The Leipzig Corpora Collection: Monolingual Corpora of Standard Size

Chris Biemann,[1] Gerhard Heyer,[1]
Uwe Quasthoff[1] and Matthias Richter[1]

Abstract

We describe the Leipzig Corpora collection (LCC), a freely available resource for corpora and corpus statistics covering more than 20 languages at the time being. Unified format and easy accessibility encourage incorporation of the data into many projects and render the collection a useful resource especially in multilingual settings and for small languages. The preparation of monolingual corpora of standard sizes from different sources (web, newspaper, Wikipedia) is described in detail.

1. The Leipzig Corpora Collection

1.1 Purpose of the Collection

Open access to basic language resources is a crucial requirement for the development of language technology, especially for languages with few speakers and scarce resources. With our corpora, we aim at providing a data basis for the development and testing of (mainly language-independent) algorithms for various NLP applications, mainly to build language models from unlabeled data. For comparative language studies, corpora of standard size are ideal for measuring and systematically comparing non-linear corpus parameters such as vocabulary growth rates, large-scale distributions and other typological characteristics.

1.2 Corpus in German and standard size corpora for 15 languages

Collecting German wordlists and texts by the Natural Language Processing group at the University of Leipzig since the 1990s has lead to the production and publication of constantly growing corpora of German in 1998, 2000 and 2003, 2005 and 2007, available via our website[2]. The methods for corpus compiling, cleaning and processing have evolved since then, recent versions of these have been published in (Biemann *et al.*, 2004). (Quasthoff *et al.,* 2006) introduces an application of this language-independent technology and the notion of standard sized corpora for 15 languages, namely Catalan, Danish, Dutch, English, Estonian, Finnish, French, German, Italian, Japanese, Korean,

[1] Department of Natural Language Processing, Faculty of Mathematics and Computer Science, University of Leipzig
e-mail: biem@informatik.uni-leipzig.de, heyer@informatik.uni-leipzig.de, quasthoff@informatik.uni-leipzig.de, mrichter@informatik.uni-leipzig.de

[2] http://wortschatz.uni-leipzig.de

Norwegian, Sorbian, Swedish and Turkish. For the international version of the Website[3], see Table 3 in the appendix for a list of sizes and sources.

1.3 Comparable resources for 50+ languages

For a corpus project covering 50 or more languages, we now propose and implement the following guidelines. All text for different languages should

1. have comparable origin (for instance newspaper texts),
2. be processed in a similar way, and hence
3. offer equivalent possibilities for the application of statistical parameters.

The processing steps are described below in more detail.

The available electronic material for different languages varies in size. In contrast to that, many numeric features (like the number of significant word co-occurrences) depend on the size of the corpus in a non-linear way. Thus, for exact numerical language comparison and to detect these dependencies, corpora of similar size are required. Hence, we defined standard sizes with reference to a certain number of sentences. Measuring corpus size in number of sentences rather than in number of words is motivated by the amount of information: While isolating languages like English tend to exhibit sentences with more words than e.g. polysynthetic languages like Greenlandic (resulting in the fact that the average English sentence length is higher), we assume that by average the amount of information per sentence is comparable.

For each language, we produce corpora of fixed sizes up to the limit given by the availability of resources. These standard sizes are defined by 10,000, 30,000, 100,000, 300,000, 1 million, 3 million sentences and so on. The difference between size steps is a factor of roughly 3. This allows a comparison of parameters for different sizes for corpora of each language.

For comparison of different kinds of text, we collect three types of corpora for a language: Newspaper texts, randomly selected web text and Wikipedia articles. There are several reasons for collecting these three kinds of text separately: First, they differ in availability. Second, before one compares different languages using statistical parameters the different kinds of text in one language give a good indication of the variance of that parameter within one language. Moreover, corpora of various genres can be relevant for different applications such as terminology extraction. Also, quality and topic coverage of the material varies.

1.3 Release Plan for 2007

In the first half of 2007, a web corpus comprising 14 million Icelandic sentences has been launched[4]. The corpus, named Íslenskur Orðasjóður, was collected by the National and University Library of Iceland. For the second half of 2007, a number of corpora is due for release: Basque, Chinese, Hungarian[5], Russian, Mexican Spanish and a freely available alternative to LDC's English Gigaword corpus.

