THE WORD CLASS AGENT MACHINE[1]

Hermann Helbig and Andreas Mertens

FernUniversität Hagen
Praktische Informatik VII/Artificial Intelligence
D–58084 Hagen
Germany

E–mail:
hermann.helbig@fernuni–hagen.de
andreas.mertens@fernuni–hagen.de

ABSTRACT

Natural language processing (NLP) is often based on declaratively represented grammars with emphasis on the competence of an ideal speaker/hearer. In contrast to these broadly used methods, we present a procedural and performance–oriented approach to the analysis of natural language expressions. The method is based on the idea that each word–class can be connected with a functional construct, the so–called word agent, processing its own part of speech. The syntactic–semantic analysis performed by the word–class agents is surveyed by a *Word Class Agent Machine* (WCAM) used in the bibliographic information retrieval system LINAS at the University of Hagen.

The language processing is organized into four main levels: on the first and second level, elementary and complex semantic constituents are created which have their correspondence in mental kernels built during human language understanding. On the third level, these kernels are used to construct the semantic representation of the propositional kernel of a sentence. During this process, the subcategorization features of verbs and prepositions play the main part. The task of the last level consists in the syntactic analysis and semantic interpretation of modal operators in the broadest sense and of the coordinating or subordinating conjunctions which connect the propositional kernels.

[1]This paper has been presented at the *Eighth Twente Workshop on Language Technology 1994 (TWLT8)*. See [Boves, Nijholt 94].

CONTENTS

LIST OF FIGURES

LIST OF TABLES

THE WORD CLASS AGENT MACHINE

Hermann Helbig and Andreas Mertens

FernUniversität Hagen
Praktische Informatik VII/Artificial Intelligence
D–58084 Hagen
Germany

E–mail:
hermann.helbig@fernuni–hagen.de
andreas.mertens@fernuni–hagen.de

ABSTRACT

Natural language processing (NLP) is often based on declaratively represented grammars with emphasis on the competence of an ideal speaker/hearer. In contrast to these broadly used methods, we present a procedural and performance–oriented approach to the analysis of natural language expressions. The method is based on the idea that each word–class can be connected with a functional construct, the so–called word agent, processing its own part of speech. The syntactic–semantic analysis performed by the word–class agents is surveyed by a *Word Class Agent Machine* (WCAM) used in the bibliographic information retrieval system LINAS at the University of Hagen.

The language processing is organized into four main levels: on the first and second level, elementary and complex semantic constituents are created which have their correspondence in mental kernels built during human language understanding. On the third level, these kernels are used to construct the semantic representation of the propositional kernel of a sentence. During this process, the subcategorization features of verbs and prepositions play the main part. The task of the last level consists in the syntactic analysis and semantic interpretation of modal operators in the broadest sense and of the coordinating or subordinating conjunctions which connect the propositional kernels.

1 INTRODUCTION

Although word–based natural language processing is not a widely used method today, it has been of growing interest for a couple of years. Some examples of word–based approaches are the following: the *Word Expert Parsing* [Small 81], the *Word–Class Controlled Functional Analysis* [Helbig 86], the *Word–Oriented Parsing* [Eimermacher 86] and the analysis with *Word Actors* [Bröker et al. 93].

The theory of word expert parsing (WEP) approaches the understanding of natural

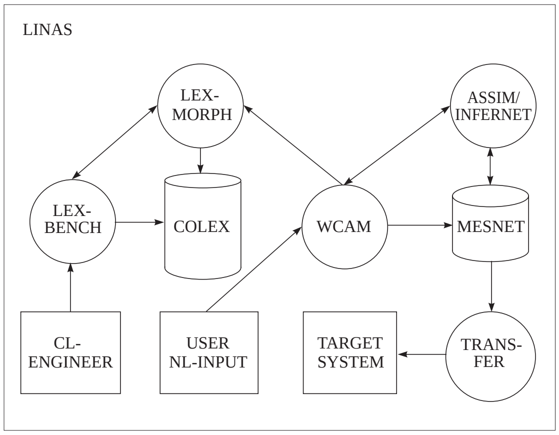

Figure 1: Overview of the NL system LINAS

language as a distributed process of interacting words. Each word is connected with an expert process which "actively pursues its intended meaning in the context of other word experts and real–world knowledge" [Small 87].

