Language-independent Methods for Compiling Monolingual Lexical Data

Christian Biemann[1], Stefan Bordag[1], Gerhard Heyer[1], Uwe Quasthoff[1], Christian Wolff[2]

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

[2]University of Regensburg
PT 3.3.48
93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Abstract: In this paper we describe a flexible, portable and language-independent 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 applications 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. Several language independent methods are used to detect semantic relations between the words of a language. These methods differ in productivity and precision for different languages, but there are highly productive and accurate methods for all languages tested. The process starts with the collection of monolingual text corpora from the Web. Next, we identify collocations, i.e. words that occur significantly often together. These collocations form a network that is analyzed further to identify semantic relations. Because semantic features are often reflected in morphosyntactic structures, we apply classifiers that use the sequence of its characters for classification. Moreover, we use context information and POS-information, if available.

While it is clear that the above mentioned methods can be used to find semantic relations or, can be used to verify the corresponding hypotheses, we want to present abstract methods, specific application and results for different languages.

Since 1995, we have accumulated a German text corpus of more than 500 Million words with approx. 9 Million different word forms in approx. 36 Million sentences. The Project - originally called "Deutscher Wortschatz" (*German Vocabulary*) - has been extended to include corpora of other European languages (Dutch, English, French, Italian). Recently, we incorporated the processing of Unicode, giving rise to a Korean Corpus as well, with more languages to follow in the near future (see table 1)

Table 1: Basic Characteristics of some of our corpora.

	German	English	Italian	Korean
Word tokens	500 Mill.	260 Mill.	140 Mill.	38 Mill.
Sentences	36 Mill.	13 Mill.	9 Mill.	2,3 Mill.
Word types	9 Mill.	1,2 Mill.	0,8 Mill.	3,8 Mill.

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 error-free. In the long run we aim at representing a large portion of current-day word usage (see [Quasthoff et al. 2003]) for a discussion on daily fluctuation of 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 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 (optimization tasks, information extraction etc.). Hence, we provide data for very different purposes.

3 Integrating Diverse Data Resources

3.1 Data Sources

The main (and in most cases only) source is text in the corresponding language taken from the web. The amount of text varies from 2 to 36 million sentences. The text is taken from web pages with the corresponding domain ending.

If available, we include information taken from electronic dictionaries. Multiword dictionary entries are especially of interest because they are needed as a seed to find more.

3.2 Text Processing

In this section we describe the construction of a text database for a given fixed language.

The preprocessing steps include format conversion, i. e. HTML-stripping, and sentence separation.

Sentence separation is done with the help of an abbreviation list. We assume the text being written in a single and previously known language. In this case we can prepare a list of abbreviations (only abbreviations ending in a period are relevant for sentence separation). If no such abbreviations are available, a preliminary list of common abbreviations that are found in multiple languages can be used.

The next step performs language verification. Here we can sort out sentences that are not in the language under consideration. The language detection module uses lists of about 1000 most frequent words of different languages. The language of a sentence is then identified comparing its words with those lists.

3.3 Indexing

Lexical analysis consists of the separation of words and multiwords and indexing of the whole text corpus. While word separation is usually simple, multiwords have to be supplied in advance to be recognized.

We maintain a complete full-text index for the whole corpus, making analysis of typical word usage a simple task.

3.4 Collocations

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: 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:

$$sig(A,B) = x - k\log x + \log k!$$
$$\text{with } n = \text{number of sentences,}$$
$$x = \frac{ab}{n}.$$

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 *Daewoo* taken from the English corpus, number in brackets indicate the relative strength of the collocation measure.

