Exploiting the Leipzig Corpora Collection

Matthias Richter*, Uwe Quasthoff*, Erla Hallsteinsdóttir*, Christian Biemann*

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

Abstract

In this paper the Leipzig Corpora Collection is introduced as a contribution to the idea that there is need for standardization of multilingual language resources. We explain the steps of building, processing and presenting corpora of comparable sizes and in a uniform format. Results from intra- and interlingual comparisons of corpora are given and methods that can build upon these corpora are shown.

Uporaba lepiziške korpusne zbirke

V članku je leipziška korpusna zbirka predstavljena kot prispevek k ideji o standardizaciji večjezičnih jezikovnih virov. Razložimo postopke gradnje, procesiranja in predstavitve korpusov primerljive velikosti in v enovitem formatu. Podani so rezultati znotraj- in medjezikovne primerjave korpusov ter predstavljene metode, ki lahko zrasejo na njihovi osnovi.

1. Introduction

Corpora are important linguistic resources. We have released a collection of standard sized corpora in 17 different languages in a uniform format that is free of charge for scientific use. Large corpora can be accessed online and downloaded from `http://corpora.uni-leipzig.de/` including a software for offline corpus exploration. This data has been prepared in order to ease and foster corpus research and as a contribution to the standardization of language resources. In this paper we describe the details of the collection and its format, explore possibilities of research given standard sized corpora and present selected results that we have already obtained. Because of the variety of topics covered in this paper we mention and discuss related works in the respective contexts instead of prepending a related work section.

After elaborating on the collection itself in Section 2. we present intra and inter language statistics in Section 3. and examples of usage of our corpora in Section 4..

2. The Leipzig Corpora Collection

2.1. Goals of the Project

The Leipzig Corpora Initiative was started during the 1990s because at that time there were no freely accessible resources available for NLP in German. Since then techniques for processing and presenting corpora have been developed which are not depending on features of specific languages. Some are described in (Biemann et al., 2004b). Having collected text resources in many different languages, it is now possible to provide access to data and statistics on these languages which are 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. As for such, more sophisticated corpus query systems are available, e.g. (Kilgarriff et al., 2004). Our focus, however, is on methods that work in absence of linguistic knowledge. And nevertheless, as discussed in detail e.g. in (Bordag, 2006) the resources we are discussing here 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 see the literature on semi-supervised learning (Zhu, 2005). Possible usage of corpora as a resource includes, but is not limited to (Baroni and Ueyama, 2006):

- monolingual lexicography (which will be a more detailed example in Section 4.1.)
- 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 (Faulstich et al., 2002)
- selecting balanced word sets for experiments e.g. in psycholinguistics

2.2. The Corpus Building Process

Our corpus building process consists of mainly four steps: collecting, pre-processing, cleaning and, eventually, calculating. The steps of the process have been described in detail in (Quasthoff et al., 2006).

Unless there already is a large text collection at hand, texts have to be collected for each language. During the

	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
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.

last years it has become a common practice to use the web as corpus or for corpus acquisition (Kilgarriff, 2001). Corpus acquisition from the web often includes seeding and crawling of web sites (Baroni and Kilgarriff, 2006). One modification of seeding that we employ is to search for current news articles with a news search engine for a very long period of time in order to ensure that certain types of text get collected.

Pre-processing is done by stripping HTML-tags from the collected texts and separating the content from boilerplates. Then a sentence boundary detection is performed and ill-formed sentences fragments get removed as well as sentences in foreign languages (Quasthoff and Biemann, 2006) and (near) duplicates.

Before scrambling the corpus on sentence level and reducing it to pre-defined sizes, further cleaning is performed. This is done to ensure there are actually properly formed sentences which are not obviously containing non-standard language. Scrambling sentences and downsampling in a way that the original documents cannot be restored ensures that the texts can be distributed without hurting copyright protection, as single sentences are too short to be regarded as intellectual property.

2.3. 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 largest unit. This is sufficient for a vast variety of applications in statistical NLP and lexicography.

For each language a full form dictionary with frequency information for each word is calculated. Further we provide co-occurrence statistics: words that co-occur significantly often with a given word. For the calculation of the significance, the log-likelihood measure (Dunning, 1993) is used as described in (Biemann et al., 2004b). Two kinds of co-occurrence data are pre-computed: Words occurring together in sentences and words found as immediate (left or right) neighbors. Only co-occurences that are above a certain significance level (p=5% for neighbors, p=1% for sentence-windows) are kept. Co-occurrence data is meant to be used extensively as a building block for further applications (cf. Section 4.2. for some ideas).

