In: Knorz, Gerhard; Kuhlen, Rainer (Hg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft (ISI 2000), Darmstadt, 8. – 10. November 2000. Konstanz: UVK Verlagsgesellschaft mbH, 2000. S. 163 – 177

Aiding Web Searches by Statistical Classification Tools

Gerhard Heyer, Uwe Quasthoff, Christian Wolff

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

Abstract

We describe an infrastructure for the collection and management of large amounts of text, and discuss the possibility of information extraction and visualisation from text corpora with statistical methods. The paper gives an overview of processing steps, the contents of our text databases as well as different query facilities. Our focus is on the extraction and visualisation of *collocations* and their usage for aiding web searches.

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*	*French*
word tokens	300 Mill.	250 Mill.	22 Mill.	15 Mill.
sentences	13,4 Mill.	13 Mill.	1,5 Mill.	860.000
word types	6 Mill.	1,2 Mill.	600.000	230.000

Table 1: Basic Characteristics of the Corpora

The intent of the project is to collect large amounts of textual data for experimenting with corpus based semantic processing. 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.

Methodological Approach

Our collection is comprehensive rather than error-free. 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 optimisation routines (see Quast-hoff 1998a for details).

In addition, only a high data volume of the corpus allows for the extraction of information like sentence-based 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.

The backbone of our project is a relational database system. We have chosen *mySQL* (cf. *http://www.mysgl.com)* 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, while Web access to the corpus-related information services is provided by an Apache Web Server running on a Windows NT machine. 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.

In contrast to work by Rieger at the University of Trier, or Paprotte at the University of Münster, the primary concern of the project is the collection of a very large amount of textual and lexical data that we intend to make publicly available. In this respect, the project follows similar intentions as CISLEX (Guenthner 1996), but in comparison covers a much larger set of data. Also, CISLEX seems to focus on morphological analyses, while for us there textual and lexical data are used as raw material for further semantic processing such as analysing definitions, calculating collocations, disambiguating terms, deriving co-hyponyms, etc..

Processing Model

Our corpus-processing infrastructure is based on two major design objectives:

- acceptance of a wide variety of digital text and
- definition of a chain of processes for the automatic setup of data corpora, extraction mechanisms, and access services.

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 CDROMs 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.

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 identifcation 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 co-operation with other NLP groups (e. g. the TNT tool for part-of-speech tagging, cf. Brants 2000).

Information Categories in the Database

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 dass 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 dass 0, while an entry like *Acropolis* with an absolute frequency of 20 belongs to frequency dass 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. Table 2 gives an overview of the most import information categories in our corpus and their relative amount (German corpus database):

Information Category	*Number of Entries*
word list	ca. 6 Mio. word forms
example sentences	ca. 13 Mio.
grammatical information	ca. 3 Mio.
morphological information	ca. 3 Mio.
descriptions	ca. 150.000
subject categories	ca. 1,5 Mio.
semantic relations	ca. 500.000
pragmatics (e. g. usage)	ca. 35.000
collocations (at sentence level)	ca. 3,5 Mio.
collocations (immediate left and right	ca. 1,5 Mio.
full text index	ca. 30 Mio.

Table 2: Information Categories in the German corpus database

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

Word *(word number:* 95400): Weltanschauung
Frequency class: 14 (Absolute count: *387)*
Subject Area: General, Chemistry, Natural Science, Science, Culture, Education, Leaming, Chemie -> Naturwissenschaft -> Wissenschaft -> Kultur
Erziehung Bildung Wissenschaft)
Morphology:welt l an | schau l 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 dass 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)

Types of Queries

Besides querying for single word entries, the SQL-based 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 pur-pose 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.

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 well- defined 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. In the literature, quite a number of different collocation measures can be found (for an in-depth discussion of various collocation measures and their application cf. RUGE 1994 and LEMNITZER 1998). Given two words A, B, each occurring *a, b* times in *n* sentences, and *k* times together, the following table shows different measures for collocation significance:

Tanimoto (Percentage of double in relation to single hits)	$sig_T(A,B) = k / (a + b - k)$
Mutual Information Index (Digression from statistical independence)	$sig_r(A, B) = log(kn / (ab))$ [= $\log(p_{AB} / (pAPB))$]
G-Test (Test for Poisson distributions)	$sig(A, B) = x - k \log x + log\ k!$ with n = number of sentences, $x = \frac{ab}{n^{\bullet}}$