Significant sentence-based collocations for Daewoo:
Leading (272), Edge (253), Motor (132), Korean (108), Co (85), Telecom (83), Korea's (82), Hyundai (67), Mo-tors (66), Shipbuilding (66), Kia (62), Korea (52), South (49), Heavy (48), Corp (46), GM (46), Samsung (44), conglomerate (39), Group (38), Ltd (37), Kim (34), LeMans (31), owned (31), Edge's (29), Products (27), group (27), Fason (26), General (25), Machinery (25), PCs (25), bankruptcy (22), venture (21), Industries (20), Electronics (19), contract (19), joint (18), shipyard (18), Goldstar (17), Okpo (17), Seoul (17), workers (17), Woo (16), cars (15), subsidiary (15), Lucky-Goldstar (14), dealers (14), industrial (14), conglomerates (13), manufacturer (13), strike (13), supplier (13), Choong (12), auto (12), Agbay (11), Koje (11), Pontiac (11), Telecommunications (11), plant (11), 50-50 (10), Dae-woo's (10), Woo-choong (10), factory (10), Joong (9), joint-venture (9), Pupyong (8), giant (8), signed (8), vehicles (8), Inchon (7), Motor's (7), Precision (7), Yoon (7), agreement (7), car (7), chaebol (7), exports (7), logo (7), multibillion-dollar (7), sell (7), units (7)

Significant left neighbors of Daewoo:
Korea's (46), conglomerate (15), Korea-based (7), manufacturer (4)

Significant right neighbors of Daewoo:
Motor (124), Telecom (110), Shipbuilding (73), Group (44), Corp (34), Heavy (25), group (23), Telecommunications (21), Electronics (14), officials (8), Precision (7), Motors (5), Securities (5), Shipbuilding's (5), industrial (5)

Figure 2: Collocation Sets for *Daewoo* (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 that allow for a collocation analysis of the complete corpus in feasible time.

For a given word, its collocation set very much reflects human associations. For this reason, collocation sets are used as input for some of the information extraction algorithms described in section 4.

3.5 POS-Tagging

If available, we use POS-tagging for the following tasks:

1. Disambiguation: If different meanings of a word differ by its POS-tag, we get different collocation sets according to the different POS-tags.

Regard the following example in Figure 3, illustrating the disambiguation of *wish* as noun and as verb in two POS-tagged English sentences.

- For[IF] six[MC] months[NNT2] ,[YC] the[AT] young[JJ] family[NN1] physician[NN1] got[VVD] her[APPG] **wish[NN1]** ,[YC] developing[VVG] close[JJ] relationships[NN2] with[IW] her[APPG] mostly[RR] single[JJ] and[CC] minority[NN1] women[NN2] patients[NN2]
- I[PPIS1] am[VBN] trying[VVG] to[TO] lead[VV0] a[AT1] different[JJ] life[NN1] now[RT] and[CC] I[PPIS1] just[RR] **wish[VV0]** all[DB] that[CST] stuff[NN1] hadn't[VVD] been[VBN] dredged[VVN*] up[RP] again[RT] ,[YC] said[VVD] the[AT] 52-year-old[NN1*]

Figure 3: Disambiguation of *wish* with POS.tags.

2. For several applications, one looks for collocations with a certain POS-tag. In example, when looking for synonyms the candidate set for a given word reduces to those candidates having the same POS-tag.

We use TNT, a freely available POS-Tagger based on Cascaded Markov Models (cf. [Brants 2000]).

3.6 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.

4 Tools for Information Extraction

4.1 Morphology related similarity

Denoted with morphology related similarity are relations between words that can first be noticed due to its regularity in word formation. Especially we find

- inflection
- derivation
- compound formation

Secondly we can identify groups of words due to low-frequent n-grams of characters, which might be considered as a weaker form of morphology.

Those methods can be used to identify words to belong to some sublanguage. Examples are

- names of chemical substances
- highly technical terms in general
- proper names of regional origin, for instance, Italian surnames compared to English words.

For classifying words based on morphological similarity, we use a trie-like data structure, the *affix-compression trie* that is trained on the characters of a word read from the beginning or reversed.

For training, a pair wise list of words and their classes is needed. A toy example for base form reduction can be found in table 2:

Table 2: training set for base form reduction. The semantics of the instruction: cut n characters (number) away and add the following text characters.

