CUCWeb: a Catalan corpus built from the Web

G. Boleda[1] **S. Bott**[1] **R. Meza**[2] **C. Castillo**[2] **T. Badia**[1] **V. López**[2]
[1]Grup de Lingüística Computacional
[2]Cátedra Telefónica de Producción Multimedia
Fundació Barcelona Media
Universitat Pompeu Fabra
Barcelona, Spain
{gemma.boleda,stefan.bott,rodrigo.meza}@upf.edu
{carlos.castillo,toni.badia,vicente.lopez}@upf.edu

Abstract

This paper presents CUCWeb, a 166 million word corpus for Catalan built by crawling the Web. The corpus has been annotated with NLP tools and made available to language users through a flexible web interface. The developed architecture is quite general, so that it can be used to create corpora for other languages.

1 Introduction

CUCWeb is the outcome of the common interest of two groups, a Computational Linguistics group and a Computer Science group interested on Web studies. It fits into a larger project, The Spanish Web Project, aimed at empirically studying the properties of the Spanish Web (Baeza-Yates et al., 2005). The project set up an architecture to retrieve a portion of the Web roughly corresponding to the Web in Spain, in order to study its formal properties (analysing its link distribution as a graph) and its characteristics in terms of pages, sites, and domains (size, kind of software used, language, among other aspects).

One of the by-products of the project is a 166 million word corpus for Catalan.[1] The biggest annotated Catalan corpus before CUCWeb is the CTILC corpus (Rafel, 1994), consisting of about 50 million words.

In recent years, the Web has been increasingly used as a source of linguistic data (Kilgarriff and Grefenstette, 2003). The most straightforward approach to using the Web as corpus is to gather data online (Grefenstette, 1998), or estimate counts (Keller and Lapata, 2003) using available search engines. This approach has a number of drawbacks, e.g. the data one looks for has to be known beforehand, and the queries have to consist of lexical material. In other words, it is not possible to perform structural searches or proper language modeling.

Current technology makes it feasible and relatively cheap to crawl and store terabytes of data. In addition, crawling the data and processing it off-line provides more potential for its exploitation, as well as more control over the data selection and pruning processes. However, this approach is more challenging from a technological viewpoint. [2] For a comprehensive discussion of the pros and cons of the different approaches to using Web data for linguistic purposes, see e.g. Thelwall (2005) and Lüdeling et al. (To appear). We chose the second approach because of the advantages discussed in this section, and because it allowed us to make the data available for a large number of non-specialised users, through a web interface to the corpus. We built a general-purpose corpus by crawling the Spanish Web, processing and filtering them with language-intensive tools, filtering duplicates and ranking them according to popularity.

The paper has the following structure: Section 2 details the process that lead to the constitution of the corpus, Section 3 explores some of the exploitation possibilities that are foreseen for CUCWeb, and Section 4 discusses the current architecture. Finally, Section 5 contains some conclusions and future work.

[1] Catalan is a relatively minor language. There are currently about 10.8 million Catalan speakers, similar to Serbian (12), Greek (10.2), or Swedish (9.3). See http://www.upc.es/slt/alatac/cat/dades/catala-04.html

[2] The WaCky project (http://wacky.sslmit.unibo.it/) aims at overcoming this challenge, by developing “a set of tools (and interfaces to existing tools) that will allow a linguist to crawl a section of the web, process the data, index them and search them”.

2 Corpus Constitution

2.1 Data collection

Our goal was to crawl the portion of the Web related to Spain. Initially, we crawled the set of pages with the suffix `.es`. However, this domain is not very popular, because it is more expensive than other domains (e.g. the cost of a `.com` domain is about 15% of that of an `.es` domain), and because its use is restricted to company names or registered trade marks.[3] In a second phase a different heuristic was used, and we considered that a Web site was in Spain if either its IP address was assigned to a network located in Spanish land, or if the Web site's suffix was `.es`. We found that only 16% of the domains with pages in Spain were under `.es`.

