extent. The reason for this is that in these POS the base form is more frequently used than in the others, and so the baseline (the recall obtained by querying for the base form) is already higher, thus leaving less room for improvement, as we can see in Figure 3.7.

Comment on Figures 3.6 and 3.7: by looking at the Elebila logs, we have noted that for verbs, adjectives, and nouns, more than one form of the word is used indistinctly when searching for the word, so the leftmost column shows an average of the recall of those inflections usually used, whereas place names and proper nouns are almost exclusively searched for using the nominative form, which is also the most frequent inflection, thus accounting for the non-existent improvement from the baseline or leftmost column to the next for proper nouns and place names.

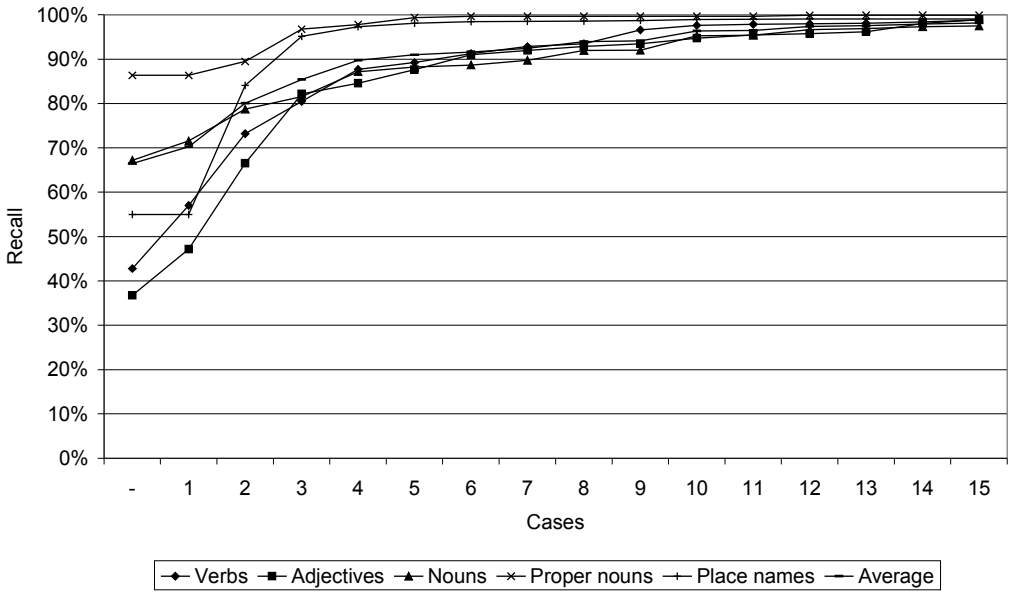


Figure 3.7: Recall obtained in the Web Corpus by including more inflections in the queries, for each POS

3.4 Additional problems and solutions

3.4.1 Post-query language filtering

Although the language-filtering words method ensures high language precision, a non-negligible number of pages that are not in Basque are still returned by the API (see Figure 3.3), and a search service should filter out these results before displaying them to the user. However, this filtering should not be done in terms of accepting or rejecting whole pages. In a web-as-corpus tool, it is not the whole page that we want to leave or filter out, but each occurrence of the search term. With the language-filtering words method, we ensure that almost all of the pages downloaded will have Basque in them, but not that they will be exclusively in Basque. There are many bilingual pages on the web and, due to the Basque language being co-official with Spanish in the Basque Autonomous Community and in some parts of the Charter Community of Navarre, there are a great many web pages and documents in both Spanish and Basque, *e.g.*, many local and regional government gazettes. So bilingual pages in which the search term can be in a non-Basque part are returned at times, and we need to show only the contexts of the words that are in a piece of text in Basque.

To solve this, we apply LangId to some context around each occurrence of the search term. Choosing the right length of the context was no small matter: 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. By performing some experiments we found that the best result was obtained if we tried initially with a fairly broad context; then, if LangId said that the text was not Basque, which would normally be due to parts in other languages being included, more attempts were made by reducing its length progressively until a minimal length was reached; the occurrence would be included in the result if any of the attempts said that the language was Basque.

3.4.2 Variant suggestion

We already said in the introduction that the late start of Basque standardization and the only very recent introduction of the Basque language in the educational process have been responsible for the fact that written production, the Internet included, is rich in errors, different versions or spellings of words, etc. So in our web-as-corpus service, one could ask for an incorrect word –without even knowing that there is a correct form with many more results– and find enough evidence to consider it correct.

There is another problem, caused by the fact that the web is not linguistically tagged. In linguistically tagged and manually disambiguated corpora, different variants of a word (archaic spellings, common errors, etc.) or even typing errors have their correct lemmas assigned, so searching for a certain lemma would also return occurrences of the variants, but not in our tool.

We solve these problems by means of variant suggestion. Expanding the query using variants of the search term to improve the results has been suggested in the literature, either by automatic expansion (Spärck Jones and Tait, 1984) or interactive suggestion (Belkin, 2000). The expansion is usually done with synonyms obtained from a thesaurus or related words extracted by statistical measures over corpora, relevance feedback, etc. In our case, the query is not automatically expanded with variants; the user is informed about the existence of the variants and given the option of looking for them at a simple click. And the variants we suggest are aimed at solving the problems mentioned above: known variants, common errors, deprecated forms, and archaic spellings. This implementation makes use of the **EDBL**, a lexical database developed by the IXA Group of the UPV/EHU-University of the Basque Country and used by all

the linguistic tools made for Basque (Aduriz et al., 1998). This database links each word to its linguistic variants (common errors, archaic spellings, deprecated words, etc.). So if the terms entered by the user have some variant or correct form in the EDBL, they are suggested and can be looked for at a click. If, for example, we are interested in the collocations or terms in which the noun *jarduera* (*activity*) is the head, the system offers the possibility of also retrieving the occurrences of *iharduera*, a now deprecated spelling widely used until 1998, and vice versa.

3.4.3 Ambiguous word forms

Another problem coming from the non-linguistically-annotated nature of the web is that there are cases where an inflected form of a word forms another completely different word. For example, the dative form of the lemma *pilota* (*ball*) is *pilotari*, which also means *pelota player*. If this form is sent in the morphologically expanded query, many pages that contain the inflected form of the word might appear, but which are completely unrelated to the original word that we wanted to look for. In our example, a search for the lemma *pilota* will logically 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*. To avoid this, before an inflection of a word is included in the morphological query expansion, it can be looked up in a dictionary to make sure it does not exist as another lemma.

3.5 Implementation as a web service

With the methodology explained in this chapter, we have implemented a web service that allows the Internet to be queried as a Basque corpus, called **CorpEus** (Leturia et al., 2007a). It makes use of the APIs of search engines with language-filtering words and morphological query expansion, suggests variants of words, more than one search term can be entered (with the possibility of performing an exact phrase search by enclosing them in double quotes), shows KWiCs from pages in many formats (HTML, XML, RSS, RDF, TXT, DBF, PDF, DOC, RTF, PPT, PPS, and XLS), concordances are only shown if LangId says the context is Basque, KWiCs can be ordered following different criteria, ordering is made on the fly as they come in, the possible analysis of lemma and POS are shown in the KWiCs (with different colours for only one analysis, various ambiguous analysis or no analysis), and, finally, shows different charts with counts of word forms, possible lemmas or POS, word before, word after, etc.

4 filter words are used by default, thus the language-precision usually obtained in a search in Basque is raised from 15% to around 95%. The loss in recall is around 50%, but when few or unsatisfactory results are returned, fewer filter words can be used. With 3 words we would achieve a better precision-recall compromise (86-87% precision and 68-65% recall). And morphological query expansion obtains an increase in recall of 47% on average.

3.6 Conclusions

The work we carried out that has been explained in this chapter shows that applying the combination of some NLP techniques (morphological query expansion and language-filtering words, alongside some other small improvements and tweaks) to the APIs of search engines is a valid method for building a cost-effective web-as-corpus tool for Basque that significantly improves the performance of major search engines. This has been proven both theoretically (by performing corpora-based precision and recall measurements for Basque) and practically (by building and successfully launching the Basque web-as-corpus tool CorpEus).

We also believe that the precision and recall data and frequency lists obtained in this work will constitute very valuable documentation for future IR projects for Basque.

Moreover, we are of the opinion that the steps followed here for specifying the implementation details of the methodology and for measuring the improvements obtained with it, could be very valuable for developing similar tools for other languages with similar features and problems (morphologically rich and/or minority languages), of which there are several and that are in much need of such tools. Currently, major search engines cover only about forty languages (the most widely spoken ones) appropriately, while the tools needed for implementing the methodology described in the chapter (N-gram based language detection tools and lexical processing tools) exist for many others, even regional and minority languages.

4 Corpus cleaning

In the chapter dedicated to the state of the art, we said that when pages are downloaded for the web to be put into a corpus –a corpus of any kind and built by any method–, they necessarily have to undergo some cleaning and filtering stages (namely length filtering, language filtering, spam filtering, porn filtering, boilerplate removal, near-duplicate detection and containment detection), and described how these have been done in projects of this kind.

We have also implemented these filters in all our corpus collection systems: the search engine system and the crawling system to collect general corpora described in Chapter 5, the system developed to collect specialized corpora described in Chapter 6 and the system developed to collect comparable corpora described in Chapter 7. We explain how we deal with each of these problems in this chapter.

4.1 Language filtering

In any work to collect Basque pages from the web, language filtering after downloading is a must. As we have already stated, search engines do not offer language filtering for Basque, and even if we apply the language-filtering words method to the query, its results, although good, are not perfect. And in the crawling method it is absolutely necessary: a page that is in Basque can have links to pages that are in other languages; we have to detect whether the downloaded pages are in Basque not only to decide whether to include them in the corpus, but also to avoid queuing the links found in pages that are not in Basque, otherwise we would be downloading too many pages for no purpose.

For the language detection, 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. This does not mean that we remove every non-Basque paragraph; if we did, we might also remove some short quotes important for the understanding of a text. As our intention is to eliminate sufficiently large amounts of noise, we remove sequences of non-Basque paragraphs that exceed 10% of the length of the document, and individual paragraphs only if the total amount of the language of the paragraph in the document exceeds 40%.

4.2 Length filtering

Filtering documents by length and leaving out pages that are too small (usually error pages or with not much textual content) or too large (usually spam or lists) is a proven and effective way of improving the quality of the corpus, as Fletcher (2004) showed. But unlike him and others, who apply a length filter based on the size of the downloaded file (they keep those which are over 5KB and less than 200KB), we reject documents the length of which after conversion to plain text is under 1,000 characters or over 100,000 characters. That is to say, we remove documents that are roughly shorter than half a page (not enough continuous text to be interesting) or longer than 50 pages (likely to be spam or list). Although we might be missing many interesting documents by this maximum size filter, we prioritize precision over recall.

4.3 Spam and porn filtering

Although it is necessary for other languages to implement such filters in any web-corpus collecting processes, 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.

4.4 Boilerplate removal

Boilerplate removing is essential to avoid ugly concordances and skewed frequencies and help the other filters. The tool we use for this is **Kimatu** (Saralegi and Leturia, 2007), a tool we built for the aforementioned CleanEval competition (Baroni et al., 2008). Of the 10 systems that took part in it, ours scored second in both text-only (TO) and text and markup (TM) scoring systems and on the average score, in all of them very close to the winners: in TM our system obtained a precision of 65.3% whereas the best system in this category, Htmcleaner (Girardi, 2007) achieved 65.6%; in TO we got 83.4% and the best system (Marek et al., 2007) scored 84.1%; and on average, we got 74.3% whereas the winners (Marek et al., 2007) obtained 74.7%.

The CleanEval competition is one of the references for boilerplate cleaning. The other one is the previous and simpler BTE method (Finn et al., 2001), based on tag density, and which is also used in many corpus collection projects. Ferraresi et al. (2008) and Pomikálek et al. (2009) reported that the BTE algorithm performed better than any CleanEval system. But Pomikálek recognized in his Ph.D. thesis (Pomikálek,

2011) that such a claim was made based on his own work and that further research by himself revealed that BTE did so well because it obtains good recall, but not such good precision, and the CleanEval collection and scoring method favoured recall over precision.

Thus, by employing the system that scored second –very close to the first– in the reference competition for cleaning web pages, we believe we are using state-of-the-art techniques for boilerplate removal.

The process that Kimatu uses is the following:

- Web pages are converted to XHTML using HTML Tidy (Raggett, 1998).
- The page is divided into blocks, which are sets of continuous paragraph-level elements with the same tag and class attribute.
- Blocks are assigned a content relevance ratio, which is a weighted combination of block length, average sentence length, punctuation signs ratio, and link density.
- Blocks that are above a threshold for this ratio are considered content.
- A bootstrapping process is applied to regain other shorter blocks that did not pass the first threshold but whose content relevance ratio is above another threshold, if they are next to a content block (to detect titles, etc.).
- Some heuristics are applied to detect repeated quotes in fora, blogs, etc.

4.5 Near-duplicate detection

In our corpus collecting processes we have included a near-duplicate detection module based on Broder's shingling, fingerprinting, and supershingling algorithm (Broder, 2000).

We explained in the state of the art chapter that this algorithm was developed with IR objectives in mind and aimed at detecting almost exact documents, but that in the context of corpus building it is interesting to detect smaller similarities. Broder's algorithm can be adapted to detect these smaller similarities.

Broder's technique consists of taking all the **shingles** (sequences of n tokens of c characters) of a document and fingerprinting them by using Rabin's fingerprints (Rabin, 1981). If the set of shingles of documents A and B are called respectively S_A and S_B, then the resemblance $r(A,B)$ of the two is defined as

$$r(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \tag{4.1}$$

According to Broder (*ibid.*), "experiments seem to indicate that high resemblance (that is, close to 1) captures well the informal notion of "near-duplicate" or "roughly the same"."

Storing all the shingles of the documents and calculating the resemblance between every pair of documents is infeasible on a large scale. The optimization contributed by Broder is based on the fact that we do not need to calculate the resemblances, only to detect high resemblances. For achieving that, Broder orders these fingerprinted shingles (which are numbers) from smallest to largest and takes the first t of them (t being the multiplication of another two numbers k and s), which will be the **sketch** of the document. Afterwards k groups of s elements are made with the sketch, and these k elements (which are called **supershingles**) are again fingerprinted. For the detection of near-duplicate documents, it suffices to keep these few fingerprinted supershingles, that is, a few numbers, and a coincidence in few of them is enough to ensure high resemblance. This has been proven because the probability that two documents A and B, having a resemblance ρ and whose shingles have been grouped into k groups of s elements, share more than r groups is given by the following formula:

$$P_{k,s,r} = \sum_{r \le i \le k} \binom{k}{i} \rho^{s \cdot i} (1 - \rho^s)^{k-i} \tag{4.2}$$

Broder found that, for appropriate choices for k, s, and r, the probability function behaves as a very sharp high-band pass filter even for small values of k and r. For example, for values of k, s, and r of 6, 14 and 2, the probability function shows a graph like the one shown in fig 4.1.

In the graph we see that, for a k value of 6 (6 supershingles), if two documents have a resemblance of 95%, the probability that they share 2 or more supershingles is 90%, whereas it is very improbable if the resemblance is less than 80%. This means that it is enough to store 6 numbers for each document and look for coincidences in 2 of them to be almost sure that they have a great resemblance. The efficiency in terms of storage and calculation is remarkable, and it can be used to detect almost duplicates even on a web scale.

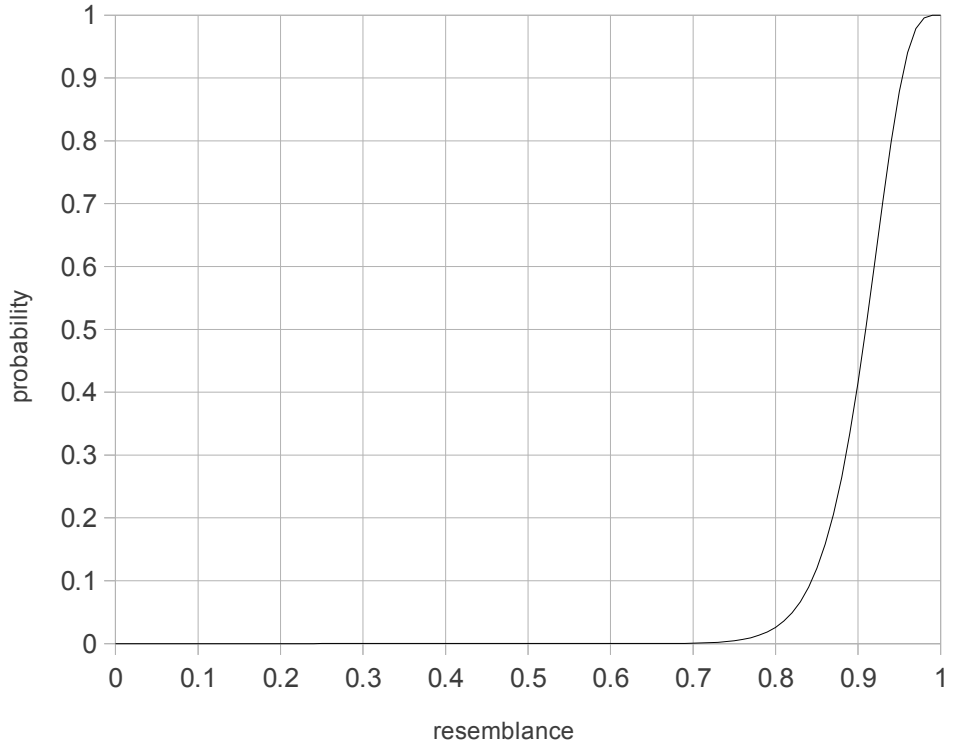


Figure 4.1: Probability of coincidence of 2 supershingles of length 14 (out of 6) for any given resemblance

However, we can also find suitable values of *k*, *s*, and *r* for our case, that is, to detect smaller resemblances. For example, for values of 20, 5 and 1, the probability graph obtained is shown in fig 4.2.

We can observe by looking at the graph that, using 20 supershingles, if two documents have a resemblance of more than 60%, the probability that they share 1 or more supershingles is very high, but if it is less than 40% it is highly improbable. The cut-off is not so vertical, which means that documents with a resemblance of between 40-60% will sometimes be detected and other times not. However, the choice of the threshold for our resemblance filter was not absolute; in fact, we could have put it at 60%, or 30%, depending on what we value more: including some duplicate content in the corpus or discarding some content without any particular reason. Thus, a cut-off at roughly 50% is enough for our case.

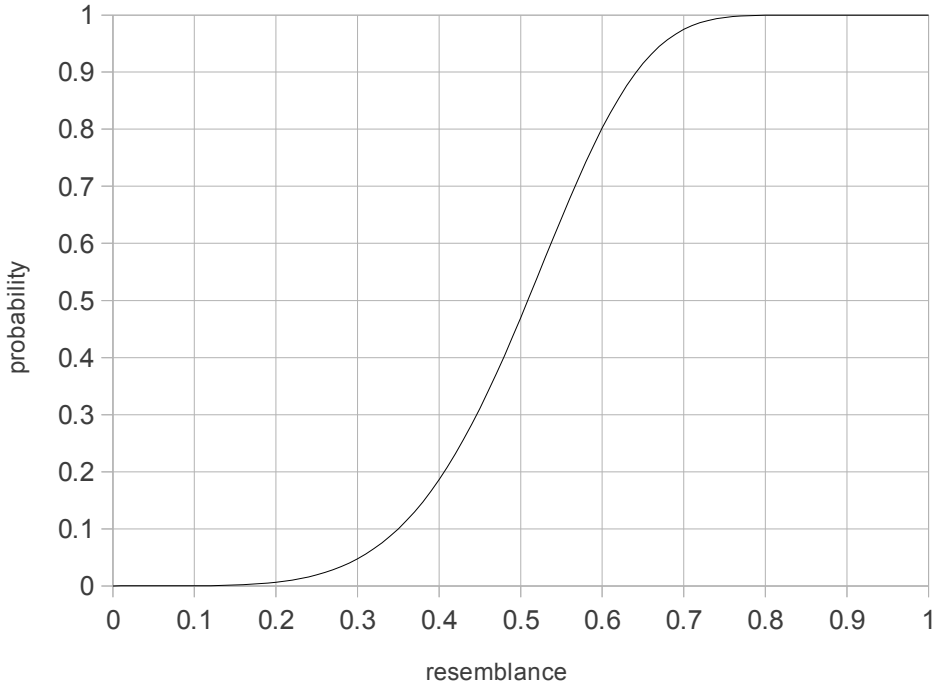


Figure 4.2: Probability of coincidence of 1 supershingle of length 5 (out of 20) for any given resemblance

This is the filter we have implemented in our system. We store 20 supershingles per document and look for pairs of documents where any of them coincide –which is very cheap computationally speaking–, and when they do, we can be sure that they share about half of the content.

We have not implemented the detection of exact duplicates by means of hashing techniques that some people use, because they are removed all the same, either by the length filter (they are usually error pages) or the near-duplicate filter.

4.6 Containment detection

In our downloading process, we included a containment detection method also based on Broder's previous works (1997), because although it is not as optimized as near-duplicate detection, it is possible to use it for small- and medium-sized corpora building, which is our case. We have to keep more shingles of each document, but we are again able to detect whether more than half of a document is contained inside another one.

If the set of shingles of documents A and B are called respectively S_A and S_B, then the containment $c(A,B)$ of A in B is defined as

$$c(A,B)=\frac{|S_A \cap S_B|}{|S_A|} \tag{4.3}$$

Again, it is impractical to store all the shingles of a document. Instead, for each document we will only keep a sketch L_i, composed of those shingles that are divisible by 2^i, i being dependent on the document's length l, such that

$$100*2^i \leq l < 100*2^{i+1} \tag{4.4}$$

This way, we can obtain an estimate of the containment with the following formula:

$$c(A,B) \simeq \frac{|L_i(A) \cap L_i(B)|}{|L_i(A)|} \tag{4.5}$$

The i number is taken from the longer document. If the i number was different for the two documents, then we will have to calculate the L_i of the smaller, which is simple: L_{i+1} is calculated by taking from L_i only those that are divisible by 2^{i+1}.

So, for containment detection this is the method we use. For each document only between a hundred and two hundred numbers are stored. And we do not have to calculate the containment for every pair. When we need to know if one particular document is to go into the corpus, we look at the ones that have already been included to see whether there is any with which at least one of the stored numbers coincides, which is fast if we have them indexed. And then we calculate the resemblance for them only, which are usually not so many.

5 Obtaining a large general Basque corpus using the web as source

We have already said that solutions to query the web live for linguistic evidence like the one we have described in Chapter 3, although still interesting for some cases –the case of Basque being one of them–, have in recent years been gradually replaced by the approach that makes use of the web to obtain from it texts that will be used to build classic corpora. Using this approach, gigaword corpora (billions of words) have been collected for many languages (English, German, Italian...).

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 them –or even to previous generation ones like the 100 million-words BNC. The largest corpus in Basque contains just about 25 million words. And 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 chapter we present the research we have done to build a large general corpus of Basque from the web. We have tested and evaluated which of the two methods mentioned in the literature, that is, by crawling or by using search engines (see Subsection 2.2.1 of the state of the art), best suits Basque, in terms of parameters such as speed, cost, size or quality.

As we said in that subsection, it seems to be sufficiently widely accepted that the crawling method can obtain much larger corpora than the search engine method. However, it is unclear whether the same will happen with Basque. Search engines do not treat Basque properly, and this could affect the results obtainable by the search engine method. Or, since the Basque web is not as large as that of other languages, maybe the search engine method would suffice to obtain corpora as large as by means of crawling.

Thus, the main objective of the research described in this chapter was to build a general Basque corpus that was as large as possible (comparable with 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. We describe here the work carried out and the results of the evaluation of the different methods, which were previously published in (Leturia, 2012).

5.1 Search engine method

5.1.1 Methodology description

The main references for building large general corpora using search engines are Sharoff's (2006) and Kilgarriff et al.'s (2010) works. Sharoff uses the search engine method to build 100-200 million-word corpora for 6 languages (Chinese, English, German, Romanian, Ukrainian, and Russian) and Kilgarriff et al. for another 8 (Dutch, Indonesian, Norwegian, Swedish, Thai, Vietnamese, Hindi, and Telugu). The methodology we will follow will be similar to theirs. We will explain the particularities or details of our implementation here.

Scannell (2007) and Ghani, Jones, and Mladenić (2003) also used search engines to build corpora for minority languages, so, theoretically, these works are closer to ours. However, the sizes of the corpora they were aiming at and obtained are quite small, and we are interested in obtaining large corpora, so we decided to take the works referred in the previous paragraph as a base instead.

Basically, the method is the following: starting from a list of seed words, combinations of these are sent to the APIs of search engines and the resulting pages are downloaded.

Regarding the list of seed words, Sharoff uses one 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.). They obtained the most-frequent-words lists from reference corpora like BNC. Kilgarriff et al. did not have corpora for their languages available, so they downloaded the Wikipedia dumps (Wikimedia Foundation, 2001) of those languages and extracted the most frequent words from there, but they used those whose frequency ranking was between 1,000 and 6,000. The words they use are surface forms, not lemmas.

In our case, we also took the list of frequent words from XX. mendeko Euskararen Corpusa (see Section 1.3), and we removed those Sharoff described as non-desired. 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 and Kilgarriff et al.'s regarding lemmatization. As Sharoff points out, general search engines do not perform lemmatization, so their seed word lists are formed by surface forms. But in our case, we use a list of lemmas and apply the morphological query expansion we described in Subsection 3.2.1 when calling the API of the search engine, because this method has proven to be effective to obtain lemma-based searches.

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. Kilgarriff et al. only used words longer than 5 characters to avoid the possibility of their existing in other 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. What we do is to apply the technique of the language-filtering words described in Subsection 3.2.2, like Sharoff for Ukrainian, by appending the most frequent words in Basque to the query (*... AND eta AND da AND ez AND ere*), because this is the most effective method for obtaining results in Basque alone from search engines, although it means a loss in recall.

In his work, Sharoff also suggests that more than 500 seed words can be used. And so do Kilgarriff et al., who use a seed list of 5,000 words. It would indeed be interesting to test the effect of the length of the seed word list on the corpus collection process and assess the corpora obtained; so, we tried with seed word lists of 500, 1,000, 2,000, 5,000 and 10,000 words.

In Sharoff's 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. According to him, “the presence of one-two com-

mon words also does not guarantee an instance of connected text" and "using queries longer than four words the number of pages returned gets smaller, so that the result will not qualify as a random snapshot of the Internet." 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). Kilgarriff et al. conducted experiments to see which could be the optimal combination length for each language (according to their criterion, optimal is long enough to avoid documents in other languages from being returned and short enough to get at least 10 hit counts in 90% of the queries), and they sent 2-, 3- or 4-word combinations to the API, depending on the language.

We were also interested in seeing the effect of combination length. 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 (Kilgarriff et al. used Yahoo's and Bing's). From the results returned by the API, Sharoff and Kilgarriff et al. download 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 word 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.

Finally, regarding the number of queries, Sharoff made 5,000 while Kilgarriff et al. 30,000. We made 12,000 queries for each variation of seed word list length and combination length.

5.1.2 Quantitative evaluation

5.1.2.1 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 5.1. We will now analyse various aspects of the corpora obtained.

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

Table 5.1: Sizes of the collected corpora for each length of seed word list

First we will take a look at the hit counts returned by each query to the API (Figure 5.1). To compensate for the great fluctuations in hit counts and produce an observable graphic, the average hit counts of the last 200 queries are shown. We can observe that the larger the seed word list, the considerably smaller the hit counts returned are (the graph is in logarithmic scale). This is completely logical, because in a longer seed word list, the words will be less frequent and it is normal for them to yield fewer results.

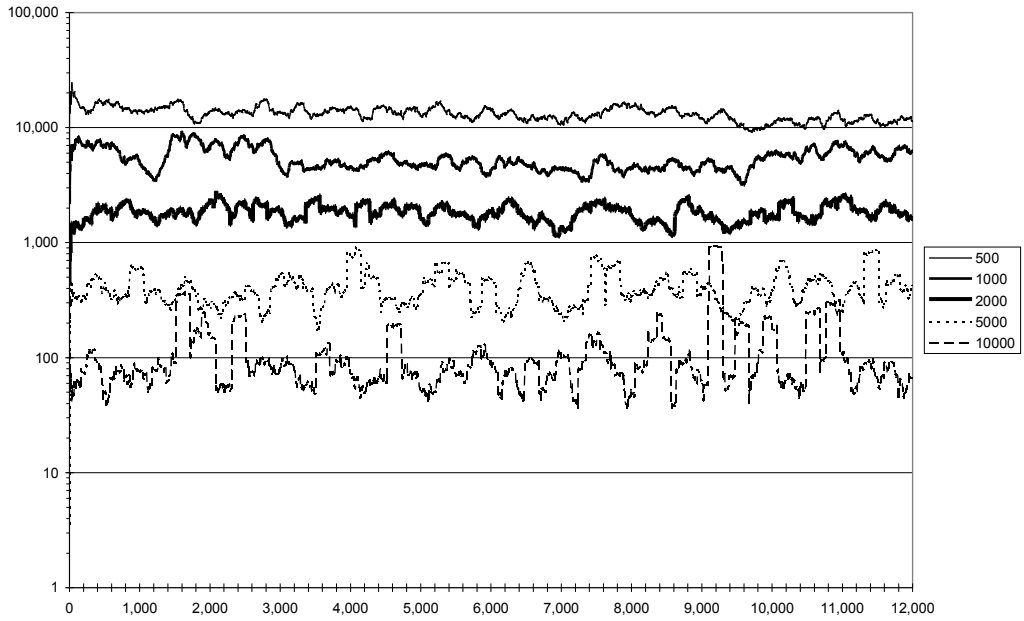


Figure 5.1: Hit counts returned by the search engine APIs for each length of seed word list

However, this does not necessarily imply that the corpora obtained with longer seed word lists will be smaller. There are other factors that affect the size, such as the number of duplicate URLs returned in the sum of the queries, the length of the pages, etc. Besides, even though the APIs report a big hit count, we only download the first 50 pages, so smaller hit counts might result in smaller corpora only for hit counts below 50; and the only seed word list that obtains less than 50 results is often the 10,000-word one. So, hit counts are not expected to be an important factor in corpus size, except for probably the 10,000-word list corpus.

This is corroborated by the sizes of the corpora (Table 5.1). There we observe that the smaller the seed word list we use, the smaller the resulting corpora are, although, as we have seen in the previous figure, 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-word list, the words are so rare that very few pages contain 3 of them, and very often less than 50 results (even 0) are returned and so the corpus obtained is smaller, and will probably be likewise for seed word 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 Figure 5.2. 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 one intended for edition (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 document lines get all messed up, header and footer text are repeated for every page and inserted into other paragraphs... As Fletcher (2007a) 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." For all these reasons, 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.