Word form	Reduction instruction
Price	0
Vice	0
Mice	3ouse
splice	0

Now we train a suffix compression trie on the reversed word forms. Nodes in the trie get all the reduction instructions assigned to them that can be found in the subnodes, together with their number of occurrences. The leaves of the trie correspond to one word form. Figure y shows the suffix compression trie for the training set of table 2:

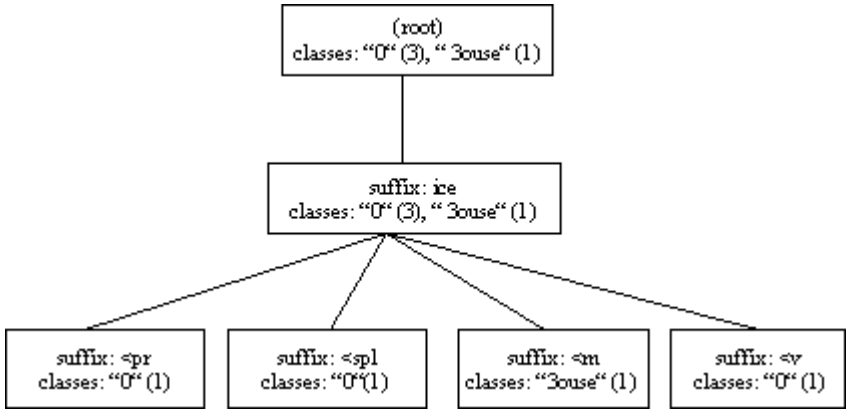

Figure 3: suffix compression trie. Note that the intermediate suffix node "ice" bears a bias of 3:1 for the reduction instruction "0". The start-of-word character is "<"

When classifying words with the suffix tree, the path through the trie is followed until no more node matches. In case of a word in the training set, the resulting node is

a leaf, for unknown words the node is usually at intermediate position from root to the leaves. For nodes with multiple reduction instruction classes, the instruction having the highest number of occurrences is selected.

In the example, when classifying *voice*, the resulting node is "ice" with the winning instruction "0". The suffix path for *police* is "ice"-"<spl" (partial match) with the same reduction instruction, while *half-mice* is correctly reduced to *half-mouse*.

4.2 Splitting subgraphs for lexical disambiguation

As the collocation analysis results in one large, connected graph with words as nodes and the collocation relationship as links between such nodes it is worth-while to have a closer look at the structure of this graph. First of all, standard measurements can be performed like the distribution of node degrees, the average shortest path length, clustering coefficient and others, see [Steyvers & Tenenbaum 2002] for a comprehensive work. Though from all these measurements it is possible to infer that the graph must have the small world property, described by [Strogatz 1998] and further developed by [Barabasi 2000], this knowledge can also be of practical use.

From the graph having the small world property it is possible to assume that the whole graph is structured into local clusters of words or 'communities' to borrow terminology from research on the Web as a graph [Kleinberg et al. 1999]. These clusters consist of words which belong together due to various semantic relations like cohyponymy, synonymy, antonomy and other as opposed to other words which are not in this cluster because there is no semantic relationship between most of the words from the cluster and the given other word.

Now it is obvious that if a word is ambiguous, it will occur in two or more such clusters at the same time, whereas there will be no other connections between those clusters as they otherwise denote different topics, see the visualized subgraph (simulated annealing, [Schmidt 1999]) example of the word *king* in figure 6. It is then possible to formulate the two following assumptions:

- The whole graph consists of clusters
- There is no such triplet of words which is still ambiguous

There are only very rare exceptions to the second assumption like 'gold', 'silver' and 'bronze' where having three words still is ambiguous.

Using these assumptions it was possible to formulate an algorithm, described in greater detail in [Bordag 2003], which can split the subgraph around a given word according to the different clusters, thus giving a semantic disambiguation of the word. The results of such a disambiguation can then be used for various purposes like information retrieval, other classification tasks and word sense disambiguation.

Figure 4 gives the fully unsupervised disambiguation of *driver*, having on the one hand the *conductor* reading, on the other hand being a piece of *software* for accessing connected hardware.

