Appeared in *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, pp. 312–317, 1997.

# Word Class Functions for Syntactic-Semantic Analysis

**Hermann Helbig** and **Sven Hartrumpf***
Applied Computer Science VII (AI)
University of Hagen
58084 Hagen, Germany
{Hermann.Helbig,Sven.Hartrumpf}@FernUni-Hagen.De

## Abstract

In this paper, Analysis with Word Class Functions (WCFA) is presented as a paradigm for syntactic-semantic analysis of natural language. The main characteristics of this approach are: word-orientation, the central role of word class functions, two phases of the activity of words, semantic orientation (priority of performance and acceptability vs. grammaticality), and incremental processing. The increments in this processing correspond to semantic kernels representing constituents which have been already *understood* during analysis. After a short comparison with other word-oriented approaches to language analysis, we present a WCF machine (WCFM) which has been implemented in Lisp and is used as an NLI to bibliographic databases and multimedia databases. The four levels of processing which the machine distinguishes are explained. Finally, we present some directions for further research.

## 1 Introduction

### 1.1 Analysis with word class functions (WCFA)

The paradigm of **analysis with word class functions** (**WCFA**) can be characterized in its current state by five aspects: First, WCFA is word-oriented, i. e., the word plays the central role in language understanding. So we avoid stipulating a grammar in the traditional sense. This has advantages for incremental parsing and for robust parsing as the description of a word class inherently includes information about how to deal with incomplete or ungrammatical sentences.

Second, the concept of **word class functions** (**WCFs**) is important. A WCF corresponds to a large degree to the traditional parts of speech like adjective and adverb. However, the classification in WCFA must be finer in order to account for differences in syntax and semantics. For example, there are WCFs for positive, comparative, and superlative forms of adjectives. A WCF may even be restricted to one lexeme.[1]

Third, a WCF consists of two phases or acts that are psycholinguistically motivated. When a word in a sentence is first encountered, it leads to certain expectations about what may follow in the sentence. This *opening* of expectations is handled by the **opening act** of a WCF. The second act of a WCF, the **completing act**, completes expectations and refines the syntactic and semantic information the word contributes to a larger phrase. This is done after the expected words or constituents have been analyzed and semantically structured.[2]

Fourth, WCFA is semantically oriented. Therefore it is more important to build a semantic representation of acceptable sentences than to realize a perfect grammaticality check. As acceptability is more useful for applications like NLIs and spoken language processing, this priority seems to be sufficiently justified.

Fifth, WCFA completes the semantic representation of a **semantic kernel** immediately after reaching its end. A semantic kernel is any phrase which is a syntactic phrase and can be understood as a semantic unit, e. g., NPs and relative clauses. The extent of increments in WCFA are these semantic kernels. Incrementality is an aim taken from cognitive linguistics, which (Hemforth 93, pp. 127–170) has investigated with respect to case and semantic rôle information, for instance. One practical advantage is that incrementality avoids many ambiguities during analysis that can be disambiguated locally on semantic grounds.

### 1.2 Other word-oriented approaches to language analysis

We restrict our overview to three word-oriented paradigms for language analysis. One difference between WCFA and these other approaches is that, in WCFA, a word is active during two phases (acts) which open expectations and saturate expectations, respectively.

*Word Expert Parsing* (WEP) developed by

[*] This research has been supported by the *Virtuelle Wissensfabrik* (*Virtual Knowledge Factory*), a project of the German state Nordrhein-Westfalen.

[1] In this paper, we restrict ourselves to WCFs and examples for German. However, WCFA can be transferred to other languages, cf. section 7.

[2] This is one reason why opening and completing acts do not correspond to active and inactive edges in chart parsing. We want to thank an anonymous reviewer for pointing out to us this possible misunderstanding.

(Small 81; Small & Rieger 82; Small 87) views language understanding as a distributed process of interacting words. Each word is connected with an expert process that *actively pursues its intended meaning in the context of other word experts and real-world knowledge* (Small 87, p. 161). In contrast, WCFA describes the grammatical functions of whole classes of words.

The approach presented by (Eimermacher 88) is based on the word expert paradigm. It distinguishes between word class experts representing general grammar rules and word experts for analyzing the relations between single lexemes. In WCFA, there are WCFs for single words, too. However, they are not only used for the analysis of idioms, but also for the (semantic) interpretation of single words, e. g., specific prepositions and conjunctions.

