High Quality Word Lists as a Resource for Multiple Purposes

Uwe Quasthoff[1], Dirk Goldhahn[1], Thomas Eckart[1],
Erla Hallsteinsdóttir[2], Sabine Fiedler[3]

[1]Natural Language Processing Group, Computer Science Department, University Leipzig, Germany
[2]Department of Language and Communication, University of Southern Denmark, Odense, Denmark
[3]Department of English, University of Leipzig, Germany
E-mail: {quasthoff, dgoldhahn, teckart}@informatik.uni-leipzig.de, erla@sdu.dk, sfiedler@uni-leipzig.de

Abstract

Since 2011 the comprehensive, electronically available sources of the Leipzig Corpora Collection have been used consistently for the compilation of high quality word lists. The underlying corpora include newspaper texts, Wikipedia articles and other randomly collected Web texts. For many of the languages featured in this collection, it is the first comprehensive compilation to use a large-scale empirical base. The word lists have been used to compile dictionaries with comparable frequency data in the Frequency Dictionaries series. This includes frequency data of up to 1,000,000 word forms presented in alphabetical order. This article provides an introductory description of the data and the methodological approach used. In addition, language-specific statistical information is provided with regard to letters, word structure and structural changes. Such high quality word lists also provide the opportunity to explore comparative linguistic topics and such monolingual issues as studies of word formation and frequency-based examinations of lexical areas for use in dictionaries or language teaching. The results presented here can provide initial suggestions for subsequent work in several areas of research.

Keywords: word lists, corpora, frequency data

1. Introduction – the *Leipzig Corpora Collection*

Word lists are an important product of corpora and have been used in quantitative linguistics for a long time (for example as a support to stylometry and authorship attribution). But for most of the quantitative results, the underlying word lists are not made available, even though the distribution of word lists is less restricted according to copyright than the distribution of corpora. In the following, the production and possible usage of wordlists is described. The *Leipzig Corpora Collection*[1] *(LCC)* provides corpora in more than 230 languages and different genres: newspaper texts, random web texts and Wikipedia articles (Goldhahn et al., 2012). All corpora are segmented in single sentences and all data are provided for download with sentences in random order. Their use is granted free of charge for all non-commercial, personal and scientific purposes and new corpora are added on a yearly basis. All corpora are identified by a strict naming convention that includes the language in the form of the three letter code according to ISO 639-3, the genre and the production year.

2. Generation of the word lists

The word lists together with their frequencies is generated from the disjoint union of multiple corpora in the given language. Moreover, a word pattern-based definition is used to remove non-words.

2.1 Tokenization

The segmentation of texts into words is crucial for any subsequent analysis dealing with words. In a first step, the text is segmented into a sequence of possible words and some non-words, which will be removed later.

The text is broken up into elements using white space (such as blanks or line breaks). The resulting parts are usually words with possible additional punctuation marks on the left or right. Only in the following cases can they be confused with parts of words:

- A period occurs at the end of the word in abbreviations, initials or ordinal numbers. Such an abbreviation is recognized by regular expressions if
  a) it consists only of uppercase letters or
  b) it contains additional inner periods.

  For other abbreviations ending in a period, an abbreviation list is necessary. If no abbreviation list is available for a certain language, a general abbreviation list containing the most frequent abbreviations of some large languages is used.
- A word contains an apostrophe. Whether a word-internal apostrophe should be considered a white space, may depend on the language or even on individual words. In French, for example, the forms *aujourd'hui* and *quelqu'un* are to be considered words, whereas it might be useful to treat the articles *l'* and *le'* as separate words.
- In some languages, apostrophes are allowed at the beginning or the end of a word. They can occur in a very limited number of special words (for example, *'n, 't* in Dutch) and are then treated by means of an exception list. In the case of ordinary words, apostrophes should not be removed from words.