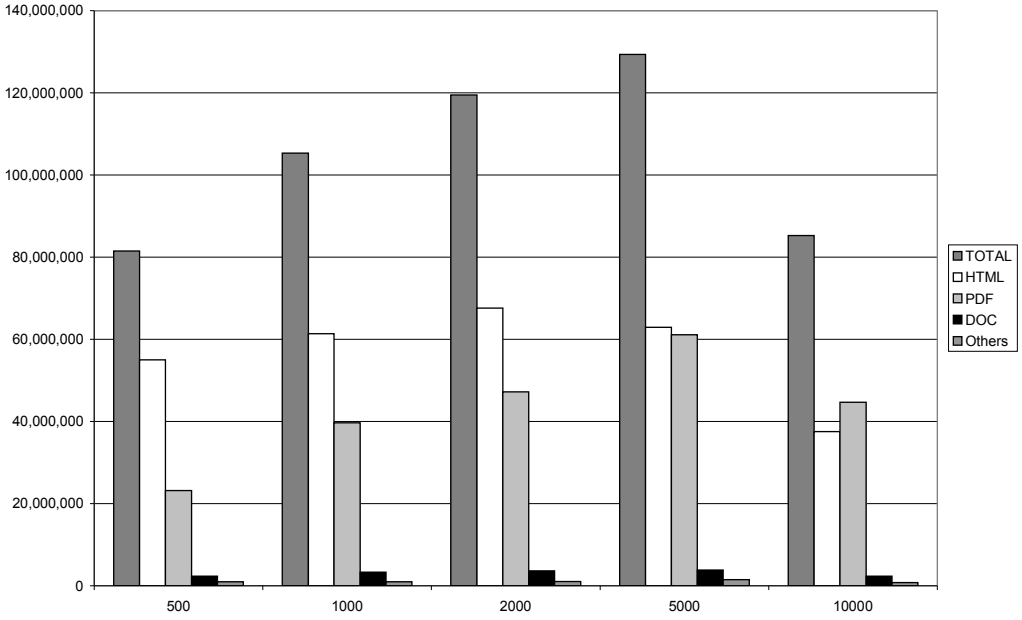


Figure 5.2: Size in words of the corpora obtained for each seed word list length,total and by type of document

Another clear difference in the corpora is the average size of the documents (Table 5.1), 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 shown in Table 5.2. 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.

Seed word list length	Different websites
500	4,452
1,000	4,849
2,000	4,675
5,000	4,398
10,000	3,021

Table 5.2: Website variety of the collected corpora for each seed word list length

Finally, there is one more point worth mentioning: using the search engine 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 over again and are discarded by the duplicates detection, so the bigger the corpus is, the lower its growth rate becomes, as the graph in Figure 5.3 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.

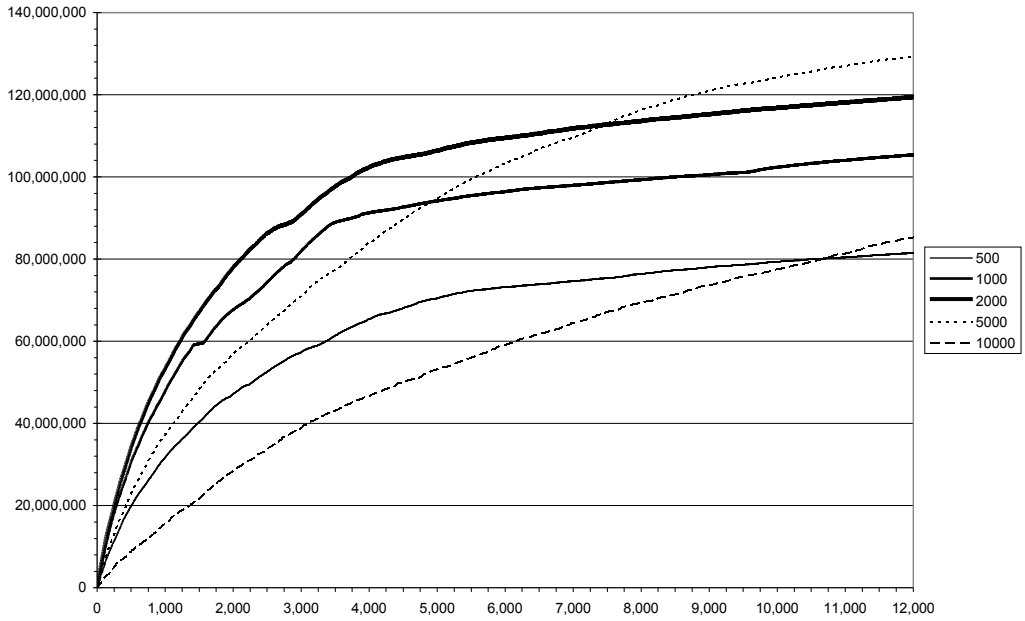


Figure 5.3: Growth rate of the corpora obtained for each seed word list length

So, it is not clear whether by using the search engine 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 of 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.

5.1.2.2 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 5.3. Again, we will take a look at some features of them.

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

Table 5.3: Sizes of the collected corpora for each combination length

If we look at the hit counts (Figure 5.4), we can see that the longest combinations yield the fewest hit counts. This is completely normal, the more words we send to the search engine, the fewer pages there will be that contain them all.

But again, there is not a direct correlation between hit counts 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 (Figure 5.5).

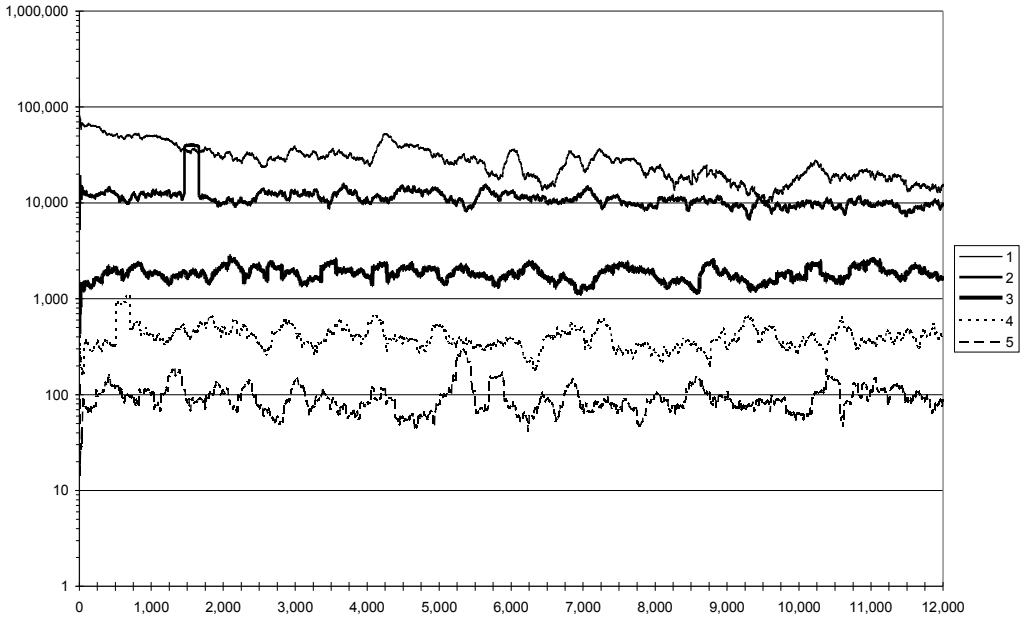


Figure 5.4: Hit counts returned by the search engine APIs for each combination length

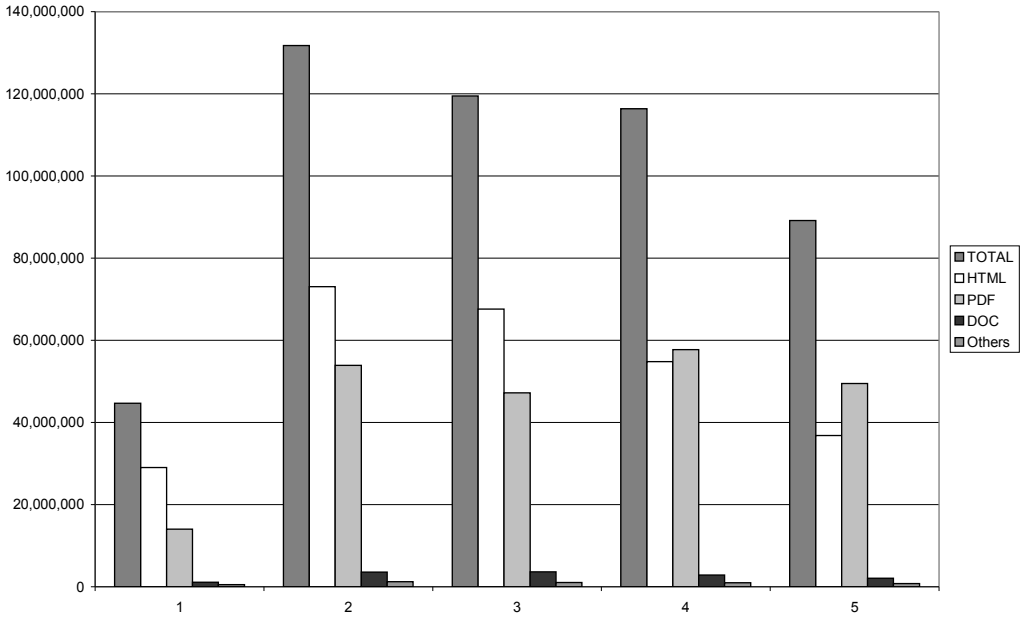


Figure 5.5: Size in words of the corpora obtained for each combination length, total and by type of document

Regarding the document length (Table 5.3), the same phenomenon as with the seed list length happens: for longer combinations, the size of the documents grows. But the website variety (Table 5.4) in this case falls for combinations longer than 2. And the dramatic fall in the growth rate also occurs in all these cases (Figure 5.6)

Combination length	Different websites
1	4,089
2	6,095
3	4,675
4	3,824
5	2,547

Table 5.4: Website variety of the collected corpora for each combination length

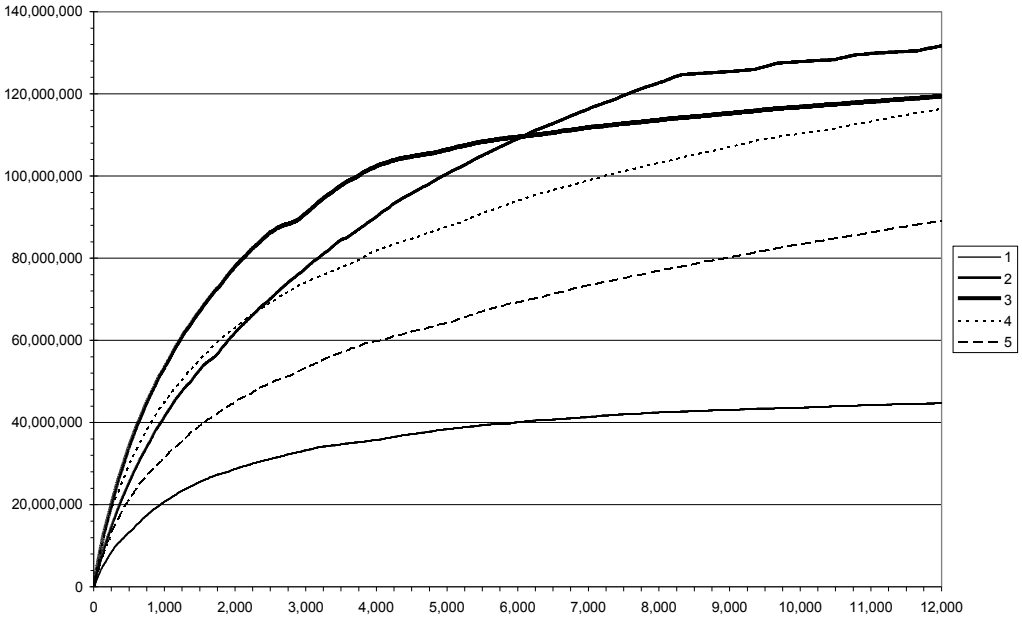


Figure 5.6: Growth rate of the corpora obtained for each combination length

5.2 Crawling method

5.2.1 Methodology description

Among the many projects that build large general corpora by crawling, in Subsection 2.2.1 of the state of the art we mentioned some of them: the WaCky! initiative (Baroni et al., 2009), the WebCorp Linguist's Search Engine (Kehoe and Gee, 2007), the COW

corpora (Schäfer and Bildhauer, 2012), ClueWeb09 (Callan et al., 2009) and BiWeC (Pomikálek et al., 2009). Our crawling for a Basque large general corpus is based on the techniques used by these authors.

The crawling method needs a list of seed URLs as a starting point. The corpora collected by the WaCky! initiative (Baroni et al., 2009) and the BiWeC corpus (Pomikálek 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). The COW corpora (Schäfer and Bildhauer, 2012) also use seeds obtained from queries to search engines. ClueWeb09 (Callan et al., 2009) used many seeds, some were URLs obtained from a previous crawl and others were obtained from search engines; the words to make the queries were obtained from AOL (Sogou Inc., 2004), Yahoo! (Yahoo! Inc., 1994) and Sogou (Sogou Inc., 2004) query logs, and also from categories of the Open Directory Project (ODP, 1998).

In our case, we took the almost 1,300 URLs of the *Euskara* (Basque language) section of the Open Directory Project (ODP, 1998). 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, it is almost certain that ultimately we could reach the whole SCC and OUT (see Subsection 2.2.1 from the state of the art). However, this is one of the points that we wanted to test in our experiments, and we will see if it proves correct in the results.

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

However, there is one point worth noting. Normally, web corpus building projects crawl a very large corpus, much larger than the final expected size, and do the filtering and cleaning afterwards, keeping only a part of the initial downloaded raw corpus. This is the case of most of the projects we referred to above, the only exception we are aware of is BiWeC (Pomikálek et al., 2009) –they implement the size filter in the crawling.

The rejected document proportion is indeed high: Schäfer and Bildhauer (2013) report that 94% of the downloaded content is discarded in the cleaning and filtering process in a crawl for a German corpus limited to the .de domain! This means that we have to download 20 times larger a corpus than the intended final size.

This rejection proportion is sure to get much higher in the case of Basque for two reasons. The first is that the mentioned crawl that obtained a reject proportion of 94% was limited to a domain (.de) in which we can expect that many of the pages will be in German. Since the Basque Country is not an independent state, we do not have a NTLD (National Top-Level Domain) and therefore cannot limit the crawl to a domain where most of the documents would be in the Basque language. However, this might change in the future: the ICANN or Internet Corporation for Assigned Names and Numbers (ICANN, 1998) –the association that grants TLDs or Top-Level Domains– accepted in 2005 to include .cat as a top-level domain not for pages from Catalonia but for pages in Catalan or about the Catalan language and culture; following that line, a petition was made to the ICANN for a .eus TLD for pages in Basque or about the Basque language and culture, which was accepted in June 2013; since there are now more than 66,000 .cat domains, even if more modest success can be expected for .eus due to the smaller number of Basque speakers, we can expect that when the .eus domains get going, it will be possible to limit a crawl for texts in Basque to that domain and expect a high percentage of them to be in Basque.

The other is that, in the above-mentioned case, the target language (German) is a major language and we can expect that not very many of the links go to pages in other languages, or at least not as many as in Basque pages. This was proved in a Google report that conducted an analysis of language connections on the web, by using the proportions of external links of pages in a language that pointed to a page in another language. That report showed the proportion of links from any language towards English (which was quite large for any case, as would be expected) and highlighted other proportions that were larger than expected from random distribution, which were usually explained by people and communities speaking both languages. In the case of German, for example, 17% of the links coming from pages in German point to pages in English, and there are no further outstanding relations. Whereas for Basque, 24% of pages in Basque point to pages in English and the report showed that another 25% point to pages in Spanish and another 2% to pages in French. The report does not

show all the data for every language pair, but any other connection not mentioned there is supposed to be low. Therefore, although we cannot know the exact percentages of introversion of Basque and German, we can affirm with a fairly high degree of confidence that the proportion of outward links to other languages from pages in Basque will at least double –or even triple: 17% vs. 51%– that from pages in German. Thus, since we cannot limit the crawl to a national or cultural domain, a crawl that started from seed pages in Basque would diverge to pages in other languages much sooner than from German –or any other major language with an NTLD–, and since the links are followed recursively, the diversion would grow exponentially.

These two reasons would make a blind crawl and subsequent filtering extremely inefficient for our case. Disk storage is a resource we do not have to spare, and we cannot afford to download a corpus that would be, say, a hundred times larger than the final intended size. That is why we do not use existing crawling software –Nutch, Heritrix...– as is normally done. We have implemented a crawler that applies all the cleaning and filtering stages to a page at the moment it is downloaded, and only queues the links found there if the page is in Basque. This process is slower, but it makes the most of disk storage and it can be left ongoing as long as one wants to, until the target corpus size is reached.

5.2.2 Quantitative evaluation

With the crawling method, and starting with the almost 1,300 seed URLs from the *Euskara* section of the ODP, we have so far queued 29,419,985 links, tried to download 2,483,284 of them, successfully downloaded 2,419,690 pages (the rest were not available at the time, or had been discontinued, or gave errors) and included 271,058 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, as we have already said), 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 210 million. Its features can be seen in Table 5.5.

Documents	Words	Words per document	PDFs	Different websites
271,058	210,243,505	775.64	23,350	3,306

Table 5.5: Size of the corpora collected by crawling

Only 23,350 documents come from PDFs, that is, only 8.61%. And the average document length is small, just 775 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 comparable to those obtained with search engines.

It is also interesting to take a look at the growth rate of the corpus (Figure 5.7). There is certainly a decrease in it, but it is not that pronounced: in the first million links followed, 106.6 million words were collected, whereas with the last million links we obtained 29 million words. It has gone down to 27.1% 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.

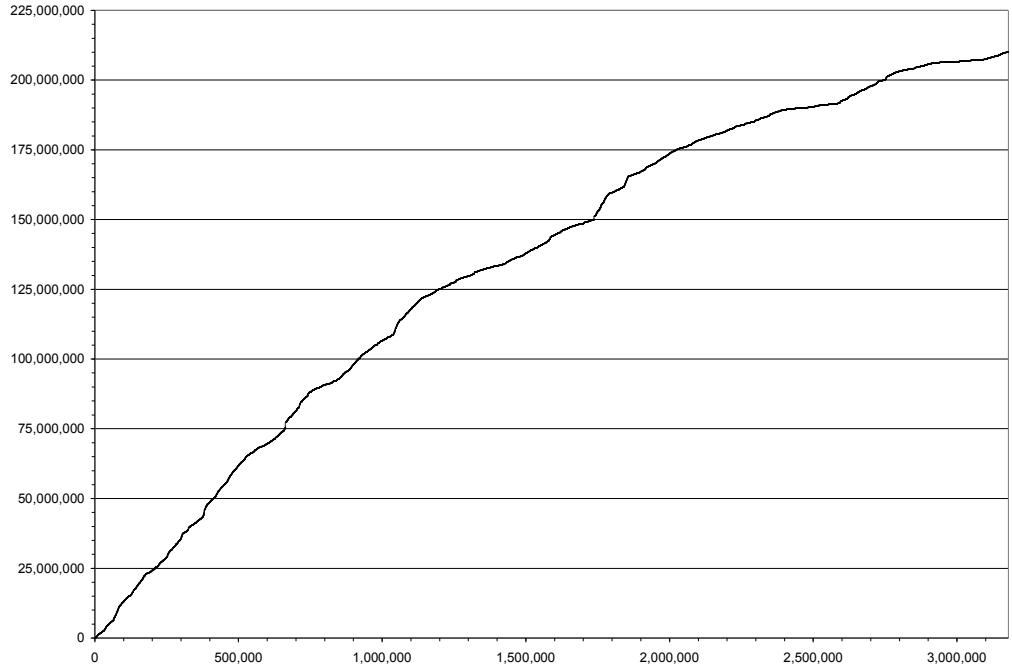


Figure 5.7: Growth rate of the corpus obtained by the crawling method

5.3 Qualitative analysis

In the previous sections, 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.

In corpora built the classical way, the composition of it is usually known beforehand, because texts are included in it following a previous selection. But in the case of corpora built using semi-automatic procedures "the possibility to control the materials that end up in the final corpus is limited," "its composition cannot be determined a priori," and "this makes *post-hoc* evaluation a crucial task," according to Ferraresi (2007). Baroni et al. (2009) express it similarly: "Automated methods of corpus construction allow for limited control over the contents that end up in the final corpus. The actual corpus composition needs therefore to be investigated through post hoc evaluation methods."

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. Kilgarriff et al. (2010) admit that "the only strategy we know of [for doing a first-pass evaluation of a corpus without waiting for its evaluation by a range of language researchers and lexicographers over time] is by comparison." This comparison is done with regard to other corpora with the same characteristics or purposes. In the case of large general corpora, they are compared with reference corpora such as the BNC (Aston and Burnard, 1998) –provided there is something of the kind for the language concerned.

To carry out our 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 **SEC** hereinafter) and the crawling method corpus (**CRC** hereinafter). At first sight, they are very different: the number of URLs coinciding in both is only 2,946 (the SEC is made up of 85,562 URLs and the CRC of 271,058).

Apart from comparing these two corpora with each other, we compared them with two reference corpora of Basque we described in Section 1.3: **XX. mendeko Euskararen Corpusa** (a 4.6-million-word balanced corpus of twentieth century literary texts), hereinafter **XXMEC**, and **Lexikoaren Behatokiko Corpusa** (a 26.5-million-word corpus of 21st century media texts), hereinafter **LBC**.

5.3.1 Most characteristic words by LLR

One way of comparing corpora is by using the **Log-Likelihood Ratio** or **LLR** association measure (Dunning, 1994) to identify the words that are more characteristic of each one with regard to the other (Rayson and Garside, 2000); this is the method used both by Sharoff (2006), Kilgarriff et al. (2010) and Ferraresi et al. (2008) for evaluating the search engine method corpora they collected in the first two cases and the ukWaC in the second.

LLR highlights the words whose frequency is above the expected frequency in the hypothesis of their having the same presence in both corpora. If we have two corpora C_1 and C_2 where $f_{i,j}$ is the frequency of a word w_i in corpus C_j and s_j the size of corpus C_j, then the expected frequency of the word w_i in corpus C_j under the equal distribution hypothesis is:

$$E_{i,j} = s_j \frac{f_{i,1} + f_{i,2}}{s_1 + s_2} \tag{5.1}$$

And the LLR or salience of that word w_i in either C_1 or C_2 is as follows:

$$LLR(w_i, C_1, C_2) = 2\left(f_{i,1} \ln\left(\frac{f_{i,1}}{E_{i,1}}\right) + f_{i,2} \ln\left(\frac{f_{i,2}}{E_{i,2}}\right)\right) \tag{5.2}$$

Or, if we develop further:

$$LLR(w_i, C_1, C_2) = 2\left(f_{i,1} \ln\left(\frac{f_{i,1}}{s_1 \frac{f_{i,1} + f_{i,2}}{s_1 + s_2}}\right) + f_{i,2} \ln\left(\frac{f_{i,2}}{s_2 \frac{f_{i,1} + f_{i,2}}{s_1 + s_2}}\right)\right) \tag{5.3}$$

The words more characteristic of C_1 regarding C_2 will be those with the highest LLR that have a greater relative frequency in C_1 than in C_2, and vice versa.

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. Owing to the fact that in the case of XXMEC 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 from the one we use, we had to discard hyphenated compounds –a common phenomenon in Basque for

the creation of new words–, proper nouns and numbers (the XXMEC 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 the XXMEC 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 the XXMEC 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 the LBC 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, the SEC, 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 education domain (*hezkuntza –education–*, *ikasle –pupil–*, *ikastetxe –school–*, *irakaskuntza –teaching–*, *irakasle –teacher–*, etc.). The cause of the prominence of administrative words might lie in the fact that regional, provincial, and local governments publish their official gazettes in PDF format; and, as we saw before, the SEC corpus has a large proportion of PDFs, so these might be mostly of an adminis-

trative nature. But this prominence of administrative words is not so great when compared with the CRC corpus (the CRC is also a web corpus with a fair percentage of PDFs, although not as large as the SEC).

Finally, the words characteristic of the corpus obtained by crawling, the CRC, 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. And regarding the first group, it is a common characteristic of web corpora to have a greater proportion of such words than classical corpora. Ferraresi (2007) also noted this phenomenon in the ukWaC and pointed out that it "is quite unsurprising, insofar as they represent meta-references to the medium of communication that hosts them." Apart from common, he thinks it is also "a welcome finding, since one of the main aims of the corpus is that of documenting recent phases of language evolution, of which the increasing importance of web- and computing-related words could be an example."

There are three things that must be noted about this word list comparison method. The first is that it highlights the words that stand out in one corpus with regard to another, but this does not mean they stand out by themselves in the corpus. This means that when we say that the word *blog* is characteristic of the CRC corpus, we mean its frequency is proportionally much greater in that corpus compared with the other corpora, not that it is one of the most frequent words of the corpus. The second is that the LLR measures are not absolute, they are affected by both corpora sizes. This means that we cannot compare two such values to say that "this word stands out more in this corpus with relation to that one than in this other corpus with relation to that other one." And also that we have listed in what aspects the corpora are different, but we do not know anything about to what extent they differ. And third, the method indicates in what way two corpora differ, not how much they resemble each other. It would be in-

teresting to have a measure to tell how similar two corpora are. Ferraresi (2007) proposes that "a way to understand how the two corpora are similar would be to also take into account all the differences that did not emerge from the analysis." It is interesting, but it is much more work to analyse all the similar words than just a handful of different ones. Besides, it is not easy to tell where we should put the threshold to consider an LLR measure as normal variation or significant difference; as we said, the measures are not absolute.

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 5.6.

Corpus	Total words	Words f ≥ 20
XXMEC	53,993	9,147
LBC	36,311	12,922
SEC	74,132	33,056
CRC	74,037	33,755

Table 5.6: Number of distinct and "useful" 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 XXMEC 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 LBC corpus

(which is almost 4 times bigger). This is due to the fact that a considerable part of the XXMEC 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 automatic taggers do not currently recognize.

It might come as a surprise that the number of words (both total and "useful") in the CRC corpus are similar to that in the SEC corpus, although the former is much larger. This is due to the already mentioned fact that only lemmas of single words have been taken into account, while disregarding proper nouns, numbers, or hyphenated compounds because the taggers used in the XXMEC corpus did not include these. And the number of lemmas in the web corpora is close to the total number of lemmas with those characteristics (not hyphenated compounds, proper nouns, or numbers) that the tagger can recognize. Once having arrived at or got close to this number, it is not possible to automatically recognize many more lemmas even if the size of the corpus doubles.

Alegria et al. (2005), when building the Zientzia eta Teknologiaren Corpusa (Corpus of Science and Technology) or the ZTC, wanted to estimate the relationship between the size of a corpus in Basque and the number of lemmas. In order to calculate that, they took two corpora –one produced from classical literature texts translated into Basque taken from the Pentsamenduaren Klasikoak collection (University of the Basque Country, 2010) and the other one comprising articles from the popular science magazine Elhuyar (Elhuyar Foundation, 2001)– and saw the evolution in the number of lemmas as the size of the corpora grew, shown in Figure 5.8.

According to Yang et al. (2000), the function that most resembles the observed lemma / corpus size proportions is given by the following formula (l being the number of lemmas and s the corpus size):

$$l = \alpha s^{\beta} \tag{5.4}$$

The values of α and β are dependent on the language, corpus type, etc. And since they can be calculated using the observed data shown in Figure 5.8 and applying the method of least squares, Alegria et al. (*ibid.*) obtained the graph shown in Figure 5.9 estimating the evolution in the number of lemmas for corpus sizes up to 20 million words.

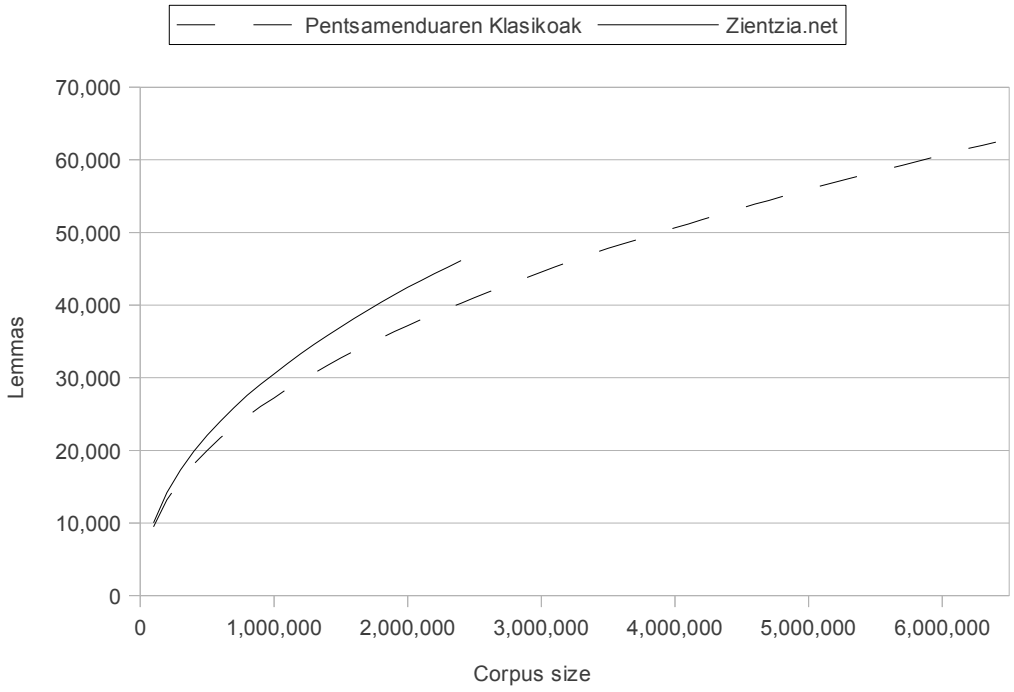


Figure 5.8: Evolution of number of lemmas in relation to corpus size

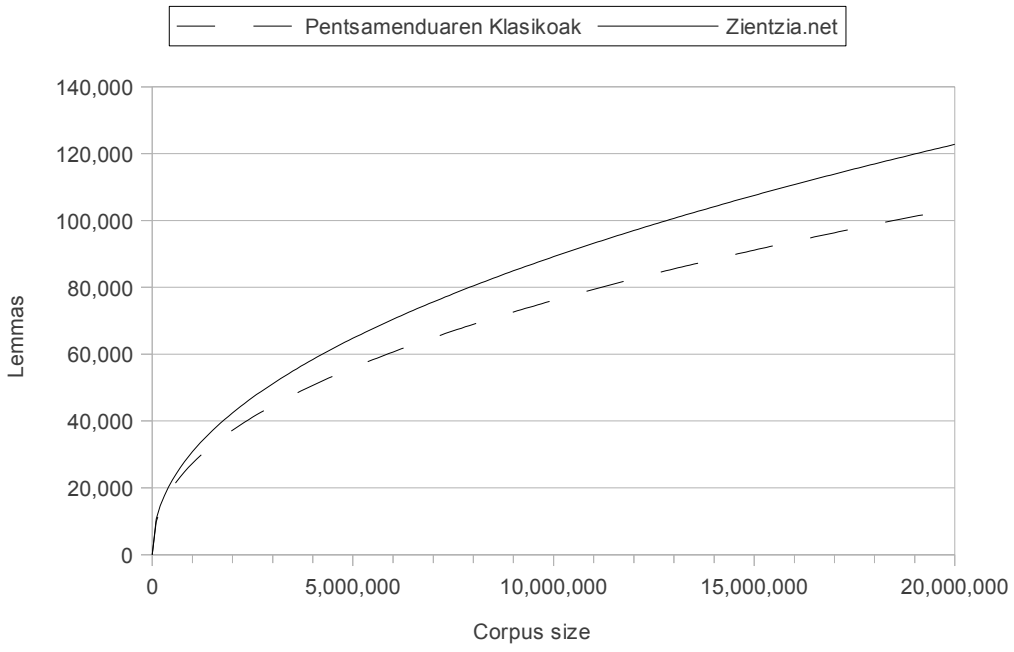


Figure 5.9: Estimate of evolution of number of lemmas in relation to corpus size

We can observe that the evolution depends greatly on the type of corpus, and the evolution of lemmas in our web-collected corpora can be different. But it is clear that, in any case, the number of lemmas that we obtain in our web corpora is reached very much sooner than the corpora sizes we have obtained, which means that we have indeed arrived close to the limit in lemmas of the tagger.