Additional data is included if available. As of now, only the German dictionary already 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 on word and sentence level with all kinds of additional data such as grammar, links and annotation.

2.4. Database Structures und Conversion Issues

The structures of the MySQL database have been kept as simple as possible with much effort having been put into short query response times with large amounts of data. One type of table is meant for storing words, sentences and sources with id, frequency (for words only) and the respective string. Another type of table is used for sentence-based and neighbor co-occurrences between two word ids, the co-occurrence's frequency and its statistical significance. Finally there are inverted lists, one for sentence id and source id and one for word id and sentence id which also contains the word's position in the sentence. There is also a table with pre-calculated meta statistics of the database.

There is at the moment no conversion script available for specific source formats, but it is only a matter of a few lines of code to transform any sane text corpus format into a database in the Leipzig Corpora Collection's format. There is a software available at request from the authors which takes a sentence segmented text and a list of multi word units as input and calculates a full text index with position and sentence-based and neighbor co-occurrences. As an example, the 21 corpora of the TEI-encoded JRC-Acquis collection (Steinberger et al., 2006) were converted with ease.

2.5. Distribution and Availability

On our web site `http://corpora.uni-leipzig.de/` corpora for the following languages can be accessed online: Catalan, Danish, Dutch, English, Estonian, Finnish, French, German, Icelandic, Italian, Japanese, Korean, Norwegian, Sorbian[1], Spanish, Swedish, Turkish. There also is a download site at `http://corpora.uni-leipzig.de/download.html` where smaller corpora[2] of these languages can be obtained free of charge in two formats: flat text files and MySQL databases. The Leipzig Corpus Browser is a tool written in Java for accessing the MySQL databases. The software provides a lot more predefined query options than the web site does and makes adding customized queries easy. This

[1] spelling is correct: Upper and Lower Sorbian are slavonic minority languages with approximately 100 000 speakers in the south of Eastern Germany.

[2] As of now the larger corpora are available only after email request.

can be used for example to add more sophisticated queries and for the integration of additional data resources. The browser should be operational on any platform that supports Java 5, however it has only been tested on Microsoft Windows, Mac OS X, Linux and Solaris. It is available free of charge from the download page as well.

3. Statistical Results

There are several problems when comparing statistical data for corpora of different types of selection, languages, and sizes. In Section 3.1.1. the effect of of the type of selection is measured. In Section 3.1.2. we compare measurements for different corpus size. The non-linear growth of some size parameters is shown to fulfill power laws. This in turn is used to combine results for different languages in Section 3.2..

3.1. Intra Language Statistics

3.1.1. Sampling

A series of experiments was conducted to quantitatively study the intra language effects of sampling sentences at random from starting sets of different size. Starting point was a corpus of 40 million German sentences that were in text order. 100 000 sentences were selected at random from one distinct segment of size 1 million, 4 million, 10 million or all 40 million sentences. Each experiment was repeated 40 times and numbers of tokens, types, sentence co-occurrences, neighbor co-occurrences as well as average type and token length and text coverage with the top n types was measured. The results are summarized in Table 2.

It turns out that the numbers of types and co-occurrences show a big variation. On the other hand the average type, token and sentence length as well as text coverage with the top n types remain extremely stable. The experiment also proposes that the amount of text from which one chooses the final sample has got a small but significant influence on the average numbers of types (the larger, the more) and co-occurrences (the larger, the less) observed, which is stronger with sentences based than with neighbor co-occurrences. This result does not defy intuition as one would expect, when looking at random samples from a corpus of infinite size, to see content words' frequencies – and therefore also the number of co-occurrences – decrease and the number of hapax legomena increase. A t-test tells that this result is highly significant (p=0.1%) only when comparing the columns for 1 million and 40 million sentences. The values for the intermediate segment sizes can not contribute statistically significant support, yet they are neither opposing the observations. On the other hand the effect is not so dramatically strong that we would need to ensure that there is very precisely the same amount of source text from which we start downsampling standard size corpora. If there are, however, several orders of magnitude this systematic skew should be taken into consideration.

3.1.2. Scaling

In the following, we compare Finnish corpora containing 100K, 300K, 1M and 3M sentences, respectively. For these, we count the number of tokens and the number of sentence and next neighbor co-occurrences (above some threshold). The result, shown in Figure 1, is typical for all languages we analyzed.

