Corpus Portal for Search in Monolingual Corpora

Uwe Quasthoff*, Matthias Richter*, Christian Biemann*

*Leipzig University, Computer Science Department
Natural Language Processing Group
Augustusplatz 11, 04109 Leipzig, Germany
{quasthoff,mrichter,cbiemann}@informatik.uni-leipzig.de

Abstract

A simple and flexible schema for storing and presenting monolingual language resources is proposed. In this format, data for 18 different languages is already available in various sizes. The data is provided free of charge for online use and download. The main target is to ease the application of algorithms for monolingual and interlingual studies.

1. Introduction

Corpora are important linguistic resources. Up to now there is no portal providing free and uniform access to large corpora of many different languages. At `http://corpora.informatik.uni-leipzig.de/` we use the approach of the Projekt Deutscher Wortschatz at `http://wortschatz.uni-leipzig.de/`, which is described in (Biemann et al., 2004a), and present similar data for 18 languages. The main benefits using this methodology are:

- uniform format and web interface for all corpora
- comparable data sets for different languages
- corpora in several pre-defined sizes
- statistical information including co-occurrence data
- standardized visualization of co-occurrence data
- access to a collection of linguistic resources free of charge
- seamless integration into applications via SOAP-based Web Services

2. Goals of the Project

An aim of our project is to provide access to data and statistics on a number of different languages available in a unified format and in standard sizes. Further, we want to provide basic linguistic services free of charge for anyone who has a use for them, without having to sign agreements, paying shipping fees and alike.

Of course, free corpora as opposed to high-quality expensive resources may not fulfill all requirements in text quality and balancing and cannot provide manually added metadata or large-scale annotation. Nevertheless, as discussed in detail e.g. in (Bordag et al., 2005) they are sufficient for a number of lexical acquisition and other NLP tasks such as extraction of knowledge, automatic calculation of semantic associations and collocations as well as word sense induction. Unlabelled data can greatly improve learning tasks in general, e.g. using co-training (Blum and Mitchell, 1998).

Possible usage of corpora as a resource includes, but is not limited to:

- preparing monolingual dictionaries
- researching linguistic questions
- comparing different languages on a statistical basis
- parameterizing language models e.g. for speech recognition
- expanding queries with statistically similar words
- extracting significant terms from documents by comparison against a reference corpus
- selecting balanced words sets for experiments e.g. in psycholinguistics

3. Languages and Corpora

Corpora in the languages listed in Table 1 are collected from the web and consist either of newspaper texts or of randomly collected web pages. The maximum sizes of the corpora offered are restricted by present availability, rather than being arbitrarily chosen. Our notion of corpus is centered around the sentence as the biggest unit. This is sufficient for a vast variety of statistical NLP applications and helps us to avoid copyright problems.

3.1. Preprocessing

The pre-processing of the data consists of the following steps:

- HTML-Stripping: We do not only need to strip HTML-Tags but also to remove embedded meta data, scripts, stylesheets, objects, tables, comments and so on.
- Sentence separation: We accept any sentence boundary marked in the HTML code. For the rest of the text it is important not to split sentences after periods which do not denote a full stop. If available, a list of common abbreviations is useful.
- Removal of foreign language sentences: As we can see in Table 4, a quite high text coverage can be achieved with the 10 000 most frequent words. We use these words for sentence-based language identification, see e.g. (Dunning, 1994) for an overview.

	language	size	source
cat	Catalan	10 million	WWW
dan	Danish	3 million	WWW
dut	Dutch	1 million	Newspaper
eng	English	10 million	Newspaper
est	Estonian	1 million	various
fin	Finnish	3 million	WWW
fre	French	3 million	Newspaper
ger	German	30 million	Newspaper
ice	Icelandic	1 million	Newspaper
ita	Italian	3 million	Newspaper
jap	Japanese	0.3 million	WWW
kor	Korean	1 million	Newspaper
nor	Norwegian	3 million	WWW
ser	Serbian	1 million	various
sor	Sorbian	0.3 million	various
spa	Spanish	1 million	Newspaper
swe	Swedish	3 million	WWW
tur	Turkish	1 million	WWW

Table 1: languages, maximum size in sentences and sources of the corpora.

