Finding viable seed URLs for web corpora: A scouting approach and comparative study of available sources

Adrien Barbaresi

Finding viable seed URLs for web corpora: a scouting approach and comparative study of available sources

Adrien Barbaresi
ICAR Lab
ENS Lyon & University of Lyon
15 parvis René Descartes, 69007 Lyon, France
adrien.barbaresi@ens-lyon.fr

Abstract

The conventional tools of the "web as corpus" framework rely heavily on URLs obtained from search engines. Recently, the corresponding querying process became much slower or impossible to perform on a low budget. I try to find acceptable substitutes, i.e. viable link sources for web corpus construction. To this end, I perform a study of possible alternatives, including social networks as well as the Open Directory Project and Wikipedia. Four different languages (Dutch, French, Indonesian and Swedish) taken as examples show that complementary approaches are needed. My scouting approach using open-source software leads to a URL directory enriched with metadata which may be used to start a web crawl. This is more than a drop-in replacement for existing tools since said metadata enables researchers to filter and select URLs that fit particular needs, as they are classified according to their language, their length and a few other indicators such as host- and markup-based data.

1 Introduction

1.1 The "web as corpus" paradigm and its URL seeds problem

The state of the art tools of the "web as corpus" framework rely heavily on URLs obtained from search engines. The BootCaT method (Baroni and Bernardini, 2004) consists in repeated search engine queries using several word seeds that are randomly combined, first coming from an initial list and later from unigram extraction over the corpus itself. As a result, so-called "seed URLs" are gathered which are used as a starting point for web crawlers. This approach is not limited to English: it has been successfully used by Baroni et al. (2009) and Kilgarriff et al. (2010) for major world languages.

Until recently, the BootCaT method could be used in free web corpus building approaches. To my best knowledge it is now passé because of increasing limitations on the search engines' APIs, which make the querying process on a low budget much slower or impossible. Other technical difficulties include diverse and partly unknown search biases due in part to search engine optimization tricks as well as undocumented PageRank adjustments. All in all, the APIs may be too expensive and/or too unstable to support large-scale corpus building projects.

API changes are combined with an evolving web document structure and a slow but inescapable shift from "web as corpus" to "web for corpus" due to the increasing number of web pages and the necessity of using sampling methods at some stage. This is what I call the post-BootCaT world in web corpus construction.[1]

Moreover, the question whether the method used so far, i.e. randomizing keywords, provides a good overview of a language is still open. It now seems reasonable to look for alternatives, so that research material does not depend on a single data source, as this kind of black box effect combined with paid queries really impedes reproducibility of research. Using diverse sources of URL seeds could at least ensure that there is not a single bias, but several.

Additionally, the lack of interest and project financing when dealing with certain less-resourced languages makes it necessary to use light-weight

[1] Note that the proponents of the BootCaT method seem to acknowledge this evolution, see for example Marco Baroni's talk at this year's BootCaTters of the world unite (BOTWU) workshop: "My love affair with the Web... and why it's over!"

approaches where costs are lowered as much as possible (Scannell, 2007). In this perspective, a preliminary light scouting approach and a full-fledged focused crawler like those used by the Spiderling (Suchomel and Pomikálek, 2012) or the COW (Schäfer and Bildhauer, 2012) projects are complementary. A "web for corpus" crawling method using a seed set enriched with metadata as described in this article may yield better results, e.g. ensure a more diverse and less skewed sample distribution in a population of web documents, and/or reach faster a given quantitative goal.

1.2 Looking for alternatives, what issues do we face?

Search engines have not been taken as a source simply because they were convenient. They actually yield good results in terms of linguistic quality. The main advantage was to outsource operations such as web crawling and website quality filtering, which are considered to be too costly or too complicated to deal with while the main purpose is actually to build a corpus.

In fact, it is not possible to start a web crawl from scratch, so the main issue to tackle can be put this way: where may we find web pages which are bound to be interesting for corpus linguists and which in turn contain many links to other interesting web pages?