The larger corpus is bound to have many more words, but these additional words will be of the kind of neologisms, specialized terminology, hyphenated compounds, proper nouns, etc. that taggers do not include. In fact, although not the XXMEC or the LBC, at least we can compare the web corpora –the SEC and the CRC– while taking into account hyphenated compounds and proper nouns as well. While doing so we observed that, logically, the larger corpus had significantly more "useful" words, as Table 5.7 shows. And the actual number of additional words in the larger corpus is bound to be higher, since this study does not include neologisms, specialized terminology, and many proper nouns (only the proper nouns the tagger recognizes).

Corpus	Lemmas		… + Hyphenated Compounds		… + Proper Nouns	
	Total	f ≥ 20	Total	f ≥ 20	Total	f ≥ 20
SEC	74,132	33,056	301,324	41,193	315,242	45,709
CRC	74,037	33,755	299,554	43,960	316,265	49,129

Table 5.7: Number of distinct and "useful" words in the web corpora, with hyphenated compounds and proper nouns

However, it may still seem striking that, although the CRC is almost double the size of the SEC, the number of "useful" lemmas is only 10% higher. But Figure 5.9 shows that the larger the corpus, the fewer the number of new lemmas that appear for each equal increase in corpus size. So if we are dealing with corpora of 100-200 million words, we cannot expect much larger numbers of new lemmas.

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 regarded 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 5.8.

Corpora type	Corpora	Coverage	Enrich-ment	Corpora	Coverage	Enrich-ment
Classical	XXMEC / LBC	57.14%	14.36%	LBC / XXMEC	80.71%	38.36%
Classical / Web	XXMEC / SEC	26.13%	0.48%	SEC / XXMEC	94.44%	83.67%
	XXMEC / CRC	25.55%	0.50%	CRC / XXMEC	94.30%	83.13%
	LBC / SEC	36.74%	0.81%	SEC / LBC	93.99%	83.85%
	LBC / CRC	35.89%	0.93%	CRC / LBC	93.76%	83.69%
Web	SEC / CRC	89.80%	22.64%	CRC / SEC	91.70%	29.80%

Table 5.8: Coverage and enrichment of each corpus with regard to each of the others

The table shows that the web corpora cover high above 90% of the classical corpora with an enrichment above them of around 80%, whereas the coverage of the classical corpora over the web ones is below 40% and their enrichment is always below 1%; these data are similar to the ones obtained in the aforementioned research by Baroni et al. with ukWaC/BNC and itWaC/La Repubblica.

The comparison between the two web corpora, SEC and CRC, shows that they are around 90% substitutable by each other (that is, 90% of "useful" words in any of them are also present in the other), that the CRC provides almost 30% more "useful" words than the SEC and that the SEC has almost 23% of new words not present in the CRC.

Again, this is counting single lemmas only; if we do the same analysis with proper nouns and hyphenated compounds (which we can only do with the SEC and the CRC), the differences between both corpora become greater. As expected, the larger corpus, the CRC, covers and enriches the SEC more than the other way round, as can be seen in Table 5.9. And again, if neologisms, specialized terminology, and all proper nouns (not just the ones the tagger recognizes) could be taken into account (thus calculating the real coverage and enrichment), the difference would be greater.

Corpora	Coverage			Enrichment		
	Lemmas	… + Hyphenated Compounds	… + Proper Nouns	Lemmas	… + Hyphenated Compounds	… + Proper Nouns
SEC / CRC	89.80%	83.81%	83.26%	22.64%	15.71%	16.49%
CRC / SEC	91.70%	88.91%	89.49%	29.80%	25.80%	28.27%

Table 5.9: Coverage and enrichment of the web corpora with regard to each of the others, with hyphenated compounds and proper nouns

5.4 Conclusions

We have proven that both crawling and using search engines are valid methods for obtaining BNC-sized corpora for Basque. With the search engine 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 significantly larger than those obtained via search engines, with far fewer PDFs and the potential to be much bigger. The size of this corpus is now 210 million words, but we expect to expand it considerably 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 contains 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, Basque linguistics, and the Basque language in general.

We also believe that the conclusions we have drawn regarding the obtaining of large general Basque corpora from the web could be applied to other minority languages and that by using the methodology described, these languages can collect large corpora from the web too, provided some basic tools exist for them (namely N-gram based language detection and morphological analysis and generation) and that there is a sufficient number of texts on the language on the web. Many languages exist that meet these conditions but which do not yet have large general corpora.

6 Using the web to build specialized corpora in Basque

In the introduction we argued why Basque needs corpora as much as, or more than, many other bigger languages. And it is in need of specialized corpora above all, because terminology is the area with least *de jure* normalization. The only specialized corpus in Basque is the ZT Corpus (Areta et al., 2007), a corpus on Science and Technology that is a very valuable resource, but which does not fulfil all the needs of Basque for several reasons: first, it does not include texts on social sciences; second, it is divided into very general topics, so it is impossible to search texts dealing exclusively with anatomy or computer science, for example; and third, it is not kept up-to-date. Thus, building a system to collect specialized corpora in Basque using the Internet as a source is an interesting and necessary task we wanted to address in this thesis.

In the chapter dedicated to the state of the art we saw that there are mainly two methods for collecting specialized corpora from the web. One of them is **focused crawling** (Chakrabarti et al., 1999), which consists of crawling but with some sort of bias towards the domain we are interested in (in the form of the initial seeds, of domain-filtering of the downloaded pages on the fly, of trying to guess the domain from the URL before queuing it, etc.). Focused crawling normally needs a final filtering stage, via machine learning for example. The other is the **BootCaT method** (Baroni and Bernardini, 2004), which relies on querying search engines for combinations of words in the target domain. This second method has been more popular since its appearance, although in the last few years the restrictions that search engine APIs have been imposing on their conditions of use are limiting its wider employment.

Both methods present advantages and drawbacks. Focused crawling needs some domain filtering, which is difficult if there is no corpus in the domain available beforehand (and this is often the case with less-resourced languages like Basque); in return, the filtering allows high domain-precision. The BootCaT method is easier to apply –a list of words is enough–, but its domain-precision may not be as good (the original paper on BootCaT reports precision of 66% in an evaluation performed on a small sample of 30 texts from each of the two corpora collected). A combination of both that would achieve the former's precision with the latter's ease of implementation would be ideal.

In order to collect specialized corpora for Basque from the web, our idea is to try to develop a method that will use a method similar to BootCaT's for gathering the texts and then apply some easy filtering method that will not need a large corpus to train or compare.

However, since BootCaT uses search engines and the performance of these regarding Basque is far from satisfactory, we will have to work on techniques to improve this too.

This chapter will describe the work we have done in pursuit of these objectives, together with the results of two evaluations we have performed on the developed tool, one based on manual observation and the other on its application to the automatic terminology extraction task. The experiments have been previously published in (Leturia et al., 2008b; Gurrutxaga et al., 2009).

6.1 Adapting the BootCaT method to Basque

6.1.1 Problems of BootCaT for Basque

Some features of the Basque language and the Basque web cause topic precision to fall dramatically when using the standard BootCaT methodology for collecting corpora in Basque, as the experiment we describe below shows.

We used BootCaT to gather some small corpora on geology and computer science: we made 20 queries with 2, 3 and 4 n-gram combinations and downloaded the first 10 pages. Then we looked at all of the documents to see if they were appropriate for the corpus (desired topic and language, informative, not duplicates, etc.), and the results we obtained are shown in Table 6.1.

Domain	n	Total		Appropriate			
		Docs	Words	Docs	%	Words	%
Computer Science	2	65	1,282,001	33	50.77	289,259	22.56
	3	60	2,853,710	25	41.67	406,426	14.24
	4	48	2,321,888	22	45.83	355,254	15.30
Geology	2	85	2,526,820	13	15.29	379,131	15.00
	3	31	1,606,312	8	25.81	184,371	11.48
	4	3	195,246	2	66.67	101,731	52.10
Total		**292**	**10,785,977**	**103**	**35.27**	**1,716,172**	**15.91**

Table 6.1: BootCaT domain-precision results for Basque

The overall precision obtained, less than 16%, is far from the 66% reported in the BootCaT paper. The percentage of each of the reasons for a document to be considered inappropriate are shown in Table 6.2.

Domain	n	Reason					
		Wrong domain		Wrong language		Other	
		Docs	%	Docs	%	Docs	%
Computer Science	2	21	65.63	5	15.63	6	18.75
	3	17	48.57	11	31.43	7	20.00
	4	16	61.54	4	15.38	6	23.08
Geology	2	31	43.06	26	36.11	15	20.83
	3	4	17.39	2	8.70	17	73.91
	4	0	0.00	0	0.00	1	100.00
Total		**89**	**47.09**	**48**	**25.40**	**52**	**27.51**

Table 6.2: Kinds of inappropriate pages

The documents classified as *other* are duplicates, part of a much bigger document including other domains, etc.

This study is by no means exhaustive, but our objective was not to quantify the loss in precision exactly. We were just aiming to show that topic precision and general quality of a corpus obtained with BootCaT are much worse when looking for corpora in Basque. Besides, we must take into account that in this experiment we did not perform the bootstrapping process of extracting the words out of the downloaded pages to get new ones; if we had done so, the pages downloaded in the next stage would most likely have yielded even worse topic precision.

The reasons for this have partly been explained in Section 3.1, *i.e.*, that no search engine offers the possibility of returning pages in Basque alone (so when looking for technical words, as is often the case with specialized corpora, it is very probable that they exist in other languages too, and that the queries return many pages that are not in Basque) and that Basque is a morphologically rich language (thus, any lemma has many different word forms, so looking for a word's base form alone, as search engines do, brings fewer results).

But we must add to these reasons that the Basque web is not as big as those of other languages, and this means that the only pages existing for certain queries with combinations of various words are very long documents (blogs, magazines in PDF format,

books, etc.) where the desired topic is just a small part of the whole document, or where the words searched for are simply found by chance in different parts of the long document.

6.1.2 Proposed solution

The solution we propose to improve the performance of BootCaT with Basque is the same as in the two works described previously, that is, by applying morphological query expansion and language-filtering words to the queries, as described in Section 3.2.

The effect we anticipate of the increase in recall obtained by using the morphological query expansion technique is a smaller percentage of big PDFs in the downloaded documents, and more pages downloaded in some topics with 4-word combinations in the queries.

Regarding language-filtering words, we are aware that BootCaT does give the option of language-filtering by means of a list of frequent words in the language, but that filtering is done after downloading the pages. If filtering is conducted that way, many searches for words that exist in other languages will bring very few or no results in Basque and all the pages will be filtered out, thereby wasting bandwidth, time, and calls to the API of the search engine.

In order to obtain an even better performance, we also apply all the cleaning and filtering stages that we applied when collecting large general corpora described in Chapter 4: language-filtering by LangId applied at the paragraph level to avoid non-Basque parts of bilingual documents, length filtering to avoid documents that are too short and too long, boilerplate removal via Kimatu (Saralegi and Leturia, 2007), near-duplicate detection by our variant of Broder's (2000) algorithm and containment detection using Broder's (1997) technique.

6.1.3 Evaluation and results

In order to evaluate and measure the improvements of our system, we built some corpora by putting it into practice. We chose the same two topics with which we evaluated the performance of BootCaT with Basque, *i.e.*, computer science and geology. This time for each of the 2-, 3- and 4-word combinations we built three different corpora, which we then evaluated manually. The final sizes of the 18 corpora collected can be seen in Table 6.3.

Domain	Corpus	n		
		2	3	4
Computer Science	x	758	274	43
	y	745	256	56
	z	674	176	52
Geology	x	97	22	3
	y	125	14	3
	z	146	27	2

Table 6.3: Sizes of the collected corpora

The pages returned have been evaluated manually. Regarding the effectiveness of the language-filtering words method, only 2.46% of documents retrieved by search engines did not contain any Basque. As to the language identifier that is applied at paragraph level, it removes supposedly non-Basque parts from 28% of the downloaded documents. Due to the amount of work this entails, we did not evaluate the recall of this step (that is, we did not look at all the documents to see how many non-Basque parts had been left out). But we did look at a sample of the cleaned documents to see if the removed parts were really non-Basque, and although we did not measure it quantitatively, the performance can be considered to be very good.

The morphological query expansion method improves recall in Basque IR, so the number of long PDFs should go down when it is used, which in fact turns out to be the case: in the BootCaT experiments, almost 72% of the documents were PDFs, but this time, only 13% are PDFs in the computer science corpus and 41% in the geology corpus; and the average document length also went down by 25%.

The length filter left out 31% of the downloaded documents because they were too long, and 10% and 3% of the computer science and geology corpora, respectively, because they were too short. By taking a look at the rejected ones, we confirmed that the filter achieves its goal, as the great majority were uninteresting, general or multi-topic documents.

The near-duplicates filter removed 5% of the downloaded documents, and the containment filter another 5%. In the small evaluation we made for precision we found no errors; recall was not evaluated.

Apart from individually evaluating these improvements made to the process one by one, since the aim of each and every one of them is to enhance the quality of the corpora obtained mainly regarding domain precision, it is imperative to evaluate the collected corpora by looking at domain precision, to see if these tweaks had any effect and actually improved the BootCaT results. We took a random sample of 30 documents out of each of the 18 corpora built for the evaluation, and saw whether they belonged to the desired topic or not. Due to their small size (see Table 6.3), all the documents of n=4 and of geology n=3 were checked. The results we obtained are shown in Table 6.4.

Domain	Corpus	n		
		2	3	4
Computer Science	x	46.66%	63.33%	82.93%
	y	50.00%	66.66%	70.00%
	z	53.33%	63.33%	63.89%
	Avg.	**50.00%**	**64.44%**	**72.27%**
Geology	x	53.33%	40.91%	100.00%
	y	56.66%	64.29%	100.00%
	z	46.66%	56.76%	100.00%
	Avg.	**52.22%**	**53.98%**	**100.00%**
Avg.		**51.11%**	**59.21%**	**86.14%**
Avg.		**65.49%**		

Table 6.4: Domain precision obtained with the improvements for Basque

In view of these results, we can conclude that our little improvements, all together, do yield much better domain precision results when looking for corpora in Basque, and are not far from the baseline for other languages (which, it should be remembered, was around two thirds for the only evaluation reported). We can say the series of improvements that we propose do in fact improve the otherwise disastrous performance when looking for documents in Basque.

We have also observed that, without any filtering, the best topic precision results are obtained, logically, with 4-word queries, but due to the reduced amount of Basque content on the Internet, corpora obtained on some topics are extremely small with these kinds of queries. And there is no way one can know *a priori* which topics will be affected, so it is maybe better to use 3-word queries, even though the topic precision obtained will be a little lower.

6.2 General improvements in the BootCaT methodology

However, depending on the application, 33% of noise in the corpora –both in the BootCaT method in general and with our improvements for Basque– can be considered to be unacceptable. Thus, another of our objectives is to try to improve this precision by developing a method to apply a final domain filtering stage that will not need a large corpus to train like machine learning does.

The technique we propose takes, as a starting point, a sample mini-corpus of documents on the topic, instead of a list of words. This mini-corpus has two uses: first, the list of keywords to be used in the queries is automatically extracted from it; second, it is used to filter the downloaded documents according to domain by using document-similarity techniques.

6.2.1 Automatic keyword extraction from a sample mini-corpus

The basis of our system is a sample mini-corpus of documents on the target topic, which will have to be collected manually. This sample will be used for extracting the word list for the queries and in the final topic-filtering stage as well, so the criteria when collecting the sample is that it should be as heterogeneous as possible and cover as many different subjects of the domain as possible. According to our experiments, as few as 10 documents may be enough for a very specialized topic, but more might be needed for more general topics.

The words to be used in the queries are automatically extracted from this sample corpus, thus avoiding the work of finding appropriate words on the topic. This is usually more laborious than finding texts on the topic, at least for Basque, because there are many topics for which there are still no specialized dictionaries or glossaries.

The keyword extraction method is based on the work by Saralegi and Alegria (2007) in the **DokuSare** project –a project to automatically show cross-lingual related articles in a media website–, where they extract keywords for measuring document similarity. The method follows these steps:

- First, the mini-corpus is automatically lemmatized and POS-tagged
- Then, the **Relative Frequency Ratio** or **RFR** (Damerau, 1993) is calculated for all nouns, proper nouns, adjectives, verbs, entities, and multiword terms. This ratio is calculated by dividing the relative frequency of a word in the specialized mini-corpus by the relative frequency of the word in a

general corpus consisting of 450,000 words of newspaper articles. This is the formula of the RFR ratio (*f* being the frequency, *doc* the document, and *gc* the general corpus):

$$RFR(w_i, doc) = \frac{f(w_i, doc)}{f(w_i, gc)} \quad (6.1)$$

- Then, the most significant of them are chosen by applying an empirically determined threshold.

The extracted list consists of (mostly) domain-specific words, but some of them might be too specific or rare, as the RFR measure tends to promote on excess words that are not present in the general corpus. The usual way to avoid this is to use a raw frequency threshold to choose the candidate words for the RFR measure, but this is not so easy to apply in our case, because the sample mini-corpora are small (on purpose). And in any case, these undesired words are usually removed in the manual revision stage explained in the next paragraph.

In order to maximize the performance of the queries, the extracted list is revised manually. Too specific or too local proper nouns, too general words, and polysemous words that have other meanings in other areas are removed. Normally, the whole process to obtain the mini-corpus and manual revision of the word list is still less costly than trying to obtain a word list, because of the absence of specialized dictionaries explained above.

6.2.2 Domain filtering

Topic or domain detection is usually approached through machine learning methods. While these can obtain good performances, they also have their drawbacks: they need fairly big training sets and times, they are trained for a fixed set of topics, etc.

Our approach to this matter has been to try to use a small set of sample documents (*i.e.*, the sample mini-corpus out of which the keywords are extracted) and document similarity measures (Lee et al., 2005) based on keyword frequencies to say whether a document belongs to a domain or not. It is widely accepted (Sebastiani, 2002; Sharoff, 2007) that domain detection can be done using keywords.

This kinds of document similarity measures are usually applied between two documents to see if they deal with the same or a similar subject, as in the aforementioned DokuSare project; but in our case, we have a document and a corpus, which are elements of different scale, and also the level of similarity to be handled is somehow smaller, since we just need to measure whether they coincide in the domain. They have also been applied to measure similarity between two corpora (Kilgarriff and Rose, 1998), which is also a little different from our case.

However, the general idea of our project is very similar to that of DokuSare: to represent both the documents to be filtered and the sample mini-corpus through a set of features based on keywords, and to use some similarity measure to see if they share the same topic.

But as we said, we are going to measure the similarity between elements of a different scale, *i.e.*, a document and a set of documents. So we have tried by measuring the similarity between a document and the mini-corpus directly, and also by measuring the similarity of a document with each of the documents of the sample mini-corpus, and taking the maximum.

For the representation of both the downloaded documents and the sample corpus or each of the documents of the sample corpus, we use the **bag-of-words paradigm**, which models the most significant keywords, *i.e.*, nouns, proper nouns, adjectives, and verbs, in a vector. The words are selected and weighted by a certain frequency measure. We have tried two: the aforementioned RFR and a new one we have specified as **Relative Rank Ratio** or **RRR**.

We felt that this new frequency measure fitted Zipf's law (1949) better and could be better suited for comparing documents of different sizes. It is defined as the ratio between the relative frequency-ranking of a word in the document or corpus involved, and the relative frequency-ranking of a word in a general corpus. This is its exact formula (being *fr* the frequency-ranking, *r* the number of different rankings, *doc* the document, and *gc* the general corpus):

$$RRR(w_i, doc) = \frac{1 - \frac{fr(w_i, doc)}{r(doc)+1}}{1 - \frac{fr(w_i, gc)}{r(gc)+1}} \quad (6.2)$$

We have observed that this measure works better if we apply some sort of smoothing to words that are not found in the general corpus, because otherwise the formula gives them very high values, and they are often rare words or spelling errors that worsen the results.

For measuring the similarity we use the cosine, the most extended way to measure the similarity between documents represented in the vector space model.

So for comparing two documents x and y, being $w_i(i \in \{1,n\})$ the keywords present in any of the two, we prepare the vectors (x_1, x_2, … x_n) and (y_1, y_2, … y_n), where x_i and y_i are the RFR or RRR ratios of the word w_i in the documents x and y respectively, and then we calculate the cosine between the two, which is specified as follows:

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (6.3)$$

6.2.3 Evaluation and results

As an evaluation experiment, we built some corpora, again for computer science and geology. For each of these domains we built three sample mini-corpora, consisting of 10, 20 and 30 documents, the two smaller ones made up of documents chosen at random out of the bigger corpus. From each of these six sample mini-corpora we automatically extracted the word lists and revised them manually as indicated. Then out of each of the six lists we built three different corpora using 2-, 3- and 4-word combinations in the queries.

Afterwards, we manually chose a sample of appropriate documents and another one of inappropriate ones out of each of them, each made up of 15 documents (if the corpus was large enough). Then we applied the aforementioned similarity measures to these documents in the two ways explained, and for each of the 18 corpora we obtained charts like those shown in Figures 6.1 to 6.4. More precisely, these correspond to the average of the geology and computer science corpora collected using 20-document sample mini-corpora and using 2-word combinations.

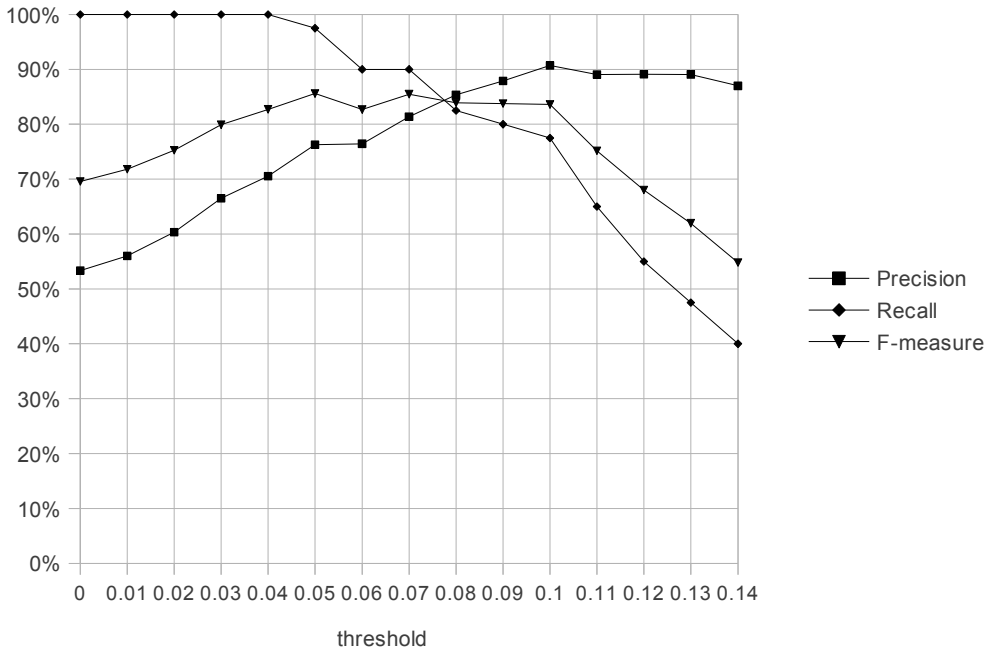


Figure 6.1: Results with RRR measure, taking the sample mini-corpus as a whole

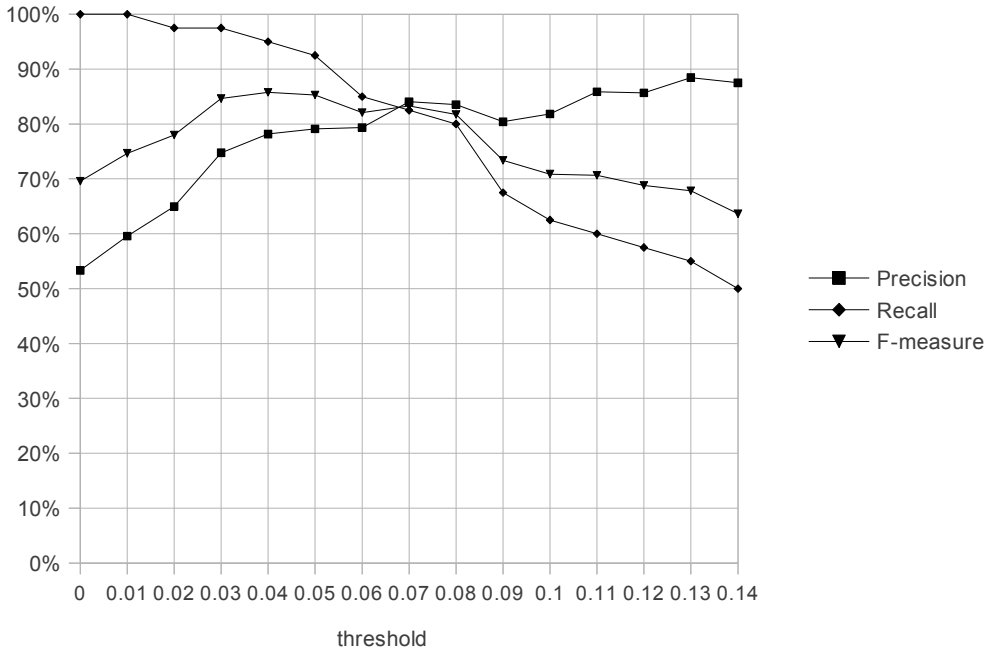


Figure 6.2: Results with RFR measure, taking the sample mini-corpus as a whole

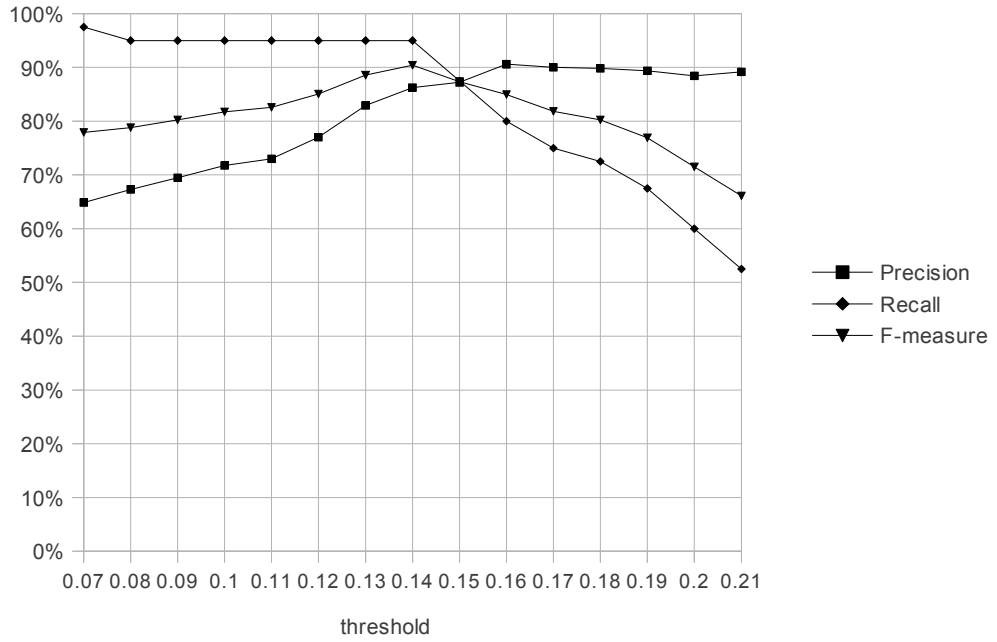


Figure 6.3: Results with RRR measure, taking each document of the sample mini-corpus individually

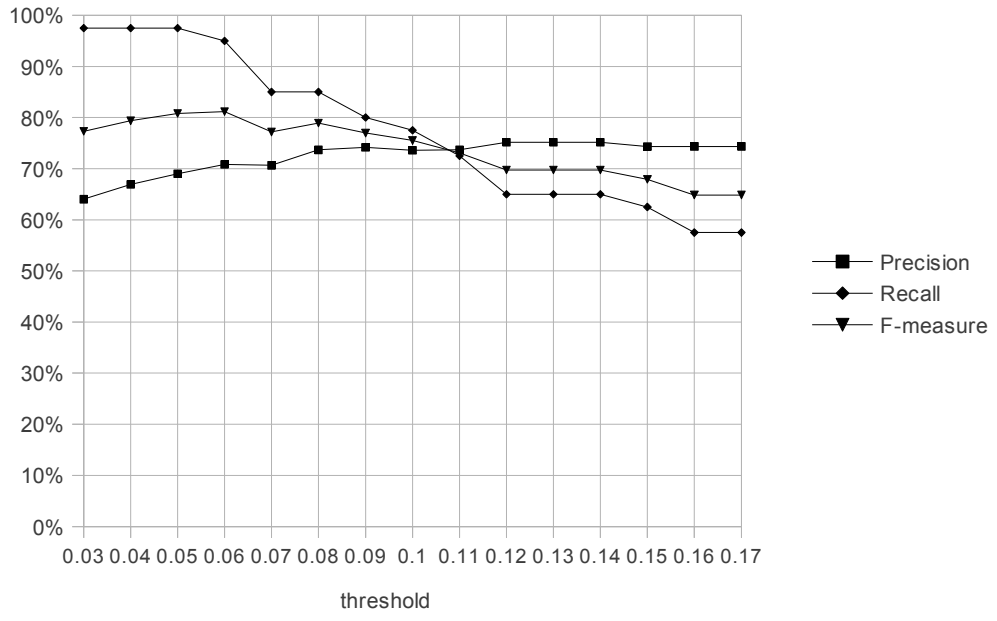


Figure 6.4: Results with RFR measure, taking each document of the sample mini-corpus individually

It is not possible to show here all the charts for all of the 18 corpora and the different averages. Instead, we will explain the conclusions we drew from their observation.