The final collection of the data was carried out in September and October 2004, using a commercial piece of software by Akwan (da Silva et al., 1999). [4] The actual collection was started by the crawler using as a seed the list of URLs in a Spanish search engine –which was a commercial search engine back in 2000– under the name of Buscopio. That list covered the major part of the existing Web in Spain at that time. [5]. New URLs were extracted from the downloaded pages, and the process continued recursively while the pages *were in Spain* –see above. The crawler downloaded all pages, except those that had an identical URL (`http://www.web.es/main/` and `http://www.web.es/main/index.html` were considered different URLs). We retrieved over 16 million Web pages (corresponding to over 300,000 web sites and 118,000 domains), and processed them to extract links and text. The uncompressed text of the pages amounts to 46 GB, and the metadata generated during the crawl to 3 GB.

In an initial collection process, a number of difficulties in the characterisation of the Web of Spain were identified, which lead to redundancy in the contents of the collection:

Parameters to a program inside URL addresses. This makes it impossible to adequately separate static and dynamic pages, and may lead to repeatedly crawl pages with the same content.

Mirrors (geographically distributed copies of the same contents to ensure network efficiency). Normally, these replicas are entire collections with a large volume, so that there are many sites with the same contents, and these are usually large sites. The replicated information is estimated between 20% and 40% of the total Web contents ((Baeza-Yates et al., 2005)).

Spam on the Web (actions oriented to deceive search engines and to give to some pages a higher ranking than they deserve in search results). Recognizing spam pages is an active research area, and it is estimated that over 8% of what is indexed by search engines is spam (Fetterly et al., 2004). One of the strategies that induces redundancy is to automatically generate pages to improve the score they obtain in link-based rankings algorithms.

DNS wildcarding (domain name spamming). Some link analysis ranking functions assign less importance to links between pages in the same Web site. Unfortunately, this has motivated spammers to use several different Web sites for the same contents, usually through configuring DNS servers to assign hundreds or thousands of site names to the same IP address. Spain's Web seems to be quite populated with domain name spammers: 24 out of the 30 domains with the highest number of Web sites are configured with DNS wildcarding (Baeza-Yates et al., 2005).

Most of the spam pages were under the `.com` top-level domain. We manually checked the domains with the largest number of sites and pages to ban a list of them, mostly sites containing pornography or collections of links without information content. This is not a perfect solution against spam, but generates significant savings in terms of bandwidth and storage, and allows us to spend more resources in content-rich Web sites. We also restricted the crawler to download a maximum of 400 pages per site, except for the Web sites within `.es`, that had no pre-established limit.

[3] In the case of Catalan, additionally, there is a political and cultural opposition to the `.es` domain.

[4] We used a PC with two Intel-4 processors running at 3 GHz and with 1.6 GB of RAM under Red-Hat Linux. For the information storage we used a RAID of disks with 1.8 TB of total capacity, although the space used by the collection is about 50 GB.

[5] http://www.buscopio.net

	Documents	(%)	Words	(%)
Language classifier	491,850	100	375,469,518	100
Dictionary filter	277,577	56.5	222,363,299	59
Duplicate detector	204,238	41.5	166,040,067	44

Table 1: Size of the Catalan corpus

2.2 Data processing

The processing of the data to obtain the Catalan corpus consisted of the following steps: language classification, linguistic filtering and processing, duplicate filtering and corpus indexing. This section details each of these aspects.

We built a language classifier with the Naive Bayes classifier of the Bow system (Mccallum, 1996). The system was trained with corpora corresponding to the 4 official languages in Spain (Spanish, Catalan, Galician and Basque), as well as to the other 6 most frequent languages in the Web (Anonymous, 2000): English, German, French, Italian, Portuguese, and Dutch.

38% of the collection could not be reliably classified, mostly because of the presence of pages without enough text, for instance, pages containing only images or only lists of proper nouns. Within the classified pages, Catalan was the third most used language (8% of the collection). As expected, most of the collection was in Spanish (52%), but English had a large part (31%). The contents in Galician and Basque only comprise about 2% of the pages.

We wanted to use the Catalan portion as a corpus for NLP and linguistic studies. We were not interested in full coverage of Web data, but in quality. Therefore, we filtered it using a computational dictionary and some heuristics in order to exclude documents with little linguistic relevance (e.g. address lists) or with a lot of noise (programming code, multilingual documents). In addition, we performed a simple duplicate filter: web pages with a very similar content (determined by a hash of the processed text) were considered duplicates.