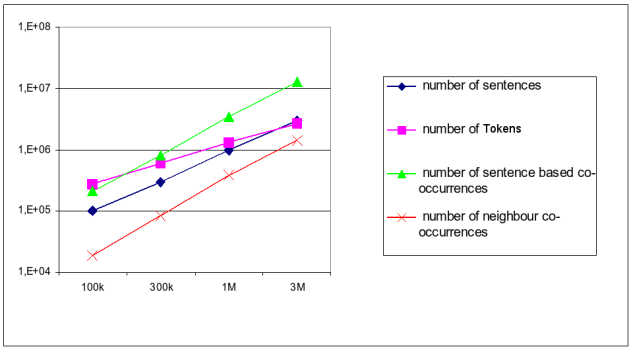

Figure 1: Effects of different corpus size on numbers of types, tokens and co-occurrences for Finnish

In this doubly logarithmic plot we can observe the following:

1. The increase is in all cases nearly linear.
2. The number of tokens increases clearly more slowly than the number of sentences.
3. The number of sentence co-ocurrences increases at a similar rate as the neighbor co-occurrences.

3.2. Inter Language Statistics

3.2.1. Basic statistics

Language statistics such as Zipf's law (Zipf, 1935; Sigur et al., 2004) have been researched in intra language basis for many years, e.g. by (Meier, 1967) for German. We are now presenting inter language basic characteristics in Tables 3 and 4 and exemplified in Figures 2 and 3:

- 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.

3.2.2. Comparing growth rates

In Figure 4 we compare Figures like 1 for different languages. For simplicity's sake, always one language is compared to the average of all languages. Here we compare Finnish, French, Italian, and Norwegian (bold lines) with the language average (thin lines).

As can be seen, there are considerable differences from the average. These differences are stronger than the intra language variation observed in Section 3.1.1.. We find both parallel and non-parallel behavior. For instance, we find:

	1 million	4 million	10 million	40 million
num. tokens	2 020 882 (98 465)	2 021 002 (81 693)	2 020 851 (65 396)	2 021 958 (2 964)
num. types	154921 (19910)	162324 (21030)	162576 (11424)	166350 (373)
num. s. co-occurrences	438459(26003)	413683 (14460)	405442 (8005)	395641 (1718)
num. n. co-occurrences	169308 (4636)	167248 (3446)	167000 (2180)	166408 (461)
coverage top 10	26.70 (0.18)	26.71 (0.19)	26.69 (0.18)	26.71 (0.18)
coverage top 100	48.42 (0.12)	48.40 (0.16)	48.41 (0.12)	48.43 (0.12)
coverage top 1000	65.81 (0.58)	66.02 (0.89)	66.09 (0.58)	65.03 (0.61)
coverage top 10000	82.62 (1.03)	82.53 (1.60)	82.68 (1.04)	82.57 (1.06)
avg. token length	5.72 (0.0077)	5.73 (0.014)	5.72 (0.0080)	5.72 (0.0077)
avg. type length	11.19 (0.026)	11.22 (0.032)	11.25 (0.026)	11.28 (0.025)

Table 2: Sampling Statistics. Arithmetic means and (standard deviations) for each set of experiments. Very stable features are marked **bold**, less stable features are marked in *italics*.