Since our primary objective is to improve topic precision, we are interested in finding a measure and a threshold that will maximize the F-measure but which will prime precision. This is usually obtained somewhere to the right and near the crossing point of the precision and recall series. On average, the highest crossing points are found with the RRR measure when compared with each document of the sample corpus individually.

We have also tried to improve the results by combining more than one of them. For example, we have tried first measuring the similarity with the whole sample mini-corpus and, if the measure is not above the threshold, trying again with the one-by-one comparison. But the only effect of this was that more documents were accepted, both good and bad ones, thus augmenting recall but at the cost of precision.

If we are to significantly improve the baseline of 66% topic precision, we would need a minimum precision of 80%, without a great loss in recall. The RRR-individually method can obtain precision and recall above 80% for most of the corpora, but with different thresholds. In other words, there is no threshold that maximizes F-measure and obtains a precision above 80%, and which works for all of the corpora.

In any case, for higher thresholds we usually obtain a higher precision (at least until it falls at some point), so it is possible to assure high precision (80-90%) if recall is not an issue. This might not be the case of Basque, since, as we have observed before, some topics already yield very small corpora and a recall of 60-40% may not be acceptable. But for English or other bigger languages, with the RRR-individually method and a threshold from 0.18 to 0.20 we can obtain a topic precision of 80-90%.

However, there is an important aspect to point out regarding this method. Obtaining high topic precision does not imply that the corpus obtained will be highly representative of the domain. In fact, since we are filtering by applying similarity measures using the documents of the sample mini-corpus, if this is not wide enough, that is, if not all the sub-areas of the domain are represented there, we might be missing areas without ever knowing it. So the quality and heterogeneity (and also size) of the sample mini-corpus is a key issue in the method proposed. But it is not easy to say what a minimum or optimum size of the sample mini-corpus is to ensure good representativeness, since it greatly depends on whether the topic is very specialized, or quite general, etc.

Although our keyword extraction is done by using the RFR measure and our experiments to test precision and recall of the domain filtering were carried out using the RFR and RRR measures, later experiments in the DokuSare project (Saralegi and Alegria, 2007) reported that Log-Likelihood Ratio worked better in their case.

In this case, the LLR of a word w_i in a document d (or the mini-corpus) with regard to a general corpus gc is calculated as follows (being $f_{i,j}$ the frequency of the word w_i in j and s_j the size of j):

$$LLR(w_i, d, gc) = 2\left(f_{i,d}\ln\left(\frac{f_{i,d}}{E_{i,d}}\right) + f_{i,gc}\ln\left(\frac{f_{i,gc}}{E_{i,gc}}\right)\right) \quad (6.4)$$

Where $E_{i,d}$ and $E_{i,gc}$ are the expected frequencies of the word in the document or the general corpus, being defined thus:

$$E_{i,j} = s_j \frac{f_{i,d} + f_{i,gc}}{s_d + s_{gc}} \quad (6.5)$$

However, as we explained before, the LLR measure is two-sided, that is, it highlights with a larger positive value words that have a frequency over the expected in either the document (or mini-corpus) or the general corpus. But in this case we are only interested in those that are prominent in the first. We solve this and convert the ratio in a one-sided measure by changing the sign of the LLR measure when the frequency of the word in the document is smaller than its expected frequency, that is:

If $f_{i,d} \geq E_{i,d}$ then

$$LLR(w_i, d, gc) = 2\left(f_{i,d}\ln\left(\frac{f_{i,d}}{E_{i,d}}\right) + f_{i,gc}\ln\left(\frac{f_{i,gc}}{E_{i,gc}}\right)\right) \quad (6.6)$$

Else

$$LLR(w_i, d, gc) = -2\left(f_{i,d}\ln\left(\frac{f_{i,d}}{E_{i,d}}\right) + f_{i,gc}\ln\left(\frac{f_{i,gc}}{E_{i,gc}}\right)\right)$$

A small experiment we conducted by collecting two specialized corpora followed by a visual examination of a sample showed that LLR worked well for our case too. The system we now use for collecting domain corpora is configurable and the user can choose among RFR, RRR or LLR for both the initial extraction of keywords and the domain filtering.

In the same conference as our method was published (Leturia et al., 2008b), Nazar, Vivaldi, and Cabré (2008) published a paper that introduces a method that shares some similarities with ours. They also highlight the insufficient quality or domain precision of BootCaT, and they also developed a tool, **Jaguar** –which is provided as a web service (IULA, 2008)–, that makes use of specialized terms and search engines to collect specialized corpora, but that applies a further filtering stage to the BootCaT method. In their case, they make some initial searches with one or a few terms and they carry out an unsupervised clustering of the downloaded documents using terms weighted by the Mutual Information or MI measure with regard to a general corpus. The user is then offered the choice of selecting one of the clusters. Further searches will then be performed with terms extracted from that cluster, and the downloaded documents will only be inserted into the corpus if they share a minimum similarity with the selected cluster.

6.3 Evaluation on an automatic terminology extraction task

With our methodology, we can collect specialized corpora in Basque with a high domain precision. But we thought it was a good idea to carry out a practical task-oriented evaluation. Kilgarriff and Grefenstette claim that the quality of a corpus should not be stated in absolute terms but in terms relative to its appropriateness for a certain use, that is to say, by asking the question “is corpus *x* good for task *y*?” (2003). Kilgarriff et al. also ask “what does it mean for a corpus to be good? It depends what we want to use the corpus for. The straightforward answer to the question is "if it supports us in doing what we want to do"” (2010).

Automatic terminology extraction is a task where specialized corpora can be used, as in (Daille, 1995; Smadja, 1993) for example. We wanted to test the suitability of corpora collected with our method for this task. Thus, we have applied an automatic terminology extraction tool to some corpora we collected automatically and evaluated

the results using a manually built terminological dictionary. The results obtained have been compared to those obtained by applying the same terminology extraction tool to a manually built corpus.

6.3.1 Corpora collection

With the specialized corpus collection method described above we built three corpora in Basque on three domains: Computer Science, Biotechnology, and Atomic & Particle Physics. The measure used for the keyword extraction in both the initial phase and the domain filtering was LLR. The seed terms automatically extracted were manually validated as indicated above. The collection of the corpora from the Internet did not have a target size, because the Internet in Basque is not as big as that in other languages, and the number of pages we would want to collect for a particular domain might not exist. So we simply launched the collecting processes and stopped them when the growing speed of the corpora fell to almost zero, thus obtaining corpora that were as large as possible. The corpora obtained and the seeds used are detailed in Table 6.5.

		Corpus		
		Atomic and Particle Physics	Computer Science	Biotechnology
Sample corpus size	Docs	32	33	55
	Words	26,164	34,266	41,496
Obtained corpus size	Docs	258	1672	399
	Words	320,212	2,514,290	578,866

Table 6.5: Seeds and obtained corpora sizes in docs and words

There is a clear imbalance between the domains. While Computer Science achieves a significant size –this is normal because, as Thelwall (2005) noted, the web contains disproportionately large quantities of computer-related texts due to the fact that typical web authors tend to be young people with above average computing skills–, Biotechnology and Atomic and Particle Physics are relatively small –as we feared, the web simply does not contain many documents in some domains– and might not be enough for some tasks. We could have lowered the domain-threshold so that they would grow larger, but the precision would have fallen and that might not be good for our termino-

logy extraction task. If we just wanted to find examples of use of terms specific to the domains, maybe the non-domain documents might be acceptable; but for the terminology extraction task, the noise might affect the results.

6.3.2 Terminology extraction

Term extraction is carried out using **Erauzterm**, an automatic terminology extraction tool for Basque (Alegria et al., 2004a; Alegria et al., 2004b), which combines both linguistic and statistical methods. The tool also offers a graphical interface which allows the user, if necessary, to explore, edit, and export the extracted terminology.

Let's describe the procedure that the extractor follows. First, a lemmatizer and POS tagger for Basque (Aduriz et al., 1996) is applied to the corpus. Then the most usual Noun Phrase structures for Basque terms are detected (Urizar et al., 2000; Alegria et al., 2004b) to obtain a list of term candidates. Term variants are linked to each other by applying some rules at syntagmatic and paradigmatic level. Afterwards, statistical measures are applied in order to rank the candidates. Multiword terms are ranked according to their degree of association or unithood using Log-Likelihood Ratio or LLR (Dunning, 1994). Single word terms are ranked according to their termhood or divergence with respect to a general domain corpus, also using LLR. Then those candidates that reach a threshold are chosen. A manual evaluation of the tool reported a precision of 65% for multiword terms and 75% for single word terms for the first 2,000 candidates.

Erauzterm was applied to the collected corpora and obtained term lists with the sizes detailed in Table 6.6.

	Corpus		
	Atomic and Particle Physics	Computer Science	Biotechnology
Extracted term list size	46,972	163,698	34,910

Table 6.6: Sizes of the extracted term lists

6.3.3 Evaluation

The candidate term lists were automatically validated against a recently compiled Basque terminological dictionary, which contains 25,000 terms, **Zientzia eta Teknologiaren Hiztegi Entziklopedikoa** or **Basic Dictionary of Science and Technology** (Elhuyar Foundation, 2009), hereinafter **BDST**. The best ranked ones of the remaining candidates were manually evaluated by experts to decide if they were terms or not.

Table 6.7 shows the number of terms validated manually or by the dictionary, for each of the three domains.

		Corpus		
		Atomic and Particle Physics	Computer Science	Biotechnology
Term candidates		46,972	163,698	34,910
Dictionary validated		6,432	8,137	6,524
Manually evaluated	Total	1,147	905	628
	Terms	887	513	432
	Not terms	260	392	196

Table 6.7: Number of terms validated by the dictionary or manually

The domain precision of the term lists was evaluated by analysing the distribution of the terms across the domains, taking the domains of the dictionary as a reference. The results of this evaluation are shown in Figure 6.5, where we can observe that all three lists show peaks in or around their respective domains, which proves that the corpora are indeed specialized and that the term lists automatically extracted belong mainly to the desired domains.

The precision of the extracted term lists, that is, the percentage of the extracted terms that in fact belonged to the desired domain, was also evaluated. Figure 6.6 shows the evolution of this precision as the number of candidate terms grows. Here we can observe that the results are different for each of the domains. As a general rule, we can say that pure sciences perform better than technologies, which might indicate that these domains are more *terminologically dense*, although we cannot be sure about this, because it could also be due to the different nature –extension, diversity, production– of the domains. Besides, we believe that the seed document selection might also affect the quality of the resulting corpora and term lists.

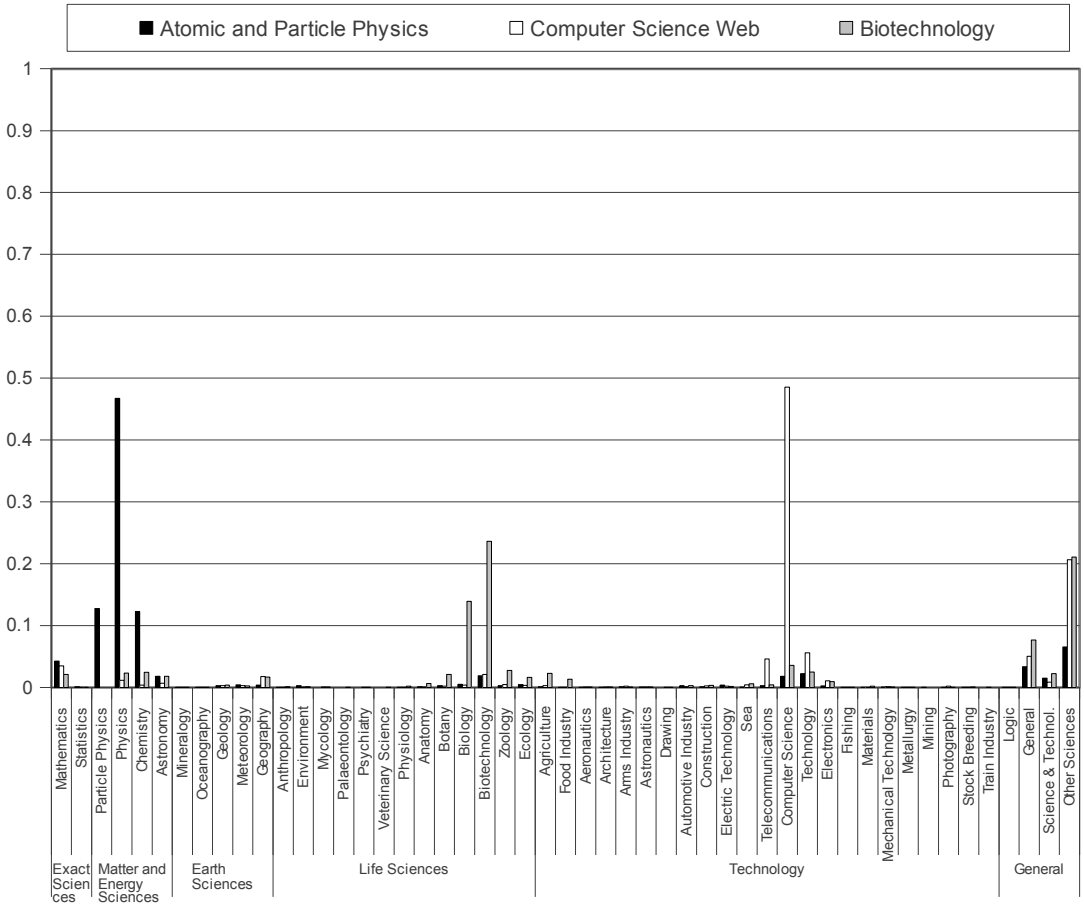


Figure 6.5: Domain distribution of the extracted term lists

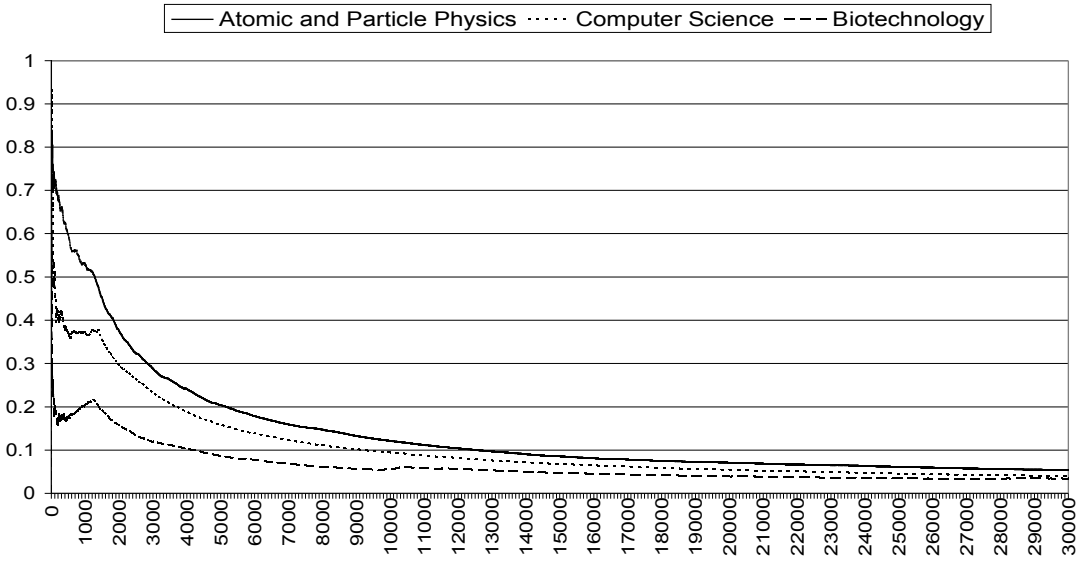


Figure 6.6: Domain precision of the extracted term lists

Also, the sizes of the collected corpora do not seem so important as far as the term extraction task is concerned: the Atomic and Particle Physics corpus achieves better results than the Biotechnology one, the former being almost half the size of the latter (Table 6.5). As we have already pointed out, the nature of the domain is more important.

We also compared the extracted term lists with the terms on the domains of the BDST and looked at the recall, that is, the percentage of the dictionary achieved, and the number of new terms extracted that were not in the dictionary. These two pieces of data are shown in Figures 6.7 and 6.8. By looking at the recall, we could draw the conclusion that the corpus building process is not good enough for compiling a quality dictionary, but we will see later that a traditional corpus does not do better. The use of corpora lacking representativeness could be put forward as a reason for that flaw. But another possible explanation for this fact could lie in the current situation of Basque terminology and text production. Although Basque began to be used in Science and Technology thirty years ago, it cannot be denied that there is a given amount of highly specialized terminology that is published *ex novo* in dictionaries, with little document support if any. That could be the reason why several terms chosen by experts and published in the dictionary do not occur in either of the two corpora. However, we can see in Figure 6.8 that many new terms appear, so the process proposed is definitely interesting for enriching or updating already existing specialized dictionaries.

6.3.4 Comparison with a manually built corpus

In order to know whether a corpus built automatically from the web could obtain better or worse results than a manually built one in the automatic terminological task, we extracted the sub-corpus of the traditionally built Computer Science domain from the Basque Corpus on Science and Technology or ZT Corpusa (Areta et al., 2007) –hereinafter ZTC–, and terminology was extracted with the same method used with the Computer Science web corpus. Then both lists were compared. Table 6.8 shows data on these two corpora and their respective term lists.

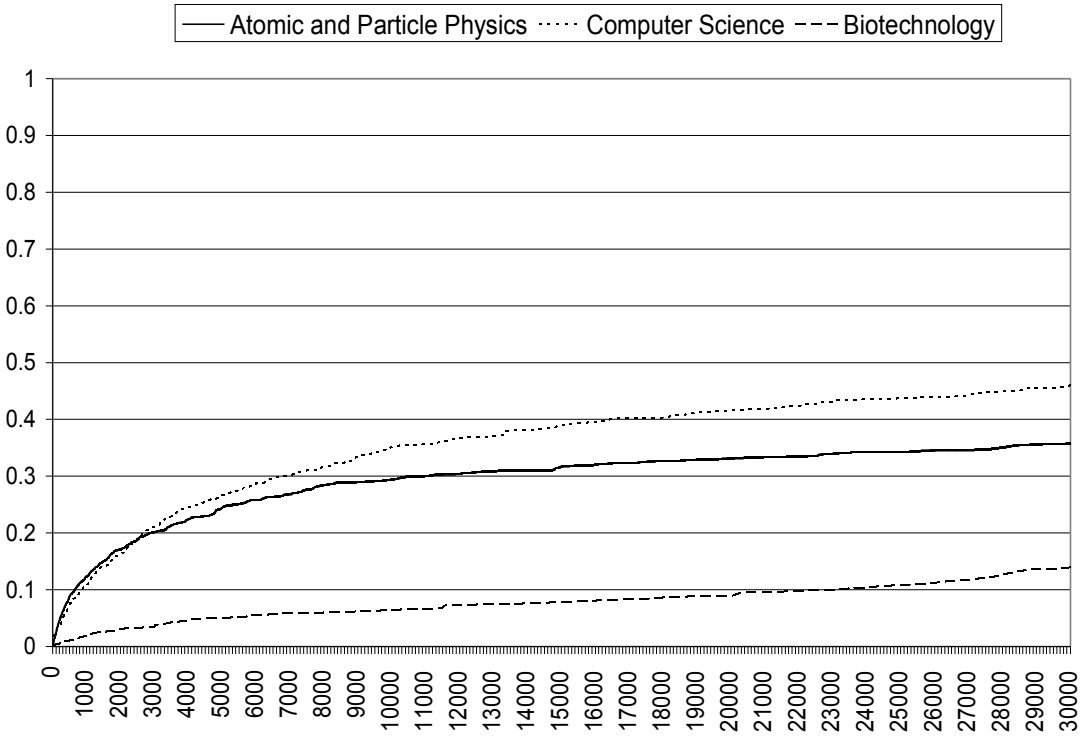


Figure 6.7: Recall of the extracted term lists compared with the dictionary

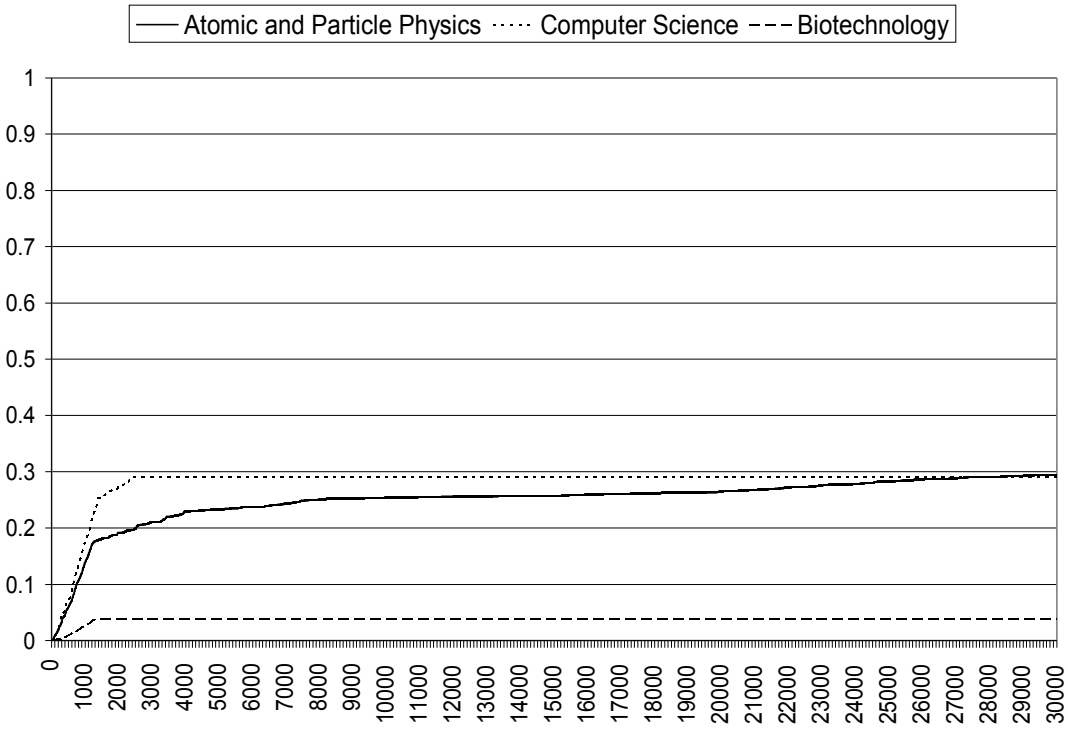


Figure 6.8: New terms in the extracted term lists that were not in the dictionary

		Corpus	
		Computer Science Web	Computer Science ZTC
Corpus size		2,514,290	332,745
Extracted term list size		163,698	24,283
Dictionary validated		8,137	3,389
Manually evaluated	Total	905	1,022
	Terms	513	479
	Not terms	392	543

Table 6.8: Corpus and term list sizes obtained for the web and traditional corpora

Figures 6.9, 6.10, 6.11 and 6.12 show, respectively, the domain distribution, domain precision, recall compared with the dictionary, and new terms that were not in the dictionary of the two extracted term lists. They prove that we can obtain similar or, in some aspects, even better results with the automatic corpus collection process. As the cost is much lower, we believe that the process proposed here is valid and very interesting for terminological tasks.

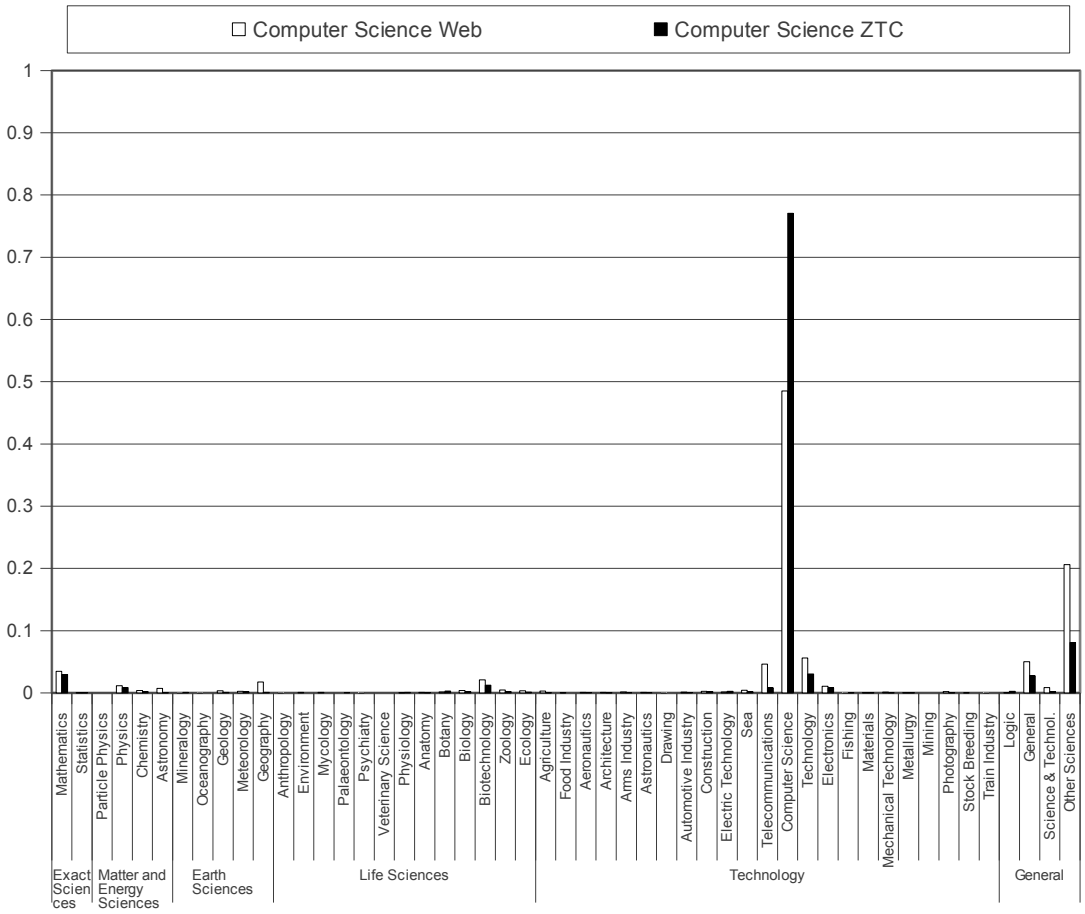


Figure 6.9: Domain distribution of the extracted term lists

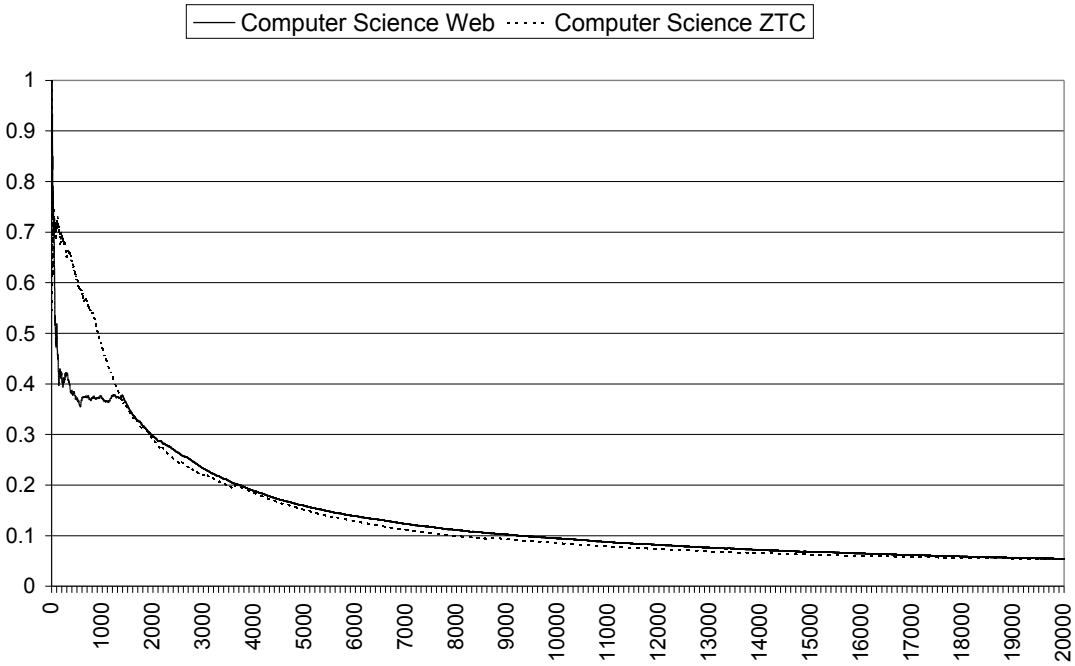


Figure 6.10: Domain precision of the extracted term lists

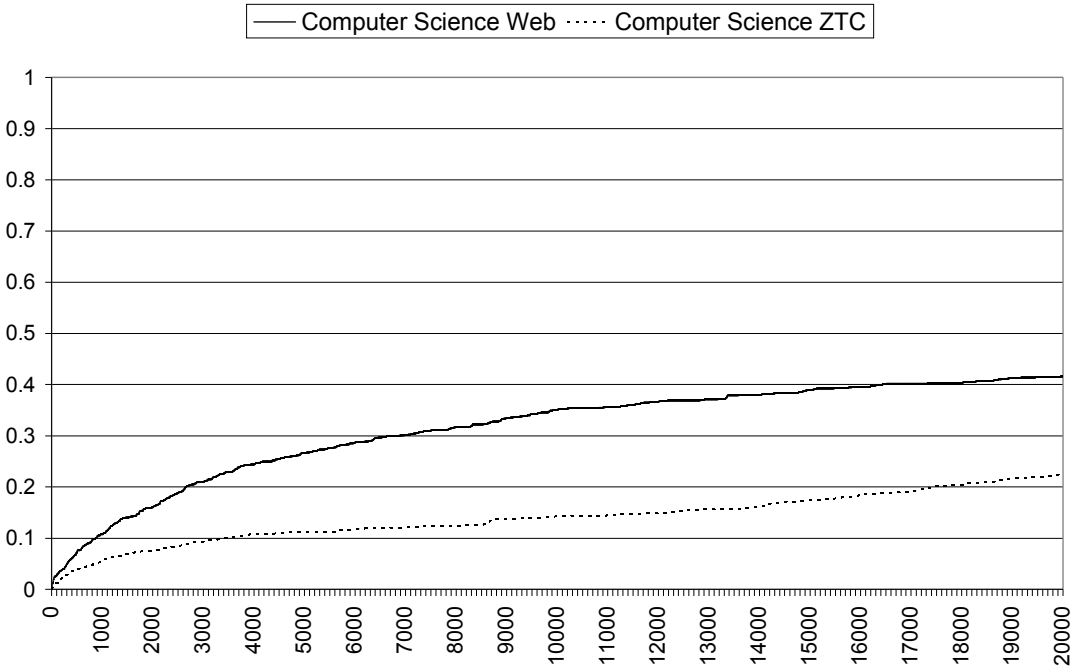


Figure 6.11: Recall of the extracted term lists compared with the dictionary

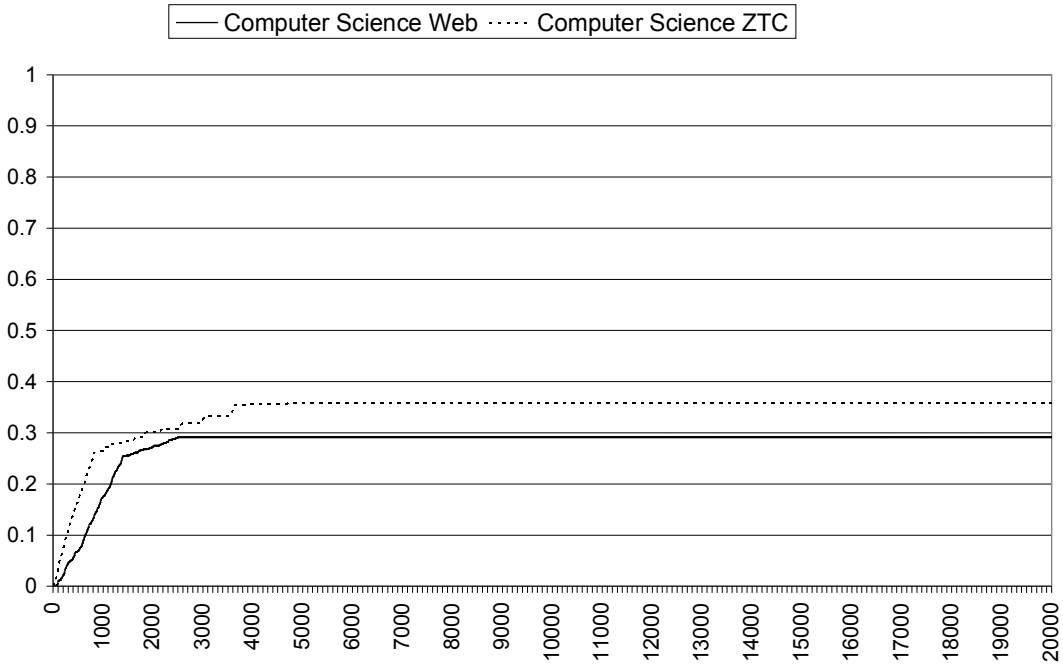


Figure 6.12: New terms in the extracted term lists that were not in the dictionary

6.4 Conclusions

In this chapter we have proposed a method for improving the very bad performance of the BootCaT method with Basque (because of its inappropriate treatment by search engines), based again on morphological query expansion and language-filtering words, and proved that it works and can achieve the same precision as other languages do.

