Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff

Natural Language Processing Group, University of Leipzig, Germany
Johannisgasse 26, 04103 Leipzig
E-mail: { dgoldhahn, teckart, quasthoff } @informatik.uni-leipzig.de

Abstract

The Leipzig Corpora Collection offers free online access to 136 monolingual dictionaries enriched with statistical information. In this paper we describe current advances of the project in collecting and processing text data automatically for a large number of languages. Our main interest lies in languages of "low density", where only few text data exists online. The aim of this approach is to create monolingual dictionaries and statistical information for a high number of new languages and to expand the existing dictionaries, opening up new possibilities for linguistic typology and other research. Focus of this paper will be set on the infrastructure for the automatic acquisition of large amounts of monolingual text in many languages from various sources. Preliminary results of the collection of text data will be presented. The mainly language-independent framework for preprocessing, cleaning and creating the corpora and computing the necessary statistics will also be depicted.

Keywords: corpus creation, text acquisition, minority languages

1. Introduction

The *Projekt Deutscher Wortschatz* (Quasthoff, 1998) started more than 15 years ago by creating a corpus-based monolingual dictionary of the German language available at http://wortschatz.uni-leipzig.de. Since June 2006 *Leipzig Corpora Collection* (*LCC*) can be accessed at http://corpora.uni-leipzig.de (Biemann 2007; Quasthoff, 2006a). It offers corpus-based monolingual full form dictionaries of several languages and a Web interface for general access. These standard sized corpora and their corresponding statistics are created from newspaper texts or Web pages. For each word the dictionaries contain:

- Word frequency information (no lemmatization, each word form is treated separately)
- Sample sentences
- Statistically significant word co-occurrences (based on left or right neighbours or whole sentences)
- A semantic map visualizing the strongest word co-occurrences

So far, dictionaries of 136 languages depicted in table 1 are available. Corpus sizes are presented in figure 1. They are plotted as a function of language rank, languages are ordered by the number of available resources.

Afrikaans	Estonian	Lithuanian	Scots
Albanian	Faroese	Lushai	Scottish Gaelic
Amharic	Fijian	Luxemburgian	Serbian
Arabic	Finnish	Macedonian	Sicilian
Aragonese	French	Malay	Sinhala
Armenian	Ganda	Malayalam	Slovak
Asturian	Georgian	Maltese	Slovenian
Azerbaijanian	German	Maori	Somali
Bashkir	German (CH)	Marathi	Sorbian
Basque	Gilaki	Min Nan Chinese	Spanish
Belarusan	Goan Konkani	Mongolian	Spanish (MX)
Bengali	Greek	Nahuatl	Sundanesian
Bicolano	Greenlandic	Nepali	Swahili
Bishnupriya	Gujarati	Newari	Swedish
Bosnian	Haitian	Norwegian (Bokmål)	Tagalog
Bretonian	Hebrew	Norwegian (Nynorsk)	Tajik
Bulgarian	Hindi	Occitan	Tamil
Catalan	Hindi, Fiji	Ossetian	Tatar
Cebuano	Hungarian	Pampanga	Telugu
Chinese	Icelandic	Panjabi	Thai
Chuvash	Ido	Papiamento	Turkish
Corsican	Indonesian	Pashto	Ukrainian
Croatian	Interlingua	Pennsylvanian Dutch	Urdu
Czech	Italian	Persian	Uzbek
Danish	Japanese	Piemontese	Venetian
Dimli	Javanese	Polish	Vietnamese
Dutch	Kannada	Portuguese (Brazil)	Waray
Egyptian Arabic	Kazakh	Portuguese (Macao)	Welsh
English	Kiswahili	Portuguese (Portugal)	Western Frisian
English (AU)	Korean	Romanian	Western Mari
English (CA)	Kurdish	Russian	Western Panjabi
English (NZ)	Kyrgyz	Rusyn	Yakut
English (UK)	Latin	Sami	Yiddish
Esperanto	Latvian	Samogitian	Yoruba

Table 1: Monolingual dictionaries accessible online at http://corpora.uni-leipzig.de.