[3] http://corpora.informatik.uni-leipzig.de/

[4] http://wortschatz.uni-leipzig.de/ws_ice/

[5] based on the web corpus from http://mokk.bme.hu/resources/webcorpus, see (Halácsy *et al*., 2004)

2. Collecting Data

The process of corpus production uses only very limited language-specific knowledge. For collecting different kinds of text, different collection methods are employed. Later, these different kinds of text will not be merged into one corpus per language, but different corpora will be produced instead.

2.1 Crawling newspapers

Getting hand at newspaper texts can be done in several ways: One can:

1. ask the publishers to supply material,
2. use releases of newspaper collections from CD/DVD,
3. or crawl newspaper content from the web.

The latter approach allows the collection of large amounts of text with rather limited resources.

For obtaining large amounts of text in a specific language, stop word queries to news search engines can be used to cover virtually all material visible to the search engine. Alternatively, collections of RSS feeds[6] provided by newspapers are a veritable source. In our approach we combine both options.

The use of crawling for a research project raises legal and ethical questions. While it is clear that storing whole texts and allowing retrieval on them would be an unacceptable violation of copyright, search engines do in fact crawl the web, store the obtained data and allow searches on this data, including text snippets in their output. To avoid copyright restrictions, we partition the collected text into sentences and scramble these up in order to destroy the original and coherent structure that would be needed to reproduce the copyrighted material. With respect to the German *Urheberrecht*, an equivalent of copyright, this approach has been considered safe.

2.2 Using Wikipedia

The Wikipedia community aims at compiling encyclopaedias in all major languages of the world. As of now, Wikipedias in 253 languages have been started, with 88 of these containing more than 5.000 articles[7]. Recent research has already exploited the structured and semantic portions of Wikipedia in several ways (see e.g. (Milne *et al.* 2006) and (Gabrilovich and Markovitch, 2007)). We take advantage from this huge collection of (un)structured textual data. When collecting corpora we take only the plain text portion of the article namespace and exclude the user‘s private pages, discussions on articles and also all kinds of meta data. Of course, meta data could be extracted and used to enrich the results easily, but exceeds the scope of the current work.

Wikipedia‘s content can be downloaded safely as a whole in at least two forms. There are XML-dumps made for setting up a fully working Wikipedia mirror. These dumps, however, contain very complex Wiki markup and the only complete parser for this markup known so far is deeply integrated in the MediaWiki engine. So it seems more

[6] E.g. http://www.newsisfree.com
[7] http://meta.wikimedia.org/wiki/List_of_wikipedias (accessed: 30 July 2007)

feasible to start with the HTML dumps[8] and to extract the article content of all files that are not in a special namespace.

The compressed dump files for the April 2007 static versions of all Wikipedias are approximately 20 Gigabytes in size and the extracted plain text files are in the same order of magnitude. An overview for smaller languages is given in Table 4 in the appendix. For most Wikipedias, only a fraction of this amount is text in the language supposed to be actually covered. Starting with word lists for 26 already known languages from the Leipzig Corpus Collection and the Acquis Communautaire corpus version 2.2 (Steinberger *et al*. 2006) we clean sources from undesired content by language identification and extract word lists for a substantial number of the remaining languages. This is a very important step when trying to separate closely related languages such as Afrikaans and Dutch, Sicilian and Italian, Bokmål and Nynorsk. As a rule of thumb, derived from the ratios of already known languages, we can expect to obtain a pure language corpus sized between a quarter and half the number of sentences identified as "non foreign" in pass 1.

2.3 Crawling the web

The *Findlinks* project was started in 2003, see (Heyer and Quasthoff, 2004). The original purpose of the project was to discover the structure of the web and make this available as a web guide via the *Nextlinks* browser companion. Findlinks implements a distributed webcrawler in a client-server architecture. The client runs on standard PCs and utilizes a computer‘s spare bandwidth and processing resources. It is extensible by plug-ins to perform various tasks, among them language separation by specific trigrams and extending this text collection for specific or unknown languages. Even though most of the online material is in the major languages, a substantial amount of text gets retrieved by the crawler for less widespread languages. We encourage to download the crawler[9] and to take part in the collection of corpora.