We have also proposed, implemented, and evaluated a methodology to improve BootCaT's domain-precision by using a small sample of documents from the domain from which to automatically extract the seed words and which will be used for evaluating similarity in a final domain-filtering stage. We compare the keyword vectors extracted from downloaded pages using a measure such as RRR or LLR with those of the documents from the small sample using the cosine, and accept the pages if the maximum cosine is above a threshold. The method can attain high domain precision (80%-90%) but sometimes at the expense of recall, which might be an issue for a language like Basque where there are possibly not many documents in the web in some domains. In this method, preparing an adequate initial sample is key for the performance: all sub-areas of the domain should be present in the sample if we want them to be present in the collected corpus too.

We have employed some Basque specialized corpora collected using the methodology in an automatic terminology extraction task, and evaluated the domain distribution of the extracted terms, their domain precision, their recall with regard to a Basque terminological dictionary, and the new terms they contribute to it. The results prove that the corpora and term lists obtained are indeed specialized in the desired domains and lead us to believe that our automatic corpus collection method, in combination with our automatic term extraction process, can be valid for terminology tasks.

We have also evaluated one of our corpora against a traditionally built corpus, and the evaluation shows that we can obtain results almost as good as with a traditional corpus regarding precision or new terms, and even better in the case of recall.

Overall, the evaluation results are encouraging and indicate that more than acceptable results can be obtained with much less work than by means of a completely manual corpus building process.

And again, we think that the methodology can be applied to other less-resourced languages as well, since the techniques used in it (morphological query expansion, language-filtering words and final domain filtering stage) need no more than the same basic tools (N-gram-based language detection and morphological analysis and generation), which many languages have available. But they will most likely also have the same problems as Basque, that is, that there might not be enough texts in some domains to obtain corpora of the sizes that some tasks require.

7 Collecting domain-comparable corpora in Basque and another language from the web

Multilingual corpora are a very valuable resource for many tasks, like translator training, machine translation, terminology extraction... Traditionally, the type of multilingual corpora most commonly used have been parallel corpora, which are collections of texts and their translations, aligned at the sentence level –each sentence is paired with its translation. This alignment level makes it ideal for the above-mentioned tasks.

However, these resources are not abundant, especially for some language pairs and domains. For Basque, for example, the only parallel corpora available are few, small, all general or from the administrative domain and all in the Spanish-Basque pair.

For this reason, in recent years comparable corpora have been used more and more. These are multilingual corpora where the texts have some feature in common, *e.g.*, domain, genre, timespan, etc. They are easier to collect than parallel corpora and can be used in similar tasks, with similar results if larger corpora are used. For Basque it is indeed a very interesting alternative to parallel corpora.

Comparable corpora can be obtained from news agencies, because a large proportion of the news is often the same in different places and languages. Or they can be obtained from the web by crawling web sites specialized in a domain. But the corpora obtained this way are usually from few sources, and it is not a valid choice for small languages like Basque, because in many domains there are just no sources with enough texts in the domain.

In this chapter we detail the work we have done and the results we have obtained when exploring ways of collecting domain-comparable corpora from the web by using search engines. We use this approach because it is the most used strategy for collecting specialized monolingual corpora –and collecting domain-comparable corpora can be considered an extension or a particular case of this– and we have proved that it can be used to obtain specialized corpora in Basque. The works detailed here have been previously published in (Leturia et al., 2009).

7.1 Proposed approaches

The first condition –necessary but not sufficient on its own– for two corpora in different languages to be considered domain-comparable is, obviously, that they belong to the same domain. The BootCaT (Baroni and Bernardini, 2004) tool and method can be used to obtain two such domain-specific corpora in different languages. But any loss or non-perfection in the domain-precision obtained in each of them affects the quality of the comparable corpus, and BootCaT's domain precision is not at all perfect (up to one third of the texts may be inappropriate, according to its authors).

Thus, we obtain the domain-specialized sub-corpora of each language with the method described in Sections 6.1 and 6.2, with which we attain a higher domain-precision (over 90% and higher can be achieved). The steps involved in the process are as follows: we start from a sample mini-corpus on the domain, we automatically extract keywords from it (as described in Subsection 6.2.1), combinations of these are sent to the search engines (using morphological query expansion and language-filtering words in the case of Basque, described in Section 3.2, and using the search engines' *filter by language* option for English), the returned pages are downloaded, the pages go through various cleaning and filtering stages (language-filtering at the paragraph level, length filtering, boilerplate removal, near-duplicate detection and containment detection; they are all explained in detail in Chapter 4) and a final domain-filter is applied using the mini-corpus as reference (described in Subsection 6.2.2).

With the method described above and a domain-filtering threshold that is high enough, we can obtain monolingual specialized corpora with very high domain precision. Higher thresholds ensure 80-90% or higher precision rates, if recall is not an issue. Otherwise, the threshold can be set to achieve the desired precision/recall balance.

However, even if we attained 100% domain-precision in each of the sub-corpora of each language, this is not enough to guarantee good comparability. Out of two corpora strictly on computer science, one could be mostly made out of texts on hardware and databases and the other on programming and networks; they can not be considered very comparable, and they would most likely not be very practical for any of the aforementioned tasks. Therefore, we are not only interested in obtaining two specialized corpora, but also that these be as comparable as possible. To achieve this we propose two different variants of applying this method to two different languages, which are explained below.

7.1.1 Two sample corpora method

The most obvious way is to use a sample mini-corpus for each language and launch the corpus collecting process independently for each of them. If the sample mini-corpora used are comparable or similar enough (ideally, a parallel corpus would be best), the corpora obtained will be comparable to some extent, too. Figure 7.1 shows a diagram of this process.

7.1.2 Dictionary method

The other method uses only a sample mini-corpus in one of the languages, and uses dictionaries for translating the extracted seed words (this is manually revised) and the domain-filtering vectors for the other language. Figure 7.2 shows the process this method follows. Both Figures 7.1 and 7.2 highlight in grey the parts that are different from the other method.

This method, theoretically, presents two clear advantages: first, the sample mini-corpora are as similar as can be (it is only one), thus we can expect a greater comparability in the end; and second, we need only collect one sample corpus.

But in reality, it presents some problems too, mainly the following two: first, because dictionaries do not cover all existing terminology, we could have some **OOV** (**Out Of Vocabulary**) words and the method may not work so well –in our experiments we have found quite a few, although we use a combination of a general dictionary and a specialized one to maximize translation coverage–; second, we have to deal with the ambiguity derived from dictionaries, and selecting the right translation of a word is not so easy. These difficulties, which are by no means insignificant, lead us to expect worse results from this method; nevertheless, we have also tried and evaluated it. To reduce the amount of OOV words, the ones that have been POS-tagged as proper nouns are included as they are in the translated lists, since most of them are named entities. And for resolving ambiguity, for the moment, we have used a naïve *first translation* approach, widely used as a baseline in NLP tasks that involve translation based on dictionaries. The basic idea this relies on is that many dictionaries order their translations according to frequency of use.

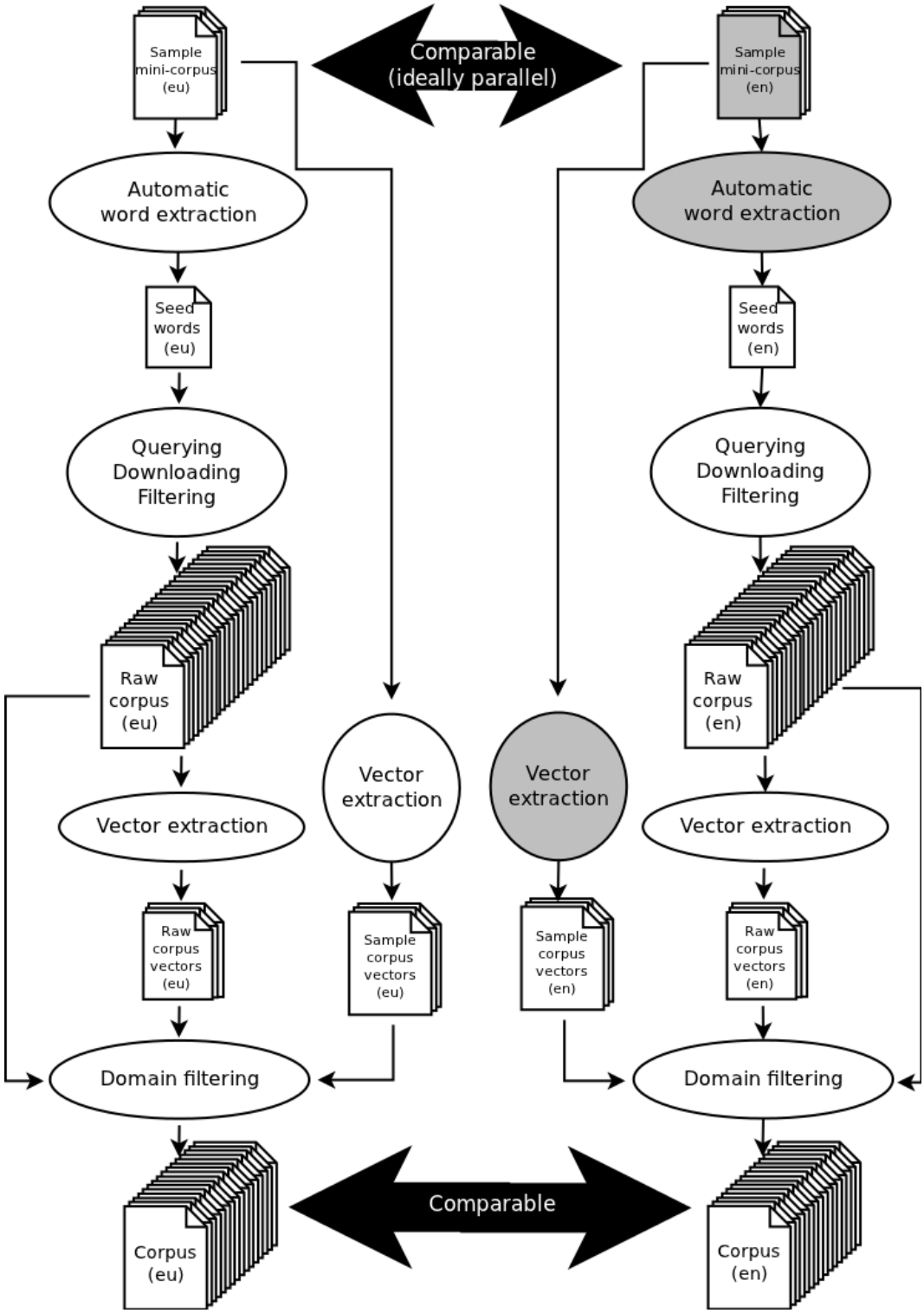


Figure 7.1: Diagram of the different sample corpora method

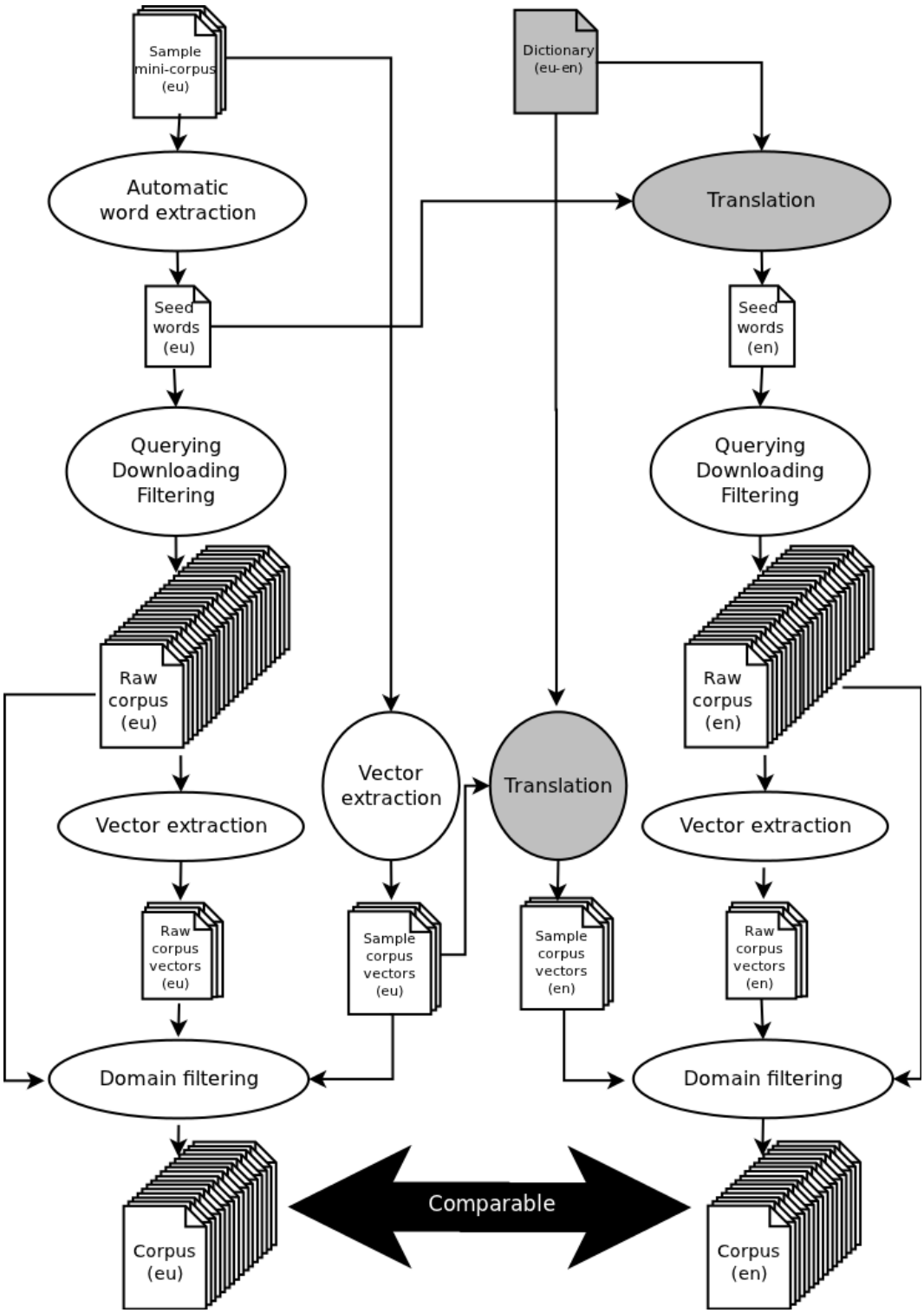


Figure 7.2: Diagram of the dictionary method

7.2 Evaluation

In order to see which of the two methods obtains a higher degree of comparability, we collected two Basque-English comparable corpora, one on Computer Science and the other on Tourism, with each of the two methods mentioned above. The sample mini-corpora used for Computer Science are 33 short articles (about 40,000 words) obtained from popular science magazines, and for Tourism 10 short articles (about 7,000 words) obtained from tourist office websites. The English versions of the sample mini-corpora are comparable in the case of Computer Science, and parallel in the case of tourism. The final size of the Computer Science corpora amounts to 2.5 million words in each language, and in the case of Tourism, 1.5 million words.

Evaluating the comparability of the obtained corpora is not an easy task, simply because there is no standard measure for measuring the comparability of multilingual corpora, and there is little literature on the topic. There are some works on measuring corpus similarity based on word-frequency lists (Kilgarriff and Rose, 1998; Kilgarriff, 1997) and others that also use POS and semantic tag frequencies (Rayson and Garside, 2000), but they deal with monolingual corpora. An option would be to apply these to multilingual comparable corpora using dictionaries.

Morin et al. (2007) suggest that, for the task of terminology extraction, the quality of a comparable corpus might be more important than its size, and show that they obtain better results with a smaller corpus if both sub-corpora belong to the same register. So the genre or register could be another criterion to weigh the comparability. But word-frequency lists are not valid features for genre identification; punctuation marks and POS trigrams should be used for this task (Sharoff, 2007; Argamon et al., 1998). In any case, domain similarity is more important for terminology extraction than genre or size, so at the moment we are more interested in the former kind of comparability (keyword frequency-based).

Finally, Saralegi, San Vicente, and Gurrutxaga (2008a) propose measuring the comparability of a corpus by computing the semantic similarities at the document level. The hypothesis behind this is that the containment of many document pairs with a fairly high semantic similarity improves terminology extraction based on context similarity.

Thus, for evaluating our two methods, we used two different ways to measure the comparability of the four corpora obtained, both based on keyword frequencies: one is by calculating the cosine distance between the vectors containing all the keywords of each corpora weighted by LLR; the other is by calculating the Chi Square (χ^2) statistic for the *n* most frequent keywords, as described by Kilgarriff and Rose (1998). But it must be taken into account that, unlike any other corpora similarity measures mentioned in the literature, the corpora we compare are in different languages, so our measurement necessarily uses dictionaries; again, we resolve ambiguities with a first-translation approach for the sake of simplicity.

The results of the evaluation are shown in Table 7.1. For the cosine, higher values are better; for χ^2, a lower value indicates greater similarity. Best results are shown in bold.

Corpus	Method	Cosine, LLR, all keywords	χ^2, n most frequent keywords				
			500	1,000	5,000	50,000	All
Computer Science	Two sample corpora	0.4102	700.61	481.57	148.70	17.60	16.55
	Dictionary	**0.4396**	**685.95**	**471.64**	**145.20**	**17.25**	**15.51**
Tourism	Two sample corpora	0.1164	382.80	**256.29**	**83.23**	**12.82**	**12.82**
	Dictionary	**0.1511**	**380.62**	261.78	86.35	13.00	13.00

Table 7.1: Evaluation results

Although the dictionary method might *a priori* appear to be a worse method –owing to OOV words and ambiguity–, the evaluation does not confirm this. In fact, the dictionary method proved to be better in most of the measures. However, this evaluation cannot be considered conclusive, for the following reasons:

- The evaluation was done with only two corpora, which show different results for some of the measures.
- We now believe that tourism might not have been a good domain choice for the evaluation, because it does not completely fit into what we know as a specialized domain (interdisciplinary terminology, etc.). Evaluations with more corpora and more domains are needed before making any definite assertions.
- There is not much literature on corpora similarity methods. Some measures have been proposed –mostly based on word frequency measures–, but

they have not been sufficiently evaluated, and there is no standard measure indeed. And regarding corpora in different languages, there is no precedent for measuring similarity. We have employed some of the proposed measures using dictionaries, and they show different results. We believe there is an urgent need for research on and standardization of multilingual corpora similarity methods.

- There might be a bias towards the dictionary method since we are using a dictionary to measure the similarity, too. To illustrate this, we can imagine an extreme case, in which using the dictionary method all the seed words have been disambiguated incorrectly and the corpora obtained has nothing to do with the desired topic, but since the same dictionary and disambiguation method is applied to the keyword vectors when evaluating the similarity, the measure obtained might still be high. However, we do not see a solution to this.

For future work, it would be interesting to try to improve the dictionary-based approach; as we have already mentioned, the preliminary work needed to obtain a comparable corpus with this method is considerably reduced (only one sample mini-corpus needs to be collected); besides, there is still much room for improvement. One of the things to be tried is to see whether manual revision of the translated vectors to be used in the domain filtering yields a better performance. Another one is to try more complex translation selection techniques –instead of the first-translation approach–, and also synonymy expansion.

7.3 Evaluation on an automatic terminology extraction task

Again, we thought it would be interesting to test the quality –in the form of domain-comparability, in this case– obtained by our method by subjecting some corpora collected with it to the automatic terminology extraction task and evaluating the results.

Since there are two automatic processes involved (the corpus collection and the terminology extraction), the performance of the whole process and the quality of the final bilingual terminology lists are affected by both, and it is not easy to tell to what extent each of them has influenced the result. So in order to measure the performance of the comparable corpora collection tool more effectively, we also collected similar compar-

able corpora manually (same domain, similar size) and performed the same terminology extraction out of them. This way, we ended up with a reference against which to measure the performance of the automatic corpora collection tool.

The works described in this section were published in (Gurrutxaga et al., 2013).

7.3.1 Corpora collection

7.3.1.1 Corpora collected automatically

The corpora were collected using the two sample mini-corpora method (see Subsection 7.1.1). The domains chosen were two: Computer Science and Physics. We prepared sample mini-corpora for each of them. The mini-corpus of Computer Science in Basque consisted of the 41 articles about Computer Science in the Basic Dictionary of Science and Technology (Elhuyar Foundation, 2009) –hereinafter BDST–, a dictionary containing 23,000 concepts from all the sub-domains of Science and Technology; all concepts are provided with a definition in Basque and equivalents in Spanish, English, and French, and 600 of them with an encyclopedic article in Basque. In addition to that, 33 news items from Zientzia.net (Elhuyar Foundation, 2001), a popular science website in Basque, were included in the mini-corpus. The sample mini-corpus of Physics in Basque consisted of the 76 articles on physics from the BDST. For the mini-corpora in English, the articles in Wikipedia that defined the same concepts as in the articles of the BDST were taken, and we googled for news that dealt with the same subjects as the news items from Zientzia.net.

We did not have a target size for the automatically collected corpora: for bilingual terminology tasks based on context similarity, the larger the corpora, the better the results are; and the Internet in Basque is not as big as that in other languages and the number of pages we would want to collect for a particular domain might not exist. So we simply launched the collecting processes for the Basque part and stopped them when the growing speed of the corpora fell to almost zero, thus obtaining corpora that were practically as large as possible. Then we obtained English corpora that were roughly 40% larger in words than the Basque corpora (Basque appends articles and prepositions to content words, so for the same texts in Basque and English, the English ones are about that percentage larger in terms of words). The sizes of the sample mini-corpora and the obtained corpora are shown in Table 7.2.

		Corpus			
		Computer Science		Physics	
		Basque	English	Basque	English
Sample corpus size	Docs	74	74	76	76
	Words	66,461	193,406,266	73,760	306,263
Obtained corpus size	Docs	1,310	1,051	780	442
	Words	2,506,049	3,506,218	1,155,995	1,710,219

Table 7.2: Sizes of the sample mini-corpora and the obtained corpora

7.3.1.2 Corpora collected manually

The corpora collected manually are roughly the same size and were obtained from different sources: books, media, websites, etc. We are aware that the manual corpus building implemented cannot be compared with a standard reference corpus building process, which is designed to guarantee or achieve to a great extent a high level of representativeness and balance. Nevertheless, we made a careful selection of the Basque books on Physics and Computer Science freely available, and tried as much as possible to include texts from the different sub-domains. In some sub-domains, there are few Basque publications (for example, Optics in Physics, or Artificial Intelligence in Computer Science). This fact must be taken into account in the selection of English texts for the corresponding manual corpora. So we endeavoured to collect English texts that, at least *a priori*, would ensure that the manual corpora were as comparable as possible.

The sub-domain distribution of each of the corpora is shown in Figures 7.3 and 7.4 and, as can be seen there, they can be considered to be comparable enough.

7.3.2 Terminology extraction

The terminology extraction tool we use is **AzerHitz** (Saralegi et al., 2008b; Saralegi et al., 2008a). This system is based on the idea that translation equivalents tend to co-occur within similar contexts, the same hypothesis that is used for the identification of synonyms. In order to do this, AzerHitz extracts the candidate terms of each language, then models their contexts, translates them, and calculates their degree of similarity. Alternatively, translation equivalents are also detected by means of string similarity or cognate detection. We will describe each of the steps in more detail.

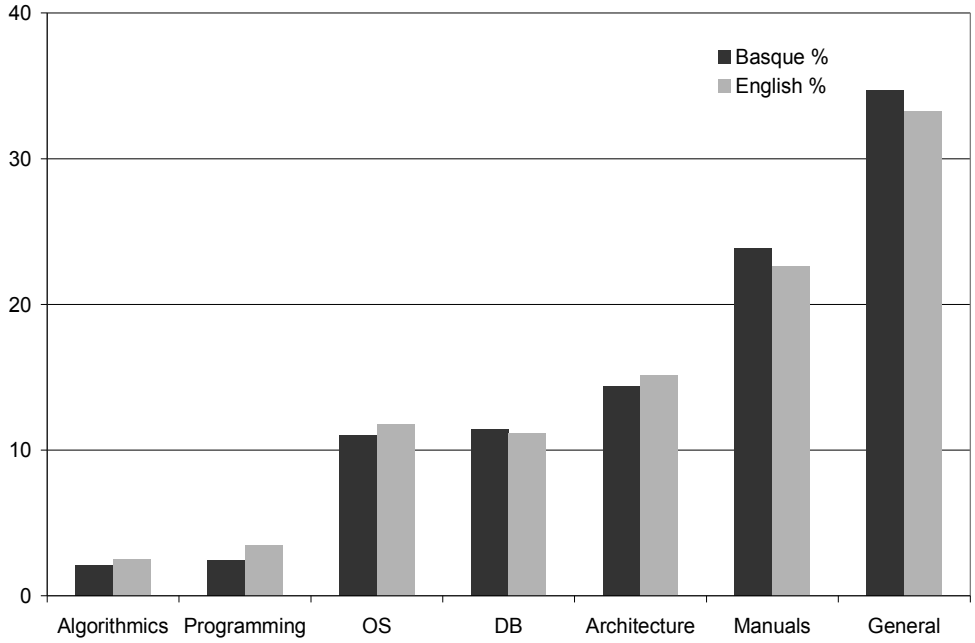


Figure 7.3: Subdomain distribution of the Computer Science manual corpus

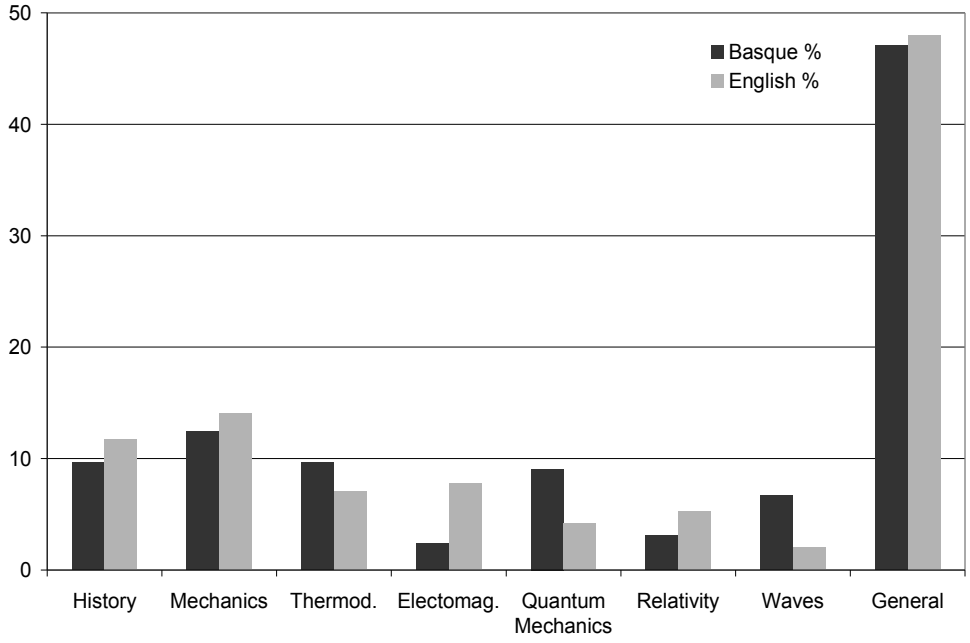


Figure 7.4: Subdomain distribution of the Physics manual corpus

The monolingual term candidate extraction, both in Basque and English, is done by using a hybrid approach that first looks for some linguistic patterns and afterwards applies some statistical processing to them. For Basque, the term extraction tool **Erauzterm** is used (Alegria et al., 2004b; Alegria et al., 2004a) (which is described in Subsection 6.3.2). For English, the corpus is parsed with the chunker of the **TreeTagger** tool (Schmid, 1994) and then the most used morphosyntactic patterns of terms are looked for. Then, different statistical methods are applied to one-word and multi-word term candidates. For one-word terms, LLR is used as termhood measure for calculating the domain relevance of the term with respect to an open domain corpus; for multi-word terms, unithood is used as a clue to termhood; to measure unithood, LLR is employed as the association measure. The ones with the highest measures are taken as the final term candidates.