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

2.2 Combining different corpora

In preparation of a Frequency Dictionary the corpora to be considered are selected. There is a standardized quality check resulting in a technical report (Quasthoff et al., 2013) describing the quality of the different corpora for one language. Usually the quality is similar, but older corpora can have some shortcomings (like character set problems, failures in extraction and segmentation procedures) which cannot be repaired without the original data. Hence, some corpora might be excluded. All the selected corpora are aggregated to a so-called mixed corpus. As the LCC corpora are based on different text acquisition strategies, corpora of the same language are not necessarily disjointed. Due to de-duplication and additional cleaning mixed corpora are also slightly smaller than the sum of their parts.

2.3 Word lists

The frequency-ordered word lists of the corpora can be considered as raw word lists. They are produced and their quality is checked by automated means (Quasthoff et al., 2011). Those lists also contain non-words of different kinds, but the quality of the most frequent words in a list increases with the size of the corpus. Those lists have standalone uses for some tasks like language identification, see section 3.6 below.

2.4 Word definition

In the preceding, the term ‘word’ has been used to refer to character strings that occur in the corpus between blank spaces and possible additional punctuation marks. However, not all such strings can be considered to be words and therefore, we apply the following additional restrictions: A word can include the ordinary letters of the corresponding alphabet. Moreover, a word can include numerals, apostrophes, hyphens and full stops, provided that

- full stops only occur in abbreviations,
- hyphens only occur word-internally,
- numerals do not occur word-initially and a word contains a maximum of two numerals.

For some languages like English or Indonesian which have compounds consisting of multiwords, the following rule is applied: Such a compound A B (consisting of two words A and B separated by a blank) is considered as a word if we either find the word A-B (with a hyphen instead of the blank) or the word AB (continuously written) with reasonable frequency.

2.5 Maximal word list size depending on corpus size

For a longer word list, only the high-frequent words (here: top-10,000) can be checked manually. For the remaining word list, pattern-based criteria apply. Usually the quality of a word list drops with frequency. But even the low-frequent entries contain many correct words. Hence, there is a trade-off between length of the list and quality in the low-frequent range. At present there is no automatic quality testing for different frequency ranges of a word list, but usually minimum frequencies of 5 or 10 are considered to be appropriate.

The following table gives the frequencies at different positions in the word list for different corpus sizes. The numbers are averages for different languages and genres chosen from the LCC.

Corpus size in sentences	***freq@ 10K***	***freq@ 30K***	***freq@ 100K***	***freq@ 300K***	***freq@1 M***
10K	2	0-1	0	0	0
30K	**4-6**	1-2	0	0	0
100K	**7-20**	**1-5**	0-1	0	0
300K	30-45	**7-12**	1-3	0-1	0
1M		29-36	**4-8**	0-1	0
3M			**12-30**	2-4	0-1
10M				**5-25**	1-3
30M				15-45	**2-6**
100M					**10-25**

Table 1: Word frequencies for different positions in the wordlists.

3. Applications

3.1 Frequency Dictionaries

The book series *Frequency Dictionaries* (Quasthoff et al., 2011) is based on the word lists described above. These printed books contain some statistical information about the words of the corresponding language, the most frequent 1,000 words ordered by frequency, and an alphabetically ordered 10,000 word list with frequency information. The lists have been checked carefully by hand to identify and mark incorrect forms or misspelled words and an accompanying CD-ROM contains the same information for a word list up to 1,000,000 words. The actual size of the word list depends on the corpus size as described above.

3.2 Word lists

Linguo-statistical results using the word lists throw light on various aspects of language, see (Köhler, 2008). They do not just show strong regularities in a particular language but also provide data for different languages thus facilitating language comparisons. These comparison criteria include the following:

- the alphabet and its letter frequencies,
- word length distribution,
- word structure, e.g. relationship between vowels and consonants, number and length of syllables,
- vocabulary range measured by text coverage,
- dependency of some features from the rank in the word list, using the example of frequency and word length.