2.4 Data Cleaning

While there are different character encodings for different languages, all data is converted to UTF-8. Before doing so, one has to identify the character set of the source. In the case of Wikipedia, we already have UTF-8. In all other cases we trust the character set entry in the corresponding HTML tag. If this character set entry turns out to be wrong, the corresponding text will be eliminated during the cleaning process.

- Sentence splitting. For sentence boundary detection we use
  - HTML tags for detecting the end of headlines and block level elements such as paragraphs,
  - punctuation marks,
  - special rules for numbers and dates, and
  - a general abbreviation list for the detection of non-boundaries. The problem of varying abbreviations for different languages will be dealt with by a forthcoming abbreviation detector, inspired by (Kiss and Strunk, 2006).

[8] available from http://static.wikipedia.org/

[9] http://wortschatz.uni-leipzig.de/nextlinks/index_en.html

- Word segmentation. For Chinese and Japanese, freely available word segmentation tools are applied. We use HLSegment[10] for Chinese and MeCab[11] for Japanese.
- Cleaning by foreign language identification. All corpora collected from the web contain undesired material. First, we want to remove foreign language sentences. For this we use a language identifier based on the most frequent 5000 words for each of the known languages. With the help of this list, we get a probability for the sentence to belong to a language. A sentence is assigned to the language of maximal probability, if the following conditions are fulfilled:
  - The result is reliable, i.e. the probability for the first language is above some threshold and the probability for the second language is much less than for the first language.
  - The sentence contains at least two words from the list of the chosen language.

  On average, for a corpus in a language other than English, about 10% or more of different language material can be anticipated.
- Pattern based cleaning. Due to the collection methods, the sentence splitter usually returns non-sentences having different sources. With pattern based methods, most of the non-sentences can be removed. Among the rules we apply, the ones listed in Table 1 with Icelandic examples are the most productive ones.
- Removal of duplicate sentences. Copies of sentences need to be removed because many texts are available in parts or as a whole from more than one URL.
- Random selection for corpora of standard sizes. In the last step each sentence is assigned a random number thus introducing a new order for all sentences of the whole corpus. From this randomly numbered corpus, the desired number of sentences is taken in this new ordering. This method ensures that a corpus of standard size includes all corpora of smaller standard sizes.

3. Data storage and access

3.1 Corpus Processing

The resulting sentences are processed with the *tinyCC* corpus production engine[12]. A full text index for words and their numeric position in sentences is built. The number of occurrences of each type is counted and two types of word co-occurrences are calculated with the log-likelihood ratio (Dunning, 1993): at sentence level (1% error threshold) and as immediate neighbours (5% error threshold).

3.2 Database structure

All data is produced in two formats, first a plain text format suitable for immediate access with the text editor of choice and the standard text oriented tools, then as a *MySQL* schema in cross platform binary compatible *MYISAM* format for access by database queries and with the corpus browser (see below). Both formats contain exactly the same data (except the table meta) listed in Table 2.

[10] http://www.hylanda.com/cgi-bin/download/count.asp?id=8&url=1

[11] http://mecab.sourceforge.net

[12] Available at http://wortschatz.uni-leipzig.de/~cbiemann/software/TinyCC2.html

Rule	Description	Examples	Hits
too many periods	unseparated sentences gluing words together or incomplete sentences ending with "…"	Upp í flugvél, burt úr kuldanum......	1,300,000
link artifacts or |	navigation boilerplates	Example: Forsíða > Túlkanir og þýðingar > Þýðingar Heim | Hafa samband | Veftré Leitarvél: Alþjóðahús Gagnlegar upplýsingar Algengar	220,000
begins with number dot blank	enumeration items	1. innkaup hlutu: Gláma/Kím arkitektar ehf., Laugavegi 164.	200,000
too many capital letters or digits in a row	headlines glued together with sentences or enumerations	LEIÐBEININGAR UM NOTKUN Gríptu um borðana og togaðu niður og í sundur. 7.3.2005 Tilkynning frá Högum hf. 7.3.2005 Verslunarrekstur Skeljungs komin til 10-11 25.10.2004 Tilkynning frá Högum hf. 22.6.2004 Tilkynning (...)	198,000
contains too many ":"s	Lists, e.g. of sports results	steini :: Comment :: 10 hugmyndir af bloggi.	166,000
too many {/&:}s	itemizations	Ferðaönd - Svara - Vitna í - Stelpið 31/10/05 - 0:25 Soffía frænka - Svara - Vitna í - aulinn 31/10/05 - 8:39 Kona í bleikum slopp með rúllur í hárinu.	153,000
expression too short	incomplete sentences	10. Valur ? _\åv,c ?	100,000
too many "_"s in a row	clozes	a) ________________, b) __________________ og c) __________________ Hvað myndast í kynhirslunum að lokum?	58,000