Eimermacher's word–oriented parsing relies upon the word expert paradigm, but distinguishes between word–class experts representing general grammar rules and word experts analysing the relations between single words [Eimermacher 88].

The *ParseTalk* model is a concurrent, object–oriented parsing system with grammatical knowledge completely integrated into the lexicon. Each word of a sentence the parser is currently working on activates a word actor which communicates with other initialized word actors and other components of the system [Bröker et al. 94].

Within the model of word–class controlled functional analysis (WCFA), experts are introduced for each word–class and the grammatical function of a word–class is given by a procedural specification. The characteristic feature of the word–class functions discerning them from other approaches is their partition into two different functional components which are activated at different times of the sentence analysis. In this paper, the essential aspects of the word–class agent machine (WCAM), which is a further development of the WCFA, will be presented. For more detailed information of this approach see [Helbig, Mertens 94]. A formal description of the four main levels of the word agent based language processing can be found in [Helbig 94].

2 THE LINAS PROJECT

The LINAS project aims at developing a natural language understanding system and, as a practical side effect, at providing a natural language interface to bibliographic databases. An architectural overview of those system components which are relevant to NLP is shown in Figure 1. The following major components are distinguished: an interactive lexicographer's workbench (LEXBENCH), a morphological processor (LEXMORPH), a word–class agent machine (WCAM), a knowledge assimilation and inference component (ASSIM/INFERNET), and an interface module to database systems (TRANSFER). In addition, there are two main knowledge sources: a computer lexicon (COLEX) and a network knowledge base (MESNET).

For each word of a sentence the WCAM is currently working on the morphological analysis (LEXMORPH) is activated. If the current word form has no lexical entry, the corresponding entry (the stem) can be determined by cutting off inflectional syllables.

The information included in the basic entry such as stem, grammatical specification and the accompanying word–class agent will be returned to the WCAM.

word-class agent	characterization
*ART	articles
*ADJ	adjectives
*NOM	nouns
*VB	verbs
*PREP	prepositions
*AV	auxiliary verbs
*IPATTR	interrogative pronouns (in attributive use)
*IPNOM	interrogative pronouns (in nominal use)
*ADV	adverbs
*GRAD	graduators (adverbs of degree)
*POSSPR	possessive pronouns
*RELPR	relative pronouns
*PART	participles
*COMP	comparative forms
*NUM	numerals
*NEG	negative forms
*CONJ	conjunctions
*COMMA	commas
*STOP	full stop

Table 1: Some examples of word–class agents

The aim of the WCAM is to analyse the NL input and to represent it by means of a multi–layered extended semantic network (MESNET). The analysis is done by a set of word–class agents (cf. Table 1) and the result produced by the agents is stored in the network knowledge base MESNET. The assimilation of the knowledge is organized by the ASSIM/INFERNET component. In the case of a question to a database management system as target sytem (not to the knowledge base of LINAS itself), the semantic description of the NL expression is translated into the database query language. This is done by the translation and interface module TRANSFER.

The lexical entries of the computer lexicon (COLEX) are organized as feature structures. For typical entries see Figures 3 and 4. LEXBENCH is an interactive workbench supporting the process of creating and checking lexical entries.

3 THE WORD CLASS AGENT MACHINE

3.1 WORD CLASS AGENTS

As well as in the theory of WEP, the basic idea of the WCAM is that words play the main part in NLP. In contrast to WEP, which provides one *expert* for each word, we present a word–class oriented approach. For each word–class its grammatical function is defined by a procedural specification, the so–called word–class agent, and each element of a word–class will be associated with the corresponding procedure. In LINAS more than 35 word–class agents are distinguished. Some of them are listed in Table 1. In addition to agents such as *ART (article), *ADJ (adjective), *NOM (noun) and *VB (verb), which are suggestive of the word–classes in traditional grammars, there are other agents such as *IPATTR and *IPNOM, which represent interrogative pronouns in attributive and nominal use, respectively. Furthermore, there are special agents responsible for the grammatical function of punctuation marks such as *COMMA (comma) and *STOP (full stop).