As an example, the examination of *word length* can reveal a number of interesting results. Figure 1 and table 2 provide the average word length within the most frequent N words for an Icelandic word list. Using logarithmic scaling for the frequencies, the increase in word length appears almost linear. In this case the slope of the line of best fit in the frequency range between 100 and 100,000 equals 1.64. Due to the logarithmic scaling of the N-axis, this means that the average word length is increasing by 1.64 characters, when the number of known words increases by a factor of 10.

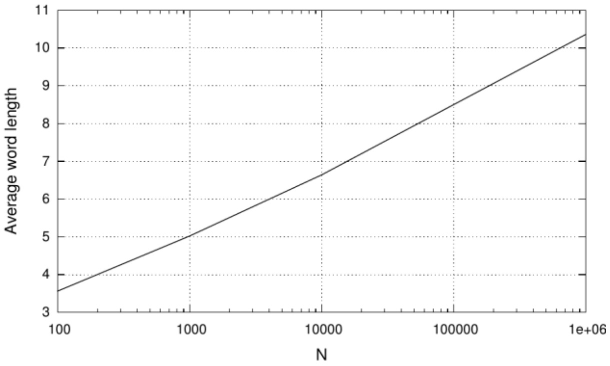

Figure 1: Average word length in various frequency ranges for an Icelandic word list.

N	*Word length*
100	3.56
1,000	5.03
10,000	6.64
100,000	8.50
1,000,000	10.36

Table 2: Average word length in various frequency ranges for an Icelandic word list.

3.3 Language comparison

The detection of similar languages is a concern across different scientific disciplines. Language typology is interested in general similarity or, in a more restricted sense, in the similarity between languages according to certain properties (Greenberg, 1963).

Using the most frequent words, automatic measurements of a distance between languages can be conducted. Furthermore, the most frequent character n-grams, which are also possible features of language comparison, are computable based on word lists.

The underlying idea is that similar languages share common vocabulary. By examining n-grams mutual constituents such as affixes can also be considered.

Utilizing measures such as

- Kendall tau rank correlation (Kendall, 1938) with an extension for lists with unequal sets of elements (Goldhahn, 2013),
- Cosine similarity (Singhal, 2001) or
- Dice coefficient (Dice, 1945), the number of common elements,

similarity values for word or n-gram lists of language pairs are determined.

When evaluating these values against known language classifications such as genealogical relationships, a high agreement can be achieved (Goldhahn, 2013) as can be seen in figure 2. In doing so, the use of n-grams is beneficial compared to words, concerning overall quality and properties such as independence of subject area. Weighting of list elements according to rank or frequency can also have positive effects on the results, since it reduces the influence of random matches. When enhancing this approach with transliteration, script boundaries can be overcome, resulting in a hierarchical similarity classification close to language genealogy.

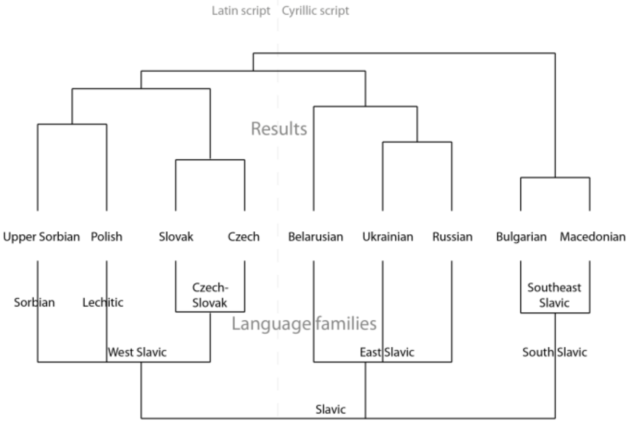

Figure 2: Results of hierarchical clustering of language similarities for Slavic languages based on the comparison of transliterated lists of character trigrams (in contrast to genealogical relationships)

3.4 Corpus comparison

While language dependent parameters are expected to vary for different languages, their behaviour for different genres within one language is difficult to predict. Measured intra-language variation can help to decide whether differences between languages can be considered as significant (Eckart & Quasthoff, 2013; Goldhahn, 2013).