Table 1: Text cleaning rules used for dropping undesired sentences, their rationale and impact on an Icelandic corpus of 19,112,187 sentences, c.f. (Hallsteinsdóttir *et al*. 2007)

table name	fields	Content
meta	attribute, value	meta data about the corpus, needed by the corpus browser, only in the database version
words	w_id, word, freq	words and their frequency counts
sentences	s_id, sentence	sentences full text
sources	so_id, source	names of sources
inv_w	w_id, s_id, pos	positions of words in sentences
inv_so	s_id, so_id	index for sentences in sources
co_n	w1_id, w2_id, freq, sig	left word, right word, neighbour frequency and log-likelihood ratio
co_s	w1_id, w2_id, freq, sig	word1, word2, co-occurrence frequency and log-likelihood ratio

Table 2: Structure of the database: table names, their fields and functionality

3.3 Web-based access

The corpora released on the LCC-DVD version 1.0 can also be browsed via our portal[13]. For any word in the corpus, the following information is displayed:

- The word and its frequency
- Three sample sentences
- co-occurring words
- within the same sentence and
- as immediate left and right neighbour
- a co-occurrence graph displaying co-occurrences at sentence level

All information, as well as further data available only for some languages like synonyms or base form reduction, is also accessible as SOAP-based web services[14] for a seamless integration into customized applications.