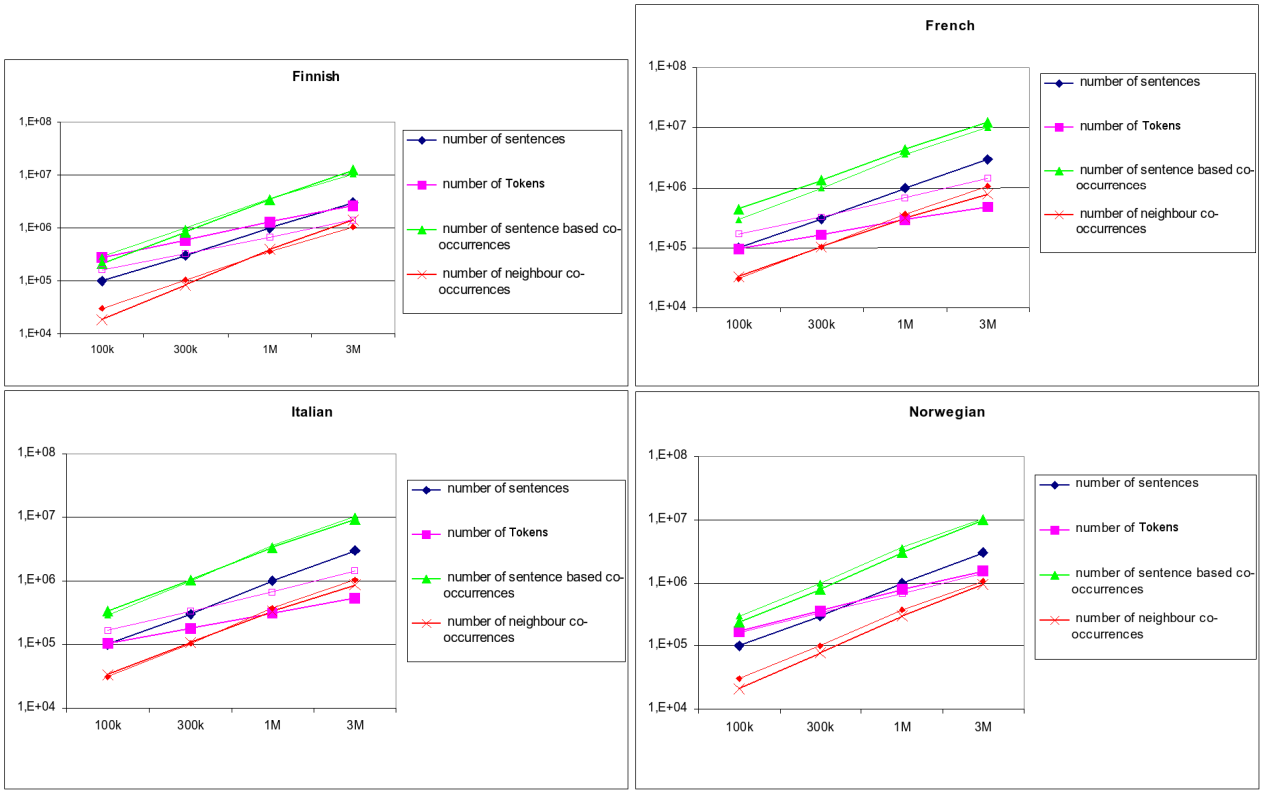

Figure 4: Effects of different corpus size on numbers of types, tokens and co-occurrences for Finnish, French, Italian and Norwegian

- Finnish has more word forms than average, This corresponds to strong morphology and huge average word length.
- In contrast, French and Italian have much less words. Moreover, the increase of the number of tokens is less then average.
- For Norwegian, the number of tokens behaves average. But it seems to have less co-occurrences of both kinds.

4. Using Corpora and Co-occurrences

4.1. Research in Phraseology

To illustrate the possible usage of corpora as a linguistic resource we discuss the usage and the usefulness of the corpora in a research project on phraseology and lexicography in this Section. Since there are no homogeneous definitions of phraseological units, the term is here to be understood in a broad sense covering heterogeneous lexicalized multi word units.

In order to select highly frequent phraseological units for the compilation of a bilingual phraseological database we determined the frequency of over 5000 phraseological units extracted from existing dictionaries for German as a Second Language in the German corpus. The frequency test was carried out in the corpus in April 2002 by using constructed search forms that correspond to possible usage forms of the phraseological units. Furthermore, we analyzed the corpus examples to extract lexicographic relevant data such as frequent syntactic and semantic usage patterns, meaning and semantic variation, external valency, syntactic and morpho-semantic restrictions and any complementary

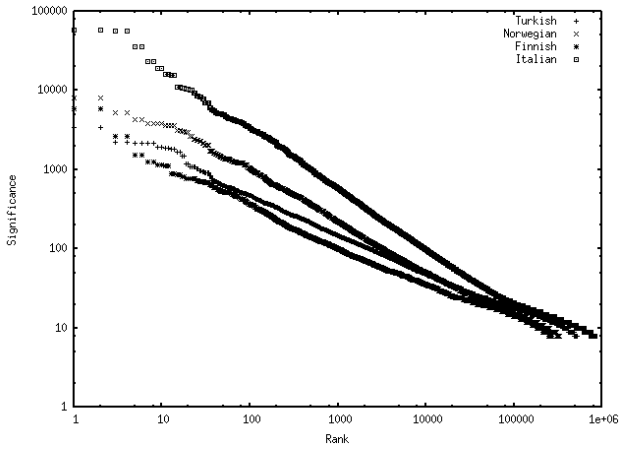

Figure 2: log-log rank - co-occurrence significance diagramm for Turkish, Norwegian, Finnish and Italian

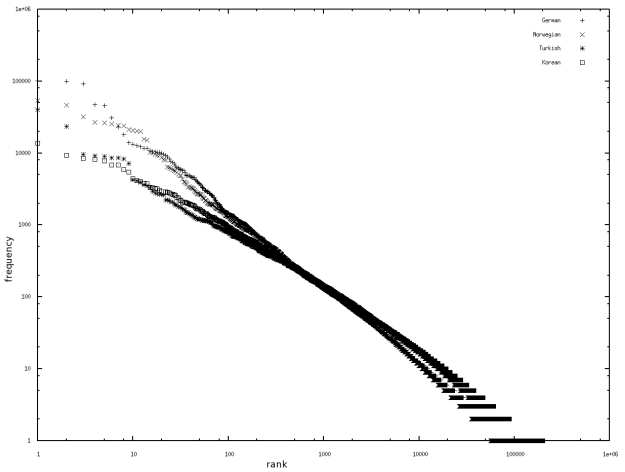

Figure 3: log-log rank - frequency diagramm for German, Norwegian, Turkish and Korean

	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
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
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

grammatical, lexical and pragmatic information needed to enable the potential non-native users of the database to correctly use the phraseological units (Hallsteinsdóttir, 2005).