3.2 OPEN ACTS AND COMPLETE ACTS

The word–class agents are divided into an OPEN–act and a COMPLETE–act.[2] This division corresponds to the two major activity modes of a word during NLP: on the one hand, a word opens certain valencies which may be satisfied later on by other words or constituents of the current sentence. On the other hand, the valencies opened in the first activity mode must be satisfied by suitable candidates. These are the words or constituents which have been already included in the analysis and which saturate the opened valencies. In Artificial Intelligence (AI) the specifying of conditions (slots) which must be filled by specific data (filler) satisfying these conditions is known as the *slot–filler–mechanism* of frames.

The two activity modes of a word are represented by different procedural specifications: the OPEN–act and the COMPLETE–act of a word–class agent. According to the special tasks of OPEN–act and COMPLETE–act, the two acts are triggered by the WCAM at different moments in the analysis.[3] The division into an OPEN–act and a COMPLETE–act is one of the main differences to the other word–oriented approaches mentioned in Section 1. Another characteristic feature of the WCAM approach is the integration of semantic interpretation processes into the COMPLETE–acts of the word–class agents.

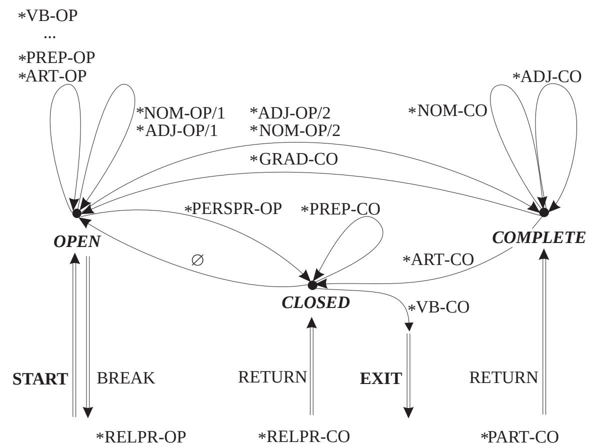

Figure 2: WCAM transition diagram

3.3 WCAM STATES

In Figure 2, the WCAM transition diagram is shown. The three states of the WCAM given in this Figure have the following meaning:

- *OPEN* – if a word is analysed by the WCAM, first the morphological component is activated in order to generate the feature structure of the corresponding lexical entry. Then this structure is passed on to the WCAM. If the current word opens valencies, the semantic de-

[2] This subdivision, however, is not made in each case. There are some word–class agents such as *ADV (adverb) and *IPNOM (the interrogative pronoun in nominal use) without a COMPLETE–act. The elements of these word–classes do not open valencies. Their part is to saturate the valencies of other words in the current sentence.

[3] The OPEN–act and the COMPLETE–act of a word–class agent are marked by extending the name of the agent with one of the two endings –OP or –CO, respectively (cf. the WCAM transition diagram in Figure 2).

scription of the word gets the marker OP (open) before it is stored in the working memory. Otherwise, if the word does not open valencies (for instance, in the case of adverbs, proper names, etc.), the description of the word is marked by CL (closed).