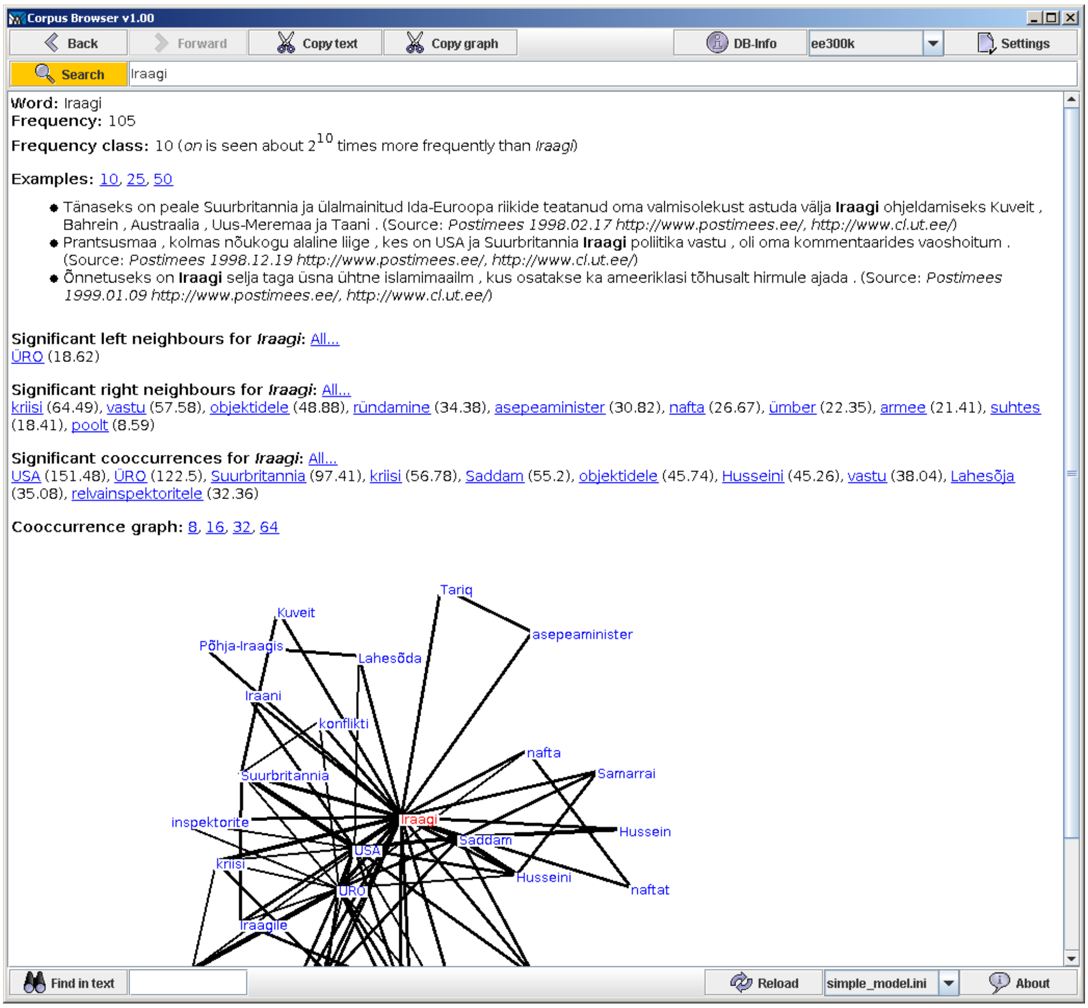
3.4 Using the Corpus Browser

There is a stand-alone corpus browser available for download. In the default configuration it shows all information as described in the previous section. But in contrast to the web interface, the browser can be tailored completely to the needs of a user. Both, the SQL statements for selecting the data to be shown, and the presentation style (for instance, one item per line or all items comma separated on one line) can be defined in a configuration file with a simple, XML-based language which is explained in the browser documentation[15]. This allows user-defined views on the database. As an example, the MySQL full text index can be used to turn the Corpus Browser into a search engine.

[13] http://corpora.informatik.uni-leipzig.de/

[14] List of web services at http://wortschatz.uni-leipzig.de/axis/servlet/ServiceOverviewServlet, ask for more

[15] http://corpora.uni-leipzig.de/download/LCCDoc.pdf



Word: Iraagi
Frequency: 105

Frequency class: 10 (*on* is seen about 2^{10} times more frequently than *Iraagi*)

Examples: 10, 25, 50

- Tänaseks on peale Suurbritannia ja ülalmainitud Ida-Euroopa riikide teatanud oma valmisolekust astuda välja **Iraagi** ohjeldamiseks Kuveit , Bahrein , Austraalia , Uus-Meremaa ja Taani . (Source: *Postimees 1998.02.17 http://www.postimees.ee/, http://www.cl.ut.ee/)*
- Prantsusmaa , kolmas nõukogu alaline liige , kes on USA ja Suurbritannia **Iraagi** poliitika vastu , oli oma kommentaarides vaoshoitum . (Source: *Postimees 1998.12.19 http://www.postimees.ee/, http://www.cl.ut.ee/)*
- Õnnetuseks on **Iraagi** selja taga üsna ühtne islamimaailm , kus osatakse ka ameeriklasi tõhusalt hirmule ajada . (Source: *Postimees 1999.01.09 http://www.postimees.ee/, http://www.cl.ut.ee/)*

Significant left neighbours for *Iraagi*: All...
ÜRO (18.62)

Significant right neighbours for *Iraagi*: All...
kriisi (64.49), vastu (57.58), objektidele (48.88), ründamine (34.38), asepeaminister (30.82), nafta (26.67), ümber (22.35), armee (21.41), suhtes (18.41), poolt (8.59)

Significant cooccurrences for *Iraagi*: All...
USA (151.48), ÜRO (122.5), Suurbritannia (97.41), kriisi (56.78), Saddam (55.2), objektidele (45.74), Husseini (45.26), vastu (38.04), Lahesõja (35.08), relvainspektoritele (32.36)

Cooccurrence graph: 8, 16, 32, 64

Figure 1: CorpusBrowser showing Iraagi (Iraq) in Estonian corpus ee300k.