The sizes of the corpus (in documents and words[6]) after each of the processes are depicted in Table 1. Note that the two filtering processes discard almost 60% of the original documents. The final corpus consists of 166 million words from 204 thousand documents.

Its distribution in terms of top-level domains is shown in Table 2, and the 10 biggest sites in Table 3. Note that the `.es` domain covers almost half of the pages and `com` a quarter, but `.org` and `.net` also have a quite large share of the pages. As for the biggest sites, they give an idea of the content of CUCWeb: they mainly correspond to university and institutional sites. A similar distribution can be observed for the 50 biggest sites, which will determine the kind of language found in CUCWeb.

	Documents	(%)
es	89,541	44.6
com	49,146	24.5
org	35,528	17.7
net	18,819	9.4
info	5,005	2.5
edu	688	0.3
others	2,042	1.4

Table 2: Domain distribution in CUCWeb

The corpus was further processed with CatCG (Àlex Alsina et al., 2002), a POS-tagger and shallow parser for Catalan built with the Connexor Constraint Grammar formalism and tools.[7] CatCG provides part of speech, morphological features (gender, number, tense, etc.) and syntactic information. The syntactic information is a functional tag (e.g. subject, object, main verb) annotated at word level.

Since we wanted the corpus not only to be an in-house resource for NLP purposes, but also to be accessible to a large number of users. To that end, we indexed it using the IMS Corpus Workbench tools[8] and we built a web interface to it (see Section 3.1). The CWB includes facilities for indexing and searching corpora, as well as a special module for web interfaces. However, the size of the corpus is above the advisable limit for these tools. [9] Therefore, we divided it into 4 subcorpora

[6] Word counts do not include punctuation marks.

[7] http://www.connexor.com/

[8] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

[9] According to Stefan Evert –personal communication–, if a corpus has to be split into several parts, a good rule of thumb is to split it in 100M word parts. In his words "depending on various factors such as language, complexity of annotations

Site	Description	Documents
upc.es	University	1574
gencat.es	Institution	1372
publicacions.bcn.es	Institution	1282
uab.es	University	1190
revista.consumer.es	Company	1132
upf.es	University	1076
nil.fut.es	Distribution lists	1045
conc.es	Insitution	1033
uib.es	University	977
ajtarragona.es	Institution	956

Table 3: 10 biggest sites in CUCWeb

and indexed each of them separately. The search engine for the corpus is the CQP (Corpus Query Processor, one of the modules of the CWB).

Since CQP provides sequential access to documents we ordered the corpus documents by PageRank so that they are retrieved according to their popularity on the Internet.

3 Corpus Exploitation

CUCWeb is being exploited in two ways: on the one hand, data can be accessed through a web interface (Section 3.1). On the other hand, the annotated data can be exploited by theoretical or computational linguists, lexicographers, translators, etc. (Section 3.2).

3.1 Corpus interface

Despite the wide use of corpora in NLP, few interfaces have been built, and still fewer are flexible enough to be of interest to linguistic researchers. As for Web data, some initiatives exist (WebCorp [10], the Linguist's Search Engine [11], KWiCFinder [12]), but they are meta-interfaces to search engines. For Catalan, there is a web interface for the CTILC corpus[13], but it only allows for one word searches, of which a maximum of 50 hits are viewed. It is not possible either to download search results.

From the beginning of the project our aim was to create a corpus which could be useful for both the NLP community and for a more general audience with an interest in the Catalan language. This includes linguists, lexicographers and language teachers.

We expected the latter kind of user not to be familiar with corpus searching strategies and corpus interfaces, at least not to a large extent. Therefore, we aimed at creating a user-friendly web interface which should be useful for both non-trained and experienced users.[14] Further on, we wanted the interface to support not only example searches but also statistical information, such as co-occurrence frequency, of use in lexicographical work and potentially also in language teaching or learning.

There are two web interfaces to the corpus: an example search interface and a statistics interface. Furthermore, since the flexibility and expressiveness of the searches potentially conflicts with user-friendliness, we decided to divide the example search interface into two modalities: a simple search mode and an expert search mode.