Table 3: Different Significance Measures for Collocations

Based on an evaluation of these measures, we haven chosen the G-Testrelated measure for the calculation of collocation significance, as this measure guarantees a good scalability of results in relation to the absolute occurrence frequency of the collocation terms. Two different types of collocations are generated: Collocation based on occurrence *within the same sentence* as well as *immediate left and right neighbours* of each word. Fig. 2 shows an example listing of the top 50 collocations for the term *retrieval* taken from the English corpus, number in brackets indicate the relative strength of the collocation measure. (As the basis for calculating the collocations are inflected word forms, individual word forms, such as *text* and *Text* for instance, are case sensitive. Difference in spelling is indicative of proper names in English, in Ger-man it even indicates difference in syntactic category.)

Top 50 significant sentence-based collocations for *retrieval:*
storage (625), text (406), data (390), information (349), search (259), document (211), full-text (204), database (149), Topic (136), indexing (129), software (123), systems (106), documents (103), image (97), CD-ROM (87), optical (82), management (78), Text (75), Verity (71), content-based (71), an-line (62), file (61), capabilities (60), query (60), access (58), processing (58), engine (51), databases (50), electronic (47), Provides (46), archival (46), files (42), hypertext (42), stored (39), archiving (38), users (38), searching (37), Boolean (35), records (35), Gescan (33), applications (33), functions (33), user (33), images (31), queries (30), relational (30), fast (29), searches (29), Information (28), disk (28), Fulcrum (27),
Significant left neighbours of *retrieval:*
text (401), information (293), data (190), full-text (161), document (96), contentbased (77), image (50), Text (44), file (31), Topic (25), concept (20), fast (18), later (18), rapid (18), an-line (16), Concept (15), database (15), quick (15), Full-text (14), easy (13), Information (12), free-text (12), interactive (12), storage (12), message (11), Data (10), computer-assisted (9), subsequent (9), Boolean-based (7), faster (7), news (7), record (7), remote (7), semantic (7), DiscPassage (6), associative (6), index-only (6), allows (5), efficient (5), instant (5), knowledge (5), literary-quote (5), quote (5), sequential (5), Content-based (4), archival (4), conceptbased (4), legal-information (4), search (4)
Significant left neighbours of *retrieval:*
software (175), systems (119), engine (83), capabilities (58), program (31), functions (25), package (23), service (19), services (17), capability (15), tool (13), methods (10), packages (10), process (10), engines (9), programs (9), technology (9), utility (9), stations (8), times (8), clients (7), mechanism (7), operations (7), facilities (6), mechanisms (6), product (6), subsystems (6), techniques (6), time (6), method (5), performance (5), products (5), speed (5), strategies (5), tools (5), client (4), purposes (4), speeds (4)

Figure 2: Collocation Sets for *retrieval* (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-ofspeech 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.

Visualisation

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 intention is to display selected relationships from the set of collocations in the resulting graph. In effect, the graphs can be used for representing different meanings of homonyms: In fig. 3 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:

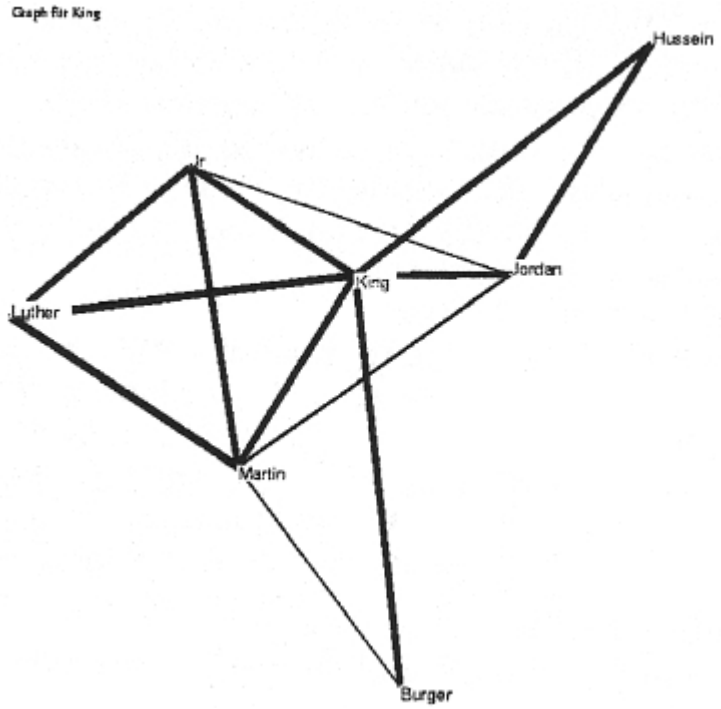

Figure 3: Collocation graph for King (English Corpus)

A second example for the word Rice taken from the English corpus, shows a nice division in the collocation set for different persons and institutions with the proper name Rice:

- an American Secretary of Defense *(Donald Rice),*
- a famous baseball player *(Jerry Rice),*
- *Rice University* and
- *Donna Rice* (ex-lover of presidential candidate Gary Hart).