Regarding the modelling of contexts, only content words are included in the contexts, that is, nouns, adjectives, and verbs. To delimit the contexts of the words, a distance-based window is established: 10 words for Basque (plus and minus 5 around the word) and 14 for English (plus and minus 7). The smaller window for Basque is due to the agglutinative nature of the Basque language, where articles, prepositions, etc. are appended to content words. Punctuation marks narrow the window when they are found. The contexts are modelled using the Okapi probabilistic model offered by the Lemur Toolkit (University of Massachusetts Amherst and Carnegie Mellon University, 2000). The context words of a word are indexed in this toolkit as if it were a document, that is, the words that make up the context of a word throughout the collection are included in the document that is indexed, referred to hereinafter as the **context document** of a word.

To compute context similarity, the Basque contexts are translated into English. A bilingual dictionary is used for this purpose. The first translation approach is taken in the case of ambiguity. For all the OOV words, cognates are looked for among all the content words in the target language. The identification of these cognates is carried out by calculating the **LCSR** or **Longest Common Subsequence Ratio** between the Basque and English content words (after processing some typographic rules to normalize equal phonology n-grams, *e.g.*, *c* → *k*, *factor* = *faktore*, or regular transformation ones, *e.g.*, *-tion* → *-zio*, *reaction* = *erreakzio*). The candidates that exceed the threshold of 0.8 are taken as translations.

Since IR systems rank documents according to the topic similarity with respect to a query, AzerHitz is based on the *topic similarity = context similarity* equivalence. A retrieval is performed by sending the translated context document of the source word as query to the Lemur Toolkit, using the Okapi probabilistic relevance model. The highest-ranked documents returned (which are actually contexts of words) are the most similar contexts, and therefore, the corresponding words are the most probable translations. To prevent noisy candidates, those that have a different grammatical category from that of the word to be searched are pruned.

In addition to context similarity, string similarity between source words and equivalent candidates is also used to rank candidates. LCSR is calculated between each source word and its first 100 translation candidates in the ranking obtained after context similarity calculation. The candidates that exceed the threshold of 0.8 are ranked first, while the position in the ranking of the remaining candidates remains unchanged. A drawback in this method is that cognate translations are promoted over translations based on context similarity.

The process that AzerHitz follows can be seen in Figure 7.5.

7.3.3 Evaluation

Out of these comparable corpora, we first extracted the monolingual terms out of the sub-corpora of each language and evaluated their domain precision against a dictionary reference. Then we tried to find the English translations for the most relevant Basque terms using the context similarity and cognate detection method. The extracted translation pairs were automatically evaluated using the dictionary reference. Those terms not found in the dictionary were evaluated manually.

The dictionary reference is made up of two sources: a) the BDST (27,084 English terms and 25,143 Basque terms) and b) a terminological database which includes terms from terminological dictionaries published by Elhuyar, specialized terms published in Elhuyar's general dictionaries, and terms extracted from translation memories. This second resource includes terms from a large variety of domains, not only from Science and Technology. In order to simplify the presentation of the results, and bearing in mind that the experiments are dealing with Physics and Computer Science

corpora, we gathered together all the domains not belonging to Science and Technology into *Other domains*. The total number is 62,043 for English terms and 101,479 for Basque terms.

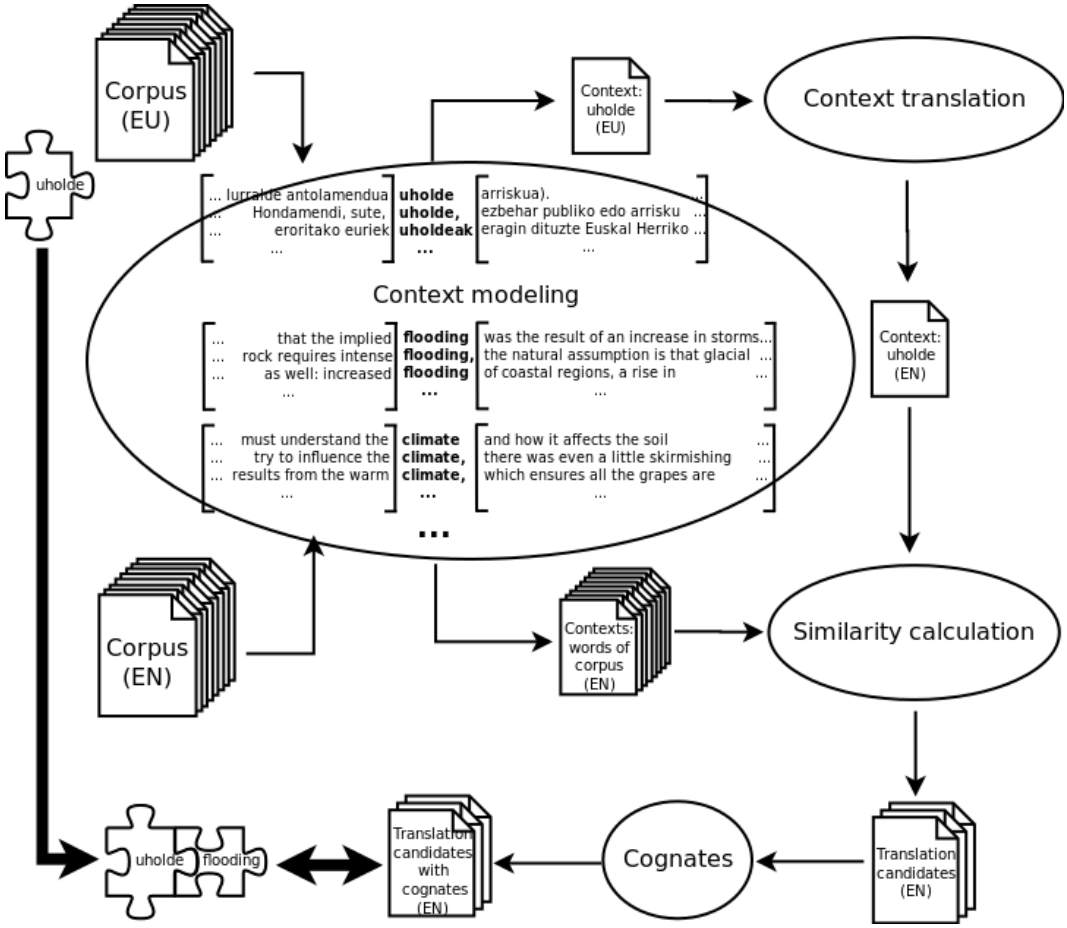


Figure 7.5: Diagram showing the process of searching for the translation of a word by context similarity and cognate detection

7.3.3.1 Monolingual terms

We first performed an evaluation of the domain precision of the corpora by evaluating the monolingual terms extracted from them. Taking into account that most terms are single words and bigrams, in the interest of simplification, the evaluation focused only on single word and bigram terms. For that purpose, all extracted term candidates with an LLR score above 10 and with a frequency over 25 in Computer Science and 10 in Physics were evaluated against the dictionary reference, in order to measure the overall term precision and the domain precision. The difference in the minimum frequency

required for the different domains is due to the smaller size of the corpora of the Physics domain. The domain distribution of the extracted terms validated against the dictionary reference can be seen in Figures 7.6 and 7.7.

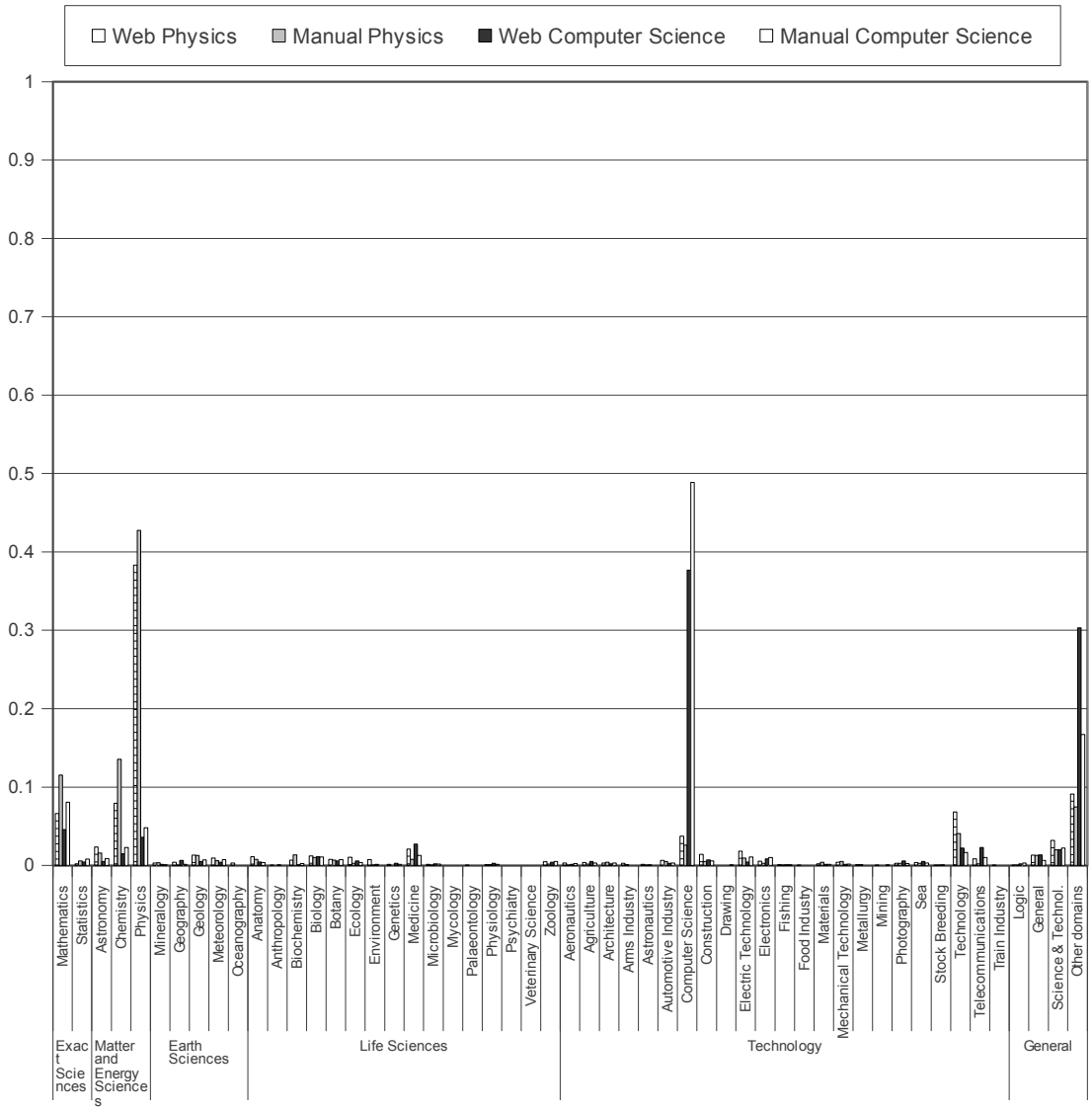


Figure 7.6: Domain distribution of extracted Basque terms

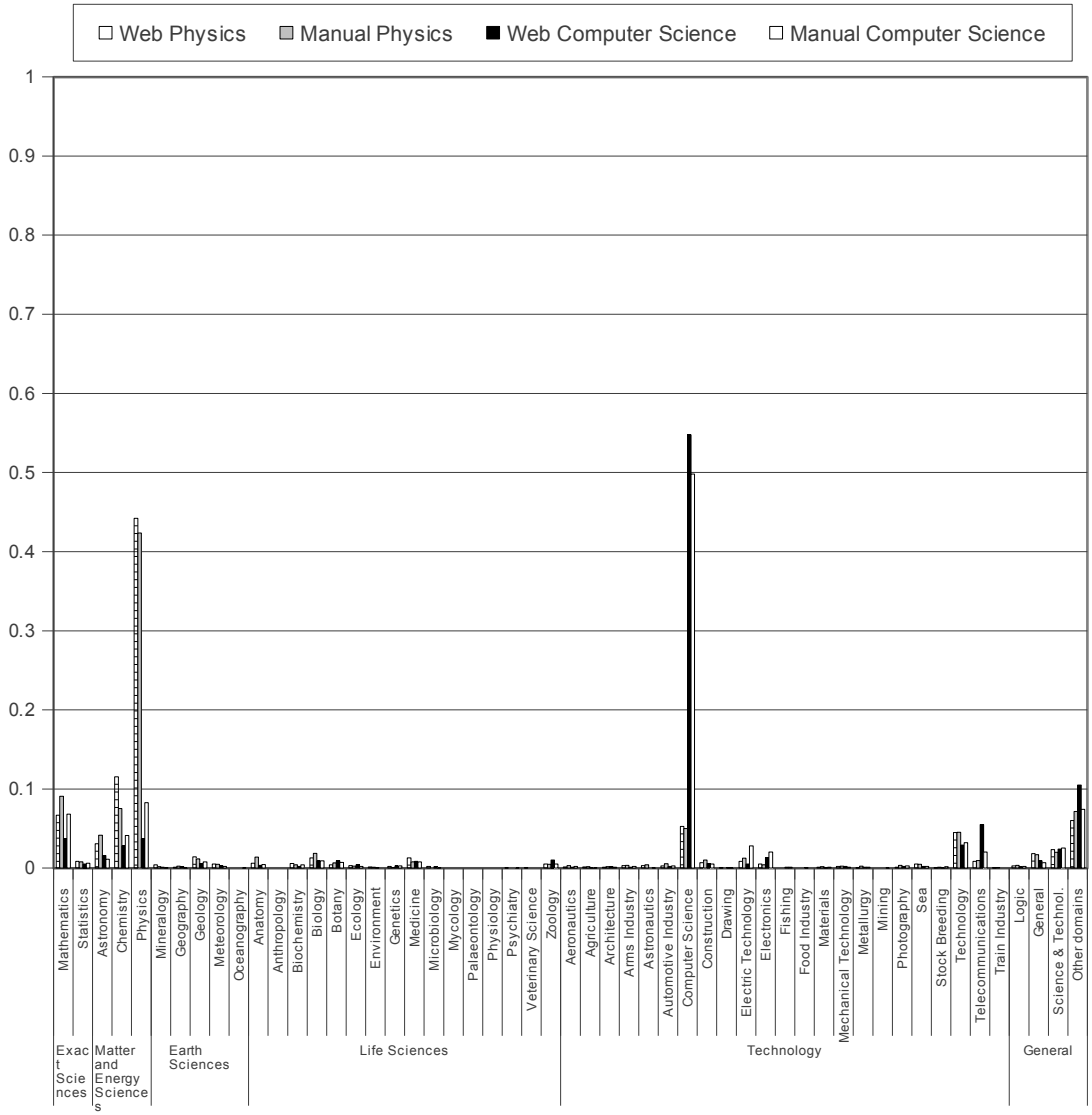


Figure 7.7: Domain distribution of extracted English terms

We can observe that all corpora logically show peaks in their respective domains or in domains closely related to them. This is true for the automatically collected web corpora too (in fact, in the English extraction, the web corpora behave better than the manual ones). The proportion of dictionary-validated terms belonging to the desired domain is generally over 40%. We can compare those figures with the proportions of Physics and Computer Science terms in the dictionary (6.08% and 3.64% respectively), and conclude that even though there is no objective reference to draw a comparison with, the corpora have a clear domain-profile and can be regarded as specialized. Nevertheless, a more detailed analysis of the domain-distribution results pre-

viously presented, alongside the domain precision results for n-best candidate lists selected by LLR and ordered according to decreasing frequency shown in Figures 7.8 and 7.9, reveal that the corpora have some different characteristics. In general, domain precision close to 0.4 is achieved in every extraction when the 1,000 best candidates are selected. The precision is slightly lower in Basque extractions perhaps due to the worse quality of the Basque corpora. This difference decreases as the amount of selected candidates increases.

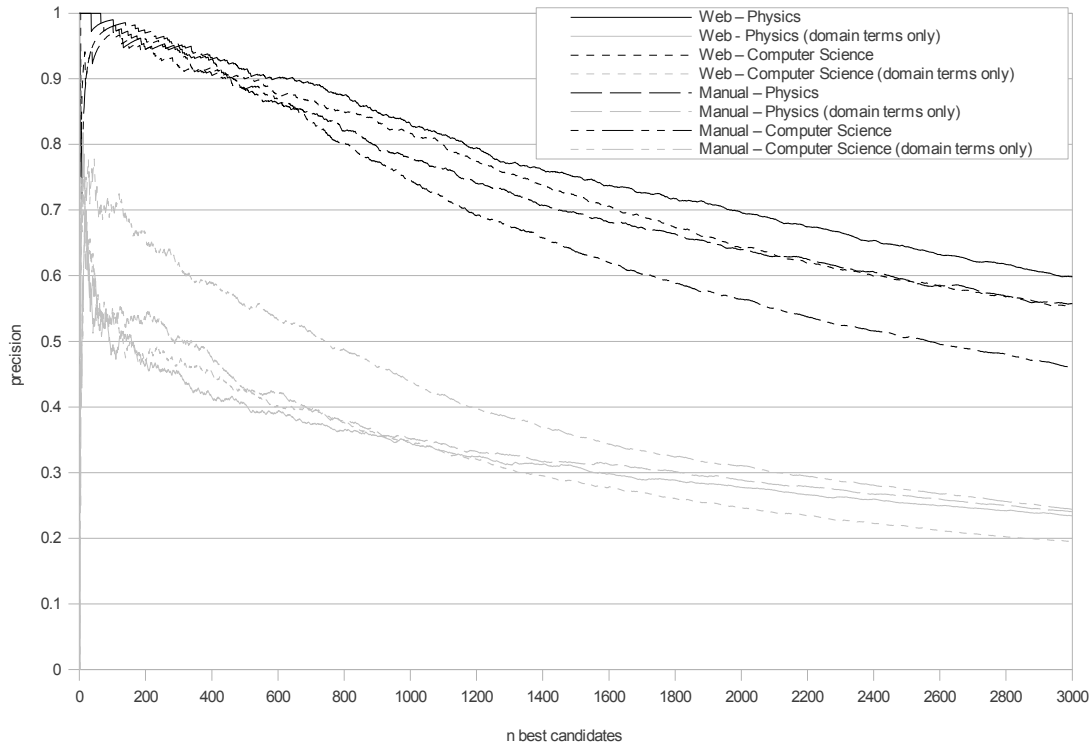


Figure 7.8: Term precision and domain precision of terms extracted from Basque corpora

Another observation we can make by looking at the domain distribution of the Basque terms (Figure 7.6) is that the Computer Science web corpus peaks sharply in *Other domains*. A possible explanation for this is that computer science is an applied science that is used in many areas such as linguistics, education, etc., and so texts and terms from these domains appear, and in the manual corpus too (Milios et al., 2003). But the proportion of these in the Basque web corpus is higher most likely because we

tried to make the corpus as large as possible and, therefore, we might have lowered the domain precision requirements too much (the Basque web is not as rich in specialized content as we would wish).

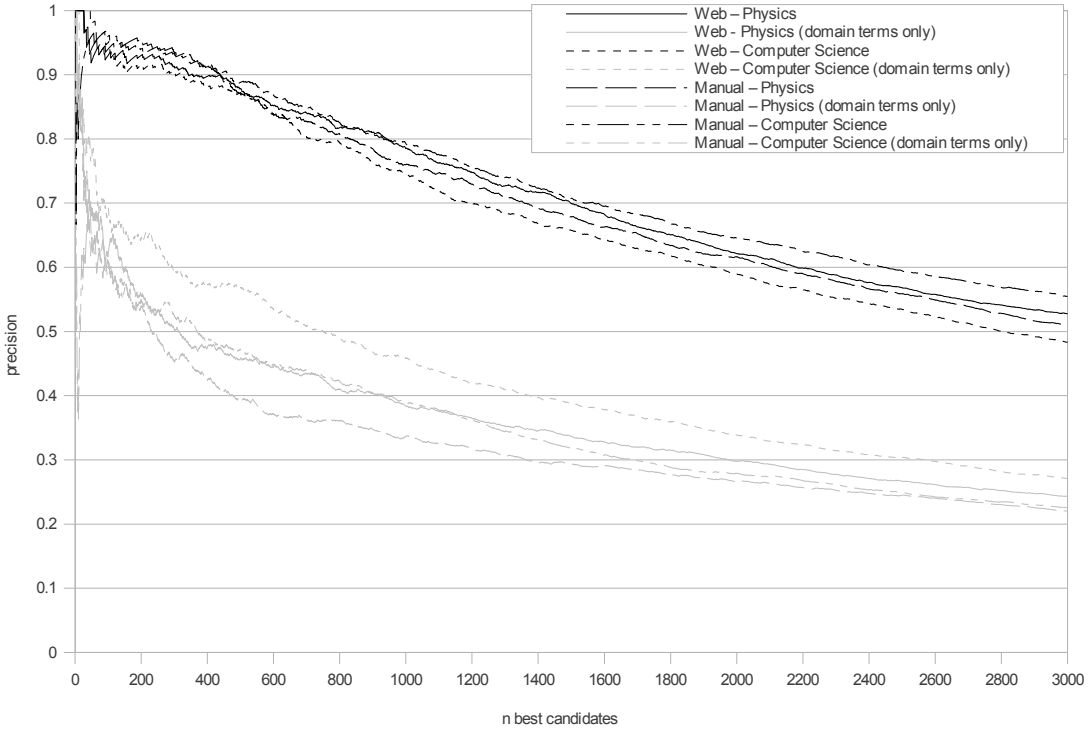


Figure 7.9: Term precision and domain precision of terms extracted from English corpora

Furthermore, the domain precision of the terms in the web corpora is lower for Basque than for English, both in Computer Science and in Physics (Figures 7.8 and 7.9). This might again be due to the scarcity of specialized content in the web in Basque, compared to English.

If we compare the domain precision of the Basque terms from the web corpora with those of the manual corpora, we can observe that the Physics corpora are more similar to each other than the Computer Science ones. The reason for this could be, as we have already pointed out, that we might have forced the size of the Basque Computer Science web corpus too much and to the detriment of domain precision.

Looking at English terms, surprisingly the domain precisions of the web corpora are better than those of the manual corpora. This phenomenon is also reflected in the bilingual extraction, which we will be discussing in the next subsection.

Finally, it must be noted that due to the completely automatic evaluation using the dictionary, it is possible that the actual precisions are higher, because the automatic process will most likely have found terms that are not present in the dictionary reference, as we observed in the similar experiment referred to in the previous chapter (see Subsection 6.3.3) where a manual evaluation was performed.

7.3.3.2 Multilingual terms

We also evaluated the precision of the multilingual term extraction. As for the source term candidates, we took the 600 single-word Basque terms and the 400 multi-word Basque terms with the highest LLR, all of them having a frequency of over 25 in computer science and 10 in physics, due to the smaller size of the physics corpora. For the English translation candidates we took all the extracted terms with a minimum frequency of 10. We applied the context similarity and cognate detection method for translation term extraction to these source candidates.

The translation pairs obtained were automatically evaluated using the dictionary reference, and the pairs that were not present in this dictionary were manually evaluated by a professional lexicographer. Tables 7.3 and 7.4 show the precision of the bilingual term extraction for different top *n* candidates, using cognate detection and without using it, and separating the results for domain-specific terms and those of other domains, too.

We consider that the results obtained in the bilingual terminology extraction task are acceptable. It must be noted that performance is better when dealing with terms in the desired domain (improvement between 11% and 28%). The cognate detection method also offers a significant improvement. In the Physics domain, our method obtains a precision of 0.31 in the Top-1 (0.47 using cognate detection) and 0.63 in the Top-30 (0.68 with cognate detection). In the Computer Science domain, it achieves 0.25 precision in the Top-1 (0.40 with cognate detection) and 0.60 in the Top-30 (0.65 with the help of cognate detection).

Computer Science		Top 1		Top 5		Top 10		Top 20		Top 30	
		-	Cog	-	Cog		Cog		Cog		Cog
Web	Domain terms	0.25	0.40	0.42	0.51	0.47	0.55	0.56	0.61	0.60	0.65
	Non-domain terms	0.11	0.19	0.16	0.24	0.19	0.26	0.24	0.30	0.26	0.32
	All terms	0.18	0.30	0.30	0.38	0.34	0.41	0.40	0.46	0.44	0.49
Manual	Domain terms	0.08	0.21	0.21	0.31	0.26	0.35	0.33	0.41	0.37	0.44
	Non-domain terms	0.03	0.08	0.06	0.11	0.08	0.12	0.10	0.14	0.11	0.15
	All terms	0.07	0.02	0.19	0.30	0.23	0.33	0.30	0.39	0.34	0.41

Table 7.3: Precision of bilingual term extraction for the Computer Science corpora

Physics		Top 1		Top 5		Top 10		Top 20		Top 30	
		-	Cog	-	Cog		Cog		Cog		Cog
Web	Domain terms	0.31	0.47	0.51	0.60	0.57	0.64	0.61	0.66	0.63	0.68
	Non-domain terms	0.17	0.29	0.27	0.37	0.32	0.41	0.37	0.45	0.40	0.48
	All terms	0.23	0.36	0.38	0.47	0.43	0.51	0.47	0.54	0.50	0.56
Manual	Domain terms	0.18	0.32	0.29	0.40	0.34	0.43	0.40	0.48	0.43	0.50
	Non-domain terms	0.09	0.20	0.14	0.23	0.18	0.26	0.22	0.29	0.24	0.31
	All terms	0.16	0.31	0.25	0.37	0.31	0.41	0.37	0.45	0.40	0.48

Table 7.4: Precision of bilingual term extraction for the Physics corpora

Nevertheless, we cannot really compare these results with other results from the literature, because the experimental setups vary greatly from one to another: the language pair, the size and domain specificity of the corpora, the size and methodology to create the reference term list to translate, etc. Experiments conducted with large news corpora (several million words) report results of up to 90% precision for the top 10-20 candidates (Fung and Yee, 1998; Rapp, 1999). Research carried out with domain specific corpora have been mostly conducted with small medical corpora (several hundred thousands words). Reported precisions range from 50% (Morin et al., 2007) to 74% (Chiao and Zweigenbaum, 2002) for the top 10-20 candidates. Moreover, few works deal with multi-word terms (Daille and Morin, 2005; Morin et al., 2007; Sharoff et al., 2007), and the accuracy is below the results achieved with single word terms. With respect to the language pair, the experiments conducted with the Basque-English pair (Saralegi et al., 2008a) obtained a maximum precision of 79% for high frequency words (frequency > 50) and 43% for lower frequency words, when dealing only with single word terms, and with a 2-million-word popular science corpus.

However, if we compare them with the results corresponding to manually collected corpora, the results of the web corpora are surprisingly better. And this happens despite the fact that, for Basque, the precision results of monolingual term extraction are better for the manual corpora than for the web corpora. One interpretation for this could be that, although the Basque manual corpora are better (more specialized), the multilingual web corpora are more comparable. This reinforces the idea that the Internet, thanks to its large size, can automatically generate corpora that are more comparable than ones produced by working manually.

It is true that the manually built corpora are not reference corpora, that is, fully balanced, randomly chosen, and representative of the domain, but for the domains we are working with and the sizes needed for this task, there are no corpora in Basque of this kind, and in English we are not aware of any either. Reference corpora would most likely outdo our automatically collected web corpora. Nevertheless, we believe that we have built good quality corpora with a reasonable amount of effort and cost. Even with an opportunistic approach, the job has not been easy and, particularly in the Basque part, might not be a possible alternative in many cases (we were lucky to have access to the library of the Elhuyar Foundation, which is a media group for popular science in Basque and which also undertakes a considerable amount of scientific and technical book translations into Basque).

In order to find the reasons for the lower performance of the manual corpora, we manually analysed the domain precision of the first 1,000 monolingual candidates extracted from both web and manual corpora, and we found that English manual corpora are richer in terms corresponding to other science and technology domains. For example, in the Computer Science manual corpus we detected a high presence of terms corresponding to Medicine and Chemistry. These terms came from articles whose topics are applied computer science such as bioinformatics. The manually collected English corpora happened to contain a non-negligible quantity of such texts. This shows that manually built corpora are also prone to errors and not necessarily better than automatically collected ones.

In view of the results, we have no other choice but to conclude that, for bilingual terminology extraction, the method of using search engines for automatic comparable corpora collection from the web, and the randomness and width of scope this involves, is at least comparable with an opportunistic manual collection produced with a reasonable amount of effort.

Another aspect to note is that, in all cases, the domain-specific terms obtained much better results than those that belonged to other domains. This confirms that all the corpora obtained can be considered as belonging to the domain to a high degree.

We can also observe that the results for Physics are better than those for Computer Science, although the latter corpora are much larger. This phenomenon was also observed in previous evaluations of monolingual term extraction (Gurrutxaga et al., 2009), and might be attributable to the more applied nature of Computer Science, which leads to the appearance of more non-domain and general terms and therefore polysemy.

The manual evaluation of the extracted candidate pairs that were not in the dictionary reference was also useful to show that the method is also valid for obtaining new terminology that is not in dictionaries: in the Physics web corpus, out of the 220 term candidates that were not in the reference, 18 new terms were found (23 with cognate detection) in the Top-1, and 39 in the Top-30; and in the Computer Science web corpus, out of 413, 31 new terms were found in the Top-1 (46 with the help of cognate detection) and 76 in the Top-30. In both cases, about 10% of the terms not in the dictionary are new valid terms in the Top-1 and about 18% in the Top-30.

7.4 Conclusions

In this chapter, we have proposed two methods for collecting domain-comparable corpora from the web. They are both based on search engines, and employ the technique described in the previous chapter for collecting specialized corpora (by using a sample mini-corpus, extracting keywords, sending combinations to search engines, downloading and filtering by domain using similarity measures with the sample mini-corpus), with a slight difference: one uses a sample mini-corpus for each language and launches the collection processes independently; the other uses just one sample mini-corpus and dictionaries to translate the keywords and the vectors for the domain filter.

They both obtain good results in an evaluation they were subjected to, but the results of the dictionary method might have benefited from the fact that the dictionary also has to be used in the evaluation.

Some corpora compiled with the two sample-mini corpora method were evaluated on an automatic terminology extraction task, and showed good precision results, especially for terms belonging to the domain of the corpus. Moreover, these were compared to those obtained in the same task by similar manually built corpora, and the automatically collected corpora ones were better. Thus, we must say that, for automatic bilingual terminology extraction, domain-comparable corpora collected automatically by our method behave at least as well as corpora of the same kind collected manually in an opportunistic way with a reasonable amount of effort.