3.5 Inserting and browsing customised data

Because of the loose coupling of the Corpus Browser with the underlying database by externally kept database queries, it is straightforward to modify the underlying database. Especially, if additional information is available at word or at sentence level, it is possible to include it in the presentation. The database structure given in Table 2 can be easily adopted to include more relevant information, for instance:

- second-order co-occurrence: Here, words are similar if they share many (first-order) co-occurrences
- sentence similarity: Sentences are similar if they share many content words.
- sentences with POS-tagging or chunking
- sentences with any other annotation like proper names, disambiguation *etc.*
- subject areas for words or sentences
- a thesaurus structure for words and data like WordNet

4. Sample language statistics

Figure 2 below illustrates the number of distinct word forms, neighbour-based and sentence-based word co-occurrences for different corpus sizes and different languages. The values for Finnish (bold) are shown in comparison to the average of 12 European languages (thin lines).

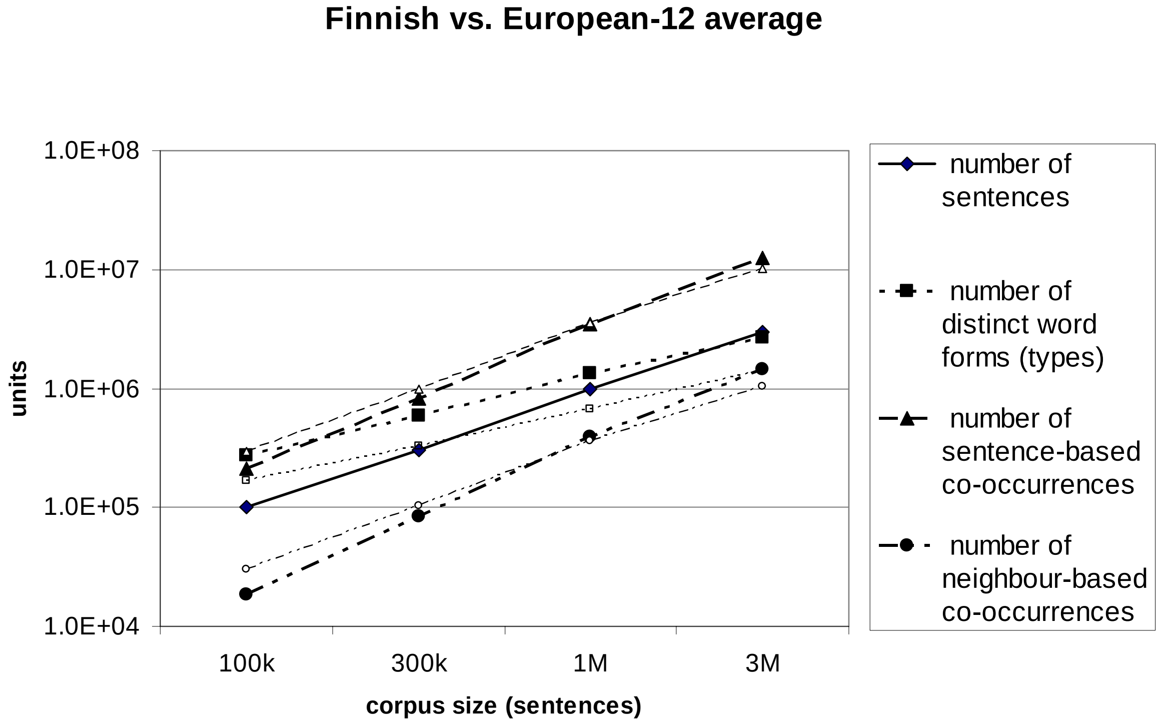


Figure 2: Comparative corpus statistics for Finnish and the mean of 12 European Languages

Different properties are clearly perceivable:

- The growth shown in Figure 2 is linear for all parameters in the log-log-plot. This means we have exponential growth for the actual parameters.
- We have nearly linear growth for the number of distinct word forms and co-occurrences compared to the corpus size measured in sentences.
- Both neighbour and sentence co-occurrences exhibit a slope close to 1. The slope for the number of distinct word forms is smaller.
- For different languages, these lines differ slightly by slope and by some constant. Different slopes in the log-log-plot correspond to exponential growth with different growth rates.

For Finnish we have:

- The number of word forms is slightly larger then average.
- The growth of the number of neighbour co-occurrences is slightly larger than average.