The frequency data was then combined with data from a research project on native speakers knowledge about the same phraseological units, whereby we have compiled a list including highly frequent and well known phraseological units that should be integrated in the basic vocabulary of German as a foreign language. This list provides a solid basis for further lexicographic and language teaching work on phraseology, e.g. in relation to the reference levels of the Common European Framework of Reference for Languages (CERF) (Hallsteinsdóttir et al., 2006).

4.2. Co-occurrences as building blocks

On the web site and in the Corpus Browser, we show co-occurrence graphs that depict associations of a target word graphically. Figure 5 makes obvious the idea of how to obtain word senses from co-occurrence graphs, see (Bordag, 2006) for details. The basic idea is to partition the co-occurrences graph into clusters each of which represents one sense.

Other applications include semantic class and taxonomy learning: words have been compared by their co-occurrences, yielding paradigmatic relations, by e.g. (Rapp, 2002).

Promising initial results have been achieved also in the attempt to separate syntagmatic and paradigmatic relations from co-occurrences sets based on typical distances between co-occurring words (Büchler, 2006). This is of course highly language specific and will need further research. A way to refine word sets is to intersect co-occurrence sets as in (Biemann et al., 2004a). 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.

5. 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. We have also presented some promising results from corpus

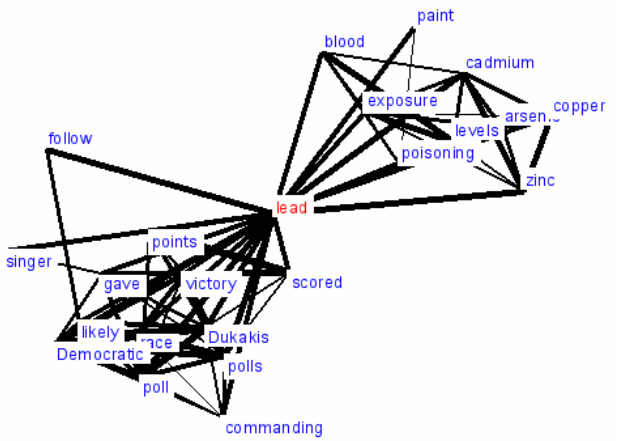

Figure 5: co-occurrence graph for "lead" from English Corpus: two meanings as metal and verb are visually perceivable

use and from inter- and intra-language comparison. Our resources are meant to be 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 cooperations and for donations of text in any language.

6. References

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Proceedings of EACL-06, Trento, Italy*.

Marco Baroni and Motoko Ueyama. 2006. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, pages 31 – 40.

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

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

Marco Büchler. 2006. Flexible Computing of Co-occurrences on Structered and Unstructered Text. Master's thesis, Leipzig University.

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

Lukas Faulstich, Uwe Quasthoff, Fabian Schmidt, and Christian Wolff. 2002. Concept Extractor - Ein flexibler und domänen-spezifischer Web Service zur Beschlagwortung von Texten. In *Proceedings of 8. Intl. Symposium für Informationswissenschaft (ISI 2002)*.

Erla Hallsteinsdóttir, Monika Sajánková, and Uwe Quasthoff. 2006. Phraseologisches Optimum für Deutsch als Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen. *Linguistik online: Neue theoretische und methodische Ansätze in der Phraseologieforschung*.

Erla Hallsteinsdóttir. 2005. Vom Wörterbuch zum Text zum Lexikon. *Zwischen Lexikon und Text - lexikalische, stilistische und textlinguistische Aspekte*.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proc. EURALEX 2004, Lorient, France*.

Adam Kilgarriff. 2001. Web as Corpus. In *Proceedings of Corpus Linguistics*.

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

Uwe Quasthoff and Chris Biemann. 2006. Measuring Monolinguality. In *Proceedings of LREC-06 workshop on Quality assurance and quality measurement for language and speech resources*.

Uwe Quasthoff, Matthias Richter, and Chris Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of LREC-06*.

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*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the LREC-06*.

Xiaojin Zhu. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

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