# A Flexible Infrastructure for Large Monolingual Corpora

## **Uwe Quasthoff & Christian Wolff**

Leipzig University Computer Science Institute, NLP Dept. Augustusplatz 10/11 04109 Leipzig, Germany {quasthoff, wolff}@informatik.uni-leipzig.de

#### Abstract

In this paper we describe a flexible and portable infrastructure for setting up large monolingual language corpora. The approach is based on collecting a large amount of monolingual text from various sources. The input data is processed on the basis of a sentence-based text segmentation algorithm. We describe the entry structure of the corpus database as well as various query types and tools for information extraction. Among them, the extraction and usage of sentence-based word collocations is discussed in detail. Finally we give an overview of different application for this language resource. A WWW interface allows for public access to most of the data and information extraction tools (*http://wortschatz.uni-leipzig.de*).

### 1. Introduction

We describe an infrastructure for managing large monolingual language resources. Since 1995, we have accumulated a German text corpus of more than 300 Million words with approx. 6 Million different word forms in approx. 13 Million sentences. The Project - originally called "Deutscher Wortschatz" (*German Vocabulary*) - has recently been extended to include corpora of other European languages (Dutch, English) as well, with more languages to follow in the near future (see table 1)

	German	English	Dutch
word tokens	300 Mill.	250 Mill.	22 Mill.
sentences	13,4 Mill.	13 Mill.	1,5 Mill.
word types	6 Mill.	850.000	600.000

Table 1: Basic Characteristics of the Corpora.

The approach is based on the extraction of sentences from various types of texts. The sentence is chosen as the basic structuring unit due to copyright restriction on the one hand, as a feasible level of linguistic representation adequate for giving examples for word tokens on the other. As we aim at developing an infrastructure for corpus processing rather than a single corpus of text, there (almost) no restriction on the type of text to be included in the corpus.

Starting off from a rather simple data model tailored for large data volumes and efficient processing using a relational data base system as storage we employ a simple yet powerful technical infrastructure for processing large amounts of texts to be included in the corpus.

Beside basic procedures for text integration into the corpus we have developed tools for post-processing our linguistic data. The corpus is available on the WWW (http://www.wortschatz.uni-leipzig.de) and may be used as a large online dictionary.

# 2. Methodological Approach

Our collection is comprehensive rather than errorfree. In the long run we aim at representing a large portion of current-day word usage available from various sources. While this does not prevent inclusion of errors (like typos in newspaper text), we are able to eliminate typical sources of erroneous information by statistical as well as intellectual optimization routines (see Quasthoff 1998a for details).

In addition, only a high data volume of the corpus allows for the extraction of information like sentencebased word collocations and information about low frequency terms. At the same time, the infrastructure should be open for the integration of various knowledge sources and tools: We strongly believe that there is no single linguistic or statistical approach for all operational needs (optimisation tasks, information extraction etc.). Hence, we provide data for very different purposes.

## 3. Technical Infrastructure

The backbone of our project is a relational database system. We have chosen *mySQL* as a low cost DBMS with high performance and availability on several major platforms. Currently, the corpus databases are stored on UNIX/Linux database servers. Using SQL as data definition and manipulation language, we can make sure that standardised APIs for access and extraction tools are available for all major programming languages.

# 4. Integrating Diverse Data Resources

#### 4.1. Data Sources

Data acquisition for our corpora is based on the analysis of available electronic text from various sources. These include

- General newspaper text (major German newspapers, English newspaper text from the TREC and TIPSTER collection, cf. Voorhees & Harman 1999).

- Electronic dictionaries (general knowledge dictionaries as well as technical and domain-related like medical dictionaries).

- Electronic books and journals, mostly CD-ROM-based collections.

- Web resources with a minimum level of language quality.

In the starting phase of corpus setup, text was primarily extracted from CD-ROMs provided by various publishers. With more and more high quality text coded in declarative markup formats like HTML being made available via the world wide web the collection strategy in our approach has changed: We employ configurable search agents for collecting texts which also do basic feature extraction like coding main subject areas in newspaper texts on the WWW.

## 4.2. Text Processing

The processing of input data is done in several steps which may roughly be divided into the necessary routines for the extension of the corpus by including new data, and postprocessing of information for the whole database.

The pre-processing steps include format conversion, i. e. extraction of raw text from various formats like PDF, MS-WinWord or HTML, the partitioning of documents into sentences, lexical analysis (word and phrase recognition as well as identification of special phrase types like multi-word proper names) and indexing of the whole text corpus.