The *ParseTalk* approach developed by (Bröker *et al.* 94) is a concurrent, object-oriented parsing system. It employs a different paradigm for analysis control using *word actors.*

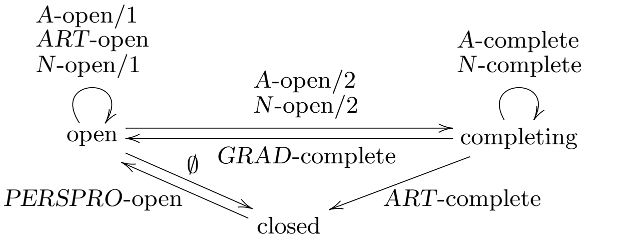
## 2 Basic concepts of WCFA

WCFs are controlled by a central unit, the **WCF machine** (**WCFM**). This machine is described in an earlier version by (Helbig 94) who bases some ideas for word class functions on (Helbig 86). The WCFM is a kind of nondeterministic automaton with three states: *open*, *completing*, *closed*. When the machine starts to analyze a sentence, it is in state *open*. In this state, the WCF for the next word to be processed is determined (after a morphological-lexical analysis) and the opening act of this WCF is called. The result of this act (a typed feature structure representing the expectations or valencies of the word under analysis) is stored in the **analysis memory** (**AM**).

If the WCFM is in state *completing*, it saturates grammatical and/or semantic expectations. This is done by calling the completing act of the WCF of the word whose expected constituents/words (e. g., subcategorized constituents) have been provided by the analysis.

The third state of the WCFM, the state *closed*, indicates that the analysis of a constituent has just been completed. In this state, the result is marked as *completed +* and can be used as a filler for valencies of other words or constituents.

The state of the WCFM is changed by opening and completing acts of WCFs. Figure 1 shows the main state transitions within simple NPs. Nota bene: the WCFM is *not* an augmented transition network; this figure just illustrates how the acts of WCFs change the states of the WCFM.



**Notes**
In case of an adjective under analysis, transition *A*-open/2 is applicable if AM's top-most element is a *GRAD* (graduator, adverb of degree); in all other cases, *A*-open/1 is applicable. Similarly, transition *N*-open/1 is used if an apposition follows the noun; otherwise, *N*-open/2 is used.

Figure 1: Transition diagram for WCFM's states during NP analysis

WCFA analyzes sentences on four levels, which will be explained in turn in sections 3 to 6:

**Level 1** analysis of elementary kernels (e. g., simple NPs containing only one nominal head)

**Level 2** analysis of complex kernels (e. g., complex NPs containing more than one nominal head or including a comparative or superlative construction)

**Level 3** analysis of simple propositions (phrase containing only one main verb)

**Level 4** analysis of complex propositions (phrase with more than one simple proposition)

Figure 2 indicates how these levels correspond to different sets of WCFs and which components can be distinguished in our Lisp implementation of WCFA. We will concentrate on the WCFM and WCFs in the following sections; the dialog model, discourse model, and morphological-lexical analysis will not be treated in this paper.

## 3 Analyzing simple NPs

On level 1, simple NPs are analyzed syntactically and described by a semantic structure (which may incrementally be changed later on). The basic algorithm for level 1 of the WCFM, which is centered around the WCFM's three states, is given in Figure 3. To exemplify the working of the algorithm, we consider sentence (1).

(1) *Der unbekannte Autor schrieb ein*
The unknown author wrote a
*sehr gutes Buch über die Berge*
very good book about the mountains
*in Georgia.*
in Georgia.