Sense Nr 1 : Car · Greyhound · Highway · Interstate · Mahoney · Patrol · Sgt. · Taxi · Trooper · accident · airbags · apparently · authorities · automatic · bags · belts · bomb · brake · brakes · bus · buses · cab · car · cars · chase · chased · collided · collision · crash · crashed · critically · crossing · crushed · dead · door · driven · driver's · drivers · driving · drove · drunk · drunken · engine · exploded [40 more]

Sense Nr 2 : 1-2-3 · 16-bit · 24-bit · 8-bit · 8514/A · ADI · ANSI · API · AUTOEXEC · Adapter · Adobe · Apple's · Apple-Talk · AutoCAD · Autodesk · BAT · BIOS · BallPoint · Ballpoint · BitBlt · Bus · CD-ROM · COM · CON · CONFIG · CTL · Chooser · DEVHLP · DEVICE · DEVICEHIGH · DEVLOD · DLL · DMA · DOS · DeskJet · DeskWriter · Device · Display · Drivers · EGA · EMM · EMS · EXE [200 more]

Figure 4: lexical disambiguation of *driver*.

4.3 The Pendulum

For extending sets of words that bear a certain semantic relation, [Quasthoff et al. 2002] describes a method that extracts first names and last names of persons from indexed, unannotated text. Using fuzzy pattern rules on very flat features, like "if there is a capitalized word behind a first name, it is likely to be a last name", the algorithm is able to extract in a bootstrapping fashion several thousands of person names from a small start set (20-50 examples are sufficient). High accuracy (about 98%, depending on language, rules and features) is assured through the iteration of a search and a verification step, resulting in accepting a name candidate only if the fuzzy rules match at a certain rate for all occurrences of the name candidate.

The Pendulum algorithm is applicable to word sets whose elements show up in certain patterns and has been successfully applied to all kinds of Named Entity subclassification, i.e. company names (see [Biemann et al. 2003a]) or island names.

Through the flatness of the features used the method is language independent in a way that most patterns are reflected in several languages and names already learnt on other language sources can be used as start sets for an new language. Moreover, patterns do not have to be handcrafted, but can be inferred from small training texts. The principle bootstrapping by search and verification can not only be applied on text as data source – experiments on POS-filtered collocations determining related concepts for given word sets are very promising.

4.4 Collocation set disjunction

The calculation of collocations can be iterated to obtain collocations of higher order in the following way: while the first-order calculation operates on sentences, the second order calculation operates on the outcome of the first order calculation and so forth. Intuitively, second-order collocations happen to be strong if two words appear in the same context and can be roughly compared to de Saussure's paradigmatic relations (cf. [de Saussure 1916], [Rapp 2002]).

However, second- or higher order collocations are in general not restrictive enough to derive synonymy or cohyonomy directly from them. But they can serve as a data source where other methods build upon. In [Biemann et al. 2003b] we describe and evaluate a method that yields good candidates for extending hierarchical synset-based lexicons like WordNet [Miller 1990] by performing set disjunction on the collocation sets of several input words that are close together in the hierarchy.

Figure 5 shows a German example, performing disjunction on the third-order collocation sets of two words after applying a word class filter.

start set: [warm, kalt] *[warm, cold]*
result set: [heiß, wärmer, kälter, erwärmt, gut, heißer, hoch, höher, niedriger, schlecht, frei] *[hot, warmer, colder, warmed, good, hotter, high, higher, lower, bad, free]*

start set: [gelb, rot] *[yellow, red]*
result set: [blau, grün, schwarz, grau, bunt, leuchtend, rötlich, braun, dunkel, rotbraun, weiß] *[blue, green, black, grey, colorful, bright, reddish, brown, dark, red-brown, white]*

start set: [Mörder, Killer] *[murderer, killer]*
result set: [Täter, Straftäter, Verbrecher, Kriegsverbrecher, Räuber, Terroristen, Mann, Mitglieder, Männer, Attentäter] *[offender, delinquent, criminal, war criminal, robber, terrorists, man, members, men, assassin]*

Figure 5: disjunction of third-order collocations. The original language in the experiment was German, English translations are marked in *italics*.

The introduction of part-of-speech information additionally allows a more precise selection of collocation sets: Using the sets of immediate left and right neighbor 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 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]). Possible remedies for this are query expansion techniques and collocation sets can be used for this.

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.

6 Further Research

After five years of being online, we register now more than 170'000 monthly visits and 4,7 Mio. page hits at 50% of yearly growth. Due to increasing access counts, we are currently developing a clustered storage and access infrastructure that will not only provide higher throughput for Web access but also a structural separation of production and presentation databases.