Researchers in the machine translation field have started another attempt to outsource competence and computing power, making use of data gathered by the CommonCrawl project[2] to find parallel corpora (Smith et al., 2013). Nonetheless, the quality of the links may not live up to their expectations. First, purely URL-based approaches are a trade-off in favor of speed which sacrifices precision, and language identification tasks are a good example of this phenomenon (Baykan et al., 2008). Second, machine-translated content is a major issue, so is text quality in general, especially when it comes to web texts (Arase and Zhou, 2013). Third, mixed-language documents slow down text gathering processes (King and Abney, 2013). Fourth, link diversity is a also problem, which in my opinion has not got the attention it deserves. Last, the resource is constantly moving. There are not only fast URL changes and ubiquitous redirections. Following the "web 2.0" paradigm, much web content is being injected from other sources, so that many web pages are now expected to change any time.[3] Regular exploration and re-analysis could be the way to go to ensure the durability of the resource.

In the remainder of this paper, I introduce a scouting approach which considers the first issue, touches on the second one, provides tools and metrics to address the third and fourth, and adapts to the last. In the following section I describe my methodology, then I show in detail which metrics I decided to use, and last I discuss the results.

2 Method

2.1 Languages studied

I chose four different languages in order to see if my approach generalizes well: Dutch, French, Indonesian and Swedish. It enables me to compare several language-dependent web spaces which ought to have different if not incompatible characteristics. In fact, the "speaker to website quantity" ratio is probably extremely different when it comes to Swedish and Indonesian. I showed in a previous study that this affects greatly link discovery and corpus construction processes (Barbaresi, 2013a).

French is spoken on several continents and Dutch is spoken in several countries (Afrikaans was not part of this study). Indonesian offers an interesting point of comparison, as the chances to find web pages in this language during a crawl at random are scarce. For this very reason, I explicitly chose not to study English or Chinese because they are clearly the most prominently represented languages on the web.

2.2 Data sources

I use two reference points, the first one being the existing method depending on search engine queries, upon which I hope to cast a new light with this study. The comparison grounds on URLs retrieved using the BootCaT seed method on the meta-engine E-Tools[4] at the end of 2012. The second reference point consists of social networks, to whose linguistic structure I already dedicated a study (Barbaresi, 2013b) where the method used to find the URLs is described in detail. I chose to adopt a different perspective, to re-examine the URLs I gathered and to add relevant metadata

[2] http://commoncrawl.org/

[3] This is the reason why Marco Baroni states in the talk mentioned above that his "love affair with the web" is over.

[4] http://www.etools.ch/

in order to see how they compared to the other sources studied here.

I chose to focus on three different networks: FriendFeed, an aggregator that offers a broader spectrum of retrieved information; identi.ca, a microblogging service similar to Twitter; and Reddit, a social bookmarking and microblogging platform. Perhaps not surprisingly, these data sources display the issues linked to API instability mentioned above. The example of identi.ca is telling: until March 2013, when the API was closed after the company was bought, it was a social microblogging service built on open source tools and open standards, the advantages compared to Twitter include the Creative Commons license of the content, and the absence of limitations on the total number of pages seen.

Another data source is the Open Directory Project (DMOZ[5]), where a selection of links is curated according to their language and/or topic. The language classification is expected to be adequate, but the amount of viable links is an open question, as well as the content.

Last, the free encyclopedia Wikipedia is another spam-resilient data source in which the quality of links is expected to be high. It is acknowledged that the encyclopedia in a given language edition is a useful resource, the open question resides in the links pointing to the outside world, as it is hard to get an idea of their characteristics due to the large number of articles, which is rapidly increasing even for an under-resourced language such as Indonesian.

2.3 Processing pipeline

The following sketch describes how the results below were obtained:

1. URL harvesting: queries or archive/dump traversal, filtering of obvious spam and non-text documents.

2. Operations on the URL queue: redirection checks, sampling by domain name.

3. Download of the web documents and analysis: collection of host- and markup-based data, HTML code stripping, document validity check, language identification.

Links pointing to media documents were excluded from this study, as its final purpose is to enable construction of a text corpus. The URL checker removes non-http protocols, images, PDFs, audio and video files, ad banners, feeds and unwanted hostnames like *twitter.com*, *google.com*, *youtube.com* or *flickr.com*. Additionally, a proper spam filtering is performed on the whole URL (using basic regular expressions) as well as at domain name level using a list of blacklisted domains comparable to those used by e-mail services to filter spam. As a page is downloaded or a query is executed, links are filtered on-the-fly using a series of heuristics described below, and finally the rest of the links are stored.

