Language Richness of the Web

Martin Majlíš, Zdeněk Žabokrtský
Charles University in Prague
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
majlis@ufal.mff.cuni.cz, zabokrtsky@ufal.mff.cuni.cz

Abstract
We have built a corpus containing texts in 106 languages from texts available on the Internet and on Wikipedia. The W2C Web Corpus contains 54.7 GB of text and the W2C Wiki Corpus contains 8.5 GB of text. The W2C Web Corpus contains more than 100 MB of text available for 75 languages. At least 10 MB of text is available for 100 languages. These corpora are a unique data source for linguists, since they outclass all published works both in the size of the material collected and the number of languages covered. This language data resource can be of use particularly to researchers specialized in multilingual technologies development. We also developed software that greatly simplifies the creation of a new text corpus for a given language, using text materials freely available on the Internet. Special attention was given to components for filtering and de-duplication that allow to keep the material quality very high.

Keywords: multilinguality, web corpus, under-resourced languages

1. Introduction

As statistical approaches become the dominant paradigm in natural language processing, there is an increasing demand for data. Linguistic data for many languages can be found on the Internet in a computer-accessible form, i.e. obtained very cheaply without additional digitalization costs. However, retrieving them requires special knowledge, computational resources, and tools. Even easy access to such data is one of the key issues for computational linguists, corpora of a reasonable size are available only for the major world languages.

We aim to collect, with minimal or no human intervention, a large multilingual corpus comprising of textual data available on the Internet for as many languages as possible, with the minimum total size of 10 MB of text per language. The novelty of our work lies in stressing the multilinguality dimension.

There are 6,909 known living languages according to the Ethnologue database,¹ but only about 390 of them are used by more than 1 million of native speakers. Roughly speaking, 95% of people are using only 5% of languages and 95% of languages are used by 5% of people. Furthermore, 55% of languages are used by a mere 1% of population.

Existing multilingual projects and multilingual web corpora as well as methods used for their construction are reviewed in Section 2. Methods used for constructing the W2C Web Corpus are described in Section 3. The constructed corpus is presented in Section 4., which sketches the amount of downloaded data and amount of texts remaining after processing. In the final Section 5., the W2C Wiki Corpus and the W2C Web Corpus are compared to the quality of retrieved texts.

2. Related Work

In this section, we briefly review the existing multilingual resources, then we introduce the existing multilingual corpora and at the end, we describe the procedures used for building web corpora.

2.1. Multilingual Resources

There are many publicly available projects which contain multilingual textual resources, including the following ones:

- The Rosetta Project² is an effort of native speakers and language specialists to construct a publicly accessible digital library of material on all known human languages. Some of the available materials are just scanned grammar books or simple dictionaries containing around 200 words.
- The Open Language Archives Community³ (OLAC) is also trying to create a digital library of language resources.
- Wikipedia⁴ is a free, web-based, collaborative, and multilingual encyclopaedia project.
- The Universal Declaration of Human Rights⁵ (UDHR) is a milestone document in the history of human rights, available in 379 languages.
- The Project Gutenberg⁶ is a volunteer effort to digitize and archive cultural works, especially books.
- Wikisource⁷ is an on-line library of free content textual sources, operated by the Wikimedia Foundation.

¹http://www.ethnologue.com/ethno_docs/distribution.asp?by=size
²http://rosettaproject.org/ and http://www.archive.org/details/rosettaproject
³http://www.language-archives.org/
⁴http://www.wikipedia.org/
⁵http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx
⁶http://www.gutenberg.org/
⁷http://www.wikisource.org/
The Watchtower\textsuperscript{8} is an illustrated religious magazine, published semi-monthly by Jehovah’s Witnesses. It is written in 366 languages. Texts are available as web pages or PDF files. All files have a very similar structure, which makes them as a very good source of parallel texts.

Open-source Software (OSS) is computer software that is available with source codes, that contains documentation, messages, and labels.