We maintain a complete full-text index for the whole corpus, making analysis of typical word usage a simple task. The underlying data model stores single words as well as concepts and phrases automatically extracted from the corpus. Beyond the raw data level, our data model provides for the integration of additional information of various categories:

- syntactic and morphological information at word level
- semantic information like subject areas or classification codes at word and sentence levels
- information about related words, either from knowledge sources like synonym dictionaries or thesauri or as the result of automatic extraction (word collocations, sentence classification).

This information is collected not only from various sources (dictionaries with classification codes or subject areas), but also by applying linguistic analysis tools, some of which are used in cooperation with other NLP groups (e. g. the TNT tool for part-of-speech tagging, cf. Brants 2000).

## 4.3. Entry Structure

The basic structure of entries in the corpus database includes information on the absolute word frequency for each entry (i. e. each inflected word form or each identified phrase like the proper name *Helmut Kohl*). Additional frequency class is calculated based on a logarithmic scale relative to the most frequent word in the corpus. For the English corpus, the most frequent word, *the*, has frequency class 0, while an entry like *Acropolis* with an absolute frequency of 20 belongs to frequency class 18, as *the* occurs approx.  $2^{18}$  times more often.

In addition to this basic statistical information, example sentences extracted from the texts most recently included in the corpus are given for each word.

If available, morphological and semantic information are presented. Fig. 1 shows an example for the entry *Weltanschauung* from the German corpus.

# 5. Tools for Information Extraction

# 5.1. Query Types

Besides querying for single word entries, the SQLbased approach allows for a broad range of query types. Among them are searches in database fields like word descriptions (subject areas), searches for grammatical information and querying the full-text index of the sentence database as well as special purpose queries like retrieving all words with a given length or selecting all words attributed with a given subject area. Additionally, administrative query types allow for the management of currently active database processes and the evaluation of access statistics.

Word (word number: 95400): Weltanschauung

Frequency class: 14 (Absolute count: 387)

Subject Area: General, Chemistry, Natural Science, Science, Culture, Education, Learning, Chemie -> Naturwissenschaft -> Wissenschaft -> Kultur Erziehung Bildung Wissenschaft)

Erziending Diridung wissensenart)

**Morphology**:welt|an|schau|ung (=welt+an=schau%ung)

#### **Grammatical Information:**

**Part of Speech:** Noun *Gender*: Feminine

*Inflection:* die Weltanschauung, der Weltanschauung, der Weltanschauung, die Weltanschauung, die Weltanschauungen, der Weltanschauungen, den Weltanschauungen, die Weltanschauungen (*inflection class fb*)

#### **Relations to other Entries:**

Synonyms: Anschauungsweise, Betrachtungsweise, Denkweise

Compare To: Fatalismus, Idealismus, Ideologie, Kommunismus, Nihilismus, Optimismus, Pazifismus, Realismus

Synonym of: Anschauungsweise, Denkart, Denkungsweise, Denkweise, Einstellung, Ideologie, Lebensanschauung, Meinung, Mentalität, Philosophie, Sinnesart, Standpunkt, Urteil, Weltbild

#### **Examples:**

Auch die Schulmedizin beinhaltet schließlich eine Weltanschauung - eben die rein naturwissenschaftliche. (*Source*: TAZ 1997)

Behindert die anthroposophische Weltanschauung nicht zugleich die Verbreitung solcher Heilmethoden? (*Source*: TAZ 1997)

Wenn man die Medizin zur Weltanschauung macht, ja. (*Source*: TAZ 1997)

Figure 1: Sample Entry for *Weltanschauung* (German corpus)

## 5.2. Sentence-Based Collocations

Beyond simple text processing we have developed a number of *information extraction tools* which are based on statistical methods. Among them the automatic calculation of sentenced-based word collocations stands out as an especially valuable tool for corpus-based language technology applications.

The occurrence of two or more words within a welldefined unit of information (sentence, document) is called a collocation. For the selection of meaningful and significant collocations, an adequate collocation measure has to be defined: Our significance measure is based on a function comparable to the well-known statistical *G-Test* for Poisson distributions: Given two words A, B, each occurring a, b times in sentences, and k times together, we calculate the significance sig(A, B) of their occurrence in a sentence as follows:

$$sig(A, B) = x - k \log x + \log k!$$

with n = number of sentences,

$$x = \frac{ab}{n}$$

Two different types of collocations are generated: Collocation based on occurrence *within the same sentence* as well as *immediate left and right neighbors* of each word. Fig. 2 shows an example listing of the top 50 collocations for the term and *Papandreou* taken from the English corpus, number in brackets indicate the relative strength of the collocation measure..