After setting up a language classifier, we will set up corpora in different standard sizes for all major languages on the web. See figure 7 for our web interface for Korean.

References

[Biemann et al. 2003a] Biemann, C., Quasthoff, U., Böhm, K., Wolff, C. (2003): Automatic discovery and Aggregation of Compound Names for the Use in Knowledge Representations, Journal of Universal Computer Science (JUCS), Volume 9, Number 6, Pp. 530-541, Juni 2003

[Biemann et al. 2003b] Biemann, C., Bordag, S., Quasthoff, U (2003): Lernen von paradigmatischen Relationen auf iterierten Kollokationen, Proceedings of GermeNet Workshop 2003, Tübingen, Germany

[Barabasi 2000] A.L. Barabasi et al . Scale-free characteristics of random networks: the topology of the World-wide web, Physica A (281)70-77, 2000

[Bordag 2002] Bordag, S. (2003): Sentence Co-occurrences as Small-World Graphs: A solution to Automatic Lexical Disambiguation, A. Gelbukh (Ed.): CICLing 2003, LNCS 2588, pp. 329-332, Springer-Verlag Berlin Heidelberg.

[Brants 2000] Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

[Davidson & Harel 1006] Davidson, R., Harel, D. (1996): Drawing Graphs Nicely Using Simulated Annealing, ACM Transactions on Graphics 15(4), 301-331.

[Jansen et al. 2000] 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.

[Kleinberg et al. 1999] Kleinberg, J., M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., S. (1999): The web as a graph: Measurements, models and methods. Lecture Notes in Computer Science, vol. 1627, pp.1-18.

[Läuter & Quasthoff 1999] 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.

[Miller 1990] Miller, G.A. (1990): Wordnet - an on-line lexical database, International Journal of Lexikography 3(4):235-312

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

[Quasthoff et al. 2002] Quasthoff, U., Biemann, C., Wolff, C. (2002): Named Entity Learning and Verification: EM in large Corpora, Proceedings of CoNNL-2002, Taipei, Taiwan

[Quasthoff et al. 2003] Quasthoff, U., Richter, M., Wolff, C., Medienalalyse und Visualisierung – Auswertung von Online-Pressetexten durch Text Mining, in Uta Seewald-Heeg (Ed.), Sprachtechnologie für die multilinguale Kommunikation, Proceedings of GLDV-03, Sankt Augustin

[Rapp 2002] Rapp, R. (2002): The Computation of Word Association: Comparing Syntagmatic and Paradigmatic Approaches, Proceedings of COLING-02, Taipei, Taiwan

[Saussure 1916] Saussure, F de. (1916): Cours de Linguistique Générale, Paris, Payot.

[Schmidt 1999] Schmidt, F. (1999): Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung, Diplomarbeit, Universität Leipzig.

[Silverstein et al. 1999] Silverstein, C. et al. (1999): Analysis of a Very Large Web Search Engine Query Log. In SIGIR Forum 33(1), 6-12.

[Steyvers & Tenenbaum 2002] M. Steyvers, J. B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. M. Steyvers, J. B. Tenenbaum, Cognitive Science, 2002

[Strogatz 1998] D. J. Watts, S.H. Strogatz. Collective dynamics of 'small-world' networks, Nature 393:440-442, 1998.

[Voorhees & Harman 1999] 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].