The number of languages and items available in the individual multilingual projects is presented in Table 1. We can observe that there are two language counts boundaries. The first one is around 60: Sixty languages are available in Project Gutenberg, Wikisource, and popes blessing Urbi et Orbi. This is the number of languages used in developed or newly industrialized countries countries, which covers almost 70\% of all people. The following boundary is around 300 languages which are used in the Universal Declaration of Human Rights, Wikipedia, the Watchtower, and Launchpad. This is the number of languages that is at least theoretically available and used in written form on the Internet. This covers almost 90\% of all people. In order to cover more languages, special language interest groups and linguistics specialists are required.

All these projects contain a relatively small amount of data in comparison to the total amount of data available on the Internet. Furthermore, all these resources do not provide balanced texts and special tools are needed for their extraction. Most researchers have therefore been focusing mainly on the construction of corpora from publicly available web pages.

### 2.2. Multilingual Web Corpora

Many web corpora were constructed during the last decade. Some of the most prominent are the following ones:

- Corpus Factory is a multilingual corpus constructed by Kilgarriff et al. (2010). It contains texts in 8 languages – Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai, and Vietnamese.
- Crúbádán 1.0 is a multilingual corpus introduced by Scannell (2007), containing texts in 487 languages.
- Crúbádán 2.0 is a successor of Crúbádán 1.0, published in December 2011 and containing texts in 1023 languages.
- WaCky was introduced by Baroni et al. (2009). This corpus contains texts in 6 languages – English, German, Italian, Russian, Spanish, and French.
- Crúbádán 3.0 is a successor of Crúbádán 2.0, published in June 2011 and containing texts in 2500 languages.
- I-X – Sharoff (2006) introduced BNC-like multilingual web corpus. This corpus contains texts in 6 languages – English, German, Italian, Russian, Spanish, and French.
- Wikipedia is a multilingual encyclopedia project.
- Wikisource is an illustrated religious magazine, published semi-monthly by Jehovah’s Witnesses.
- Corpus Factory is a multilingual corpus constructed by Kilgarriff et al. (2010). It contains texts in 8 languages – Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai, and Vietnamese.
- Crúbádán 1.0 is a multilingual corpus introduced by Scannell (2007), containing texts in 487 languages.
- Crúbádán 2.0 is a successor of Crúbádán 1.0, published in December 2011 and containing texts in 1023 languages.

Table 1: Multilingual resources – summary. Launchpad and Gnome are examples of Open-source Software projects.

| Projects        | Languages | Size            |
|-----------------|-----------|-----------------|
| Rosetta Project | over 2,500| over 100,000    |
| OLAC            | 4,575     | 82,051 items    |
| Wikipedia       | 281       | 19,034,746      |
| UDHR            | 379       | at most 379     |
| Project Gutenberg| 60       | 34,000 documents|
| Wikisource      | 62        | 1,028,303 pages |
| Watchtower      | 366       | thousands pages |
| Launchpad       | 323       | 1,730,383 strings |
| Gnome           | 173       | about 1 million of strings |

Table 2: Sizes of prominent web corpora in million of words.

| Corpus          | Lang | Median | Mean | Total |
|-----------------|------|--------|------|-------|
| Corpus Factory  | 8    | 102.0  | 85.5 | 684   |
| Crúbádán 1.0    | 487  | 0.068  | 1.6  | 769   |
| Crúbádán 2.0    | 1023 | 0.127  | 1.5  | 1556  |
| I-X             | 3    | 126.0  | 136.0| 409   |
| WaCky           | 3    | 1500.0 | 1592.0| 4778  |

The number of languages represented in the listed corpora as well as their data sizes are presented in Table 2.

### 2.3. Building Web Corpus

The process of building a web corpus is very similar across all the existing corpora and can be divided into several steps as follows:

1. Building an initial corpus from reliable text sources.
2. Generating n-tuples of words and using them as search queries.
3. Downloading the found web pages.
4. Removing boilerplate code.
5. Identifying language.
6. Removing duplicate content.