- *COMPLETE* – in this state the COMPLETE–act of a word–class agent is activated in order to satisfy the valencies being opened by the corresponding OPEN–act. Thus, in this state the decision that a group of words forms a particular constituent is made. For example, coming to the end of a nominal phrase (NP), the *ART–CO–act is triggered in order to saturate the valencies being opened before in the *ART–OP–act. As a result, an NP–constituent will be constructed and its complex syntactic–semantic structure will be stored in the working memory.
- *CLOSED* – this state of the WCAM indicates the fact that the current constituent is completely analysed with all its valencies being saturated. The semantic structure of this constituent is marked by CL, now being permitted to satisfy the valencies of other constituents. At the end of a sentence, indicated by a triggered *VB–CO–act, the valencies of the verb will be satisfied. Otherwise, if there are no COMPLETE–acts being left to perform, the WCAM immediately changes into the OPEN–state (cf. the Ø–transition in Figure 2).

$$\begin{bmatrix} \mathit{MORPH} & \begin{bmatrix} \mathit{WCA} & \mathit{*NOM} \\ \mathit{GEN} & \mathit{N} \\ \mathit{FLEX} & \mathit{S11} \end{bmatrix} \\ \mathit{SELECT} & \begin{bmatrix} \mathit{SEMREL} & \mathit{THM} \\ \mathit{WCA} & \mathit{*PREP} \\ \mathit{LEX} & \textit{ÜBER} \\ \mathit{CAS} & \mathit{4} \end{bmatrix} \\ \mathit{SEMREL} & [\,\mathit{SUB} \;\; \mathit{DRUCKERZEUGNIS}\,] \\ \mathit{SEMSORT} & \mathit{IF} \end{bmatrix}$$

Figure 3: Lexical entry of *Buch* (*book*)

3.4 AN EXAMPLE

The word–class agents are triggered and inspected by the WCAM. Depending on the word the WCAM is currently analysing, the corresponding word–class agent is activated. For example, consider the following sentence consisting of the articles *der* and *ein*, the adjective *jung*, the nouns *Mann* and *Buch*, and the verb *schreiben*:

- *Der junge Mann schrieb ein Buch.*
- ○ *The young man wrote a book.*

The initial state of the WCAM is OPEN (cf. Figure 2). When the first word *der* of the given sentence is included in the analysis, the WCAM activates the OPEN–act of the word–class agent *ART, which is called *ART–OP. The system remains in the OPEN–state. Analysing the next word, i.e. the adjective *jung*, the WCAM triggers the *ADJ–OP–act. If there is no graduator[4] between the article and the ad-

[4] For instance, if the article is followed by an adverb of degree such as *sehr* (*very*), the *ADJ–OP/2–act is triggered and the WCAM changes to the COMPLETE–state.

jective, the *ADJ–OP/1–act is activated and the WCAM remains in the OPEN–state because at this moment not any valencies could be satisfied.

Analysing the noun *Mann*, the *NOM–OP/2–act[5] is activated and the WCAM changes to the COMPLETE–state. Now the valencies opened by the article and the adjective can be satisfied and the COMPLETE–acts of the adjective and the article are activated successively. In the course of these COMPLETE–acts, the agreement with respect to gender, case and number of the analysed word forms is checked. As a result of the *ART–CO–act, the WCAM changes to the state CLOSED. In this state, the NP is completely analysed and a semantic representation of the whole phrase has been created and stored.

If an OPEN–act of a word–class agent is activated, this information is stored in the central working memory (CWM) which is organized as a stack. Thus, as analysis proceeds, the corresponding COMPLETE–acts will be triggered just the other way round.

Because there are no COMPLETE–acts being left to perform (cf. the description of the WCAM–states given above), the state of the WCAM changes into the OPEN–state. Next the verb *schreiben* is read and its OPEN–act will be activated. The following NP is processed by analogy to the first one. The processing of the NP results in the CLOSED–state of the WCAM. At this stage, the *VB–CO–act is triggered and the items of the case frame on the syntactic and the semantic level are allocated to the valencies opened by the verb (cf. Figure 4). Expecting an agent (AGT) and an object (OBJ), among others, the valencies of the verb are satisfied by the complex semantic structures of the first and the second NP, playing the AGT– and OBJ–role, respectively. On the syntactic level, the agreement in respect of person, number and case for the AGT–role and with reference to case for the OBJ–role is checked. On the semantic level, for both roles the agreement in regard of semantic sorts is verified.

[5]*NOM–OP/1 is triggered in the case of apposition (e.g., *Peter der Große*, *Mrs Jane Smith*, *the Oxford English Dictionary*, *Archimedes' Law*).