There are two other major filtering operations to be aware of. The first concerns the URLs, which are sampled prior to the download. The main goal of this operation is strongly related to my scouting approach. Since I set my tools on an exploration course, this allows for a faster execution and provides us with a more realistic image of what awaits a potential exhaustive crawler. Because of the sampling approach, the "big picture" cannot easily be distorted by a single website. This also avoids "hammering" a particular server unduly and facilitates compliance with *robots.txt* as well as other ethical rules. The second filter deals with the downloaded content: web pages are discarded if they are too short. Web documents which are more than a few megabytes long are also discarded.

Regarding the web pages, the software fetches them from a list, strips the HTML code, sends raw text to a server instance of *langid.py* (description below) and retrieves the server response, on which it performs a basic heuristic tests.

3 Metadata

The metadata described in this section can be used in classificatory or graph-based approaches. I use some of them in the results below but did not exhaust all the possible combinations in this study. There are nine of them in total, which can be divided in three categories: corpus size metrics, which are related to word count measures, web science metrics, which ought to be given a higher importance in web corpus building, and finally the language identification, which is performed using an external tool.

[5] http://www.dmoz.org/

3.1 Corpus size metrics

Web page length (in characters) was used as a discriminating factor. Web pages which were too short (less than 1,000 characters long after HTML stripping) were discarded in order to avoid documents containing just multimedia (pictures and/or videos) or microtext collections for example, as the purpose was to simulate the creation of a general-purpose text corpus.

The page length in characters after stripping was recorded, as well as the number of tokens, so that the total number of tokens of a web corpus built on this URL basis can be estimated. The page length distribution is not normal, with a majority of short web texts and a few incredibly long documents at the end of the spectrum, which is emphasized by the differences between mean and median values used in the results below and justifies the mention of both.

3.2 Web science metrics

Host sampling is a very important step because the number of web pages is drastically reduced, which makes the whole process more feasible and more well-balanced, i.e. less prone to host biases. IP-based statistics corroborate this hypothesis, as shown below.

The deduplication operation is elementary, it takes place at document level, using a hash function. The IP diversity is partly a relevant indicator, as it can be used to prove that not all domain names lead to the same server. Nonetheless, it cannot detect the duplication of the same document across many different servers with different IPs, which in turn the elementary deduplication is able to reveal.

Links that lead to pages within the same domain name and links which lead to other domains are extracted from the HTML markup. The first number can be used to find possible spam or irrelevant links, with the notable exception of websites like Amazon or Wikipedia, which are quite easy to list. The latter may be used to assess the richness (or at a given level the suspiciousness) of a website by the company it keeps. While this indicator is not perfect, it enables users to draw conclusions without fetching all the downstream URLs.

Moreover, even if I do not take advantage of this information in this study, the fetcher also records all the links it "sees" (as an origin-destination pair), which enables graph-based approaches such as visualization of the gathered network or the assessment of the "weight" of a website in the URL directory. Also, these metadata may very well be useful for finding promising start URLs.

3.3 Language identification

I consider the fact that a lot of web pages have characteristics which make it hard for "classical" NLP approaches like web page language identification based on URLs (Baykan et al., 2008) to predict the languages of the links with certainty. That is why mature NLP tools have to be used to qualify the incoming URLs and enable a language-based filtering based on actual facts.

The language identification tool I used is *langid.py* (Lui and Baldwin, 2012). It is open-source, it incorporates a pre-trained model and it covers 97 languages, which is ideal for tackling the diversity of the web. Its use as a web service makes it a fast solution enabling distant or distributed work.

As the software is still under active development, it can encounter difficulties with rare encodings. As a result, the text gets falsely classified as for example Russian or Chinese. The languages I studied are not affected by these issues. Still, language identification at document level raises a few problems regarding "parasite" languages (Scannell, 2007).