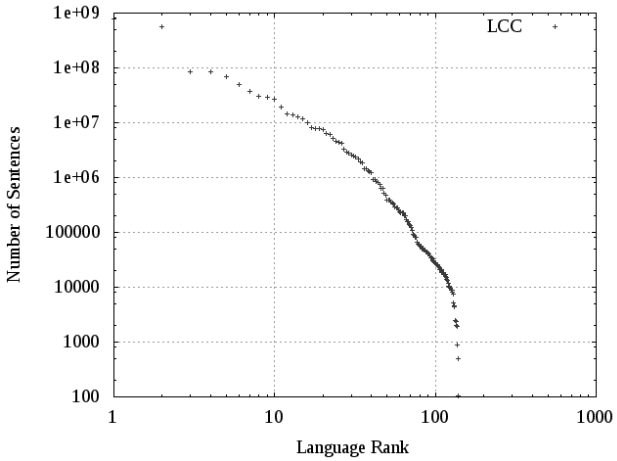

Figure 1: Number of sentences for all languages of *LCC* ordered by rank (log-log scale).

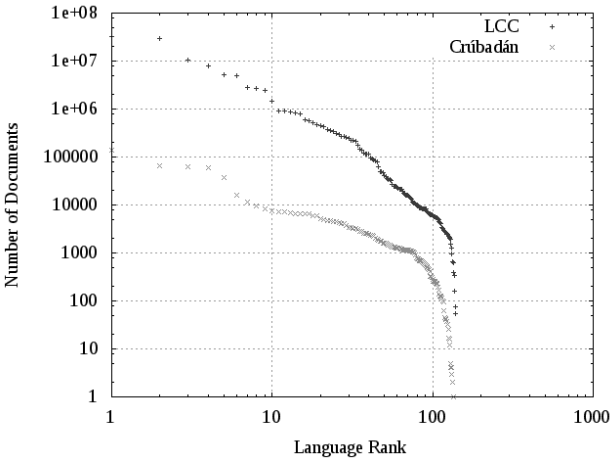

Figure 2: Number of documents for all common languages of *LCC* and *Crúbadán* ordered by rank (log-log scale).

In this paper we describe recent and current progress of *LCC* in collecting and processing text data automatically for a large number of languages. We aim mainly at languages of "low density", where only few text data exists online. Our goal is to increase the number of dictionaries and expand existing ones. Focus will be set on the infrastructure for the automatic acquisition of large amounts of monolingual text data in many languages from various Web sources. In addition the framework used for automatically preprocessing, cleaning and creating monolingual dictionaries and computing statistical information will also be presented.

Most Web corpora projects concentrate on creating dictionaries for one language, possibly offering text of different genres or sources, like *Corpuseye*[1] or *UKWeb*[2].

There are few other projects concerned with collecting text for several languages or offering access to statistics for these languages. One of them is *Web As Corpus*[3] which allows queries for concordances of words or phrases in 34 languages. Web as Corpus creates no dictionaries, all queries are directly processed using the search engine Bing. As opposed to *LCC* no cleaning, sentence segmentation, further processing or statistical evaluation of text data takes place.

One interesting venture is *Crúbadán*[4] by Kevin Scannell (2007). *Crúbadán* gathers documents for an enormous amount of languages using a bootstrapping approach. Currently ressources for 1023 languages exist. In *Crúbadán* text data for nearly 60% of all languages consist of 5 or less documents. A comparison of text size for common languages of *LCC* and *Crúbadán* can be found in figure 2.

Further differences between the projects are online access to statistics like co-occurrences, including a semantic map of the strongest co-occurrences, and example sentences offered by *LCC* and the possibility to download all these data for further academic use. An example of a co-occurrence graph for *Galatasaray*, a famous football club of Istanbul, is depicted in figure 3.

In the last 6 months the number of corpora of *LCC* has already been increased significantly (see figure 4). To add more languages to the collection, an automatic processing pipeline for collecting and processing text data has been implemented.