- Removal of ill-formed sentences: Parts of the data collected are not sentences but garbled text. In order to obtain a quality language resource the sentences included must meet certain conditions. These are explained in detail in Section 3.2..
- Removal of duplicates and near-duplicates: When doing co-occurrence statistics, duplicate and near-duplicate sentences have a large impact on the results, especially among the less frequent types occurring in them. Therefore we discard sentences that are very similar, e.g. only differing in a number contained or the form of quotation marks.
- Reduction of the corpora to pre-defined sizes: As a last step, the sentences are randomly mixed and spare sentences are dropped. Hence, it is impossible to reconstruct original documents in the resulting corpus. Legal issues of corpora collected from the web are settled by this.

3.2. Data Cleaning

Automated collection of text cannot avoid unwanted junk in the raw data. When building very large corpora with very small staff it is not an option to proof-read all the material. Instead we apply a heuristic means of separating good from bad sentences. It follows a list of formal quality requirements that each sentence in the corpus must meet:

- Sentences must begin and end in an appropriate way. In many languages this means that they begin with a number or a word in capital letters and end with a punctuation mark. Optionally there may be additional quotation marks at the begin or end of the sentence.
- In typesetting, emphasizing by using t r a c k i n g destroys words when implemented by inserting blanks instead of increasing the letter spacing. To cover this, sentences must not contain more than six single letters separated by blanks.
- If a sentence contains too many commas it is very likely to be a list, such as a soccer team. Sentences with more than nine commas are omitted for that reason. A similar restriction is applied to the number of periods (5).
- Sentences containing too many blanks compared to the length are very likely to be long sports result lists. We found a ratio of 30% and more blanks to be a good threshold for eliminating these unwanted lists.
- Sentences which contain the characters and sequences $\langle\langle$, ++, *, ˜, |, [[, & and / in abundance are likely to be code fragments and are therefore not included.
- A sign of non standard language is the presence of multiple punctuation marks such as terminating !!! or ??? in the sentence. We do not include such sentences in our standard language corpora.
- Sentences which contain very long sequences of numbers (more than 15) or capital letters (more than 20) are likely to be not interesting.

3.3. Dictionary Data

For each language, there is a full form dictionary. For any word, it provides frequency information. Further we provide co-occurrence statistics: words that co-occur significantly often with the given word. For the calculation of the significance, the log-likelihood measure (cf. (Dunning, 1993)) is used as described in (Biemann et al., 2004a). Two kinds of co-occurrences are pre-computed: Words occurring together in sentences and words found as immediate (left or right) neighbours. Co-occurrence data is meant to be used extensively as a building block for further applications. Additional data is included if available. At the moment, only the German dictionary contains grammatical information such as inflection and semantic information such as subject areas and synonyms. The open and flexible architecture, however, can easily be augmented with all kinds of additional data such as grammar, links and annotation.

4. Technical Issues

4.1. Format

The corpora are stored in a uniform schema in a MySQL database. The functionality of a database entails efficient indexing methods and allows the storage of very large resources without being limited by RAM size, as well as remote access. The choice of MySQL as a specific system was made due to its free availability and great overall performance.

As we store a lot of information about words in different tables, word numbers are used as keys for unique identification. Algorithms operating on word numbers, instead of strings, reach faster processing speed due to using a fixed length and more compact representation. In the following, the essential tables of the Wortschatz corpus format are described. We omitted several tables and some columns for the sake of clarity.

In essence, a corpus schema consists of the 5 database tables that are described in Table 2. Other tables and fields are used to include grammatical and semantic information, store base forms, parts of speech, etc.
At the time being there do not exist any conversion utilities for converting corpora from other formats into the schema described here. However, given a specific format, it is an easy task to write such a conversion and import script and subsequently let the underlying software perform the calculation of the statistical data.

4.2. Availability

All corpora are accessible from the website `http://corpora.informatik.uni-leipzig.de`. Additionally, all information that is available via the web interface can also be integrated into arbitrary applications using SOAP-based Web Services as described in (Biemann et al., 2004c). Demo clients implemented in Java can be downloaded at `http://wortschatz.uni-leipzig.de/Webservices/`. Smaller versions of the web-accessible corpora are available for download as plain text and as MySQL tables in standard sizes from 100.000 sentences on. A free platform independent Java application for accessing the corpora with a GUI is also provided. Within this framework, interlingual studies and statistics with comparable results drawing on comparable data sets can be conducted very easily.