Appendix: Figures

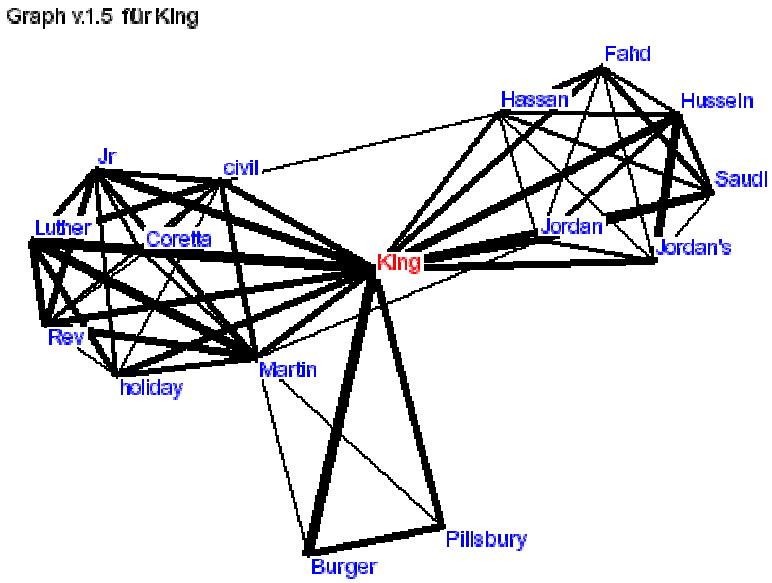

Figure 6: Collocation graph for *King* (English Corpus)

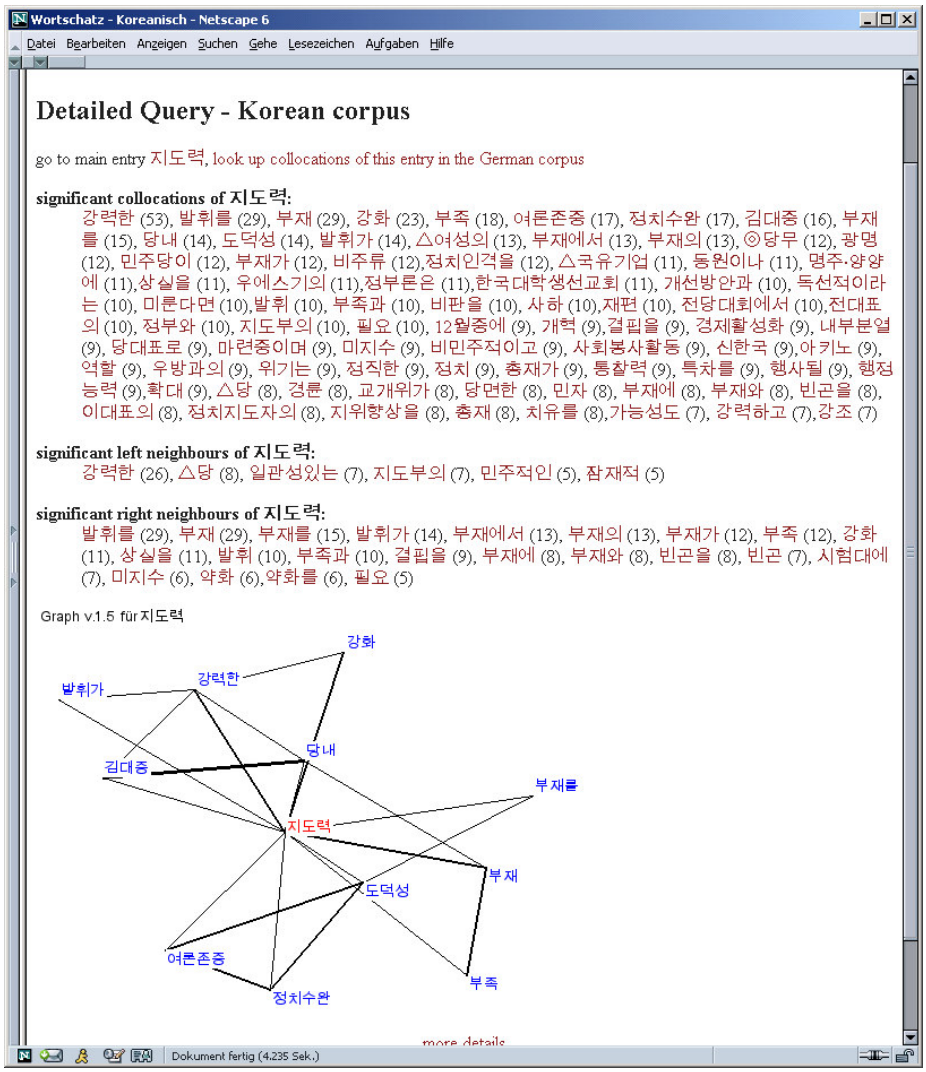

Figure 7: Web-interface for Korean: Collocations of *(leadership)*