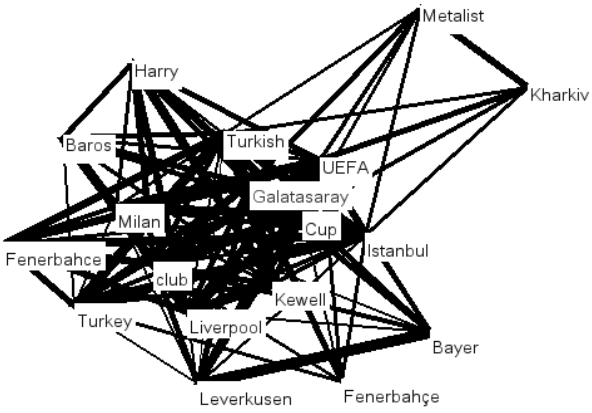

Figure 3: Semantic map for *Galatasaray*, a football club of Istanbul.

1 http://corp.hum.sdu.dk/cqp.en.html
2 http://faculty.washington.edu/dillon/csar-v02/
3 http://webascorpus.org
4 http://borel.slu.edu/crubadan/

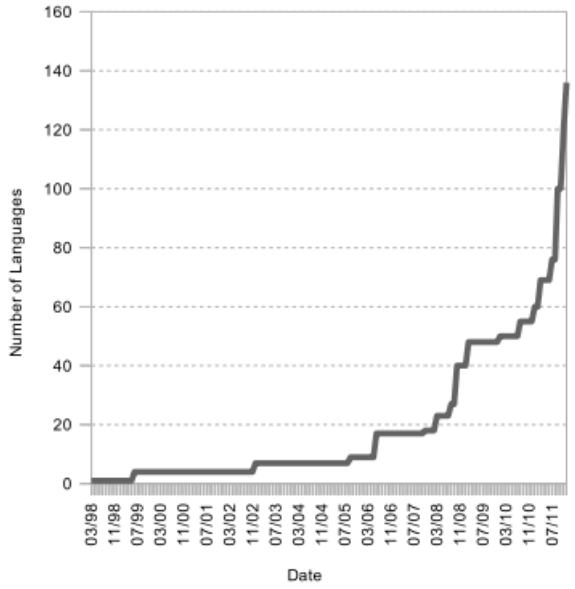

Figure 4: Number of available languages of the *Leipzig Corpora Collection*.

2. Collecting Data

In order to build monolingual corpora for different languages, first of all text data has to be collected. The World Wide Web is an obvious source for electronically available text (Kilgarriff, 2001). Different collection methods for Web sites are currently employed in the project to achieve a high coverage for each language and a high diversity concerning topics or genres. This allows creation of corpora that are a more representative sample of the language in question. Preparation of subcorpora of different genres or sizes for later comparison is also possible.

2.1 Crawling Newspapers

One approach for collecting text data is to crawl newspapers available online. Basis is a list of about 32,000 news sources in more than 120 languages provided by *ABYZ News Links*[5]. This service offers besides URLs also information regarding country and language. An overview of the distribution of *ABYZ's* resources is given in figure 5, while table 2 depicts the top-20 languages.

A framework for parallel Web crawling utilizing the standard Web site copier *HTTrack*[6] is applied. News sources of about 100 languages have been downloaded, yielding more than 250 GB of text after HTML-stripping but before language identification.

We were able to compile monolingual dictionaries for 70 languages based on this data. For the remaining languages the resources of *ABYZ News Links* either contained no text in the expected language or our crawler was excluded from downloading Web sites by the robots.txt of these Servers.

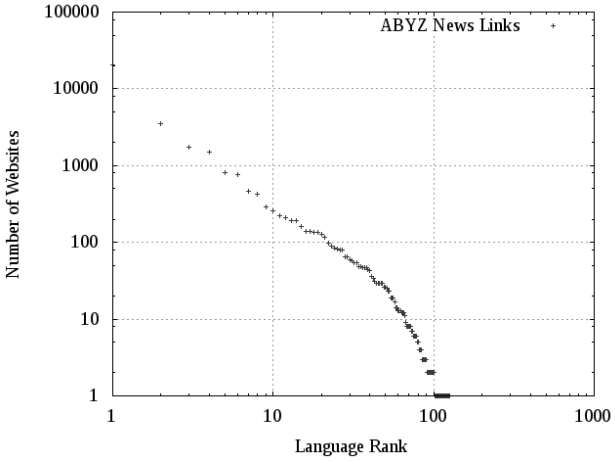

Figure 5: Number of domains for all languages of *ABYZ News Links* ordered by rank (log-log scale).