# Top 50 of 593 significant sentence-based collocations for **Papandreou**

Andreas (474), Premier (314), Socialist (214), government (212), Greece (172), Panhellenic (172), Greek (139), Movement (129), Liani (121), party (118), socialist (115), scandal (111), PASOK (95), former (92), Mitsotakis (87), Koskotas (84), elections (83), Parliament (80), premier (80), Democracy (72), ministers (71), scandals (70), coalition (63), Athens (54), Dimitra (52), June (52), Coalition (50), Minister (50), Crete (49), after (48), divorce (48), accused (47), hospital (46), Ozal (45), heart (44), Margaret (43), seats (43), George (43), London (39), American-born (38), alliance (37), financial (36), Florakis (36), Prime (35), political (35), surgery (34), bases (34), wife (33), Cabinet (33), Sunday (32)

#### Significant left neighbors of **Papandreou**:

Andreas (584), George (23), accused (15), Margaret (12), including (8), against (6), claimed (6), younger (6), if (5), indicting (5), After (4), charismatic (4), defeating (4), indicted (4) saying (4).

indicted (4), saying (4)

Significant right neighbors of **Papandreou**:

underwent (11), blackmailed (10), flew (10), came (8), told (8), married (8), government (7), met (5), failed (5), obligingly (5), left (5), arrived (4), founded (4)

# Figure 2: Collocation Sets for *Papandreou* (English corpus)

Although the calculation of collocations for a large set of terms is a computationally expensive procedure, we have developed efficient trie-based algorithms which allow for a collocation analysis of the complete corpus.

Beyond retrieving the different collocation sets for a given word, the infrastructure provides for what may be called "second order queries" on collocations: For example, the *intersection of collocation sets* for two words will contain words that have a strong relationship to both query terms. Intersecting the terms *amerikanische* (*American*) and *Präsident* (*president*) in the German corpus, yields a result set, that - among other entries - contains the names of American presidents *Bill Clinton* 

and *George Bush* with *Bill Clinton* carrying the highest significance measure for that query.

The introduction of part-of-speech information additionally allows a more precise selection of collocation sets: Using the sets of immediate left and right neighbour collocations, it is possible to retrieve typical adjectives that appear to the left of a given noun or, verbs that appear to the right of a given noun.

# 5.3. Visualization

Based on the set of collocations for any given word with a minimum number of significant sentence-based collocations we have implemented a real-time visualisation algorithm using simulated annealing (cf., Davidson & Harel 1996). The resulting graph shows selected relationships from the set of collocations (see Appendix, Fig. 3, 4). The graphs can be used for interpreting different meanings of homonyms: In fig. 4 different meanings of *King* as a proper name (*Martin Luther King, Jr, Burger King*) and as a title (head of a monarchy, *King Hussein* of *Jordan*) become apparent.

Currently, we are going to link the different language corpora both at word level and on sentence level. This will give a large multilingual dictionary and the possibility to use available multilingual aligned text in this framework.

# 6. Applications

One major advantage of the infrastructure developed for this project is its immediate portability for different languages, text domains, and application: The basic structure consisting of text processing tools, data model, and information extraction algorithms may be applied to any given corpus of textual data. This makes this approach applicable to a wide variety of basic language technology problems like

- text classification,
- document management, or
- information retrieval.

Beside the project's WWW interface and its usage as a general purpose dictionary (basic statistical, syntactic and semantic information, typical usage examples) current applications include collocation-based query expansion in Web search engines. The latter shall be illustrated by an example: Typical usage of Web Search engines is characterized by very short queries and low retrieval effectiveness (cf. Silverstein et al. 1999, Jansen et al. 2000). A possible remedy for this are query expansion techniques. We have developed a search engine interface (see Appendix, fig. 5) which allows the user to start from a single search term and select additional query terms from information available in the corpus (collocations, synonyms etc.). A simple mechanism for expanding the query is implemented using JavaScript and dynamically generated hyperlinks.

While this application makes use of our "standard" data corpus, the infrastructure can be applied to different data sets or text collection without modification. Thus, further applications like comparing special purpose document collections with the general language corpus are possible. The difference in the statistical data can help identifying important concepts ant their relations. Applications of this analysis are

- Terminology extraction and

- Support of object oriented modeling of business processes.