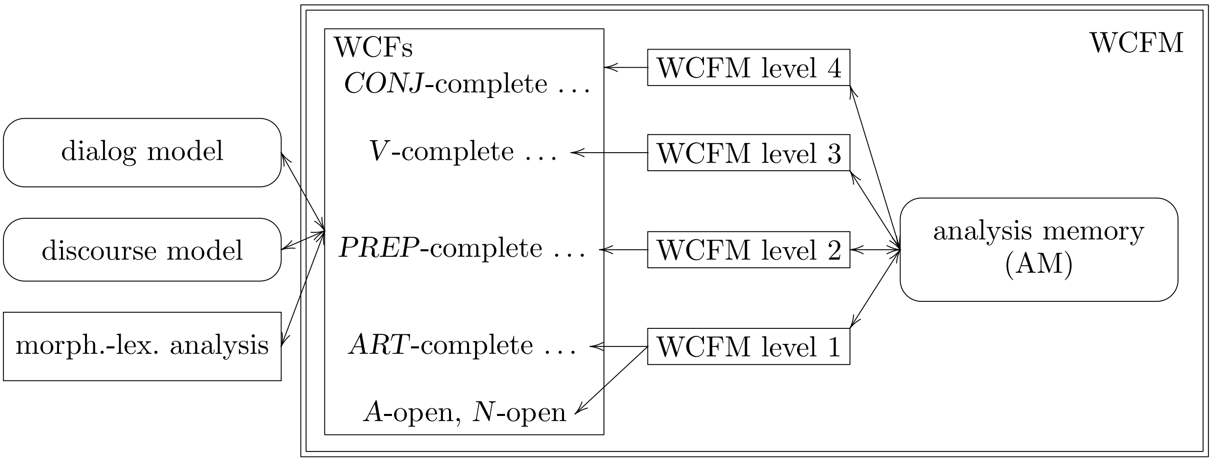


Figure 2: Architecture of a WCFA implementation

| Step | Status | New Word | Act | Work Done by Act |
|---|---|---|---|---|
| 1 | open | *der* (the) | *ART*-open | opens the expectations of an NP with specific morphosyntactic feature values for case, number, gender, declension type for possible adjectives, etc. |
| 2 | open | *unbekannte* (unknown) | *A*-open | restricts the expectations of an NP according to the complex agreement feature and semantic information of the adjective *unbekannte* |
| 3 | open | *Autor* (author) | *N*-open | generates a semantic representative $c_1$ which is an instance of the lexical concept *Autor* |
| 4 | completing | | *A-COMPLETE* | saturates the expectations raised in step 2; extends $c_1$ by the property *unbekannt* |
| 5 | completing | | *ART*-complete | saturates the expectations raised in step 1 |
| 6 | closed | | | completes the semantic kernel $c_1$ |
| 7 | open | *schrieb* (wrote) | *V*-open | opens the expectations that the verb *schreiben* raises; generates a semantic representative $c_2$ as a situation instance of *schreiben* |
| 8 | open | *ein* (a) | *ART*-open | 8–15 similar to first NP (1–6) except for the presence of an adverb of degree, 9–11 |
| 9 | open | *sehr* (very) | *GRAD*-open | opens the expectation of an AP |
| 10 | open | *gutes* (good) | *A*-open | |
| 11 | completing | | *GRAD*-complete | saturates expectations raised in step 9 |
| 12 | open | *Buch* (book) | *N*-open | generates a semantic representative $c_3$ |
| 13 | completing | | *A*-complete | 13–15 similar to 4–6 |
| 14 | completing | | *ART*-complete | |
| 15 | closed | | | |
| 16 | open | *über* (about) | *P*-open | opens the expectation of a PP |
| 17 | open | *die* (the) | *ART*-open | 17–20 similar to 1–6 except for the absence of adjectives |
| 18 | open | *Berge* (mountains) | *N*-open | generates a semantic representative $c_4$ |
| 19 | completing | | *ART*-complete | |
| 20 | closed | | | |
| 21 | open | *in* | *P*-open | 21–23 similar to 16–20 |
| 22 | open | *Georgia* | *N*-open | generates a semantic representative $c_5$ |
| 23 | closed | | | |

Table 1: Analysis process of sentence (1) on level 1

```
add a left sentence border to the AM
status := open
next_word := first(word_list)
pop(word_list)
next_word_fs := morphological-lexical analysis of
  next_word
while status ≠ done do
  current_word := next_word
  current_word_fs := next_word_fs
  next_word := first(word_list)
  pop(word_list)
  next_word_fs := morphological-lexical analysis of
    next_word
  result = call opening act for current_word
  if break recognized by opening act
    recursive call of WCFM for embedded clause
    integrate analysis result for embedded clause into
      result
    add result to the AM
  fi
  repeat
    case status of
      open:
        add result to the AM
      completing:
        set head and nonhead according to the relation
          of the top-most elements of the AM
        call completing act of nonhead with arguments
          head and nonhead
        remove nonhead from the AM
      closed:
        mark top-most element of AM as completed +
        status := open
    esac
  until status = open
  if word_list indicates end of sentence or clause
    done := true
  fi
od
add a right sentence border to the AM
```