One of the last steps in web corpus construction is the corpus quality analysis. Without comparing them to existing corpora or any other reliable source of text, it is hard to say whether high quality texts were downloaded. Bharati et al. (2000) suggested using the number of unique unigrams, entropy, word and sentence lengths to compare different corpora.

When a corpus is constructed, it is important to store and distribute in an easily accessible form. Wynne (2005) as well as E-MELD\textsuperscript{9} suggests many tips. Archival copies should be made in a format which offers LOTS (i.e., it is Lossless, Open Standard, Transparent, and Supported by multiple vendors). The corpus should also contain a proper documentation of used formats, along with information about terms of use and access restrictions.

### 3. Building the Multilingual Web Corpus

This section describes the tools and methods used for building our web corpus. While the procedures are quite similar for many different corpora, our procedure, however, differed from the already existing ones in the focus on multilinguality and the amount of collected texts at the same time. We also concentrated on an additional step aimed at preparing metadata in order to minimize human intervention in the later steps. In addition, we applied byte n-grams to identify

\textsuperscript{8}http://watchtower.org/

\textsuperscript{9}http://emeld.org/school/bpnutshell.html
the web page language instead of function words or simple domain filtering. We also used more fine-grained approach to the duplicity reduction which better reflects the nature of the web.

3.1. Overview
In the initial step we gathered metadata from Wikipedia and Ethnologue. The downloaded metadata was stored in a database. When metadata was available, a wiki corpus was built from Wikipedia articles, using which frequency lists of trigrams and quadgrams were computed. The construction of the web corpus was divided into smaller jobs executed on a computer cluster, whose results were merged into a raw corpus. This raw corpus was reprocessed using an improved language identifier. From this corpus duplicities were removed, statistics were computed, and distribution packages were prepared.

3.2. Metadata
Metadata, such as language name, its ISO codes, classification, number of speakers, writing system, etc., was automatically downloaded for each language from the Internet. The following sources were combined:

- SIL International\textsuperscript{10} – which provides an easily parsable table of all languages with their ISO codes and names.
- Wikipedia\textsuperscript{11} – with its list of all wikipedias where they use their own codes and names.
- Ethnologue\textsuperscript{12} – providing an easily parsable page with information about each language.

The extracted metadata is now available at: \url{http://ufal.mff.cuni.cz/~majlis/w2c/api/}
This website is not only a web interface accessible for humans, but serves also as a RESTful API for programmatical retrieval of this data.

The RESTful API provides access to information about all languages. It is also capable of converting language names, ISO 639-1, ISO 639-2b, or ISO 639-2t codes into ISO 639-3 codes. The following output data formats are available:

- TXT – a tab-separated plain text output, to be easily processed with Unix command-line tools.
- XML – XML output.
- JSON – JSON output which can be easily applied in programs.

3.3. Building the Initial Corpus
We needed to obtain a reliable source of texts. These texts were necessary for the construction of initial search phrases. Another intended usage was the evaluation of the quality of the web corpus.

We decided to use Wikipedia, because it contains high quality texts that can be easily retrieved and processed. Wikipedia provides Wikipedia dumps for 282 languages. We only selected the regular articles from these dumps (i.e. no user pages, images, talk pages, categories, etc.). We only used the plain text bodies of articles discarding headers, tables, and all tags. From the extracted texts, we removed duplicate lines. For each Wikipedia containing at least 5,000 articles, we created a corpus of up to 20,000 articles.

3.4. Language Identification
To identify hundreds of languages on large number of documents, we need to use a different approach to language identification as opposed to the methods used by other multilingual corpora (cf. 2.2.), which worked only for a limited number of languages or did not scale well.

We decided to use the YALI algorithm introduced by Majliš (2012). We trained this language identifier on our initial corpus.

3.5. URL Seeds
We used a trigram frequency list from the initial corpus to generate search phrases. All trigrams containing a number or a punctuation character were removed, and from the remaining ones, only lines from 2nd to 5th percentile were chosen. We omitted the most frequent trigrams because they were mostly containing short function words, that are very similar among related languages.