However, bilingual terminology extraction out of comparable corpora based on context similarity –as well as many other tasks that comparable corpora can be used for– needs a minimum size of corpora for it to work properly. Although the results for the domains chosen in this paper are satisfactory, for a language like Basque, it might not be possible to obtain enough texts for some domains, depending on their level of specificity, the production, etc.

Since the collection of comparable corpora for Basque described here is based on the collection of two specialized corpora by the method described in the previous chapter and that we there explained that that method could be used to collect specialized corpora for other minority and less-resourced languages as well, we also believe that comparable corpora for these kind of languages can be collected using this method. With the same potential problem, obviously, that there might not be enough texts in some domains to be able to produce a comparable corpus large enough for some tasks.

For future work, it would be interesting to try to improve the dictionary-based approach; as we have already mentioned, the preliminary work needed to obtain a comparable corpus with this method is considerably reduced (only one sample mini-corpus needs to be collected); besides, there is still much room for improvement. One of the things to be tried is to see whether manual revision of the translated vectors to be used in the domain filtering yields a better performance. Another one is to try more complex translation selection techniques –instead of the first-translation approach–, and also synonymy expansion.

8 Results and Conclusions

8.1 Conclusions

This thesis began, in its introductory chapter, by underlining the poor situation of corpora in Basque with regard to other languages and particularly taking into account its situation as a minority language still in a standardization process. We argued that, if using the web as corpus or source for corpora had in recent years substantially increased the amount and sizes of corpora for other languages, it was logical –if not obligatory– for Basque to employ this cheap and fast way of producing Basque corpora. We formulated the hypothesis that the Web-as-Corpus approach could be valid to make a significant change in the situation of corpora for Basque, and this thesis has aimed to try to confirm this hypothesis and, at the same time, to improve the state of corpora in Basque.

For considering the hypothesis confirmed, we set ourselves some objectives and stated that the achievement of these objectives (or at least of most of them) would mean that our hypothesis was correct. Let us now check which these objectives were and whether we have attained them or not:

- **To build a tool that would enable the web to be queried as if it were a Basque corpus**: In pursuit of this objective, we successfully built a web service to allow the querying of the web in a corpus manner, CorpEus, which solved the problems that dogged other services of this kind for Basque. We devised, implemented, and optimized the techniques of morphological query expansion and language-filtering words which have been used not only in this tool, but also in all the other search engine-based corpus collecting tools and in a search engine for Basque.
- **To develop tools that would automatically collect from the web a general corpus of Basque that would outdo existing corpora by an order of magnitude and reach a size of at least 100 million words, and which would be of a quality comparable with the other ones**: For this objective, we built two different tools, one based on search engines and the other one on crawling, for collecting large general corpora in Basque. Using them we collected corpora that increased the sizes of existing Basque cor-

pora eightfold and reached 200 million words, and we expect to build even larger ones in the near future with the crawling method.

- **To build a tool for automatically collecting domain-specialized Basque corpora from the web, of a sufficient size and quality for terminological uses (evaluated with an automatic terminology extraction task), and to collect some domain-corpora using it**: To achieve this, we built a tool which, by means of search engines, was able to collect specialized corpora without requiring much work, just the collection beforehand of a few texts in the target domain. The tool proved able to obtain very good domain-precision and was successfully used in an automatic terminology extraction task. Various corpora were collected using the tool.
- **To develop a tool for automatically collecting Basque-English domain-comparable corpora good enough and large enough to be used for automatic bilingual terminology extraction, and use it to collect some comparable corpora**: Based on the methodology to compile specialized monolingual corpora, two different techniques to collect multilingual domain-comparable corpora were devised and developed, both obtaining good results. By using one of the techniques various comparable corpora were built, which were successfully used in a bilingual terminology extraction task.
- **To make these tools and corpora publicly available to the greatest possible extent**: The web service CorpEus was put online in 2007 and has been available for public use since. One of the large corpora compiled, of 125 million words (the largest we had collected until that moment), was also put online for public consultation. In addition, some of the general and specialized corpora we collected have been used in terminological and research work.

Apart from these objectives, the development of the aforementioned corpus collection tools has required the building of different corpus cleaning tools (boilerplate removal, duplicate and containment detection, etc.), that were adapted to our needs and that contribute to the optimal operation of the corpora collection, but which can be used in other tasks as well.

Section 8.2 describes in more detail the resources and tools that the tasks performed in this thesis have produced.

In view of the fact that all of the objectives have been achieved, we can conclude that our initial hypothesis has been confirmed, which is that the Web-as-Corpus approach could make a significant change in the situation of corpora in Basque –a change which in fact has been made with the work carried out in the thesis. We have not reached a state that is comparable with the major languages, but we have developed a methodology and tools for collecting what is available and have collected much of it, thus significantly augmenting the quantity and variety of corpora that existed previously.

Moreover, since the only linguistic tools and resources that all the web-as-corpus and corpus collection tools built in this thesis require are quite basic ones (basically, N-gram based language detection and morphological analysis and generation), we believe that the methodology we applied to build the tools and collect the corpora for Basque could be applied to other languages –even inflectionally rich and minority ones– that are in a similar situation to Basque concerning corpus availability in terms of types and sizes, and thus improve their situation just as in the case of Basque.

8.2 Resources and tools produced

Throughout this thesis, we have wanted to improve the situation of the Basque language regarding corpora, by making use of the web to query it live or to obtain texts for building corpora from it. The usual methods and tools for these kinds of tasks do not work well for Basque due to various problems (the poor treatment that search engines give to Basque, for example), so we have devised ways of overcoming these difficulties and obtaining a good performance for our language. And in order to test the viability of our hypotheses, we have performed various experiments for which we have had to develop and implement diverse software tools and collected various corpora.

The experiments have generally proved successful, so the aforementioned tools and corpora can be very useful for many tasks: language technology researchers, linguists, translators, lexicographers... It has been our intention from the beginning that the lives of these tools and resources should not end with the experiments. Whenever it has

been possible, we have tried to make them available for the Basque speaking or research community to benefit from them. And when not, they are still there for future uses.

In this section we will detail these methods, tools, and resources that have been produced in the thesis.

8.2.1 Morphological query expansion and language-filtering words

One of the main contributions of this thesis are the techniques of morphological query expansion and language-filtering words, with whose aid proper search for content in Basque can be obtained from search engines. These techniques have been used throughout all the chapters of the thesis. Sections 3.2 and 3.3 provide details as to their optimal implementation (most frequent cases for the inflections, best combinations of filter words, etc.), which can be used in other projects that use search engines for Basque content retrieval.

Besides, the steps followed for obtaining the implementation details and for measuring the improvements obtained with them could also be applied to other languages with similar features and problems.

8.2.2 CorpEus, a web service to query the web as a corpus of Basque

CorpEus is a service we implemented that makes use of the methodology explained in Chapter 3 and that allows the Internet to be consulted as if it were a Basque corpus (Leturia et al., 2007a). Its main features are the following:

- It makes use of the APIs of search engines to perform a web search. To obtain the best results for Basque, it uses the methodology we described in that chapter and which we have used throughout the whole thesis: it obtains results in Basque only by means of language-filtering words and it performs a lemma-based search using morphological query expansion.
- It suggests variants of words.
- More than one search term can be entered, and it offers the possibility of performing an exact phrase search by enclosing the search terms in double quotes (applying the morphological generation only to the last word in the phrase and thus performing a proper lemma-based search for whole noun phrases or terms).

- Once the search engine has returned its results, each of the returned pages is downloaded. For the downloading, different processes are launched concurrently, so that a slow or blocked page does not stop the complete process. And the contexts are served just as the pages start arriving, so that the user does not have to wait a long time until all the pages have been downloaded and processed.
- 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, etc.). 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, CorpEus can 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.
- Each occurrence of the search terms in the pages is only shown if LangId, applied to some context around it, says it is in a piece of text in Basque.
- 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, etc. And they are ordered on the fly as they come in, without having to wait until all the results have arrived.
- 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 analyser does not recognize are shown in red.
- CorpEus can show different charts with counts of word forms, possible lemma or POS, word before, word after, etc.

With the language-filtering words method, the language-precision usually obtained in a search for Basque (that is, the percentage of pages that are actually in Basque) is raised from 15% to around 95%, because 4 filter words are used by default. This leads to a loss in recall of around 50%, but if few results are returned or the user is not happy with them he or she can try again with fewer filter words. Thus, with 3 words a better precision-recall compromise is obtained (86-87% precision and 68-65% recall)

if the user so wishes. Besides, we have to take into account that with the morphological query expansion we get an increase in recall of 47% on average; applying this increase to the results that had decreased because of the language-filtering words, the fall in recall caused by them is much smaller and even non-existent in the case of 3 filter words.

The web service CorpEus has been online since September 2007. While other web-as-corpus services of its kind have been slowly dying –as we have seen in Section 2.1, out of the 6 tools or services that have been at work sometime (that we know of), only WebCorp seems to be still working–, CorpEus is still alive and can be queried at http://www.corpeus.org. Throughout these 6 plus years, it has received about 120,000 queries, which is not bad at all taking into account the small size of the Basque language and that corpus tools are not geared towards the general public but towards language specialists (translators, dictionary makers, linguists...).

Figure 8.1 shows a screen capture of CorpEus –with its most significant sections highlighted–, where a search for *paper* (*paper*) has been performed and, as we can observe, only results in Basque are shown –although *paper* is also an English word– and inflections of the word are returned too.

8.2.3 Elebila, a Basque search engine

Although all our research has been completely driven by interest in corpora and not in IR, we have seen that we solved the problems that search engines had for Basque and built a web-as-corpus tool (CorpEus) that made an efficient use of search engine APIs to offer a good search for content in Basque. In view of this, the company Eleka showed interest in the marketing of a search engine for Basque that made use of our techniques. With our assistance, this web service was developed and put online in 2007. It is called **Elebila** (Leturia et al., 2007b) and it can be consulted at http://www.elebila.eu.

These are its most important features:

- It is an API-based search service, so it is easy and cheap to implement.
- It uses morphological query expansion for obtaining a lemma-based search.
- Optionally, only the exact form entered can be looked for.
- It makes use of language-filtering words for obtaining results in Basque alone.

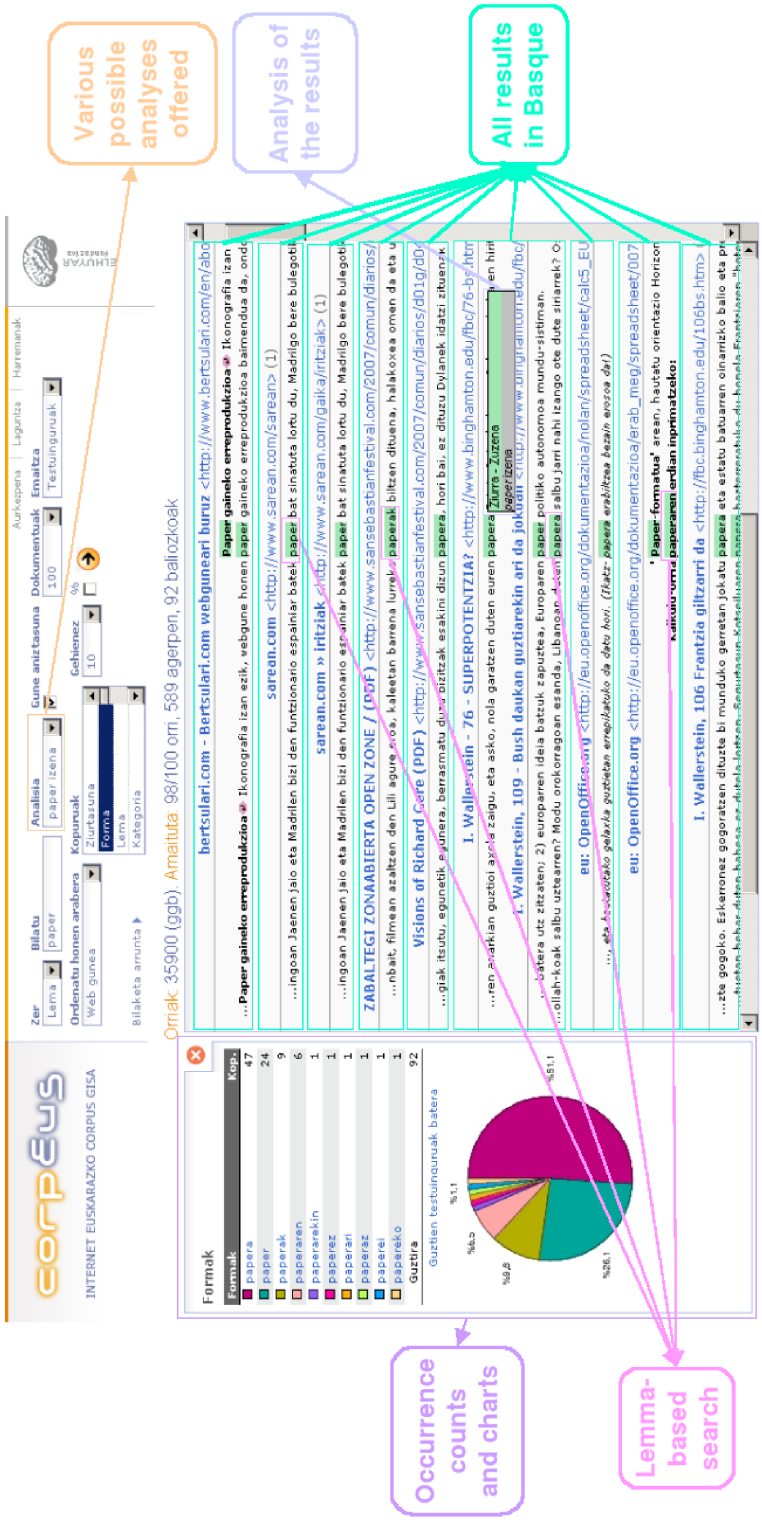


Figure 8.1: Screen capture of CorpEus

- The user can also choose to look for known variants (common errors, non-standard forms, archaic forms, etc.) of the word, which are proposed to the user based on EDBL (Aduriz et al., 1998).
- The user can enter more than one search term, and the lemma-based search is performed for all of them.
- Search engines usually offer the possibility of performing an exact phrase search by enclosing the search terms in double quotes. Elebila 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, since in Basque only the last component of the noun phrase is inflected.
- The user can enter a search term that is not a base form but a surface form of a lemma, that is, a conjugation or inflection. The search term is analysed 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 chosen for the morphological generation, but when the results are returned, the user is given the option of trying with the other analyses.
- The vast majority of the results returned by the search engine API are in Basque due to the language-filtering words, but not all. To filter out those few that are not, LangId is applied to the snippets of each result before showing them.

There is an important modification worth mentioning implemented in Elebila designed to improve navigational and transactional searches. The improvements obtained by the morphological query expansion and language-filtering words techniques are designed and optimized for finding *textual content* in Basque and, as such, these techniques perform best for informational queries, that is, queries aimed at obtaining information about something. But according to Broder (2002), informational queries account for only 39% to 48% of all the web queries. He also introduced the concepts of navigational queries (where the intention is to reach a particular site that the user has in mind, either because they visited it in the past or because they assume that such a site exists) and transactional queries (the purpose of which is to perform some web-mediated activity, such as shopping, downloading, accessing some database, etc.), and estimated that they represented more than half of the web queries. Although some oth-

er works estimate them to be less than 40% (Rose and Levinson, 2004), and although the language-oriented nature of the search service we are trying to build results in its most typical use being for informational queries (as a look at Elebila's logs confirms), a non-negligible number of transactional and navigational queries would most likely still be made to it.

But such a search service would not work so well for these kinds of queries, because of its use of language-filtering words. The inclusion of these words in the queries causes a loss in recall, as we have already shown. Many pages are left out because they do not contain one or more of the filter words, and these are mostly short pages that do not have much textual content. The pages that are the objective of navigational queries (home pages of companies or organizations) or transactional queries (entry pages of online dictionaries, social multimedia repositories such as Flickr or YouTube, online shops, etc.) are often not rich in textual content. Besides, a user might use the Basque search service to find the homepage of a Basque company whose web page is only in Spanish, French and/or English but not in Basque, so our service would not find it.

Nevertheless, major search engines work quite well for Basque navigational and transactional queries. It is mainly informational queries that they do not handle well and this is what we are trying to improve. But when looking for the address of the home page of a company or some other site, even if it is a page in Basque, the classical ranking measures (link analysis, click-through data, having the search terms in the title or URL, PageRank, etc.) usually work well, returning the desired page among the first results.

We take advantage of this fact in order to improve transactional and navigational queries. Apart from the morphologically expanded and language-filtered query, the API is also asked for the raw search terms the user entered, and the first five results are looked at to see if there are any in which the title or the URL matches the search term(s) almost exactly; if there are, they are inserted in the first positions of the other results.

The search logs of Elebila were later used to further improve and measure the performance of the morphological query expansion and language-filtering words techniques, as explained in Section 3.3 (Leturia et al., 2008a; Leturia et al., 2013). However, CorpEus and Elebila had been online before implementing these improve-

ments, with an intuitive decision about the language-filtering words and the morphological query expansion cases to be used. Elebila was evaluated on its launching to measure its performance with regard to a normal search engine (Leturia et al., 2007b). The results of each improvement of Elebila were compared 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. We are aware that hit counts returned by search engines do not constitute an exact or reliable measure (Kilgarriff, 2006), but they are used by many researchers as an acceptable approximation (Keller and Lapata, 2003). The words we chose for the evaluation were taken from the search logs spanning a whole year from the popular science portal in Basque Zientzia.net (Elhuyar Foundation, 2001). Table 8.1 shows a summary of the results of the evaluation.

Evaluation	Words	Measured variable	Result
Gain in recall due to morphological query expansion, without language-filtering words applied	Only Basque	Hit counts	89.43% increase
Gain in precision due to language-filtering words, without morphological query expansion 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, without morphological query expansion 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, with language-filtering words applied	Any kind	Hit counts	40.19% increase

Table 8.1: Summary of the results of Elebila's evaluation

A screen capture of Elebila that shows its main features is shown in Figure 8.2.

8.2.4 Large general corpora

In the pursuit of developing methods to collect large general corpora in Basque from the web (detailed in Chapter 5), we have implemented software solutions to collect such corpora by using both search engines or crawling (because we wanted to test both methods).

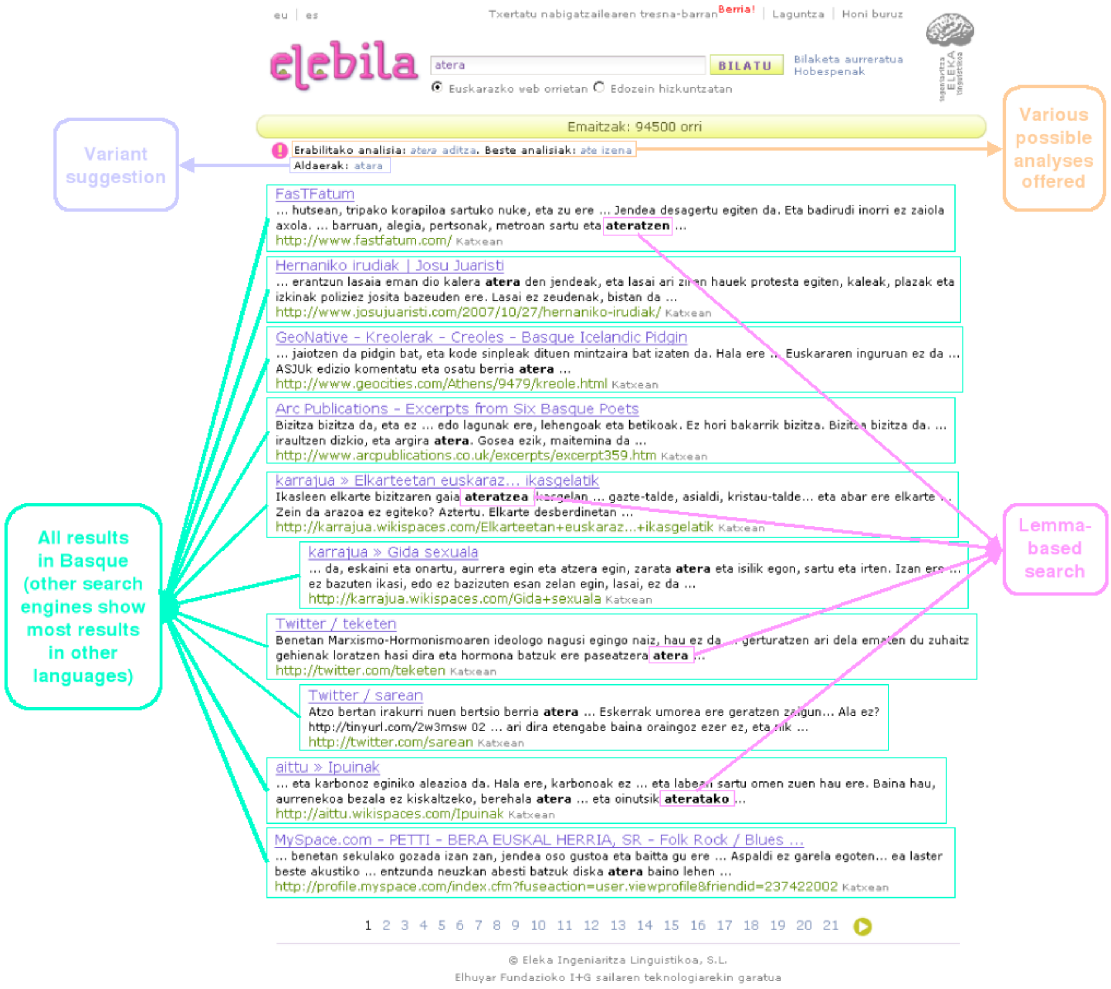


Figure 8.2: Screen capture of Elebila, showing its main features

Furthermore, various large general corpora in Basque have been collected in the comparative experiments we carried out (10 to be precise), ranging from 44 to 210 million words.

One of the corpora obtained with the search engine method has been used in a research about the quality of Basque web texts using spell-checker software (Alegria et al., 2010).

We have put online the largest of these general corpora we had at that moment for public consultation purposes; it is available at the website **Web-Corpusen Ataria** or ***Portal of Basque Web Corpora*** (Elhuyar Foundation, 2013), at the address http://web-corpusak.elhuyar.org/. Automatic collocation extraction by means of linguistic and

statistical techniques has been applied to it (Gurrutxaga and Alegria, 2011; Gurrutxaga and Alegria, 2012; Gurrutxaga and Alegria, 2013), so apart from the usual KWiCs and counts, collocations and phraseology can be queried, too (Noun-Noun, Noun-Adjective and Noun-Verb combinations). This portal also contains a Spanish-Basque parallel corpus compiled from the web by automatic methods (San Vicente and Manterola, 2012). Figures 8.3 and 8.4 show screen captures of this website, the first with a normal KWiC query over the monolingual web corpus and the second with a query for collocations.

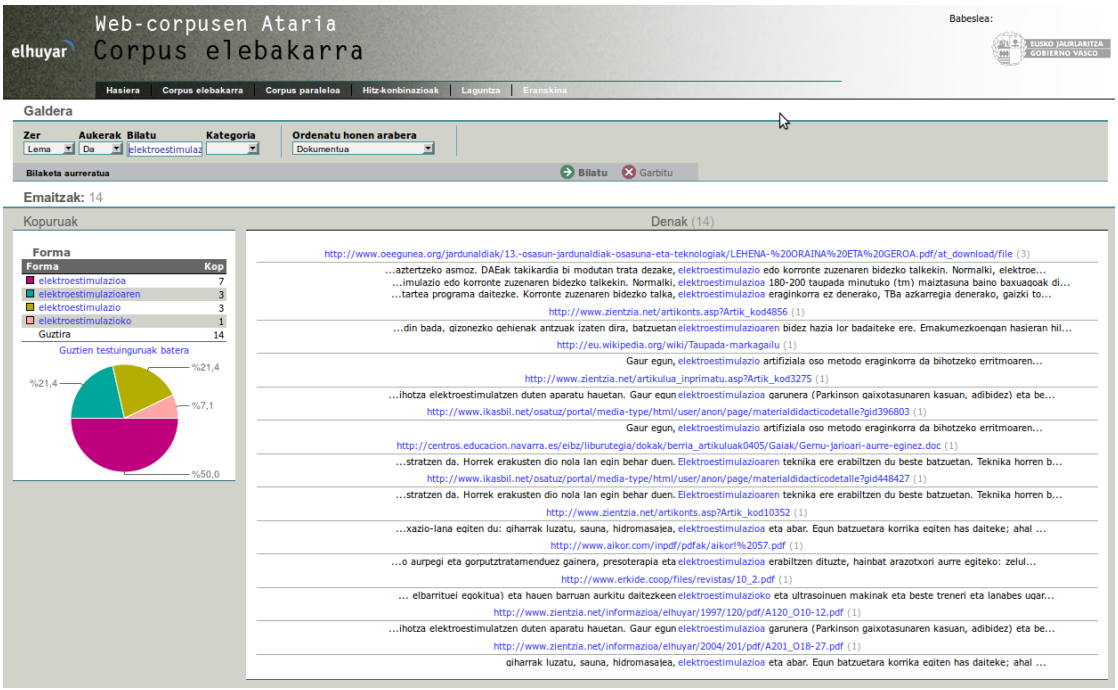


Figure 8.3: Screen capture of Web-Corpusen Ataria showing the KWiC of a query over the corpus

IZE-ADJ					
Konbinazioa	**f**	**f1**	**f2**	**t neurria**	**Adibideak**
hartzidura alkoholiko	16	31	57	4,0	**hartzidura alkoholikoa**
hartzidura laktiko	15	31	118	3,87	Horrek ondorengo **hartzidura laktikoa** erraztuko du eta produktu honek dituen bereizgarri organoleptikoak garatzen eta hezurra olibari ateratzen lagunduko du.
IZE-ADI					
Konbinazioa	**f**	**f1**	**f2**	**t neurria**	**Adibideak**
hartzidura egin	27	27	582944	4,68	3. 350 tan labean sartu suabea egon arte 4. zilar papera kendu eta erditik moztu patataren beste funtzioak patatari **hartzidura eginez** gero bodka egin dezakegu.

Figure 8.4: Screen capture of Web-Corpusen Ataria showing the results of a query for collocations

The launch of this portal meant a great milestone for Basque corpora: it was 5 times larger in size than the largest Basque corpus online so far, and Basque corpora broke the 100 million-word barrier for the first time. But just a month after, Egungo Testuen Corpusa or ETC (*Corpus of Modern Texts*) was put online (University of the Basque Country, 2013), a corpus of more than 200 million words. It is another great resource and very important in the development of the Basque language, but its objectives and features are different from our web corpus, so we can say that they complement each other as sources of evidence for Basque. Egungo Testuen Corpusa is a traditionally built corpus where texts have been obtained from a few sources with many texts each, whereas our web corpus has been collected from 4-5 thousand different websites. Another difference is that the texts from ETC all come from books or the media and so have been revised and/or written by professional writers or journalists. Our web corpus includes also the spontaneous production of texts from the general public on the web (fora, blogs, etc.), that is, non-revised texts written by non-experts or non-professionals of language, which is nowadays an undeniable linguistic phenomenon.

Our work on the collection of the large general corpus is not over. As we have seen, we have already collected a corpus of more than 200 million words, and we intend to continue with the crawling process and hope to have a much larger corpus soon to put online in the Web-Corpusen Ataria portal.

8.2.5 Specialized corpora

As explained in Chapter 6, we have built a tool to collect domain-specialized corpora in Basque that uses the search engine method of BootCaT, but corrects the problems BootCaT has with Basque and improves its domain-precision.

Using this tool, for its evaluation, corpora on various domains have been built. Different corpora on Computer Science and Geology have been collected using different collection parameters, and also some larger corpora on Biotechnology, Atomic and Particle Physics, and Computer Science.

Some of these corpora, and another one later compiled on Astronomy, have been used in the creation and updating of the Zientzia eta Teknologiaren Hiztegi Entziklopedikoa or Basic Dictionary of Science and Technology (Elhuyar Foundation, 2009).

Some of these specialized corpora have also been used in the aforementioned research about the quality of Basque web texts using spell-checker software (Alegria et al., 2010).

8.2.6 Comparable corpora

Another tool that has been created in the course of this thesis is the tool for building comparable corpora, which can use two different methods: one based on different sample corpora for each of the languages and the other based on just one sample and dictionaries (see Chapter 7).

For evaluating this tool we have collected Basque-English comparable corpora on Tourism and Computer Science (two of each, one with each method) and another one on Physics.

8.2.7 Corpus cleaning tools

In order to develop the corpus collection systems mentioned in the previous subsections, various corpus cleaning tools have been developed that have been described in Chapter 4. They achieve state-of-the-art performance and can be used in future Basque corpora collecting projects, as well as in any other work that needs the functionalities of those tools.

8.3 Publications produced

This section will list the publications in conference proceedings, specialized journals or books produced by the research carried out in this thesis, classified by chapters.

8.3.1 Querying the web directly as if it were a corpus of Basque (Chapter 3)

Igor Leturia, Antton Gurrutxaga, Iñaki Alegria, and Aitzol Ezeiza. 2007a. CorpEus, a "web as corpus" tool designed for the agglutinative nature of Basque. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 69–81, Louvain-la-Neuve, Belgium.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. 2007b. EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*, pages 47–54, Amsterdam, The Netherlands.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, and Eli Pociello. 2008a. 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 Evaluations (LREC)*, Marrakesh, Morocco.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. 2013. Morphological query expansion and language-filtering words for improving Basque web retrieval. *Language Resources and Evaluation*, 47(2):425–448.

8.3.2 Corpus cleaning (Chapter 4)

Xabier Saralegi and Igor Leturia. 2007. Kimatu, a tool for cleaning non-content text parts from html docs. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 163–167, Louvain-la-Neuve, Belgium.

8.3.3 Obtaining a large general Basque corpus using the web as source (Chapter 5)

Iñaki Alegria, Izaskun Etxeberria, and Igor Leturia. 2010. Errores ortográficos y de competencia en textos de la web en euskera. *Procesamiento del Lenguaje Natural*, 45:137–144.

Igor Leturia. 2012. Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India.

8.3.4 Using the web to build specialized corpora in Basque (Chapter 6)

Igor Leturia, Iñaki San Vicente, Xabier Saralegi, and Maddalen Lopez de Lacalle. 2008b. Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th International Workshop on the Web As Corpus (WAC)*, pages 40–46, Marrakech, Morocco.

Antton Gurrutxaga, Igor Leturia, Xabier Saralegi, and Iñaki San Vicente. 2009. Evaluation of an automatic process for specialized web corpora collection and

term extraction for Basque. In *Proceedings of eLexicography Conference 2009*, Louvain-la-Neuve, Belgium.