Figure 3: Basic algorithm for level 1 of the WCFM

$$\begin{array}{ll}
c_1 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM}\mid\text{NET} & \big(\text{sub}(c_1,\ \text{Autor})\big)\end{bmatrix}\\
\text{unbekannt} & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{a}\end{bmatrix}\\
\text{der} & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{art}\end{bmatrix}
\end{array}$$

$$\begin{array}{ll}
c_1 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{LAY}\mid\text{REFER} & \text{det}\\ \text{NET} & \left(\begin{array}{l}\text{sub}(c_1,\ \text{Autor})\\ \text{prop}(c_1,\ \text{unbekannt})\end{array}\right)\end{bmatrix}\end{bmatrix}
\end{array}$$

$$\begin{array}{ll}
c_5 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{LAY}\mid\text{REFER} & \text{det}\\ \text{NET} & \big(\text{name}(c_5,\ \text{Georgia})\big)\end{bmatrix}\end{bmatrix}\\
\text{in} & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{p}\end{bmatrix}\\
c_4 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{LAY}\mid\text{REFER} & \text{det}\\ \text{NET} & \big(\text{pred}(c_4,\ \text{Berg})\big)\end{bmatrix}\end{bmatrix}\\
\text{über} & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{p}\end{bmatrix}\\
c_3 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{LAY}\mid\text{REFER} & \text{indet}\\ \text{NET} & \left(\begin{array}{l}\text{sub}(c_1,\ \text{Buch})\\ \text{prop}(c_3,c_6)\\ \text{modp}(c_6,\ \text{sehr},\ \text{gut})\end{array}\right)\\ \multicolumn{2}{l}{(\textit{last element is actually}}\\ \multicolumn{2}{l}{\textit{at concept node } c_6)}\end{bmatrix}\end{bmatrix}\\
c_2 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{v}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{NET} & \left(\begin{array}{l}\text{subs}(c_2,\ \text{schreiben})\\ \text{temp}(c_2,\ \text{PAST})\end{array}\right)\end{bmatrix}\end{bmatrix}\\
c_1 & \begin{bmatrix}\text{SYN}\mid\text{CAT} & \text{n}\\ \text{SEMSEL}\mid\text{SEM} & \begin{bmatrix}\text{LAY}\mid\text{REFER} & \text{det}\\ \text{NET} & \left(\begin{array}{l}\text{sub}(c_1,\ \text{Autor})\\ \text{prop}(c_1,\ \text{unbekannt})\end{array}\right)\end{bmatrix}\end{bmatrix}
\end{array}$$

Figure 4: Contents of the AM after step 3 (top box), 6 (middle box), and 23 (bottom box) with parts of feature structures

Table 1 shows how this sentence is processed by the algorithm for WCFM's level 1. To illustrate the role of the AM, we show its contents after step 3 in the top box of Figure 4. Each semantic kernel is associated with a typed feature structure containing morphological, syntactic, and semantic information. Nodes in the semantic representation of constituents that are generated during analysis have indexed names like $c_2$. In our WCFA realization, we use multi-layered extended semantic networks (MESNET), a paradigm described by (Helbig & Schulz 97), to represent the meaning of natural language expressions. MESNET is based on a multidimensional classification of concepts (the nodes of the semantic network) and structural information expressed by means of a predefined set of semantically primitive relations and functions (the arcs of the semantic network). In addition, there are representational means for encapsulating concepts or subnetworks and for describing quantification and modalities (especially for negation). The representation of a semantic kernel is stored under the set-valued feature SEMSEL | SEM | NET. Some values of semantic features may have to be overwritten later on, e. g., the information that a node has a determinate reference (LAY | REFER det).

Step 4 and 5 are the completing acts of the topmost nonhead elements. Completing acts combine nonheads with their heads. After these completing acts for the adjective *unbekannt* and the definite determiner *der*, the AM has the contents shown in the middle of Figure 4. On level 1, the WCFM produces for sentence (1) the result given at the bottom of Figure 4.