5. Example Statistics

Monolingual statistics is a field of research that has got a long tradition (see, for example, (Meier, 1967)). From the plethora of possible statistics we picked a few basic ones and illustrate the power of co-occurrence data.

5.1. Basic statistics

Some basic characteristics are presented in tables 3 and 4:

- number of types (nty)
- number of tokens (nto)
- average type length (tyl): word length from each type in the corpus divided by the number of types
- average token length (tol): word length from each token in the corpus divided by the number of tokens
- coverage of text: given a text, the most frequent 10 / 100 / 1 000 / 10 000 types make up a certain percentage of this text

All data is obtained from a 100 000 sentence corpus of the respective language.
Another basic statistics is testing Zipf's law (c.f. (Zipf, 1935)) on ranked numerical data. The law states that the numerical value observed (for example the frequency of a word or the frequency of a co-occurrence) multiplied with the rank induced by the numerical order will result in an approximately constant product. In other words: The data will be arranged along a line in a log-log scaled plot. The Zipfian distribution is omnipresent in language, see e.g. (Sigur et al., 2004). Figure 1 suggests that the law is also fulfilled for co-occurrences ranked by their significance.

	nty	nto	tyl	tol
cat	110 034	2 178 029	8.04	4.57
dan	157 560	1 623 436	10.28	5.27
dut	124 986	1 588 453	9.94	5.27
est	191 225	1 401 652	10.37	6.58
fin	266 633	1 206 771	11.80	7.94
fre	101 782	2 352 542	8.54	5.03
ger	183 567	1 816 287	11.78	5.47
ice	155 903	1 787 209	9.84	5.16
ita	105 139	1 842 639	8.81	5.28
nor	165 090	1 551 530	10.26	5.25
ser	116 248	1 285 161	8.25	4.61
sor	170 917	1 764 778	8.16	4.43
swe	169 825	1 503 581	10.32	5.51
tur	200 122	1 319 398	9.21	6.58

Table 3: number of types and tokens, average type and token length

	10	100	1 000	10 000
cat	24.31	45.30	65.20	87.82
dan	19.63	42.58	62.74	83.10
dut	22.53	45.23	65.78	85.54
est	11.61	25.92	47.62	73.28
fin	10.98	20.72	37.48	62.39
fre	21.38	45.73	66.25	88.65
ger	26.69	48.45	65.97	82.54
ice	21.62	40.74	61.22	82.39
ita	17.88	40.59	62.41	85.93
kor	5.68	17.54	37.16	64.33
nor	19.42	41.96	62.05	82.05
ser	22.25	41.57	62.20	82.19
sor	15.95	35.37	58.70	79.99
swe	18.76	40.25	60.59	80.93
tur	9.75	19.69	38.80	67.12

Table 4: percentage of text coverage by the most frequent 10, 100, 1 000, 10 000 types

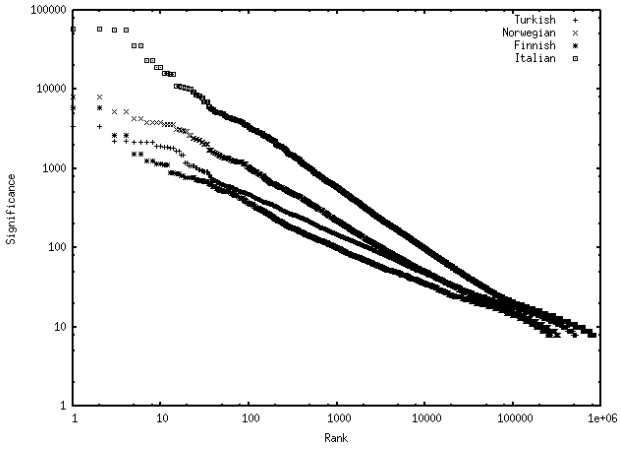

Figure 1: log-log rank - co-occurrence significance diagramm for Turkish, Norwegian, Finnish and Italian

table name	fields	functionality
Wordlist	word number, word string, frequency	Mapping from words to word numbers, stores the absolute frequency count
Sentences	sentence number, sentence	Stores the text in a sentence-based format
Index	word number, sentence number	Full text index for finding all the sentences a word occurs in
Co-occurrences (sentence)	word number 1, word number 2, co-occurrence count, significance	Stores association scores for words occurring together in sentences
Co-occurrences (neighbours)	word number 1, word number 2, co-occurrence count, significance	Stores association scores for words occurring next to each other