8.3.5 Collecting domain-comparable corpora in Basque and another language from the web (Chapter 7)

Igor Leturia, Iñaki San Vicente, and Xabier Saralegi. 2009. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of the 5th International Workshop on the Web As Corpus (WAC)*, pages 53–61, Donostia/San Sebastian, Spain.

Antton Gurrutxaga, Igor Leturia, Iñaki San Vicente, and Xabier Saralegi. 2013. Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors, *BUCC - Building and Using Comparable Corpora*. Springer, Dordrecht, The Netherlands.

8.4 Further work

In this thesis we have addressed the subject of the Web-as-Corpus in four of its variants: the live querying of the web as a corpus and its use as a source for building general, specialized, and domain-comparable corpora. However, there are two notorious gaps, two modalities that we have not dealt with: parallel corpora and genre-specific corpora (either monolingual or comparable).

The subject of parallel corpora with Basque as one of the languages and their automatic collection from the web is a very interesting subject indeed, but it has been addressed by San Vicente and Manterola (2012) with satisfactory results: they have collected a Spanish-Basque parallel corpus of 20 million words and an English-Basque one of 2 million words. The Spanish-Basque parallel web corpus has been put online for querying at the aforementioned Web-Corpusen Ataria or *Portal of Basque Web Corpora* (Elhuyar Foundation, 2013).

The subject of genre-specific corpora is one of the pending issues. Punctuation marks and POS trigrams are reported to be valid features for genre detection (Sharoff, 2007; Argamon et al., 1998), but many others are also found in literature (zu Eissen and Stein, 2004). However, it is not clear whether these punctuation marks and POS trigrams would be valid for Basque (for one, articles and prepositions do not exist in Basque, they are appended to the end of other words), and it is something that has scarcely been addressed. Besides, texts from the web introduce a whole new set of genres, turning even the definition of the set of genres into a non-trivial task (Mehler et al., 2011). And even if genre identification was solved, it is not clear how it would

be possible to collect pages only on a genre from the web. Search engines do not allow the features that can be used for genre identification (POS, link structure, etc.) to be requested, so crawling and post-filtering might be the only option. Nonetheless, it is a very interesting subject that we would like to address in the future.

Another very interesting piece of work would be, instead of going to the web to gather certain domains and genres, to develop genre and domain classification techniques and apply them to a large general corpus built by crawling and build a balanced corpus. This is another thing we leave as further work.

Regarding large general corpora, since the crawling corpus still has the potential to grow larger –as we saw in Subsection 5.2.2–, we intend to let the corpus building process go on to make the corpus as large as possible, and then to put it online in the Web-Corpusen Ataria or *Portal of Basque Web Corpora* (Elhuyar Foundation, 2013).

Another task we intend to carry out is the collection of a wide variety of specialized and comparable corpora, trying to cover as many areas and languages as possible, and then put them in a new section of the above-mentioned Web-Corpusen Ataria or *Portal of Basque Web Corpora*.

And finally, since we have stressed that the methods and tools used in this thesis could be applied to other languages, it would be very interesting to actually do so and collect corpora or build web-as-corpus tools for some minority languages which do not have them.

9 Bibliography

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, and Ruben Urizar. 1996. EUSLEM: A Lemmatiser / Tagger for Basque. In *Proceedings of 7th EURALEX International Conference*, pages 17–26, Göteborg, Sweden.

Itziar Aduriz, Izaskun Aldezabal, Olatz Ansa, Xabier Artola, and Arantza Diaz de Ilarraza. 1998. EDBL: a Multi-Purpose Lexical Support for the Treatment of Basque. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, volume II, pages 821–826, Granada, Spain.

Itziar Aduriz, Maxux Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Diaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World*, 56:1–15.

Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza, Aitor Sologaistoa, Aitor Soroa, Andoni Valverde, Nerea Areta, Antton Gurrutxaga, Igor Leturia, and Rafa Saiz. 2005. Zientzia eta teknologiaren corpusa. In *Euskera zientifiko-teknikoa: Normalizaziotik homologazinora*. Mendebalde Kultura Alkartea, Bilbao, Spain.

Iñaki Alegria, Xabier Artola, and Kepa Sarasola. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, II(4):193–203.

Iñaki Alegria, Izaskun Etxeberria, and Igor Leturia. 2010. Errores ortográficos y de competencia en textos de la web en euskera. *Procesamiento del Lenguaje Natural*, 45:137–144.

Iñaki Alegria, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. 2004a. An Xml-Based Term Extraction Tool for Basque. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, pages 1733–1736, Lisbon, Portugal.

Iñaki Alegria, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. 2004b. Linguistic and Statistical Approaches to Basque Term Extraction. In *Proceedings of GLAT 2004*, Barcelona, Spain.

Jacek Ambroziak and William A. Woods. 1998. Natural Language Technology in Precision Content Retrieval. In *Proceedings of the International Conference of Natural Language Processing and Industrial Applications*, Moncton, Canada.

Apache Software Foundation. 1999. Apache Project. http://www.apache.org.

Nerea Areta, Antton Gurrutxaga, and Igor Leturia. 2008. Begiratu bat corpus-baliabideei. *Bat Soziolinguistika aldizkaria*, 66:71–92.

Nerea Areta, Antton Gurrutxaga, Igor Leturia, Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza, and Aitor Sologaistoa. 2007. ZT corpus:

Annotation and Tools for Basque Corpora. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK. http://www.ztcorpusa.net.

Shlomo Argamon, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of the International workshop on Innovative Internet Information Systems (IIIS-98)*, Pisa, Italy.

Guy Aston and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh, U.K.

R. Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer, Dordrecht, The Netherlands.

Carme Bach, Roser Saurí, Jordi Vivaldi, and Maria Teresa Cabré. 1997. *El corpus de l'IULA: descripció*. Universitat Pompeu Fabra, Barcelona, Spain.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286(509).

Judit Bar-Ilan and Tatyana Gutman. 2005. How the search engines respond to some non-English queries? *Journal of Information Science*, 31(1):13–28.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, pages 1313–1316, Lisbon, Portugal.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43:209–226.

Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval: a competition for cleaning web pages. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC)*, Marrakech, Morocco.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 87–90, Trento, Italy.

Marco Baroni, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlỳ. 2006. WebBootCaT: Instant domain-specific corpora to support human translators. In *Proceedings of the 11th Annual conference of the European Association for Machine Translation (EAMT)*, pages 247–252, Oslo, Norway.

Marco Baroni and Motoko Ueyama. 2004. Retrieving Japanese specialized terms and corpora from the World Wide Web. In *Proceedings of KONVENS 2004*, pages 13–16, Vienna, Austria.

Marco Baroni and Motoko Ueyama. 2006. Building general- and special purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium*, Tokyo, Japan.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT/NAACL)*, pages 16–23, Edmonton, USA.

Basque Government. 2010. Translation memories of the Basque Government's Official Translation Service. http://www.ivap.euskadi.net/r61-vedorok/eu/contenidos/ds_recursos_linguisticos/memorias_traduccion/eu_izo/memorias_traduccion_izo.html.

Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. 2009. Purely URL-based topic classification. In *Proceedings of the 18th International World Wide Web Conference (WWW)*, pages 1109–1110, Madrid, Spain.

Eda Baykan, Monika Henzinger, and Ingmar Weber. 2008. Web page language identification based on URLs. In *Proceedings of the VLDB Endowment*, pages 176–187, Auckland, New Zealand.

Božo Bekavac, Petya Osenova, Kiril Simov, and Marko Tadić. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian & Croatian. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, pages 1187–1190, Lisbon, Portugal.

Nicholas J. Belkin. 2000. Helping People Find What They Don't Know. *Communications of the ACM*, 43(8):58–61.

Gunnar Bergh, Aimo Seppänen, and Joe Trotta. 1998. Language Corpora and the Internet: A joint linguistic resource. In Antoinette Renouf, editor, *Explorations in Corpus Linguistics*, pages 41–54. Rodopi, Amsterdam, The Netherlands.

Silvia Bernardini, Marco Baroni, and Stefan Evert. 2006. A WaCky introduction. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 9–40. Gedit, Bologna, Italy.

Tim Berners-Lee. 1989. Information management: A proposal. Technical Report Technical Report, CERN.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Longman, Harlow, UK.

Martin Braschler and Peter Schäuble. 1998. Multilingual information retrieval based on document alignment techniques. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, pages 183–197, Heraklion, Greece.

Sergei Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference (WWW)*, pages 107–117, Brisbane, Australia.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997*, pages 21–29, Salerno, Italy.

Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium*, pages 1–10, Montreal, Canada.

Andrei Z. Broder. 2002. A Taxonomy of Web Search. *ACM SIGIR Forum*, 36(2).

Andrei Z. Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference (WWW)*, pages 309–320, Amsterdam, The Netherlands.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. The ClueWeb09 Dataset. http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf.

William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

Soumen Chakrabarti, Martin Van der Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific web resource discovery. In *Proceedings of the 8th International World Wide Web Conference (WWW)*, pages 545–562, Toronto, Canada.

Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th annual ACM Symposium on Theory of Computing (STOC)*, page 388, Montreal, Canada.

Yun Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1208–1212, Taipei, Taiwan.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 293–300, Barcelona, Spain.

Béatrice Daille. 1995. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical Report UCREL Technical Papers 5, UCREL.

Béatrice Daille and Emmanuel Morin. 2005. French-English terminology extraction from comparable corpora. *Natural Language Processing - IJCNLP*:707–718.

Fred J. Damerau. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29:433–447.

Gilles-Maurice De Schryver. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies*, 11(2):266–282.

Alexis Deveria. 2013. Can I use New semantic elements? http://caniuse.com/html5semantic.

Ted Dunning. 1994. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Efthimis N. Efthimiadis, Nicos Malevris, Apostolos Kousaridas, Alexandra Lepeniotou, and Nikos Loutas. 2009. Non-english web search: an evaluation of indexing and searching the Greek web. *Information Rerieval*, 12(3):352–379.

EIZIE. 2002. Translation memories of the Association of Translators, Correctors and Interpreters of Basque Language. http://www.eizie.org/Tresnak/Memoriak.

Elhuyar Foundation. 2001. Zientzia.net (Science portal). http://www.zientzia.net.

Elhuyar Foundation. 2009. Zientzia eta Teknologiaren Hiztegi Entziklopedikoa (Basic Dictionary of Science and Technology). http://zthiztegia.elhuyar.org.

Elhuyar Foundation. 2013. Web-Corpusen Ataria (Portal of Basque Web Corpora). http://webcorpusak.elhuyar.org.

Eroski Foundation. 2010. Consumer Corpusa (Consumer Corpus). http://corpus.consumer.es.

Euskaltzaindia. 1984. Orotariko Euskal Hiztegiaren Testu-Corpusa (Text Corpus of the Universal Basque Dictionary). http://www.euskaltzaindia.net/oeh.

Euskaltzaindia. 2002. XX. mendeko Euskararen Corpusa (Corpus of Basque of the XXth Century). http://xxmendea.euskaltzaindia.net/Corpus/.

Euskaltzaindia. 2009. Lexikoaren Behatokia (Observatory of the Lexicon). http://lexikoarenbehatokia.euskaltzaindia.net.

Adriano Ferraresi. 2007. *Building a very large corpus of English obtained by Web crawling: ukWaC*. Ph.D. thesis, University of Bologna, Bologna, Italy.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large Web-derived corpus of English. In *Proceedings of the 4th International Workshop on the Web As Corpus (WAC)*, Marrakech, Morocco.

Dennis Fetterly, Mark Manasse, Marc Najork, and Janet L. Wiener. 2004. A large-scale study of the evolution of Web pages. *Software: Practice and Experience*, 34:213–237.

Aidan Finn, Nicholas Kushmerick, and Barry Smyth. 2001. Fact or fiction: Content classification for digital libraries. In *Proceedings of Personalisation and Recommender Systems in Digital Libraries Workshop*, Dublin, Ireland.

William H. Fletcher. 2004. Making the web more useful as a source for linguistic corpora. In Ulla Connor and Thomas A. Upton, editors, *Corpus Linguistics in North America 2002*. Rodopi, Amsterdam, The Netherlands.

William H. Fletcher. 2006. Concordancing the Web: Promise and Problems, Tools and Techniques. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, editors, *Corpus Linguistics and the Web*, pages 25–46. Rodopi, Amsterdam, The Netherlands. http://www.kwicfinder.com.

William H. Fletcher. 2007a. Implementing a BNC-compare-able web corpus. In Cédrick Fairon, Hubert Naets, Adam Kilgarriff, and Gilles-Maurice De Schryver, editors, *Building and exploring web corpora*, pages 43–56. Cahiers du Cental, Louvain-la-Neuve, Belgium.

William H. Fletcher. 2007b. Web Concordancer. http://webascorpus.org/searchwac.html.

William H. Fletcher. 2011. Corpus analysis of the World Wide Web. In Carol A. Chapelle, editor, *Encyclopedia of Applied Linguistics*. Wiley-Blackwell, Hoboken, USA.

Free Software Foundation. 1983. GNU Project. http://www.gnu.org.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, pages 414–420, Montreal, Canada.

Weizheng Gao and Tony Abou-Assaleh. 2007. GenieKnows web page cleaning system. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 135–140, Louvain-la-Neuve, Belgium.

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

Christian Girardi. 2007. Htmcleaner: Extracting the Relevant Text from the Web Pages. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 141–143, Louvain-la-Neuve, Belgium.

Google, Inc. 2004. Basque Branch of the Google Directory. http://www.google.com/Top/World/Euskara/.

Gregory Grefenstette. 1999. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB Translating and the Computer Conference*, London, UK.

Jiafeng Guo, Gu Xu, Hang Li, and Xueqi Cheng. 2008. A unified and discriminative model for query refinement. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 379–386, New York, USA.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the*

2011 International Workshop on Multiword Expressions (MWE), Portland, USA.

Antton Gurrutxaga and Iñaki Alegria. 2012. Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining Different Features of Idiomaticity for the Automatic Classification of Noun+Verb Expressions in Basque. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT/NAACL)*, pages 116–125, Atlanta, USA.

Antton Gurrutxaga, Igor Leturia, Iñaki San Vicente, and Xabier Saralegi. 2013. Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors, *BUCC - Building and Using Comparable Corpora*, pages 51–75. Springer, Dordrecht, The Netherlands.

Antton Gurrutxaga, Igor Leturia, Xabier Saralegi, and Iñaki San Vicente. 2009. Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque. In *Proceedings of eLexicography Conference 2009*, Louvain-la-Neuve, Belgium.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, Lisbon, Portugal.

Nevin Heintze. 1996. Scalable Document Fingerprinting. In *Proceedings of the USENIX Workshop on Electronic Commerce*, Oakland, USA.

Inma Hernáez, Eva Navas, Igor Odriozola, Kepa Sarasola, Arantza Diaz de Ilarraza, Igor Leturia, Araceli Diaz de Lezana, Beñat Oihartzabal, and Jasone Salaberria. 2012. *The Basque language in the digital age - Euskara aro digitalean*.META-NET White Paper Series. Springer, Dordrecht, The Netherlands. http://www.meta-net.eu/whitepapers/volumes/basque.

Degen Huang, Shanshan Wang, and Fuji Ren. 2013. Creating Chinese-English Comparable Corpora. *IEICE Transactions on Information and Systems*, 96(8):1853–1861.

Matthias Hüning. 2001. WebCONC. http://gandalf.uib.no/lingkurs/templates_c/%25%256E%5E6ED%5E6EDF58DD%25%25labor.html.php.

ICANN. 1998. Internet Corporation for Assigned Names and Numbers. http://www.icann.org.

Internet Archive. 1996. Internet Archive. http://www.archive.org.

IULA. 2008. Jaguar. http://jaguar.iula.upf.edu.

Andrew Kehoe and Matt Gee. 2007. New corpora from the web: making web text more “text-like.” In Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen, and Jukka Tyrkkö, editors, *Towards Multimedia in Corpus Studies*. University of Helsinki, Helsinki, Finland. http://wse1.webcorp.org.uk/.

Andrew Kehoe and Antoinette Renouf. 2002. WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In *Proceedings of the 11th International World Wide Web Conference (WWW)*, Honolulu, USA. http://www.webcorp.org.uk.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

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

Kimmo Kettunen. 2007. Managing keyword variation with frequency based generation of word forms in IR. In *Proceedings of NODALIDA Conference*, pages 318–323, Tartu, Estonia.

Rohit Khare, Doug Cutting, Kragen Sitaker, and Adam Rifkin. 2004. Nutch: A Flexible and Scalable Open-Source Web Search Engine. Technical Report Technical Report CN-TR-04-04, CommerceNet Labs.

Adam Kilgarriff. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of the 5th Workshop on Very Large Corpora (WVLC)*, pages 231–245, Beijing and Hong Kong, China.

Adam Kilgarriff. 2001. Web as corpus. In *Proceedings of Corpus Linguistics 2001*, pages 342–344, Lancaster, UK.

Adam Kilgarriff. 2003. Linguistic Search Engine. In *Proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC)*, pages 53–58, Lancaster, UK.

Adam Kilgarriff. 2006. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

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

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proceedings of the 7th International Conference on Language Resources and Evaluations (LREC)*, pages 904–910, Valletta, Malta.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–52, Granada, Spain.

Adam Kilgarriff, Pavel Rychlỳ, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of 11th EURALEX International Conference*, pages 105–116, Lorient, France.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Tailand.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 441–450, New York, USA.

Aleksander Kołcz, Abdur Chowdhury, and Joshua Alspector. 2004. Improved robustness of signature-based near-replica detection via lexicon randomization. In *Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, Seattle, USA.

Olivier Kraif. 2002. Translation alignment and lexical correspondence. In Bengt Altenberg and Sylviane Granger, editors, *Lexis in Contrast*. John Benjamins, Amsterdam, The Netherlands.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 191–202, Pittsburgh, Pennsylvania.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, USA.

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. Trawling the Web for emerging cyber communities. In *Proceedings of the 8th International World Wide Web Conference (WWW)*, Toronto, Canada.

Stefan Langer. 2001. Natural languages and the world wide web. *Bulletin de linguistique appliquée et générale*, 26:89–100.

Fotis Lazarinis. 2007. Web retrieval systems and the Greek language: do they have an understanding? *Journal of Information Science*, 33(5):622–636.

Fotis Lazarinis, Jesus Vilares, and John I. Tait. 2007. Improving non-English web searching (iNEWS07). *ACM SIGIR Forum*, 41(2):72–76.

Geoffrey Leech. 2002. The Importance of Reference Corpora. In *Hizkuntza-corpusak. Oraina eta geroa*. UZEI, Donostia/San Sebastian, Spain.

Michael David Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci)*, pages 1254–1259, Stresa, Italy.

Igor Leturia. 2012. Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India.

Igor Leturia, Antton Gurrutxaga, Iñaki Alegria, and Aitzol Ezeiza. 2007a. CorpEus, a "web as corpus" tool designed for the agglutinative nature of Basque. In

Proceedings of the 3rd International Workshop on the Web As Corpus (WAC), pages 69–81, Louvain-la-Neuve, Belgium. http://www.corpeus.org.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. 2007b. EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*, pages 47–54, Amsterdam, The Netherlands.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. 2013. Morphological query expansion and language-filtering words for improving Basque web retrieval. *Language Resources and Evaluation*, 47(2):425–448.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, and Eli Pociello. 2008a. 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 Evaluations (LREC)*, Marrakech, Morocco.

Igor Leturia, Iñaki San Vicente, and Xabier Saralegi. 2009. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of the 5th International Workshop on the Web As Corpus (WAC)*, pages 53–61, Donostia/San Sebastian, Spain.

Igor Leturia, Iñaki San Vicente, Xabier Saralegi, and Maddalen Lopez de Lacalle. 2008b. Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th International Workshop on the Web As Corpus (WAC)*, pages 40–46, Marrakech, Morocco.

Vinci Liu and James R. Curran. 2006. Web text corpus for natural language processing. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 233–240, Trento, Italy.

Anke Lüdeling, Stefan Evert, and Marco Baroni. 2007. Using the web for linguistic purposes. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, editors, *Corpus Linguistics and the Web*, pages 7–24. Rodopi, Amsterdam, The Netherlands.

Udi Manber. 1994. Finding similar files in a large file system. In *Proceedings of the USENIX winter 1994 technical conference*, pages 1–10, San Francisco, USA.

Elena Manca. 2008. From phraseology to culture: Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics*, 13(3):368–385.

Christopher D. Manning, P. Raghava, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, USA.

Michael Marek, Pavel Pecina, and Miroslav Spousta. 2007. Web Page Cleaning with Conditional Random Fields. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 155–162, Louvain-la-Neuve, Belgium.

Tony McEnery and Andrew Wilson. 2001. *Corpus linguistics*. Edinburgh University Press, Edinburgh, UK.

Alexander Mehler, Serge Sharoff, and Marina Santini, editors. 2011. *Genres on the Web*. Springer, Dordrecht, The Netherlands.

Microsoft Corporation. 2006. Bing Search API. http://datamarket.azure.com/dataset/bing/search.

Evangelios Milios, Yongzheng Zhang, Ben He, and L. Dong. 2003. Automatic term extraction and document similarity in special text corpora. In *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 275–284, Halifax, Canada.

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An introduction to Heritrix, an open source archival quality web crawler. In *Proceedings of the 4th International Web Archiving Workshop*, Bath, UK.

Gordon E. Moore. 1965. Cramming more components onto integrated circuits. *Electronics*, 38:114–117.

Fabienne Moreau, Vincent Claveau, and Pascale Sébillot. 2007. Automatic Morphological Query Expansion Using Analogy-Based Machine Learning. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR)*, pages 222–233, Rome, Italy.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 664–671, Prague, Czech Republic.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Rogelio Nazar, Jorge Vivaldi, and Maria Teresa Cabré. 2008. A Suite to Compile and Analyze an LSP Corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC)*, pages 1164–1169, Marrakech, Morocco.

ODP. 1998. Open Directory Project. http://www.dmoz.org/.

Stanislaw Osinski, Jerzy Stefanowski, and Dawid Weiss. 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Proceedings of the International Conference on Intelligent Information Systems*, pages 359–368, Zakopane, Poland.

Muntsa Padró and Lluís Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguage Natural*, 33:155–162.

Jeff Pasternack and Dan Roth. 2009. Extracting article text from the web with maximum subsequence segmentation. In *Proceedings of the 18th International World Wide Web Conference (WWW)*, pages 971–980, Madrid, Spain.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University Faculty of Informatics, Brno, Czech Republic.

Jan Pomikálek, Miloš Jakubícek, and Pavel Rychlỳ. 2012. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 502–506, Istanbul, Turkey.

Jan Pomikálek, Pavel Rychlỳ, and Adam Kilgarriff. 2009. Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41:3–13.

Provincial Council of Biscay. 2011. Translation memories of the Provincial Council of Biscay. http://www.bizkaia.net/home2/temas/detalletema.asp?tem_codigo=6130.

Provincial Council of Gipuzkoa. 2011. Translation memories of the Provincial Council of Gipuzkoa. http://www.gipuzkoa.net/imemoriak/.

William Pugh and Monika H. Henzinger. 2008. Detecting duplicate and near-duplicate files.

Michael O. Rabin. 1981. Fingerprinting by random polynomials. Technical Report Report TR-15-81, Center for Research in Computing Technology, Harvard University.

Dave Raggett. 1998. Clean up your Web pages with HTML tidy. In *Proceedings of the 7th International World Wide Web Conference (WWW)*, pages 730–732, Brisbane, Australia.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 519–526, College Park, USA.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics (ACL)*, Ann Arbor, USA.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China.

Antoinette Renouf, Andrew Kehoe, and Jay Banerjee. 2007. WebCorp: an Integrated System for Web Text Search. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, editors, *Corpus Linguistics and the Web*, pages 47–67. Rodopi, Amsterdam, The Netherlands. http://www.webcorp.org.uk.

Philip Resnik, Aaron Elkiss, Ellen Lau, and Heather Taylor. 2005. The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine. In *Proceedings of the 31st Meeting of the Berkeley Linguistics Society*, pages 265–276, Berkeley, USA.

Xavier Roche. 2004. HTTrack. http://www.httrack.com.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 13–19, New York, USA.

Iñaki San Vicente and Iker Manterola. 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.

Xabier Saralegi and Iñaki Alegria. 2007. Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*(39):71–78.

Xabier Saralegi and Igor Leturia. 2007. Kimatu, a tool for cleaning non-content text parts from html docs. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 163–167, Louvain-la-Neuve, Belgium.

Xabier Saralegi, Iñaki San Vicente, and Antton Gurrutxaga. 2008a. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st Building and Using Comparable Corpora workshop (BUCC)*, Marrakech, Morocco.

Xabier Saralegi, Iñaki San Vicente, and Maddalen Lopez de Lacalle. 2008b. Mining Term Translations from Domain Restricted Comparable Corpora. *Procesamiento del Lenguaje Natural*, 41:273–280.

Kevin P. Scannell. 2007. The Crúbadán Project: corpus building for under-resourced languages. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 5–15, Louvain-la-Neuve, Belgium.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 486–493, Istanbul, Turkey.

Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*. Morgan & Claypool, San Rafael, USA.

Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. 2003. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 76–85, San Diego, USA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 44–49.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

M. Ángeles Serrano, Ana Maguitman, Marián Boguñá, Santo Fortunato, and Alessandro Vespignani. 2007. Decoding the structure of the WWW: a comparative analysis of web crawls. *ACM Transactions on the Web*, 1(2):10.

Serge Sharoff. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 63–98. Gedit Edizioni, Bologna, Italy.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, pages 83–94, Louvain-la-Neuve, Belgium.

Serge Sharoff, Bogdan Babych, and Anthony Hartley. 2007. "Irrefragable answers" using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

Páraic Sheridan and Jean Paul Ballerini. 1996. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 58–65, Zurich, Switzerland.

Narayanan Shivakumar and Hector Garcia-Molina. 1999. Finding near-replicas of documents on the web. In Pablo Atzeni, Alberto Mendelzon, and Giansalvatore Mecca, editors, *The World Wide Web and Databases*, Lecture Notes in Computer Science, pages 204–212. Springer, Dordrecht, The Netherlands.

John McHardy Sinclair. 1996. Preliminary recommendations on Corpus Typology. Technical Report Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

John McHardy Sinclair. 2005. Corpus and text – Basic principles. In Martin Wynne, editor, *Developing linguistic corpora: A guide to good practice*. Oxbow Books, Oxford, U.K.

Frank Smadja. 1993. Retrieving Collocations from Text: XTRACT. *Computational Linguistics*, 19(1):143–177.

Sogou Inc. 2004. Sogou. http://www.sogou.com.

Karen Spärck Jones and John I. Tait. 1984. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.

Ranka M. Stanković. 2008. Improvement of Queries using a Rule Based Procedure for Inflection of Compounds and Phrases. *Research journal on Computer science and computer engineering with applications*, 37:14–20.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1–6, Genoa, Italy.

Michael Stubbs. 1996. *Text and corpus analysis. Computer-assisted studies of language and culture*. Blackwell, Oxford, UK.

Susa. 2005. Klasikoen Gordailua (Repository of Classics). http://klasikoak.armiarma.com/corpus.htm.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems*, 25(1):4.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. 2008. Focused web crawling in acquisition of comparable corpora. *Information Retrieval*, 11:427–445.

Mike Thelwall. 2005. Creating and using web corpora. *International Journal of Corpus Linguistics*, 10(4):517–541.

Mike Thelwall, Rong Tang, and Liz Price. 2003. Linguistic patterns of academic Web use in Western Europe. *Scientometrics*, 56(3):417–432.

Jörg Tiedemann. 2003. *Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Uppsala University, Uppsala, Sweden.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*, pages 491–502, Freiburg, Germany.

Twitter, Inc. 2006. Twitter. http://www.twitter.com.

Motoko Ueyama. 2006. Evaluation of Web-based Japanese reference corpora: effects of seed selection and time interval. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 99–126. Gedit Edizioni, Bologna, Italy.

Motoko Ueyama and Marco Baroni. 2005. Automated construction and evaluation of a Japanese web-based reference corpus. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK.

University of Massachusetts Amherst and Carnegie Mellon University. 2000. Lemur Toolkit. http://www.lemurproject.org.

University of the Basque Country. 2006. Ereduzko Prosa Gaur (Model Prose Today). http://www.ehu.es/euskara-orria/euskara/ereduzkoa/.

University of the Basque Country. 2010. Pentsamenduaren Klasikoak (Classic Essays). http://www.ehu.es/ehg/klasikoak/.

University of the Basque Country. 2013. Egungo Testuen Corpusa (Corpus of Modern Texts). http://www.ehu.es/etc/.

Ruben Urizar, Nerea Ezeiza, and Iñaki Alegria. 2000. Morphosyntactic structure of terms in Basque for automatic terminology extraction. In *Proceedings of 9th EURALEX International Conference*, pages 373–382, Stuttgart, Germany.

Ahmet Uyar. 2009. Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35(4):469–480.

Gertjan van Noord. 1997a. NetKwic. http://www.hit.uib.no/corpora/2000-2/0135.html.

Gertjan van Noord. 1997b. TextCat. http://www.let.rug.nl/~vannoord/TextCat/.

Minna Vihla. 1998. Medicor: A corpus of contemporary American medical texts. *ICAME Journal*, 22:73–80.

Martin Volk. 2002. Using the web as corpus for linguistic research. In Renate Pajusalu and Tiit Hennoste, editors, *Tähendusepüüdja. Hatcher of the Meaning. A Festschrift for Professor Haldur Õim*. University of Tartu, Tartu, Estonia.

W3C. 2013. HTML 5 Specification Candidate Recommendation. Technical Report HTML 5 W3C Candidate Recommendation 6 August 2013, W3C. http://www.w3.org/TR/html5/.

Wikimedia Foundation. 2001. Wikipedia. http://www.wikipedia.org.

Wikimedia Foundation. 2013a. Wikipedia article on Basque language. http://en.wikipedia.org/wiki/Basque_language.

Wikimedia Foundation. 2013b. Wikipedia article on Inflection. http://en.wikipedia.org/wiki/Inflection.

William A. Woods. 2000. Aggressive morphology for robust lexical coverage. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 218–223, Seattle, USA.

William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green. 2000. Linguistic knowledge can improve information retrieval. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 262–267, Seattle, USA.

Jinxi Xu and W. Bruce Croft. 1998. Corpus-Based Stemming Using Cooccurrence of Word Variants. *ACM Transactions on Information Systems*, 16(1):61–81.

Yahoo! Inc. 1994. Yahoo. http://www.yahoo.com.

Dan-Hee Yang, Pascual Cantos Gómez, and Mansuk Song. 2000. An Algorithm for Predicting the Relationship between Lemmas and Corpus Size. *ETRI Journal*, 22(2):20–31.

Federico Zanettin, Silvia Bernardini, and Dominic Stewart. 2003. *Corpora in translator education*. St. Jerome Publishing, Manchester, UK.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR)*, London, UK.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, USA.

Sven Meyer zu Eissen and Benno Stein. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, pages 256–269, Ulm, Germany.