Using a language identification system has a few benefits: it enables finding "regular" texts in terms of statistical properties and excluding certain types of irregularities such as encoding problems. Web text collections are smoothed out in relation to the statistical model applied for each language target, which is a partly destructive but interesting feature.

There are cases where the confidence interval of the language identifier is highly relevant, for instance if the page is multi-lingual. Then there are two main effects: on one hand the confidence indicator gets a lower value, so that it is possible to isolate pages which are likely to be in the target language only. On the other hand, the language guessed is the one with the largest number of identifiable words: if a given web page contains 70 % Danish and 30 % English, then it will be classified as being written in Danish, with a low confidence interval: this information is part of the metadata I associate with each web page. Since nothing particular stood out in this respect I do not mention it further.

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
Dutch	12,839	1,577	84.6	27,153	3,600	5,325,275	73.1
French	16,763	4,215	70.2	47,634	8,518	19,865,833	50.5
Indonesian	110,333	11,386	66.9	49,731	8,634	50,339,311	18.6
Swedish	179,658	24,456	88.9	24,221	9,994	75,328,265	20.0

Table 1: URLs extracted from search engines queries

4 Results

4.1 Characteristics of the BootCaT approach

First of all, I let my toolchain run on URLs obtained using the BootCaT approach, in order to get a glimpse of its characteristics. I let the URL extractor run for several weeks on Indonesian and Swedish and only a few days for Dutch and French, since I was limited by the constraints of this approach, which becomes exponentially slower as one adds target languages.[6] The results commented below are displayed in table 1.

The domain name reduction has a substantial impact on the set of URLs, as about a quarter of the URLs at best (for French) have different domain names. This is a first hint at the lack of diversity of the URLs found using the BootCaT technique.

Unsurprisingly, the majority of links appear to be in the target language, although the language filters do not seem to perform very well. As the adequate matching of documents to the user's language is paramount for search engines, it is probably a bias of the querying methodology and its random tuples of tokens. In fact, it is not rare to find unexpected and undesirable documents such as word lists or search engine optimization traps.

The length of web documents is remarkable, it indicates that there are likely to contain long texts. Moreover, the median length seems to be quite constant across the three languages at about 8,000 tokens, whereas it is less than half that (3,600) for Dutch. All in all, it appears to be an advantage which clearly explains why this method has been considered to be successful. The potential corpus sizes are noteworthy, especially when enough URLs where gathered in the first place, which was already too impracticable in my case to be considered a sustainable option.

The number of different IPs, i.e. the diversity in terms of hosts, seems to get gradually lower as the URL list becomes larger. The fact that the same phenomenon happens for Indonesian and Swedish, with one host out of five being "new", indicates a strong tendency.

4.2 Social networks

Due to the mixed nature of the experimental setting, no conclusions can be drawn concerning the single components. The more than 700,000 URLs that were analyzed give an insight regarding the usefulness of these sources. About a tenth of it remained as responding websites with different domain names, which is the lowest ratio of this study. It may be explained by the fast-paced evolution of microblogs and also by the potential impurity of the source compared to the user-reviewed directories whose results I describe next.

As I did not target the studied languages during the URL collection process, there were merely a few hundred different domain names to be found, with the exception of French, which was a lot more prominent.

Table 2 provides an overview of the results. The mean and median lengths are clearly lower than in the search engine experiment. In the case of French, with a comparable number of remaining URLs, the corpus size estimate is about 2.5 times smaller. The host diversity is comparable, and does not seem to be an issue at this point.

All in all, social networks are probably a good candidate for web corpora, but they require a focused approach of microtext to target a particular community of speakers.

4.3 DMOZ

As expected, the number of different domain names on the Open Directory project is high, giv-

[6] The slow URL collection is explained by the cautious handling of this free and reliable source, implying a query rate limiting on my side. The scouting approach by itself is a matter of hours.