The analysis process of simple NPs is mostly straightforward. However, it is interrupted if a complex constituent occurs inside the analyzed NP; in German, these interrupting constituents can be participle constructions (cf. sentence (2)), for example.

(2) Der ein *teures* *Rad* *besitzende*
The an expensive bike owning
*Schüler* *stürzte.*
pupil crashed.
'The pupil who owns an expensive racing bike crashed.'

The nested structures are detected in WCFA because of a sequence of words whose categories are never next to each other in an NP (in sentence (2) two consecutive determiners) or because of agreement violations. The nested phrase is analyzed by a recursive call of the WCFM and then integrated into the surrounding phrase.

In addition to such participle constructions, the coordination of constituents leads to breaks in the analysis process in WCFA. These breaks are indicated in German by commas or conjunctions like *und* (*and*). Constituent coordinations, which have to be distinguished from sentence coordinations, cause an immediate call to the completing act of the WCF *COORD*.

A third phenomenon that breaks the straight analysis process of German NPs are relative clauses as shown by sentence (3).

(3) *Der* *Junge,* *der* *das* *Fenster*
The boy, who the window
*zerschlagen* *hatte,* *nahm* *die*
smashed had, took the
*Spielsachen.*
toys.
'The boy, who had smashed the window, took the toys.'

As the comma is regarded as a special word class, the opening act of the WCF *COMMA* can handle these cases by recursively calling the WCFM for a subordinate clause. The result of this call for a relative clause is combined with the semantic analysis of the relative pronoun's antecedent by the completing act of *COMMA*. For other embedded clauses, the result will be handled by the completing act of the verb of the superordinate clause.

## 4 Analyzing complex NPs

On the second level, elementary kernels (semantic concepts roughly corresponding to simple NPs) are combined to complex kernels. These complex kernels are NPs modified by PPs (*das Buch über Napoleon/the book about Napoleon*), comparative constructions (*größer als Peter/taller than Peter*), and superlative constructions (*der größte unter den Studenten/the tallest among the students*).

The most difficult problem on this level is to decide whether to attach a PP to the verb or some preceding NP. A WCFA solution for this well-known problem of PP attachment is given by (Helbig *et al.* 94).

On level 2 of WCFA, the semantic analysis of comparative and superlative constructions is finished by finding the comparison frames that are present in all such phrases (at least implicitly). This is done by the completing acts of the WCFs *COMP* and *SUPER*, respectively. These acts have to cope with the linguistic fact that comparative and superlative phrases can be discontinuous in German as example (4) indicates.

(4) *Weil* *Maria* [am schnellsten]$_{AdvP1}$
Because Maria fastest
*rechnet* [von allen Studenten]$_{AdvP2}$,
calculates of all students,
*wird* *sie* *gelobt.*
is she praised.
'Maria is praised because she calculates fastest among all students.'

The opening acts of these WCFs are called on level 1 for all adjectives that start such comparative and superlative constructions.

Two aspects of level 2 can't be treated in this paper due to its limited length. First, the completing acts for main verbs, for auxiliary verbs, and (in German) separated verb prefixes, combine their analysis results. Second, antecedents for referential expressions like pronouns, pronominal adverbs, etc. are searched by the corresponding WCFs. If there are several possible antecedents the most likely one is chosen and the remaining alternatives are stored in a priority queue to allow backtracking.

## 5 Analyzing simple propositions

Level 3 of WCFA deals with attaching simple or complex kernels to main relators like the main verb of a sentence. As in valency theory, WCFA distinguishes (syntactically) mandatory complements, (syntactically) optional complements, and adjuncts. The algorithm for level 3 tries to classify possible complements and adjuncts so that all mandatory complements are filled (possibly by ellipsis resolution). As in HPSG ((Pollard & Sag 94)) and other lexicalized grammar formalisms, the subcategorization frame is central for this decision. The multivalent lexicon COLEX (cf. (Schulz & Helbig 96)), which we use for our WCFA implementation, contains for each element of a valency schema syntactic information (category, case/preposition, optionality, etc.) and rich semantic information (the semantic relation (cognitive rôle); semantic sort and (possibly underspecified) values of semantic features taken from a predefined ontology and a set of 16 boolean features, respectively; cf. (Helbig & Schulz 97)).