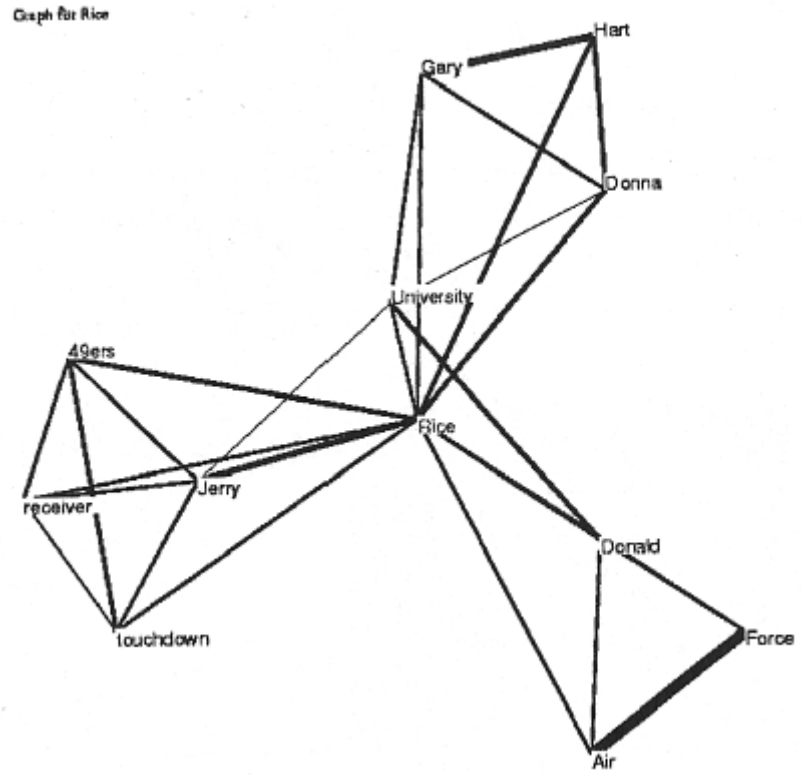

Figure 4: Visualisation of the collocation set or Rice

Separation of Semantic Relations

The calculation of collocations is based on simple statistical measures and does not explicitly *name* the type of semantic relation which holds for a given collocator and its collocates. A further processing of collocation sets is highly desirable, though.

Collocation sets and their visualisation may be employed displaying and partitioning of multiple meanings for single entries. As different meanings of a given word tend to include different subsets of collocation terms which are more closely related to each other, the resulting collocation graph can Show a distinct separation of the entire collocation set. The following example shows this for the set of collocations for *Schweine* (pigs): On the right side of the image typical co-hyponyms like *Rinder (cattle), Kühe (cows), Schafe (sheep)* and *Hühner (chicken)* are displayed, while the left side includes collocations which illustrate aspects of pigs as a food product: *Handelsklassen (grade of goods), folgender (following), Schlachtgewicht (weight at slaughtering time), abgerechnet (discounted).*