Rank	Domains	ISO 639-3	Language
1	20383	eng	English
2	3550	spa	Spanish
3	1722	por	Portuguese
4	1496	fra	French
5	800	deu	German
6	773	rus	Russian
7	470	ara	Arabic
8	428	zho	Chinese
9	291	nor	Norwegian
10	257	swe	Swedish
11	223	nld	Dutch
12	207	ita	Italian
13	194	ron	Romanian
14	191	fin	Finnish
15	160	dan	Danish
16	140	jpn	Japanese
17	137	srp	Serbian
18	134	ces	Czech
19	133	kor	Korean
20	128	tur	Turkish

Table 2: Top-20 languages of *ABYZ News Links* considering number of domains.

2.2 Crawling Generic Web Pages

Another possibility to collect text data is to crawl the World Wide Web randomly. We achieve this by different approaches.

2.2.1 FindLinks

FindLinks[7] (Heyer and Quasthoff, 2004) is a distributed Web crawler which uses a client-server architecture (see figure 6). The Java-based client runs on any standard PC and processes a list of URLs, which is distributed by the *FindLinks*-server.

Originally, its purpose was to analyze the structure of the World Wide Web by analyzing links found on the Web sites. Meanwhile, the client additionally collects the

5 http://www.abyznewslinks.com
6 http://www.httrack.com
7 http://wortschatz.uni-leipzig.de/findlinks/

analyzed Web pages and sends them to the server. The client is provided on the Web site of *Projekt Deutscher Wortschatz*. Everybody is encouraged to use it in order to help the project to collect more text data. Currently about 10 users download more than 30 million Web sites each day. So far about one Terabyte of raw text has been downloaded, and every day several Gigabytes are added. Until now no corpora have been created on the basis of this data, opening up the possibility to add a reasonable number of corpora in the near future.

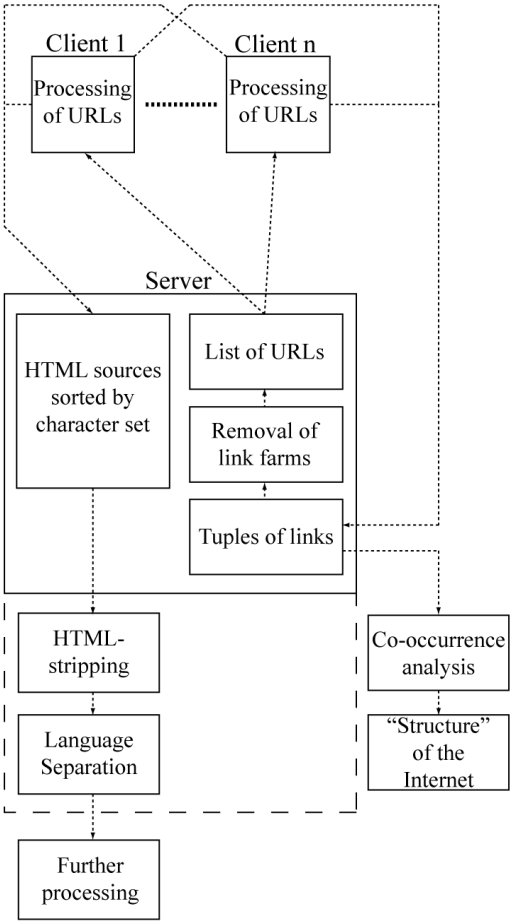

Figure 6: Schematic representation of the functionality of the distributed Web crawler *FindLinks*.

2.2.2 Standard Web Crawler

A list containing over 6.5 million internet domains has been created. By using the *HTTrack* based system complete domains are downloaded in parallel. Until now more than 80 Gigabytes of HTML-stripped text data have been collected.