We used 30 queries to Google and stored the first hundred of links.

3.6. Downloading the Data
We used our own system to download web pages, which is capable of running on multiple machines in an unstable environment with many breakages. It is also able to learn patterns of web pages worth downloading. The system consists of a single server responsible for distributing jobs and collecting results, and many clients, called workers divided into three types – crawlers, parsers, and detectors.

The crawler is responsible for downloading a specific URL and storing information about the current time, URL md5 hash, HTTP status code, base URL, charset, and size in bytes of the webpage.

The parser is responsible for extracting texts and links from web pages. It processes only HTML pages with the corresponding mime-type and a HTTP status code 200.

The detector is responsible for language identification. It processes only texts that are long enough and does not contain too many links.

When the server receives a result from any detector, it processes only URLs in the collected language. All outgoing links are added to the database and the text is appended to the corpus.

The first jobs are constructed from the initial URL seeds. The construction of the corpus with out system is relatively easy from the user point of view is, because it is sufficient to use only a single command as shown in Figure 1.

\texttt{./create-corpus.sh czech 10M}

Figure 1: Command for building a web corpus with 10 million words in Czech.

\textsuperscript{10}\url{http://sll.org}
\textsuperscript{11}\url{http://www.wikipedia.org/}
\textsuperscript{12}\url{http://www.ethnologue.com/}
3.7. Duplicity Reduction

While other multilingual corpora discard duplicities on the document level, we decided to use a more fine-grained approach in order to not throw away the whole document if it contains a duplicate passage. This decision is important especially for under-resourced languages. There are at least three reasons for such approach – spam, common passages, and incorrectly detected boilerplate code.

The spam problem is caused by the fact that a good position in search engine results is crucial for business success. There are thousands of pages trying to sell the same product, but users usually click only on the top few links. Therefore, spammers are trying to manipulate the search engine indexing (this technique is called spandexing). They build scraper sites – automatically generated tightly-knit website pages referring to each other. Their content is typically generated from Wikipedia or other publicly available resources. To trick the search engines, these websites do not contain exact copies of original texts, but rather only mixed fractions. It may fool the duplicity detection on document level. Another technique used by spammers, is spamming blogs, where bots comment blog posts. These comments contain links to the spammers’ website to increase its popularity. Spam in comments may also be the source of duplicities and therefore decrease the corpus quality. When a blogger writes a spot on his/her blog in language X, the text is valuable for the corpus. Later, when a few spam comments are attached, this article will still be identified as language X, but it will not be so valuable, because it will also contain some English sentences. When many such articles are added, the same comments may be presented many times.

The common passages problem is caused by writers who need to define terms in their articles. The general approach is using definitions from Wikipedia. For example, the following phrase (the first sentence from the Wikipedia article about Internet) is used on roughly 300,000 pages, according to Google:

The Internet is a global system of interconnected computer networks that use the standard Internet protocol suite (TCP/IP) to serve billions of users worldwide.

And the last reason is removing the boilerplate code which is repeated on every page from a certain website. After the duplicity reduction step, the corpus contains only unique paragraphs.

4. The Constructed Corpora

We have constructed 2 different corpora which are now publicly available at:
http://ufal.mff.cuni.cz/~majlis/w2c/

This web page also contains more detailed statistics and precomputed frequency lists from unigrams to fivegrams. The list of all languages included in these corpora is presented at the end of this article in Table 8.

4.1. W2C Wiki Corpus

The W2C Wiki Corpus contains texts in 106 languages with a total size of 8.53 GB. The detailed information about the

![Figure 2: The W2C Wiki Corpus – size in MB for individual languages, sorted according to their size in the W2C Wiki Corpus. English with 429 MB, German with 342 MB, and Russian with 350 MB are not displayed.]

| Languages | Raw text size
|-----------|----------------|
| 100       | 131.3 GB       |
| 200       | 120.2 GB       |
| 300       | 110.3 GB       |
| 400       | 100.4 GB       |
| 500       | 90.5 GB        |
| 600       | 80.6 GB        |
| 700       | 70.7 GB        |
| 800       | 60.8 GB        |
| 900       | 50.9 GB        |
| 1000      | 50.0 GB        |
| 1100      | 40.1 GB        |
| 1200      | 30.2 GB        |
| 1300      | 20.3 GB        |
| 1400      | 10.4 GB        |

Table 3: Download Statistics

The amount of texts, but we decided to prefer quality over quantity.