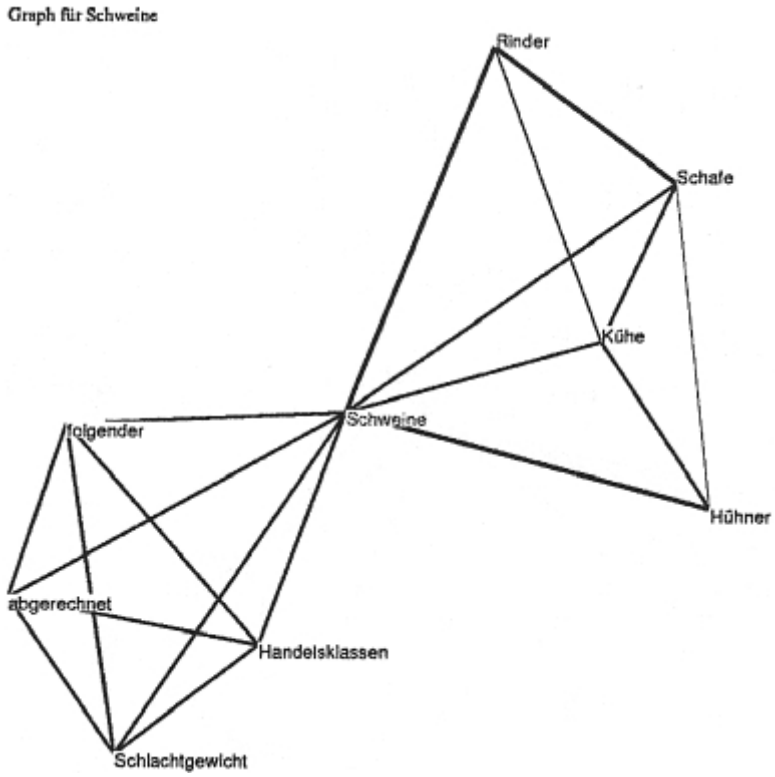

Figure 5: Collocation Graph for Schweine (pigs)

While this type of visualization doest not result in an explicit partitioning of collocation sets, extraction of meaningful subsets may be done be identifying typical representatives for the relations involved: If a typical term for a certain type of relation is known *(extractor),* the intersection of the sets of collocations for the original term and for the extractor results in a subset which contains collocation for the selected relationship. This works fine for sets of collocations in which words are polysemic or refer to concepts as well as to proper names. An example taken from the English corpus shall illustrate this method. The set of collocations for *board* (approx. 90.000 tokens in the corpus, thus frequency class 7) contains terms related to the meaning of board as a component of a Computer as well as for *board* in the sense of a set of people serving a special pupose within an institution or company:

directors (3710)	video (754)	circuit (552)
16-bit (345)	seats (319)	chip (268)
bulletin (1692)	VGA (737)	elected (529)
appointed (338)	trustees (298)	president (263)
chairman (1165)	executive (643)	RAM (462)
expansion (338)	add-in (296)	8-bit (250)
members (944)	shareholders (608)	director (462)
named (335)	slots (289)	proposal (242)
memory (910)	offer (571)	slot (443)
school (330)	shareholder (275)	accelerator (221)
member (890)	company's (570)	approved (436)
across (329)	chief (273)	authorized(220)

meeting (881)	fax (557)
Graphics (324)	Coprocessor (272)

Table 4: Top 40 of several hundred significant collocations for board

Given an adequate term representative of one of the typical meanings of board, the set of collocations can be sematically partitioned. The following tabels shows the collocation subsets for the intersection of collocation sets of *board* and *memory* and *board* and *members* (ordered in decreasing significance).

common collocations for board and memory	*common collocation for* board *and* members
upgrade	directors
includes	elected
drive	eight
expansion	meeting
card	proposal
bus	committee
monitor	join
boards	representatives
video	appointed
graphics	vote
processor	seats
cache	voted
chips	elect
PC	membership
controller	
chip	
serial	
slot	
slots	
Intel	
CPU	
sockets	
PS	
RAM	
adapter	
add-in	
VGA	
coprocessor	
motherboard	
SIMMs	

Table 5: Selection for collocation subset related to different word meanings

Applications

One major advantage of the infrastructure developed for this project is its immediate portability for different languages, text domains, and applications: 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 discussed in more detail.

Recent studies (cf. Silverstein et al. 1999, Jansen et al. 2000) have shown that information retrieval on the Web is remarkable different from the interaction with more traditional types of information retrieval engines like bibliographic databases or full text archives. The following list of properties characterises the problem of using web search engines:

- the Web contains mass data (approx. 1 billion documents by the beginning of 2000) with little or no coherent structure
- users are not knowledgeable with respect to information retrieval systems and their interfaces
- queries tend to be very short (less than three terms on average), usage of search operators is an exception and they are often used in a wrong way
- searches tend to result in large document sets which are evaluated only partially by the searchers.

While there are a number of approaches for an optimisation of this situation, like improving retrieval models, the interfaces to search engines or using clustering and filtering techniques, we concentrate an user-driven query expansion as a technique for which the results of our information extraction tools may be employed:

For every search term the user may review collocations and synonyms and add them to his query definition before submitting it to a search engine. We have developed a search engine interface (see Fig. 6) 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.