The simple mode allows for searches of words, lemmata or word strings. The search can be restricted to specific parts of speech or syntactic functions. For instance, a user can search for an ambiguous word like Catalan “la” (masculine noun, or feminine determiner or personal pronoun) and restrict the search to pronouns. Or look for word “traduccions” (‘translations’) functioning as subject. The advantage of the simple mode is that an untrained person can use the corpus almost without the need to read instructions. If new users find it useful to use CUCWeb, we expect that the motivation to learn how to create advanced corpus queries will arise.

The expert mode is somewhat more complex but very flexible. A string of up to 5 word units can be searched, where each unit may be a word

and how much RAM you have, a larger or smaller size may give better overall performance.”.

[10] http://www.webcorp.org.uk/

[11] http://lse.umiacs.umd.edu

[12] http://miniappolis.com/KWiCFinder

[13] http://pdl.iec.es

[14] http://www.catedratelefonica.upf.es/cucweb

form, lemma, part of speech, syntactic function or combination of any of those. If a part of speech is specified, further morphological information is displayed, which can also be queried.

Each word unit can be marked as optional or repeated, which corresponds to the Boolean operators of repetition and optionality. Within each word unit each information field may be negated, allowing for exclusions in searches, e.g. requiring a unit not to be a noun or not corresponding to a certain lemma. This use of operators gives the expert mode an expressiveness close to regular grammars, and exploits almost all querying functionalities of CQP –the search engine.

In both modes, the user can retrieve up to 1000 examples, which can be viewed online or downloaded as a text file, and with different context sizes. In addition, a link to a cache copy of the document and to its original location is provided.

As for the statistics interface, it searches for frequency information regarding the query of the user. The frequency can be related to any of the 4 annotation levels (word, lemma, POS, function). For example, it is possible to search for a given verb lemma and get the frequencies of each verb form, or to look for adjectives modifying the word *dona* (‘woman’) and obtain the list of lemmata with their associated frequency. The results are offered as a table with absolute and relative frequency, and they can be viewed online or retrieved as a CSV file. In addition, each of the results has an associated link to the actual examples in the corpus.

The interface is technically quite complex, and the corpus quite large. There are still aspects to be solved both in the implementation and the documentation of the interface. Even restricting the searches to 1000 hits, efficiency remains often a problem in the example search mode, and more so in the statistics interface. Two partial solutions have been adopted so far: first, to divide the corpus into 4 subcorpora, as explained in Section 2.2, so that parallel searches can be performed and thus the search engine is not as often overloaded. Second, to limit the amount of memory and time for a given query. In the statistics interface, a status bar shows the progress of the query in percentage and the time left.

The interface does not offer the full range of CWB/CQP functionalities, mainly because it was not demanded by our “known” users (most of them linguists and translators from the Department of Translation and Philology at Universitat Pompeu Fabra). However it is planned to increasingly add new features and functionalities. Up to now we did not detect any incompatibility between splitting the corpora and the implementation of CWB/CQP deployment or querying functionalities.

3.2 Whole dataset

The annotated corpus can be used as a source of data for NLP purposes. A previous version of the CUCWeb corpus –obtained with the methodology described in this paper, but crawling only the `.es` domain, consisting of 180 million words– has already been exploited in a lexical acquisition task, aimed at classifying Catalan verbs into syntactic classes (Mayol et al., 2006).

Cluster analysis was applied to a 200 verb set, modeled in terms of 10 linguistically defined features. The data for the clustering were first extracted from a fragment of CTILC (14 million word). Using the manual tagging of the corpus, an average 0.84 f-score was obtained. Using CatCG, the performance decreased only 2 points (0.82 f-score).

In a subsequent experiment, the data were extracted from the CUCWeb corpus. Given that it is 12 times larger than the traditional corpus, the question was whether “more data is better data” (Church and Mercer, 1993, 18-19). Banko and Brill (2001) present a case study on confusion set disambiguation that supports this slogan. Surprisingly enough, results using CUCWeb were significantly worse than those using the traditional corpus, even with automatic linguistic processing: CUCWeb lead to an average 0.71 f-score, so an 11 point difference resulted. These results somewhat question the quality of the CUCWeb corpus, particularly so as the authors attribute the difference to noise in the CUCWeb and difficulties in linguistic processing (see Section 4). However, 0.71 is still well beyond the 0.33 f-score baseline, so that our analysis is that CUCWeb can be successfully used in lexical acquisition tasks. Improvement in both filtering and linguistic processing is still a must, though.