	% in target	URLs retained	Length		Tokens (total)	Different IPs (%)
			mean	median		
Dutch	0.6	465	7,560	4,162	470,841	68.8
French	5.9	4,320	11,170	5,126	7,512,962	49.7
Indonesian	0.5	336	6,682	4,818	292,967	50.9
Swedish	1.1	817	13,807	7,059	1,881,970	58.5

Table 2: URLs extracted from a blend of social networks crawls (FriendFeed, identi.ca, and Reddit) with no language target. 738,476 URLs analyzed, 73,271 URLs retained in the global process.

ing the best ratio in this study between unfiltered and remaining URLs. The lack of web pages written in Indonesian is a problem for this source, whereas the other languages seem to be far better covered. The adequacy of the web pages with respect to their language is excellent, as shown in table 3. These results underline the quality of the resource.

On the other hand, document length is the biggest issue here. The mean and median values indicate that this characteristic is quite homogeneous throughout the document collection. This may easily be explained by the fact that the URLs which are listed on DMOZ mostly lead to corporate homepages for example, which are clear and concise, the eventual "real" text content being somewhere else. What's more, the websites in question are not text reservoirs by nature. Nonetheless, the sheer quantity of listed URLs compensates for this fact. The corpus sizes for Dutch and French are quite reasonable if one bears in mind that the URLs were sampled.

The relative diversity of IPs compared to the number of domain names visited is another indicator that the Open Directory leads to a wide range of websites. The directory performs well compared to the sources mentioned above, it is also much easier to crawl. It did not cost us more than a few lines of code followed by a few minutes of runtime to gather the URLs.

4.4 Wikipedia

The characteristics of Wikipedia are quite similar, since the free encyclopedia also makes dumps available, which are easily combed through in order to gather start URLs. Wikipedia also compares favorably to search engines or social networks when it comes to the sampling operation and page availability. It is a major source of URLs, with numbers of gathered URLs in the millions for languages like French. As Wikipedia is not a URL directory by nature, it is interesting to see what are the characteristics of the pages it links to are. The results are shown in table 3.

First, the pages referenced in a particular language edition of Wikipedia often point to web pages written in a foreign language. According to my figures, this is a clear case, all the more since web pages in Indonesian are rare. Still, with a total of more than 4,000 retained web texts, it fares a lot better than DMOZ or social networks.

The web pages are longer than the ones from DMOZ, but shorter than the rest. This may also be related to the large number of concise homepages in the total. Nonetheless, the impressive number of URLs in the target language is decisive for corpus building purposes, with the second-biggest corpus size estimate obtained for French.

The IP-related indicator yields good results with respect to the number of URLs that were retrieved. Because to the high number of analyzed URLs the figures between 30 and 46% give an insight into the concentration of web hosting providers on the market.

5 Discussion

I also analyzed the results regarding the number of links that lead out of the page's domain name. For all sources, I found no consistent results across languages, with figures varying by a factor of three. Nonetheless, there seem to be a tendency towards a hierarchy in which the search engines are on top, followed by social networks, Wikipedia and DMOZ. This is one more hint at the heterogeneous nature of the data sources I examined with respect to the criteria I chose.

This hierarchy is also one more reason why

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
DMOZ							
Dutch	86,333	39,627	94.0	2,845	1,846	13,895,320	43.2
French	225,569	80,150	90.7	3,635	1,915	35,243,024	33.4
Indonesian	2,336	1,088	71.0	5,573	3,922	540,371	81.5
Swedish	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8
Wikipedia							
Dutch	489,506	91,007	31.3	4,055	2,305	15,398,721	43.1
French	1,472,202	201,471	39.4	5,939	2,710	64,329,516	29.5
Indonesian	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
Swedish	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Table 3: URLs extracted from DMOZ and Wikipedia

search engines queries are believed to be fast and reliable in terms of quantity. This method was fast, as the web pages are long and full of links, which enables to rapidly harvest a large number of web pages without having to worry about going round in circles. The researchers using the BootCaT method probably took advantage of the undocumented but efficient filtering operations which search engines perform in order to lead to reliable documents. Since this process takes place in a competitive sector where this kind of information can be sold, it may explain why the companies now try to avoid giving it away for free.

In the long run, several questions regarding URL quality remain open. As I show using a high-credibility source such as Wikipedia, the search engines results are probably closer to the maximum amount of text that is to be found on a given website than the other sources, all the more when the sampling procedure chooses a page at random without analyzing the rest of a website and thus without maximizing its potential in terms of tokens. Nonetheless, confrontation with the constantly increasing number of URLs to analyze and necessarily limited resources make a website sampling by domain name useful.