4 GENERAL ASPECTS

4.1 DISAMBIGUATION PRINCIPLES

Although a variety of approaches and a lot of systems have been developed, a general natural language system is still out of sight. This desideratum is generally explained with the enormous complexity of natural language (see, e.g., [Stürmer 93]). One of the main reasons for this complexity lies in the ambiguity of natural language expressions. Therefore, NLP must be based on adequate strategies which support the disambiguation processes.[6]

In order to explain the human preferences in natural language understanding, researchers in the fields of computational linguistics (CL) and psycholinguistics have suggested several principles such as *Minimal Attachment*, *Right Association*, *Head Attachment* and *Lexical Preferences* (see, e.g., [Frazier, Rayner 82], [Ford et al. 82] or [Allen 87]). Some of them

[6]For the resolution of ambiguity see, for example, [Hirst 87], [Hindle, Rooth 93], [Schmitz, Quantz 93].

have been confirmed by psycholinguistic experiments (see, e.g., [Hemforth et al. 92] or [Hemforth et al. 94]).

In the word agent based language processing of the WCAM the following general disambiguation principles are implemented:

- *Completion* – semantic constituents, which have their correspondence in mental kernels built during human language processing, are created and completed as soon as possible.
- *Valency* – in the broadest sense, the valencies of words affect the attachment preferences. The principles of *Compatibility* and *Incompatibility* handle these preferences by making use of the subcategorization feature SELECT and the compatibility feature COMPAT in the lexicon (cf. Figure 4). Compatible feature values of two constituents permit the subordination of the second one and, on the other hand, feature values which are not compatible inhibit the attachment of constituents. Furthermore, the features are processed in an descending order (principle of *Priority*). Obligatory valencies of a word have priority over optional ones and the latter have priority over adverbial qualifications.
- *Right Association* – a new constituent tends to be interpreted as being part of the preceding constituent.
- *Preference of Reading* – this principle states that, in the case of polysemous words, one of the alternative meanings tends to be preferred. The preferred reading must be marked in the lexicon.

A detailed description of the disambiguation principles used in the WCAM and illustrative examples are given in [Helbig, Mertens 94]. The co–operation of the strategies for disambiguation and the lexical information, which plays an important part in resolving structural ambiguities, is presented in [Helbig et al. 94a].

$$
\begin{bmatrix}
\mathit{MORPH} & \begin{bmatrix} \mathit{WCA} & \mathit{*VB} \\ \mathit{FLEX} & \{\mathit{PG}\ \ \mathit{NS}\ \ \mathit{VK1}\} \end{bmatrix} \\[2ex]
\mathit{SELECT} & \left| \begin{matrix}
\begin{bmatrix} \mathit{SEMREL} & \mathit{AGT} \\ \mathit{CAS} & 1 \\ \mathit{SEMSORT} & \mathit{PS} \end{bmatrix} \\
\left\{ \begin{matrix} \begin{bmatrix} \mathit{SEMREL} & \mathit{OBJ} \\ \mathit{CAS} & 4 \\ \mathit{SEMSORT} & \mathit{IF} \end{bmatrix} \\ \begin{bmatrix} \mathit{SEMREL} & \mathit{DAT} \\ \mathit{CAS} & 3 \\ \mathit{SEMSORT} & \mathit{PS} \end{bmatrix} \end{matrix} \right\} \\
\left(\begin{bmatrix} \mathit{SEMREL} & \mathit{THM} \\ \mathit{WCA} & \mathit{*PREP} \\ \mathit{LEX} & \mathit{ÜBER} \\ \mathit{CAS} & 4 \end{bmatrix} \right)
\end{matrix} \right| \\[2ex]
\mathit{COMPAT} & \left\{ \begin{matrix} \mathit{RSLT} & \mathit{INSTR} & \mathit{ZWCK} & \mathit{DIRC} \\ \mathit{AUW} & \mathit{UMST} & \mathit{LOK} & \mathit{TEMP} \\ \mathit{BISZ} & \mathit{SEITZ} & \mathit{NACHZ} & \mathit{CAUS} \end{matrix} \right\} \\
\mathit{SEMREL} & [\mathit{SYNM}\ \ \mathit{VERFASSEN}] \\
\mathit{SEMSORT} & \mathit{V}
\end{bmatrix}
$$