# 7. Outlook

Since the project went online in spring 1998, we have registered more than 350.000 accesses from more than 40,000 users with a constant growth rate of 20 % per month. Due to increasing access counts, we are currently developing a clustered storage and access infrastructure which will not only provide higher throughput for Web access but also a structural separation of production and presentation databases.

## 8. References

- Brants, T. (2000). TnT A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA [to appear].
- Davidson, R., Harel, D., 1996. Drawing Graphs Nicely Using Simulated Annealing, *ACM Transactions on Graphics* 15(4), 301-331.
- Jansen, B. J. et al. (2000), Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. In *Information Processing & Management* 36(2), 207-227.

- Läuter, M., Quasthoff, U. (1999), Kollokationen und semantisches Clustering. In Gippert, J. (ed.) 1999. *Multilinguale Corpora. Codierung, Strukturierung, Analyse. Proc. 11. GLDV-Jahrestagung.* Prague: Enigma Corporation, 34-41.
- Quasthoff, U. 1998A. Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values." In: Proc. First International Conference on Language Resources & Evaluation [LREC], Granada, May 1998, Vol. II, 853-856.
- Quasthoff, U. 1998B. Projekt der deutsche Wortschatz. In Heyer, G., Wolff, Ch. (eds.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 93-99.
- Silverstein, C. et al. (1999), Analysis of a Very Large Web Search Engine Query Log. In *SIGIR Forum* 33(1), 6-12.
- Voorhees, E.; Harman, D. (eds.) 1999. Overview of the Seventh Text REtrieval Conference (TREC-7). In Voorhees, E.; Harman, D. (eds.), Proc. TREC-7. The Seventh Text REtrieval Conference. Gaithersburg/MD: NIST [= NIST Special Publication 500-242].

## 9. Appendix: Figures



Figure 3: Collocation graph for Papandreou (English Corpus)





Theosenegestuetric Suchs - Netscope				<b>1</b> 8
pani Baabahan Anisida Gaba Communica	a Ree			
Informationen zum Projekt Deutscher Wortschatz		Mens Abfrage		wortschatz lexikon
Desachtes Wort: Clinton F. Großerhenbung beachtes F. Großerhenbung F. mar eigenen Wortschatz anzeigen ( Wort: Clinton Kallskatiesen im Sate Pranident (443), Dale (187), US (79), Mate (683), Werken Einer (	anchachlägen -Präsident (138), Washington (15 56), oppskäsansken (65), Boh J	28), Brashikaner (120), Kongerß (117) Seis (50), Kohl (50), Bagiernag (50), Gr	), Johin (106), Ennh (91), Potnid mi (48), azamkanischen (48), S	mim (39), <u>Advance</u> (79), <u>USA</u> Valikance (44), Bornies (39),
Gouveneur (18), Resublikasers griffkante linke Nachbarn: <u>Prändent</u> (699), <u>US-Prändent</u> ( <u>Gouveneur</u> (6), <u>US-Prändent</u>	. (18), <u>Treffen</u> (17), <u>Gingrich</u> (15), 198), <u>Engineung</u> (83), <u>Dräsidenten</u> 8), <u>Wahikampfer</u> (6), <u>wersprach</u> (	<u>Weille Haur</u> (35), <u>Demokraten</u> (33), <u>D</u> (22), <u>Demokrat</u> (21), <u>arklärts</u> (19), <u>Eb</u> (6)	wnokrat (72), <u>Weillen Hauses</u> maar (12), waaf (11), Ensi (7),	77), <u>Genze Roch</u> (71) Istonic (7), Ehrpaurz (6),
igoificante rechte Nachharm kondiste (14), betonte (13), som antexten (6), appelliert (6), beker	nh (13), lean (12), midante (11), s Alante (8), ballant (8), fondent (8), fo	vrozunach (9), kóndiat (8), nominirri (8), lorderiu (6)	Good (7), unteracticitated (7), see	ln (7), <u>Darizzon</u> (6), <u>Lerka</u> (6),
	Suche Clinton US-Prasider nacht In: Alta Vista 💌 💶	nt Afkansas Sebistebilizeten Fale - Backer - Fachal (Filmare Kitte) -	? Ein Service für Sis	
	Hotist - Interneurch - Kolibri - J	Lotra - Lycor - Hathan - Hetguide		
-@- Dokument Übemät	á.			1440000

Figure 5: Search Interface for Web Search Enhancement Using Collocations