4 Discussion of the architecture

The initial motivation for the CUCWeb project was to obtain a large annotated corpus for Catalan. However, we set up an architecture that enables

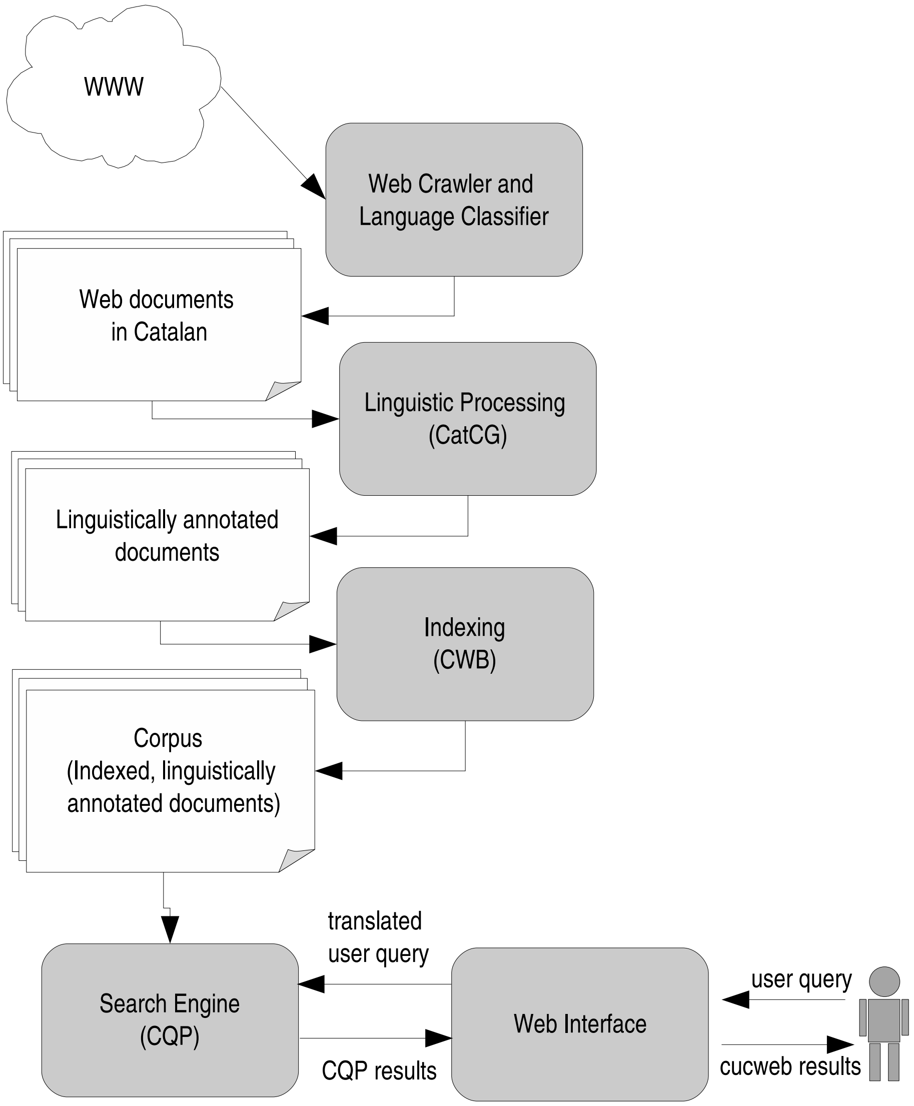
WWW
Web Crawler and Language Classifier
Web documents in Catalan
Linguistic Processing (CatCG)
Linguistically annotated documents
Indexing (CWB)
Corpus (Indexed, linguistically annotated documents)
translated user query
Search Engine (CQP)
Web Interface
CQP results
user query
cucweb results

Figure 1: Architecture for building Web corpora

the construction of web corpora in general, provided the language-dependent modules are available. Figure 1 shows the current architecture for CUCWeb.