Furthermore based on statistical analysis corpus comparison can be used to enhance quality of linguistic resources. Typical correlations or distributions can be analyzed and problematic resources identified by finding statistical anomalies (for example, based on word or sentence length, distribution of letter frequencies, etc.). Extreme points can therefore be seen as hints for problems in the corpus generation process, including used text acquisition procedures (Eckart et al., 2012).

3.5 Quantitative and qualitative word formation studies

As an example, in the following, the relevance of large corpora in linguistic research is to be illustrated by an ongoing study of word formations in Danish, Icelandic and German that is based on the *Leipzig Corpora Collection* and the word lists of the corresponding frequency dictionaries. The main goals of the study are the completion of a linguistic description of word formation principles, a description of the quantitative realization of those principles in large corpora (by analyzing the word formations on the top-10,000 list in the frequency dictionaries) and an interlingual comparison. The scope of the comparison of word formation principles in German, Danish and Icelandic and their realization in corpora is to provide an overview of differences and similarities in word formation in the three languages. In all three languages, for instance, word compounding using hyphens is allowed. But the frequencies of usage for this word formation pattern are very different.

Language	***Words in top 10,000***	***Words in top 1,000,000***	***Sample words from top-10,000***
German	32	96,199	*E-Mail, US-Dollar, Baden-Württemberg, S-Bahn, Karl-Heinz, rot-grüne, CD-ROM*
Danish	12	62,548	*e-mail,* *Jyllands-Posten,* *E-mail,* *Rosenkrantz-Theil,* *Lolland-Falster,* *EU-lande*
Icelandic	2	8,777	*e-ð, KR-ingar*

Table 3: Frequency of word compounds using hyphens in different languages.

3.6 Resource for the language industry

Some applications explicitly require frequency information for full forms, not for lemmas. In all the cases below the knowledge of low-frequent words is helpful. Such applications are

- Language identification,
- Spell checkers, OCR,
- Speech to text.

In these cases, the unknown words have to be compared with similar but known words. Large high-quality word lists facilitate the finding of replacements for more infrequent words.

4. Conclusion

Word lists are an excellent example of the utilization of corpora. The lists can be processed and exchanged without violating the copyright of the underlying texts and can therefore be made available upon request. The lists are great resources for statistical analysis, language comparison, quantitative and qualitative linguistic research as well as software applications.

5. References

Dice, L. R. (1945). *Measures of the Amount of Ecologic Association Between Species. Ecology* 26 (3): 297–302.

Eckart T. and Quasthoff, U.: *Statistical Corpus and Language Comparison on Comparable Corpora*. In: *BUCC – Building and Using Comparable Corpora*, Springer, 2013.

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

Goldhahn D., Eckart, T. and Quasthoff, U.: *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages*. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

Goldhahn, D. (2013). *Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken*. Dissertation. Leipzig: University of Leipzig.

Greenberg, J. H. (1963). *Some universals of grammar with particular reference to the order of meaningful elements. Universals of language*, 2, 73-113.

Kendall, M. G. (1938). *A new measure of rank correlation*. *Biometrika*, 30(1/2), 81-93.

Köhler, R., Altmann, G., Piotrowski, R. (Eds.): *Quantitative Linguistics, An International Handbook*, Vol. 27. Berlin: de Gruyter, 2008

Quasthoff, U., Fiedler, S., and Hallsteinsdóttir, E. (eds.): *Frequency Dictionary German*. Leipziger Universitätsverlag, 2011.

Quasthoff, U., Goldhahn, D., and Heyer, G. (2013). *Technical report series on corpus building (Vol. 1)*. Leipzig: University of Leipzig, http://asvdoku.informatik.uni-leipzig.de/corpora/index.php?id=references.

Singhal, A. (2001). *Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.