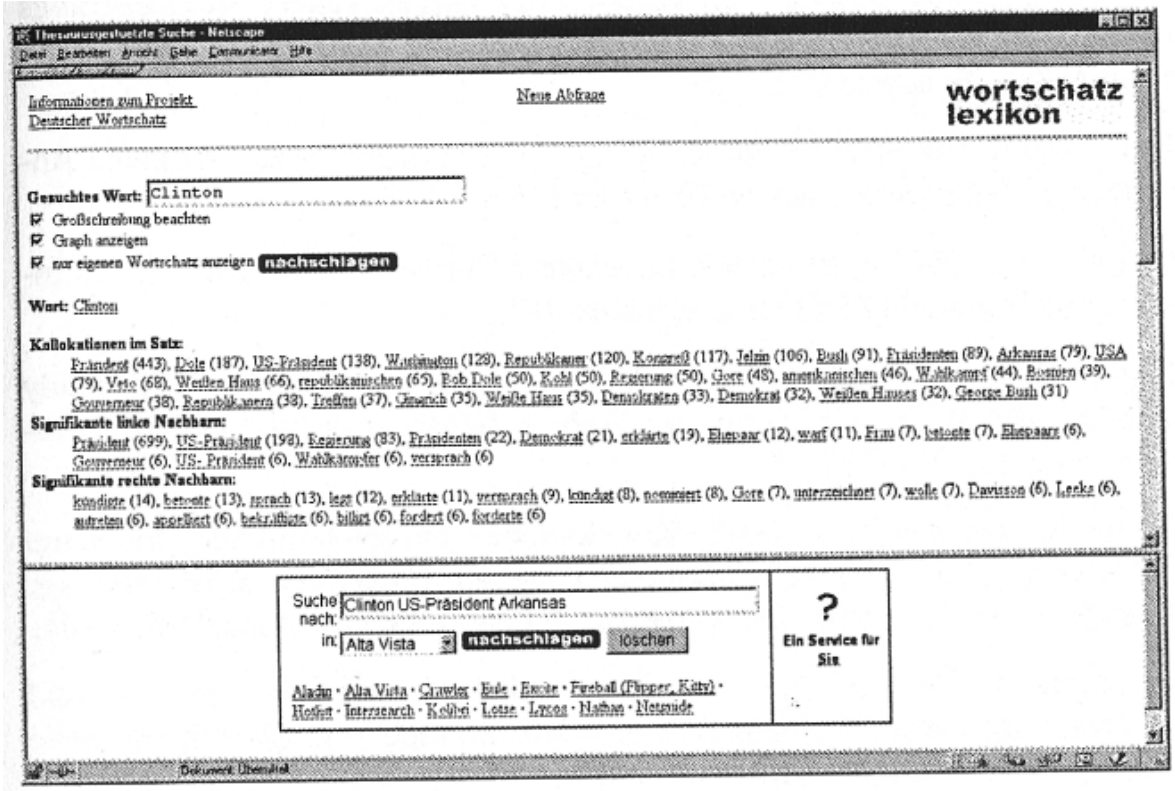

Figure 6: Search Interface for Web Search Enhancement Using Collocations

Outlook

While the web search application makes use of our "standard" data corpus, the infrastructure can be applied to new and different data sets or text collections 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 and their relations. Applications of this analysis are, amongst others,

- domain specific terminology extraction and
- support of object oriented modeling of business processes.

In the latter example, business reengineering according to the methods proposed by Ortner (cf. Ortner 1997) is supported by generating significant semantic relations from software documentation for further use in modeling object-oriented software models.

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.

Guenthner, F. (1996), Electronic Lexica and Corpora Research at CIS. In International Journal of Corpus Linguistics 1(2).

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.

Lemnitzer, Lothar (1998). "Komplexe lexikalische Einheiten in Text und Lexikon." In: Heyer, Gerhard; Wolff, Christian (edd.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 85-91.

Ortner, Erich (1997). Methodenneutraler Fachentwurf. Stuttgart & Leipzig: Teubner.

Quasthoff, Uwe. 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, Uwe. 1998B. Projekt der deutsche Wortschatz. In Heyer, G., Wolff, Ch. (eds.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 93-99.

Ruge, Gerda (1994). Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation. Hildesheim & New York: Olms.

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 RE-trieval 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].