The set of possible adjuncts characterized by their semantic rôles is also given by a feature. In contrast to the semantics of complements, the semantics of adjuncts in a sentence is not lexicalized. Instead, the completing acts for relevant WCFs like *V* interpret adjuncts based on the semantics of the main relator, the kind of preposition (in case of a PP), and the semantic sort and features of the NP.

## 6 Analyzing complex propositions

On level 4 of WCFA, complex propositions are analyzed. We subsume under this term the subordination and coordination of clauses and all kinds of modal information which can typically be realized as modal verbs (*should*, etc.), or modal adverbs (*perhaps*, etc.). We will restrict our description to subordination and coordination of clauses.

The WCFM analyzes the first clause of a clause coordination or subordination; its end is indicated by a comma and a conjunction (WCF *CONJ*) or subjunction (WCF *SUBJ*) or another intervening element. Then, the completing act of the intervening element calls the WCFM recursively to analyze the second clause. The propositions resulting from the analysis of both clauses are connected semantically by the completing act of the intervening element. The completing act for the German subjunction *weil* (*because*), a specialization of the completing act of the WCF *SUBJ*, leads (for the semantic representation language MESNET) to the semantic relation *caus* (the relation between the causing and the caused situation), for example. This act generates a semantic representation for the whole sentence which replaces the two clause representatives in the AM.

## 7 Perspectives

We see several directions for further research. One can transfer our results for German most easily to languages with mostly continuous constituents (many Germanic and Romance languages). It would be interesting to investigate how different degrees of freedom of word order can be accounted for best in WCFA.

Another linguistic aim would be to enhance the robustness of the syntactic-semantic analysis, as WCFA has already shown to be well-suited for this aim (cf. section 1.1). One specific question would be how (and on what levels) information about dealing with agreement errors can be expressed and used best.

Finally, an interesting question is how much knowledge can be shared by related WCFs (like determiners and possessive pronouns) using multiple default inheritance.

## References

(Bröker *et al.* 94) Norbert Bröker, Udo Hahn, and Susanne Schacht. Concurrent lexicalized dependency parsing: The ParseTalk model. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, 1994.

(Eimermacher 88) Michael Eimermacher. *Wortorientiertes Parsen.* Unpublished PhD thesis, TU Berlin, Berlin, 1988.

(Helbig & Schulz 97) Hermann Helbig and Marion Schulz. Knowledge representation with MESNET: A multilayered extended semantic network. In *Proceedings of the AAAI Spring Symposium on Ontological Engineering*, pages 64–72, Stanford, CA, 1997.

(Helbig 86) Hermann Helbig. Syntactic-semantic analysis of natural language by a new word-class controlled functional analysis. *Computers and Artificial Intelligence*, 5(1):53–59, 1986.

(Helbig 94) Hermann Helbig. Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung – Teil II – Die vier Verarbeitungsstufen. Informatik-Bericht 159, FernUniversität Hagen, Hagen, 1994.

(Helbig *et al.* 94) Hermann Helbig, Andreas Mertens, and Marion Schulz. Die Rolle des Lexikons bei der Disambiguierung. In Harald Trost, editor, *KONVENS-94 – Verarbeitung natürlicher Sprache*, pages 151–160, Berlin, 1994.

(Hemforth 93) Barbara Hemforth. *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens*, volume 40 of *Dissertationen zur Künstlichen Intelligenz (DISKI)*. Infix, Sankt Augustin, 1993.

(Pollard & Sag 94) Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar.* Studies in Contemporary Linguistics. University of Chicago Press, Chicago, Illinois, 1994.

(Schulz & Helbig 96) Marion Schulz and Hermann Helbig. COLEX: Ein Computerlexikon für die automatische Sprachverarbeitung. Informatik-Bericht 210, FernUniversität Hagen, Hagen, Germany, December 1996.

(Small & Rieger 82) Steven Small and Chuck Rieger. Parsing and comprehending with word experts (a theory and its realization). In Wendy Lehnert and Martin Ringle, editors, *Strategies for Natural Language Processing*, pages 89–147. Erlbaum, Hillsdale, 1982.

(Small 81) Steven Small. Viewing word expert parsing as linguistic theory. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI-81)*, pages 70–76, 1981.

(Small 87) Steven Small. A distributed word-based approach to parsing. In Leonard Bolc, editor, *Natural Language Parsing Systems*, pages 161–201. Springer, Berlin, 1987.