2.3 Bootstrapping Corpora using Search Engines

By using an approach similar to Baroni (2004) and Sharoff (2006), frequent terms of a language are combined to form Google search queries and retrieve the resulting URLs as basis for the default download system. As a small set of frequent terms is needed for each language, the *Universal Declaration of Human Rights* (*UDHR*)[8] was used as a resource. It is available in more than 350 languages. For an average language, the *UDHR* contains about 2000 running words. Recently further languages were added by utilizing *Watchtower*[9] texts. Despite its partially controversial content, it offers contemporary documents in several hundred languages. Based on the lists of seeds tuples of three to five high frequent words are generated. These tuples are then used to query Bing and to collect the retrieved URLs. In a next step these Web sites are downloaded and further preprocessed.

2.4 Wikipedia

For more than 200 languages *Wikipedia* dumps were downloaded. *Wikipedia Preprocessor*[10] was used for further processing and text extraction resulting in 34 Gigabytes of text. Figure 7 depicts the number of articles in Wikipedia as a function of language rank. The top-20 languages can be found in table 3.

So far dictionaries based on *Wikipedia* were compiled for about 70 languages. Smaller sources were excluded from further processing. This is motivated by the fact that many *Wikipedias* start off by creating many stub articles using boilerplates, resulting in mostly near duplicate sentences. In order to utilize more of *Wikipedia's* languages, future work will include improved near duplicate detection.

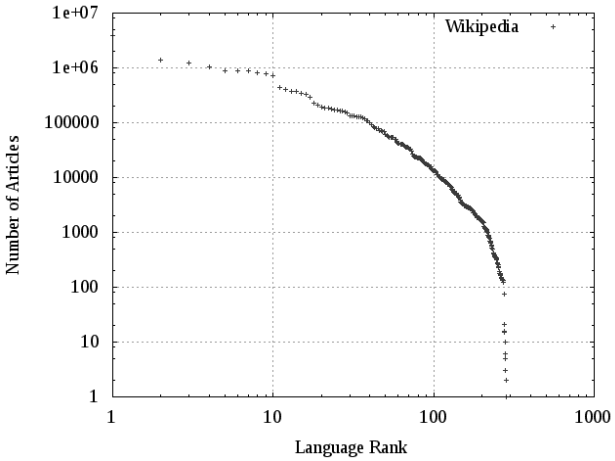

Figure 7: Number of articles for all languages of *Wikipedia* ordered by rank (log-log scale).

[8] http://www.ohchr.org
[9] http://www.watchtower.org
[10] http://sourceforge.net/projects/wikiprep/

Rank	Articles	ISO 639-3	Language
1	3885613	eng	English
2	1369865	deu	German
3	1220100	fra	French
4	1026998	nld	Dutch
5	896193	ita	Italian
6	882873	pol	Polish
7	871729	spa	Spanish
8	826909	rus	Russian
9	794673	jpn	Japanese
10	715395	por	Portuguese
11	438423	swe	Swedish
12	404454	cmn	Chinese
13	367382	ukr	Ukrainian
14	366643	cat	Catalan
15	346108	vie	Vietnamese
16	328737	nor	Norwegian
17	290049	fin	Finnish
18	223772	ces	Czech
19	211824	hun	Hungarian
20	191113	kor	Korean

Table 3: Number of articles of the top-20 languages of *Wikipedia*.

3. Processing and Cleaning of Text Data

Further steps to create dictionaries for multiple languages are HTML-stripping, language identification, sentence segmentation, cleaning, sentence scrambling, conversion into a text database and statistical evaluation.

Because of the repetition due to the number of different languages, an automatic tool chain has been implemented. It is easily configurable and only minor language-dependent adjustments, concerning e.g. abbreviations or sentence boundaries, have to be made.

Since complete evaluation by hand is not feasible, statistics-based quality assurance is necessary to achieve a satisfying quality of the resulting dictionaries (Eckart, 2012). With the help of features like character statistics, typical length distributions, typical character or n-gram distributions, or tests for conformity to well-known empirical language laws problems during corpora creation can be discovered and corrected.