The language-dependent modules are the language classifier (our classifier now covers 10 languages, as explained in Section 2.2) and the linguistic processing tools. In addition, the web interface has to be adapted for each new tagset, piece of information and linguistic level. For instance, the interface currently does not support searches for chunks or phrases.

Most of the problems we have encountered in processing Web documents are not new (Baroni and Ueyama, To appear), but they are much more frequent in that kind of documents than in standard running text.[15] We now review the main problems we came across:

Textual layout In general, they are problems that arise due to the layout of Web documents, which is very different to that of standard text. Pre-processing tools have to be adapted to deal with these elements. These include headers or footers (*Last modified...*), copyright statements or frame elements, the so-called *boilerplates*. Currently, due to the fact that we process the text extracted by the crawler, no boilerplate detection is performed, which increases the amount of noise in the corpus. Moreover, the pre-processing module does not even handle e-mail addresses or phone numbers (they are not frequently found in the kind of

[15] By “standard text”, we mean edited pieces of text, such as newspapers, novels, encyclopedia, or technical manuals.

text it was designed to process); as a result, for example, one of the most frequent determiners in the corpus is *93*, the phone prefix for Barcelona. Another problem for the pre-processing module, again due to the fact that we process the text extracted from the HTML markup, is that most of the structural information is lost and many segmentation errors occur, errors that carry over to subsequent modules.

Spelling mistakes Most of the texts published on the Web are only edited once, by their author, and are neither reviewed nor corrected, as is usually the case in traditional textual collections (Baeza-Yates et al., 2005). It could be argued that this makes the language on the Web closer to the "actual language", or at least representative of other varieties in contrast to traditional corpora. However, this feature makes Web documents difficult to process for NLP purposes, due to the large quantity of spelling mistakes of all kinds. The HTML support itself causes some of the difficulties that are not exactly spelling mistakes: A particularly frequent kind of problem we have found is that the first letter of a word gets segmented from the rest of the word, mainly due to formatting effects. Automatic spelling correction is a more necessary module in the case of Web data.

Multilinguality Multilinguality is also not a new issue (there are indeed multilingual books or journals), but is one that becomes much more evident when handling Web documents. Our current approach, given that we are not interested in full coverage, but in quality, is to discard multilingual documents (through the language classifier and the linguistic filter). This causes two problems. On the one hand, potentially useful texts are lost, if they are inserted in multilingual documents (note that the linguistic filter reduces the initial collection to almost a half; see Table 1). On the other hand, many multilingual documents remain in the corpus, because the amount of text in another language does not reach the specified threshold. Due to the sociological context of Catalan, Spanish-Catalan documents are particularly frequent, and this can cause trouble in e.g. lexical acquisition tasks, because both are Romance languages and some word forms coincide. Currently, both the language classifier and the dictionary filter are document-based, not sentence-based. A better approach would be to do sentence-based language classification. However, this would increase the complexity of corpus construction and management: If we want to maintain the notion of document, pieces in other languages have to be marked but not removed. Ideally, they should also be tagged and subsequently made searchable.

Duplicates Finally, a problem which is indeed particular to the Web is redundancy. Despite all efforts in avoiding duplicates during the crawling and in detecting them in the collection (see Section 2), there is still quite a lot of duplicates or near-duplicates in the corpus. This is a problem both for NLP purposes and for corpus querying. More sophisticated algorithms, as in Broder (2000), are needed to improve duplicate detection.

5 Conclusions and future work

We have presented CUCWeb, a project aimed at obtaining a large Catalan corpus from the Web and making it available for all language users. As an existing resource, it is possible to enhance it and modify it, with e.g. better filters, better duplicate detectors, or better NLP tools. Having an actual corpus stored and annotated also makes it possible to explore it, be it through the web interface or as a dataset.

The first CUCWeb version (from data gathering to linguistic processing and web interface implementation) was developed in only 6 months, with partial dedication of a a team of 6 people. Since then, many improvements have taken place, and many more remain as a challenge, but it confirms that creating a 166 million word annotated corpus, given the current technological state of the art, is a relatively easy and cheap issue.