4.2. W2C Web Corpus

To construct the W2C Web Corpus, we downloaded more than 100 million web pages with a total size of over 4.5 TB. The amount of unique texts is relatively small, only 54.7 GB of text, as is presented in Table 3. This amount depends heavily on the parameters used during text extraction. With more relaxed rules, we were able to retrieve twice the amount of texts, but we decided to prefer quality over quantity.

![Figure 3: The W2C Web Corpus – size in MB for individual languages, sorted according to their size in the W2C Web Corpus. English with 4601 MB, German with 342 MB, and Russian with 350 MB are not displayed.]

The W2C Web Corpus contains texts in 106 languages with a total size of 54.7 GB. The detailed information about the
Comparing W2C Wiki Corpus and W2C Web Corpus is one of the possibilities how to check whether reliable data was downloaded. Difference in certain text properties may point of the possibilities how to check whether reliable data was downloaded. The following properties are used for comparing Wikipedia to a language for which suspicious material was collected.

### Table 5: The number of languages for which has been obtained more texts than Size MB.

| Size | Languages | Size | Languages |
|------|-----------|------|-----------|
| > 10 | 100       | > 160| 72        |
| > 20 | 94        | > 320| 63        |
| > 40 | 87        | > 640| 34        |
| > 80 | 77        |      |           |

The collected size differs for various languages – for 34 languages, more than 640 MB of texts are available, for 72 languages, more than 160 MB, and for 100 languages, moreover than 10 MB of texts. More details are presented in Table 5.

### 4.3. Comparison

The comparison of the W2C Wiki Corpus and the W2C Web Corpus with existing multilingual corpora is presented in Table 6. The number of words listed for W2C corpora is underestimated, because we only use space as word delimiter, which is inaccurate especially for under-resourced languages with non-latin alphabet.

### 5. Quality Evaluation

Comparing W2C Wiki Corpus and W2C Web Corpus is one of the possibilities how to check whether reliable data was downloaded. Difference in certain text properties may point to a language for which suspicious material was collected. The following properties are used for comparing Wikipedia and the Internet:

- Average word length
- Average sentence length
- Conditional entropy

The absolute values presented here should be used with caution, as their main purpose was only the comparison of the two corpora. The numbers may change significantly with a different preprocessing. Additional figures are presented in Majliší and Zabokrtský (2011).

### 5.1. Average Word Length

The average word length may reveal problems caused by HTML parsing. From the overall statistics shown in Table 7, we assume that the downloaded data has a reasonable quality, since the ratio of word lengths on Wikipieda and on the Internet is around 1.

The values are presented in Table 9 and visualized in Figure 4. The farthest outlier is the Burmese language (mya), which has almost 3 times shorter words on Wikipedia than on the Internet.
5.2. Average Sentence Length
The average sentence length is a good text quality measure as well, since it can reveal some errors in boilerplate code removal. The sentence lengths statistics are presented in Table 7. As we can see the median and the means have been also around 1, which means that most languages are processed correctly. The average sentence lengths are visualized in Figure 5.

The farthest outlier according to this metric is the Burmese language again, which has the average sentence length of almost 1586 words on Wikipedia, whereas only 27 on the Internet. Checking any page on Burmese Wikipedia reveals that it does not contain any dot, so the whole paragraph is treated as a single sentence, while extracted segments from the Internet are much shorter, which causes the difference.