3.1 HTML-Stripping

Except for database dumps of Wikipedia, all collected data are HTML-coded. HTML-Tags, Javascript and other elements have to be removed from the documents.

Therefore *Html2Text*, a HTML-stripping-tool of the NLP-group of the University of Leipzig, was utilized. The resulting text data still contains more than just well-formed sentences, so further cleaning steps follow.

3.2 Language Identification

For most documents, only their URL is known, but it is necessary to separate the texts into different monolingual corpora. Here we use *LangSepa* (Pollmächer, 2011), a tool built at the NLP group that identifies the language of a document. *LangSepa* compares the distribution of stop-words or character unigrams and character trigrams of various languages to the distribution within the document.

Hence, statistical information concerning these distributions for a high number of languages is needed. If available, corpora of the *LCC* were used to acquire these statistics. In order to classify further languages the *Universal Declaration of Human Rights* and *Watchtower* were utilized, adding up to about 450 languages.

3.3 Sentence Segmentation

The next step is sentence segmentation. Not all languages have the same sentence boundary marks. Especially for languages using other writing systems than the Latin alphabet, information about possible sentence boundaries is needed. We utilized Web sites like www.sonderzeichen.de to generate a list which is as complete as possible.

For better sentence boundary detection, abbreviation lists can be used. For several languages these lists already exist and more lists will be prepared. A possible source for such lists is *Wikipedia*.

3.4 Cleaning

The resulting elements are cleaned in two additional steps. First, non-sentences are identified based on patterns that a normal sentence should obey. All strings that do not comply with these patterns are removed.

In a second step the purity of the text collection is enhanced by eliminating sentences that do not belong to the considered language (Quasthoff, 2006b).

Subsequently, duplicate sentences are rejected. Especially newspaper reports may appear nearly unchanged on various Web sites. The same holds for standard boilerplates. To prevent a distortion of the statistical analysis, they are discarded.

3.5 Sentence Scrambling

To avoid copyright restrictions the collected sentences are "scrambled" by destroying the original structure of the documents to inhibit the reconstruction of the original material. With respect to German copyright legislation this approach is considered safe, because in Germany there is no copyright on single sentences.

3.6 Creation of Text Databases and Statistics

The data is stored in a relational database to allow quick and generic access. Using statistical methods a full form dictionary with frequency information for each word is generated and enhanced with information about co-occurrence analysis. Also corpora of standardized sizes are created to allow for better comparison between different languages or sources.

Once the text databases have been created, they are made accessible online. Alternatively, the data can be downloaded from http://corpora.informatik.uni-leipzig.de /download.html and viewed locally with the *Java Corpus-Browser* (Richter, 2006).

3.7 Quality Assurance

Quality assurance is a major challenge as hundreds of corpora of up to millions of sentences can hardly be evaluated by hand (Eckart, 2012). So it must be a goal to do as little manual evaluation as possible. Accordingly, we identified general features of natural language texts that are possible indicators for corpus quality and can be computed in an acceptable time span. Based on these, further post-processing steps can be triggered which delete or correct the data.
Possible features to be used include, for example:

- typical length distributions (word, sentence, paragraph)
- typical character or n-gram distributions
- accordance with empirical language laws, like Zipf's law
- character statistics (rare characters)

Examples are shown in figures 8 and 9. For two corpora, sentence length distributions (in characters) were measured. Figure 8 is based on Hindi newspapers and depicts a characteristic distribution, while figure 9, based on the Sundanese Wikipedia, is atypical. On closer examination both peaks of the distribution are based on boilerplates or near duplicates which should be removed. Although work-intensive, manual evaluation is one possibility to further enhance quality. To achieve this, a questionnaire for native speakers of languages of our collection is currently being developed.