Table 2: Overview of the database schema used

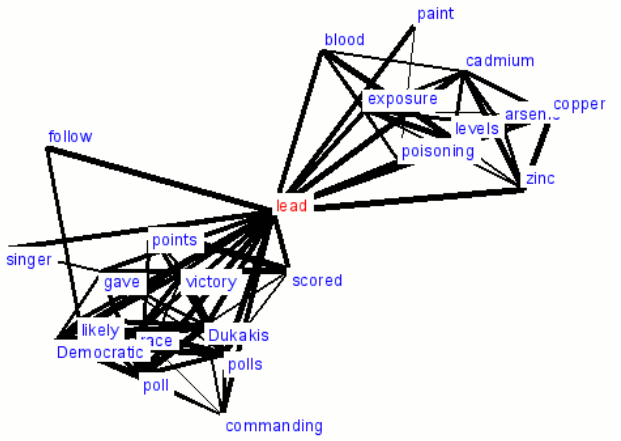

Figure 2: co-occurrence graph for "lead" from English Corpus: two meanings as metal and verb are visually perceivable

5.2. Co-occurrences as building blocks

On the web site, we show a co-occurrence graph that depicts associations of a target word graphically. Figure 2 gives an idea of how to obtain word senses from co-occurrence graphs, for details see (Bordag, 2006).

Other applications include semantic class / taxonomy learning: words can be compared by their co-occurrences, yielding paradigmatic relations, see e.g. (Rapp, 2002). Another way to arrive at refined word sets is to intersect co-occurrence sets as in (Biemann et al., 2004b). To give an example, common right neighbors of *apple* and *plum* are *fruit, trees, tree, varieties, flavors*. The highest-ranked sentence-based co-occurrences excluding neighbors are a collection of fruits and other edible things: *pear, cherry, peach, sauce, wine, spice*. While these mechanisms usually do not produce 100% pure word sets, they can serve as important selection procedures for augmenting semantic resources.

6. Conclusions and Further Work

We have presented a flexible schema of providing monolingual large natural language resources and given an insight into possible questions that may be answered by it. The available resources are growing in size and variety. In the near future, all larger languages, beginning with the official languages in the EU, will be covered. We are open for donations of text in any language.

7. References

Chris Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. 2004a. Language-independent Methods for Compiling Monolingual Lexical Data. In *Proceedings of CIcLING*, LNCS 2945.

Chris Biemann, Stefan Bordag, and Uwe Quasthoff. 2004b. Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of LREC-04*, Lisboa, Portugal.

Chris Biemann, Stefan Bordag, Uwe Quasthoff, and Christian Wolff. 2004c. Web Services for Language Resources and Language Technology Applications. In *Proceedings of LREC-04*, Lisboa, Portugal.

Avrin Blum and Toni Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madisson, WI, USA.

Stefan Bordag, Hans Friedrich Witschel, and Thomas Wittig. 2005. Evaluation of Lexical Acquisition Algorithms. In *Proceedings of GLDV-Frühjahrstagung*, Bonn, Germany.

Stefan Bordag. 2006. Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of EACL-06*, Trento, Italy.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, Volume 19, number 1*.

Ted Dunning. 1994. Statistical identification of language. In *Technical report CRL MCCS-94-273*, New Mexico State University. Computing Research Lab.

Helmut Meier. 1967. *Deutsche Sprachstatistik*. Olms, Hildesheim, 2nd edition.

Reinhard Rapp. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of COLING-02*, Taipei, Taiwan.

Bengt Sigur, M. Eeg-Olofsson, and J. van de Weijer. 2004. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica 59:1*.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language*. Houghton-Mifflin.