Resources such as CUCWeb facilitate the technological development of non-major languages and quantitative linguistic research, particularly so if flexible web interfaces are implemented. In addition, they make it possible for NLP and Web studies to converge, opening new fields of research (e.g. sociolinguistic studies of the Web).

We have argued that the developed architecture allows for the creation of Web corpora in general. In fact, in the near future we plan to build a Spanish Web corpus and integrate it into the same web interface, using the data already gathered. The Spanish corpus, however, will be much larger than the Catalan one (a conservative estimate is 600 million words), so that new challenges in processing and searching it will arise.

We have also reviewed some of the challenges that Web data pose to existing NLP tools, and argued that most are not new (textual layout, misspellings, multilinguality), but more frequent on the Web. To address some of them, we plan to develop a more sophisticated pre-processing module and a sentence-based language classifier and filter.

A more general challenge of Web corpora is the control over its contents. Unlike traditional corpora, where the origin of each text is clear and deliberate, in CUCWeb the strategy is to gather as much text as possible, provided it meets some quality heuristics. The notion of balance is not present anymore, although this needs not be a drawback (Web corpora are at least representative of the language on the Web). However, what is arguably a drawback is the black box effect of the corpus, because the impact of text genre, topic, and so on cannot be taken into account. It would require a text classification procedure to know what the collected corpus contains, and this is again a meeting point for Web studies and NLP.

Acknowledgements

María Eugenia Fuenmayor and Paulo Golgher managed the Web crawler during the downloading process. The language classifier was developed by Bárbara Poblete. The corpora used to train the language detection module were kindly provided by Universität Gesamthochschule, Paderborn (German), by the Institut d'Estudis Catalans, Barcelona (Catalan), by the TALP group, Universitat Politècnica de Catalunya (Spanish), by the IXA Group, Euskal Herriko Unibertsitatea (Basque), by the Centre de Traitement Automatique du Langage de l'UCL, Leuven (French, Dutch and Portuguese), by the Seminario de Lingüística Informática, Universidade de Vigo (Galician) and by the Istituto di Linguistica Computazionale, Pisa (Italian). We thank Martí Quixal for his revision of a previous version of this paper and three anonymous reviewers for useful criticism.

This project has been partially funded by Cátedra Telefónica de Producción Multimedia.

References

Àlex Alsina, Toni Badia, Gemma Boleda, Stefan Bott, Àngel Gil, Martí Quixal, and Oriol Valentín. 2002. CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Anonymous. 2000. 1.6 billion served: the Web according to Google. *Wired*, 8(12):18–19.

Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. 2005. Characteristics of the Web of Spain. *Cybermetrics*, 9(1).

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Association for Computational Linguistics*, pages 26–33.

Marco Baroni and Motoko Ueyama. To appear. Building general- and special-purpose corpora by web crawling. In *Proceedings of the NIJL International Workshop on Language Corpora*.

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

Kenneth W. Church and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Altigran da Silva, Eveline Veloso, Paulo Golgher, Alberto Laender, and Nivio Ziviani. 1999. Cobweb - a crawler for the brazilian web. In *String Processing and Information Retrieval (SPIRE)*, pages 184–191, Cancun, Mexico. IEEE CS Press.

Dennis Fetterly, Mark Manasse, and Marc Najork. 2004. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Seventh workshop on the Web and databases (WebDB)*, Paris, France.

Gregory Grefenstette. 1998. The World Wide Web as a resource for example-based machine translation tasks. In *ASLIB Conference on Translating and the Computer*, volume 21, London, England.

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

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

Anke Lüdeling, Stefan Evert, and Marco Baroni. To appear. Using web data for linguistic purposes. In Marianne Hundt, Caroline Biewer, and Nadja Nesselhauf, editors, *Corpus Linguistics and the Web*. Rodopi, Amsterdam.

Laia Mayol, Gemma Boleda, and Toni Badia. 2006. Automatic acquisition of syntactic verb classes with basic resources. Submitted.

Andrew K. Mccallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/∼mccallum/bow/>.

Joaquim Rafel. 1994. Un corpus general de referència de la llengua catalana. *Caplletra*, 17:219–250.

Mike Thelwall. 2005. Creating and using web corpora. *International Journal of Corpus Linguistics*, 10(4):517–541.