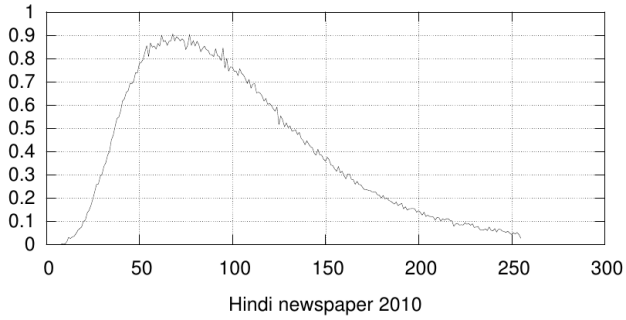

Figure 8: Sentence length distribution for Hindi newspapers (percentage for number of characters)

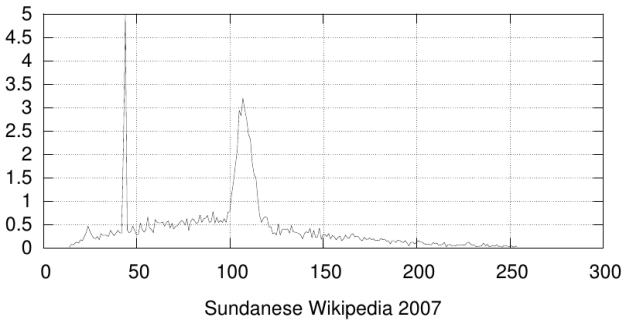

Figure 9: Sentence length distribution for Sundanese Wikipedia (percentage for number of characters)

4. Outlook

4.1 Further Work

Large amounts of text data have already been acquired in an ongoing process. The collected text data will be used to extend the collection both in the number of languages covered and in the size of resources provided. All corpora will be made freely accessible online in a uniform way.
There is still open work in fields like automatic and manual quality assurance.

4.2 Speed up of Crawling

So far the lists of URLs used for FindLinks and Web crawling contain all Top Level Domains (TLDs). Only the amount of URLs from TLDs of major countries (like .de or .com) is restricted to promote Web sites from smaller countries and hopefully rare languages. In the future this process could be adapted by favoring the TLDs of those countries where one or more languages with many speakers exist but no dictionary was created yet. A possible source of information concerning languages and the distribution of their speakers is *Ethnologue*[11].

5. Conclusion

In this paper, we have described acquisition and processing of standard-sized monolingual dictionaries of the *Leipzig Corpora Collection*. Various sources, tools for downloading and further data processing steps have been introduced.

6. References

Baroni, M.; Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

Biemann, C.; Heyer, G.; Quasthoff, U.; Richter, M. (2007). The Leipzig Corpora Collection - Monolingual corpora of standard size. Proceedings of Corpus Linguistic 2007, Birmingham, UK.

Eckart, T.; Quasthoff, U.; Goldhahn, D. (2012). Language Statistics-Based Quality Assurance for Large Corpora. Proceedings of Asia Pacific Corpus Linguistics Conference 2012, Auckland, New Zealand.

Heyer, G.; Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. Proceedings of IICS-04, Guadalajara, Mexico and Springer LNCS 3473.

Kilgarriff, A.; Grefenstette, G. (2001). Web as Corpus. In Proceedings of Corpus Linguistics 2001 Conference, Lancaster.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

Quasthoff, U. (1998). Projekt der deutsche Wortschatz. Heyer, G., Wolff, Ch. (eds.), Linguistik und neue Medien, Wiesbaden, pp. 93-99.

Quasthoff, U.; Richter, M.; Biemann, C. (2006a). Corpus Portal for Search in Monolingual Corpora. Proceedings of LREC-06.

[11] http://www.ethnologue.com

Quasthoff, U.; Biemann, C. (2006b). Measuring Monolinguality. Proceedings of LREC-06 workshop on Quality assurance and quality measurement for language and speech resources.

Richter, M.; Quasthoff, U.; Hallsteinsdóttir, E.; Biemann, C. (2006). Exploiting the Leipzig Corpora Collection. Proceedings of the Information Society Language Technologies Conference (IS-LTC), Ljubljana, Slovenia.

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Proc. WAC-3: Building and Exploring Web Corpora, Louvain-la-Neuve, Belgium.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, WaCky! Working papers on the Web as Corpus. Gedit, Bologna.