Figure 4: Lexical entry of *schreiben* (*to write*)

4.2 LEXICAL INFORMATION

A lexical entry in LINAS consists of morphological, syntactic and semantic information. The feature MORPH has as its value a feature structure which itself comprises such features as WCA, GEN and FLEX. WCA denotes the corresponding word–class agent such as *VB (verb), *NOM (noun) and *PREP (preposition). The feature FLEX contains the description of the morphosyntactic properties of a word, including type of declension (for articles, adjectives and nouns) and conjugation (for verbs), respectively. The types of declension for words in the nominal group and the types of conjugation for verbs are divided into several groups. For instance, the group S02 represents the nouns with the ending *–s* in genitive and in all plural forms (e.g., *Auto – car*). In the case of words of the nominal group, the feature GEN designates the gender. An example of a noun entry is shown in Figure 3.

The values of the feature SELECT are obligatory and optional valencies of words. For verbs it represents the case frame on the syntactic and the semantic level. For instance, the German verb *schreiben* (*to write*) requires an agent (AGT) and at least one of the two roles object (OBJ) or a dative role (DAT). Besides, it allows for an optional thematic role.[7] The shortened lexical entry of this verb is given in Figure 4. In addition, the feature structure COMPAT expresses the compatibility between verbs and adverbial qualifications.

The feature structure SEMREL of a lexical entry contains the semantic relations to other words in the lexicon. For instance, if two words have the same meaning, they are in the relation of synonymy (e.g., *schreiben – to write* and *verfassen – to pen*). Some other examples of semantic relations included in the lexicon are SUB (hyponymy), PARS (part–whole–relation),

relation	characterization
AGT	agent
ASSOC	association
CAUS	causality
DAT	dative role
DIRC	direction
EQU	equality
INSTR	instrument
LOC	location
OBJ	object involved
ORIGM	original material
PARS	part-whole-relation
POSS	possessive relation
PROP	property, attribute
RSLT	result
SUB	subordination, hyponymy
SUBA	subordination of events
TEMP	temporal relation
THM	thematic role

Table 2: Some examples of semantic relations

ASSOC (association), etc. The information about semantic relations plays a crucial part in lexical disambiguation.

The feature SEMSORT has as its value the semantic sort of the lexical entry such as *V* – Vorgang (event), *PS* – Person (person) or *L* – Lokation (location). In LINAS the classification of entities is divided into the following three levels:

- *epistemic–ontological level* – entities are divided into sorts which are philosophically motivated and which are used in defining the algebraic structure of the

[7] Another possible role – BENF (beneficiary) – is omitted here for the sake of brevity.

knowledge representation apparatus.

- *socio–cultural level* – entities are specified by a given set of features allowing for a cross classification and being motivated by the importance of these categories for the description of human activities.
- *linguistic subtlety level* – entities are characterized by connecting them to parts of a semantic network. This information is used for linguistic fine tuning of selection restrictions.

A detailed discussion of the underlying three–level classification and illustrative examples can be found in [Helbig et al. 94b].

4.3 KNOWLEDGE REPRESENTATION

The syntactic and semantic analysis aims at representing NL input by means of a multi–layered extended semantic network (MESNET) and at integrating the results of the interpretation process in a network knowledge base.