Leaving these facts unexplained in this current paper, the emphasis here is to show the usability of the corpora of standard size for language comparison.

5. Conclusions

In this paper, we have described the production process of monolingual corpora in standard sizes from various sources. Our service to the community is to provide these corpora in a cleaned and uniform way in various formats and various modes of access. Especially for languages with scarce resources, we provide an open-access basis on which any language technology can build upon. Further the majority of tools needed to build and maintain self-compiled collections have been made available. We constantly extend the collection both in the number of languages covered and in the size of resources provided.

References

Biemann, C., S. Bordag, G. Heyer, U. Quasthoff and C. Wolff (2004) Language independent Methods for Compiling Monolingual Lexical Data. In Proceedings of CicLING 2004, Springer LNCS 2945. Seoul, South Korea

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1)

Gabrilovich, E. and S. Markovitch (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of IJCAI 2007, Hyderabad, India. http://www.cs.technion.ac.il/~shaulm/papers/abstracts/Gabrilovich-2007-CSR.html

Halácsy, P., A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón (2004) Creating open language resources for Hungarian. In: Proceedings of the LREC 2004, Lisbon, Portugal

Hallsteinsdóttir, E., T. Eckart, C. Biemann, U. Quasthoff and M. Richter, M. (2007) Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In: Proceedings of NODALIDA-07, Tartu, Estonia

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

Kiss, T. and J. Strunk (2006) Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32(4).

Milne, D., O. Medelyan and I.H. Witten (2006) Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC

Quasthoff, U, M. Richter and C. Biemann (2006) Corpus Portal for Search in Monolingual Corpora. In: Proceedings of the LREC 2006, Genova, Italy

Steinberger R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş and D. Varga (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the LREC 2006, Genova, Italy

Appendix: Corpora and sizes

code	Language	Size	Source	Availability
cat	Catalan	10 million	WWW	LCC 1.0
dan	Danish	3 million	WWW	LCC 1.0
dut	Dutch	1 million	Newspaper	LCC 1.0
eng	English	10 million	Newspaper	LCC 1.0
est	Estonian	1 million	various	LCC 1.0
fin	Finnish	3 million	WWW	LCC 1.0
fre	French	3 million	Newspaper	LCC 1.0
ger	German	30 million	Newspaper	LCC 1.0
ger	German	30 million	WWW	in preparation
hun	Hungarian	10 million	WWW	in preparation
ice	Icelandic	1 million	Newspaper	online
ice	Icelandic	10 million	WWW	online
ita	Italian	3 million	Newspaper	LCC 1.0
jap	Japanese	0.3 million	WWW	LCC 1.0
kor	Korean	1 million	Newspaper	LCC 1.0
nor	Norwegian	3 million	WWW	LCC 1.0
ser	Serbian	1 million	various	in preparation
sor	Sorbian	0.3 million	various	LCC 1.0
spa	Spanish	1 million	Newspaper	online
swe	Swedish	3 million	WWW	LCC 1.0
tur	Turkish	1 million	WWW	LCC 1.0