This is part of my cost-efficient approach, where the relatively low performance of Wikipedia and DMOZ is compensated by the ease of URL extraction. Besides, the size of the potential corpora mentioned here could increase dramatically if one was to remove the domain name sampling process and if one was to select the web pages with the most out-domain links for the crawl.

What's more, DMOZ and Wikipedia are likely to improve over time concerning the number of URLs they reference. As diversity and costs (temporal or financial) are real issues, a combined approach could take the best of all worlds and provide a web crawler with distinct and distant starting points, between the terse web pages referenced in DMOZ and the expected "freshness" of social networks. This could be a track to consider, as they could provide a not inconsiderable amount of promising URLs.

Finally, from the output of the toolchain to a full-fledged web corpus, other fine-grained instruments as well as further decisions processes (Schäfer et al., 2013) will be needed. The fact that web documents coming from several sources already differ by our criteria does not exclude further differences regarding text content. By way of consequence, future work could include a few more linguistically relevant text quality indicators in order to go further in bridging the gap between web data, NLP and corpus linguistics.

6 Conclusion

I evaluated several strategies for finding texts on the web. The results distinguish no clear winner, complementary approaches are called for. In light of these results, it seems possible to replace or at least to complement the existing BootCaT approach. It is understandable why search engine queries have been considered a useful data source. However, I revealed that they lack diver-

sity at some point, which apart from their impracticality may provide sufficient impetus to look for alternatives.

I discussed how I address several issues in order to design robust processing tools which (combined to the diversity of sources and usable metadata) enable researchers to get a better glimpse of the course a crawl may take. The problem of link diversity has not been well-studied in a corpus linguistics context; I presented metrics to help quantify it and I showed a possible way to go in order to gather a corpus using several sources leading to a satisfying proportion of different domain names and hosts.

As a plea for a technicalities-aware corpus creation, I wish to bring to linguists' attention that the first step of web corpus construction in itself can change a lot of parameters. I argue that a minimum of web science knowledge among the corpus linguistics community could be very useful to fully comprehend all the issues at stake when dealing with corpora from the web.

The toolchain used to perform these experiments is open-source and can be found online.[7] The resulting URL directory, which includes the metadata used in this article, is available upon request. The light scouting approach allows for regular updates of the URL directory. It could also take advantage of the strengths of other tools in order to suit the needs of different communities.

Acknowledgments

This work has been partially supported by an internal grant of the FU Berlin as well as machine power provided by the COW (COrpora from the Web) project at the German Grammar Department. Thanks to Roland Schäfer for letting me use the URLs extracted from E-Tools and DMOZ.

References

Yuki Arase and Ming Zhou. 2013. Machine Translation Detection from Monolingual Web-Text. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 1597–1607.

Adrien Barbaresi. 2013a. Challenges in web corpus construction for low-resource languages in a post-BootCaT world. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Proceedings of the 6th Language & Technology Conference, Less Resourced Languages special track*, pages 69–73, Poznań.

[7] FLUX: Filtering and Language-identification for URL Crawling Seeds – https://github.com/adbar/flux-toolchain

Adrien Barbaresi. 2013b. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop*, pages 9–15.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC*, pages 1313–1316.

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(3):209–226.

E. Baykan, M. Henzinger, and I. Weber. 2008. Web Page Language Identification Based on URLs. *Proceedings of the VLDB Endowment*, 1(1):176–187.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and PVS Avinesh. 2010. A Corpus Factory for Many Languages. In *Proceedings of LREC*, pages 904–910.

Ben King and Steven Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 25–30.

Kevin P. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC*, pages 486–493.

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In Stefan Evert, Egon Stemle, and Paul Rayson, editors, *Proceedings of the 8th Web as Corpus Workshop*, pages 7–15.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 1374–1383.

Vít Suchomel and Jan Pomikálek. 2012. Efficient Webcrawling for large text corpora. In Adam Kilgarriff and Serge Sharoff, editors, *Proceedings of the 7th Web as Corpus Workshop*, pages 40–44.