5.3. Conditional Entropy
The conditional entropy is another measure used to compare the quality of texts retrieved from Wikipedia and from the Internet. The overall statistics are presented in Table 7. The conditional entropy ratio between the Wikipedia and the Internet is 0.88 in average, which reflects the fact that the data available on Internet has a higher variability. The plotted values are in Figure 6.

5.4. Quality Evaluation Summary
The Wiki-to-Web statistics for various metrics are presented in Table 7. This table shows that the W2C Web Corpus has similar properties as the W2C Wiki Corpus.

| Metric                      | 1st Qu. | Median | Mean  | 3rd Qu. |
|-----------------------------|---------|--------|-------|---------|
| Word Length                 | 0.948   | 0.973  | 0.973 | 1.005   |
| Sentence Length             | 0.863   | 0.960  | 1.559 | 1.068   |
| Conditional Entropy         | 0.814   | 0.894  | 0.887 | 0.962   |

Table 7: Wiki-to-Web ratios for average word length, average sentence length, and conditional entropy between texts for individual languages in the W2C Wiki Corpus and the W2C Web Corpus.

The common property of all the outliers is the fact that they are either minor languages, such as Maori (mi), Malagasy (mlg), for which only low quality texts were collected, or they are written in non-latin scripts, which are sensitive to preprocessing, such as Japanese (jpn), Chinese (zh), Nepali (nep), or Burmese language (mya).

6. Conclusions and Future Work
The W2C Web Corpus consists of texts in 106 languages available on the Internet, with a total size of 54.7 GB of text. There is more than 100 MB of text available for 75 languages, and at least 10 MB of text for 100 languages. It would be possible to achieve the same quota for more languages, albeit at the cost of decreasing corpus quality. The W2C Web Corpus is a unique data source for linguists, as it outclasses all published works both in terms of the size of the collected material and the number of languages covered. The collected data may be used for comparative analysis of related languages and construction of language models for various applications, such as machine translation, speech recognition, spell checking, etc.

We have developed tools for collecting metadata, building corpus from Wikipedia, crawling, reducing text duplicity, and statistical analysis. Along with the main W2C Web Corpus, we also constructed the W2C Wiki Corpus containing 8.5 GB of text in 106 languages from articles of the Wikipedia.
All downloaded data, more than 4.5 TB, are preserved for further investigations, so that more information about the real usage of the individual languages can be revealed, such as the distribution of character encodings or scripts for each language. Different tools for text extraction, language identification and duplicity detection may be plugged in. If the text extractor could extract text segments instead of complete pages, it would be possible to increase the corpus size for minor languages.

7. Acknowledgments

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). The research has been supported by the grant Khresmoi (FP7-ICT-2010-6-257528 of the EU and 7E11042 of the Czech Republic). We would like to thank Ondrej Dusek for his helpful comments.

8. References

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. 10.1007/s10579-009-9081-4.

A. Bharati, K. P. Rao, R. Sangal, and S. M. Bendre. 2000. Basic statistical analysis of corpus and cross comparison among corpora. Technical Report of Indian Institute of Information Technology.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and P. V. S. Avinesh. 2010. A corpus factory for many languages. In Language Resources and Evaluation.

Martin Majlíš and Zdeněk Žabokrtský. 2011. W2C - large multilingual corpus. Technical Report Prague, Czech Republic, ÚFAL, Charles University, December.

Martin Majlíš. 2012. Yet Another Language Identifier. In EACL 2012. The Association for Computer Linguistics.

Kevin P. Scannell, 2007. The Crába Dán Project: Corpus building for under-resourced languages, volume 4 of Cahiers du Cental, pages 5–15. C. Fairon and H. Naets and A. Kilgarriff and Gilles-Maurice de Schryver, Louvain-la-Neuve, Belgium.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In WaCky! Working papers on the Web as Corpus. Gedit.

Martin Wynne, 2005. Developing Linguistic Corpora: a Guide to Good Practice, chapter Archiving, Distribution and Preservation, pages 71–78. Oxford: Oxbow Books. Available online, Accessed 2011-01-01.