Table 3: Leipzig Corpora Collection: Sources and maximum standard size

Language	lang.	#articles	#kb	#unique sentences	#non foreign sentences (pass 1)
Swedish	sv	235,231	314,120	3,111,124	2,997,385
Chinese	zh	131,442	354,212	2,339,583	2,211,215
Finnish	fi	119,908	219,540	2,542,700	2,471,782
Norwegian (Bokmål)	no	116,093	192,520	2,052,158	1,966,768
Esperanto	eo	85,394	124,792	1,159,373	1,088,885
Turkish	tr	83,154	159,844	1,078,935	1,052,695
Slovak	sk	71,314	94,612	1,128,462	1,078,462
Czech	cs	70,130	161,628	1,729,946	1,628,828
Romanian	ro	67,157	101,652	813,742	692,679
Catalan	ca	65,701	109,296	1,312,394	1,288,733
Danish	da	64,558	99,944	997,886	949,555
Ukrainian	uk	63,434	85,884	1,023,615	1,016,767
Hungarian	hu	62,548	159,752	1,593,033	1,552,856
Indonesian	id	62,387	83,644	896,062	828,777
Hebrew	he	59,324	222,360	1,219,772	1,205,459
Lombard	lmo	51,296	12,540	116,667	100,791
Slovenian	sl	49,132	79,996	905,354	882,549
Lithuanian	lt	47,776	67,604	717,234	708,970
Serbian	sr	46,212	101,552	1,009,209	984,328
Bulgarian	bg	40,764	83,964	811,975	802,502
Korean	ko	38,389	68,228	529,777	518,685
Estonian	et	36,410	53,464	616,565	606,932
Cebuano	ceb	33,210	9,900	172,440	109,536
Arabic	ar	32,918	63,180	442,514	437,496
Croatian	hr	31,861	66,592	782,635	497,777
Telugu	te	28,015	14,328	128,896	118,033
Galician	gl	24,915	43,256	472,111	264,437
Greek	el	24,306	54,896	536,541	523,973
Thai	th	24,143	56,712	436,306	423,762
Norwegian (Nynorsk)	nn	23,587	40,552	375,659	170,890
Persian	fa	21,927	44,344	367,548	364,570
Malay	ms	21,483	33,956	479,084	439,627
Newar / Nepal Bhasa	new	21,410	7,660	50,894	45,165
Vietnamese	vi	20,123	66,572	674,386	631,312
Bosnian	bs	18,832	29,256	320,325	201,710
Basque	eu	18,388	24,072	213,139	206,289
Bishnupriya Manipuri	bpy	17,612	10,000	75,661	73,507
Volapük	vo	16,997	3,108	14,376	13,427
Simple English	simple	16,718	28,820	285,761	283,395
Albanian	sq	16,492	20,216	163,534	151,445
Icelandic	is	15,968	24,912	198,154	175,996

Bengali	bn	15,835	18,384	97,354	90,770
Luxembourgish	lb	15,,710	24,040	267,267	238,215
Georgian	ka	15,428	24,072	116,738	114,986
Ido	io	15,069	13,352	177,660	152,494
Breton	br	14,274	17,936	181,495	159,640
Latin	la	13,484	20,440	143,615	130,462
Neapolitan	nap	12,514	12,024	55,953	49,187
Hindi	hi	11,824	10,320	55,394	52,435
Serbo-Croatian	sh	11,411	24,580	323,581	190,526
Tamil	ta	10,871	17,860	115,449	110,638
Sundanese	su	10,673	11,080	97,407	73,958
Marathi	mr	10,254	8,992	49,300	47,997
Javanese	jv	10,228	5,824	52,846	50,907
Macedonian	mk	9,947	18,212	155,081	151,652
Welsh	cy	9,939	12,752	110,134	102,272
Sicilian	scn	9,924	9,896	78,536	68,014
Latvian	lv	9,745	19,644	183,617	179,610
Low Saxon	nds	9,597	11,824	166,022	134,918
Kurdish	ku	9,371	9,612	89,189	69,470
Walloon	wa	9,053	8,688	57,151	44,757
Asturian	ast	8,517	12,420	195,382	173,789
Piedmontese	pms	8,425	4,904	32,990	28,640
Occitan	oc	8,255	14,892	97,849	74,286
Afrikaans	af	7,714	15,084	150,299	78,308
Tajik	tg	7,680	7,288	45,077	39,868
Siberian/North Russian	ru-sib	7,205	4,328	48,417	47,651
Haitian	ht	7,053	3,640	43,587	39,246
Azeri	az	6,907	7,596	47,933	43,629
Ripuarian	ksh	6,804	7,932	39,655	33,471
Tagalog	tl	6,148	9,500	105,707	86,344
Aragonese	an	6,135	8,844	172,556	163,901
Chuvash	cv	5,876	5,220	42,448	42,054
Urdu	ur	5,869	10,132	54,659	53,739
Uzbek	uz	5,542	7,328	75,908	72,855
Corsican	co	5,408	4,300	23,333	19,486
Belarusian	be	5,309	3,068	20,927	20,756
Irish Gaelic	ga	5,141	8,876	72,605	65,464

Table 4: Wikipedias with more than 5,000 articles: size in articles, compressed kilobytes, number of unique sentences and upper bound for number of candidates for inclusion in a corpus. The Top 10 clearly exceed 1 million usable sentences and are omitted here.