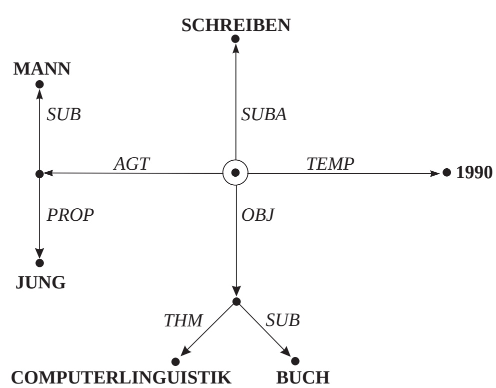

Figure 5: Semantic representation of a sentence

In LINAS the language interpretation process involves selecting the appropriate semantic relations which encode the connections between the constituents of a given NL sentence, and using them to construct a semantic network that represents the complete information of the sentence. Some examples of semantic relations and their characterizations are listed in Table 2. The semantic representation of the following example is given in Figure 5.

- *1990 schrieb der junge Mann ein Buch über Computerlinguistik.*
- ◦ *In 1990 the young man wrote a book about computational linguistics.*

The knowledge representation method used in LINAS extends the basic approach of semantic networks (SN) by accounting for the information such as the partitioning of the SN into shells and layers, the distinction between different levels and dimensions of the underlying classification, etc. The main characteristics of this new approach are the following:

- *stratification* – the SN is organized in different layers, the most prominent of them being the intensional and the preextensional layers.
- *semantic sorts* – the nodes of the SN are characterized by semantic sorts which are used in defining the network's algebraic structure. In addition, the nodes can be annotated with the semantic features and the selectional categories of the underlying three–level classification (cf. Section 4.2).
- *semantic relations and functions* – the nodes of the SN are linked to other nodes of the net by means of semantic rela-

tions and functions, which represent the semantic connections between the constituents of the sentence.

- *semantic shells* – in order to represent a complex semantic structure, a set of nodes and semantic relations of an SN can be combined in a semantic shell.
- *semantic dimensions* – the semantic dimensions are formed by a bundle of contrastive pairs: e.g., determinate vs. indeterminate reference of concepts, virtual vs. real concepts, individual vs. generic concepts, and collective vs. non–collective concepts. Furthermore, the nodes of the SN can be specified by intensional and preextensional indications of quantities.

5 THE FOUR MAIN LEVELS OF PROCESSING

The natural language processing of the WCAM is organized into four main levels (cf. Figure 6):

- on the *first level*, elementary semantic constituents are created which have their correspondence in mental kernels built during human language understanding;
- on the *second level*, complex semantic constituents are constructed by connecting the elementary constituents;
- on the *third level*, the constituents created on the first and second level are used to construct elementary propositions of a given sentence;
- in order to give a semantic representation of the whole sentence, on the *fourth level* a complex proposition is built up by connecting the elementary propositions.

The levels of the WCAM will be illustrated by analysing the following variation of the example which has been introduced in Section 4.3:

- *Der junge Mann schrieb ein Buch über Computerlinguistik.*
- ◦ *The young man wrote a book about computational linguistics.*

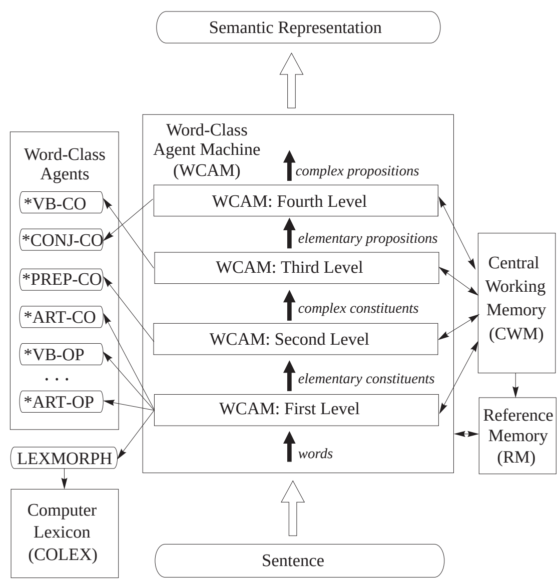

Figure 6: Overview of the WCAM levels

5.1 ELEMENTARY CONSTITUENTS

On the first level, for each word of a sentence the morphological analysis is activated in order to read the feature values of the corresponding lexical entries. If the whole sentence has been analysed on the first level

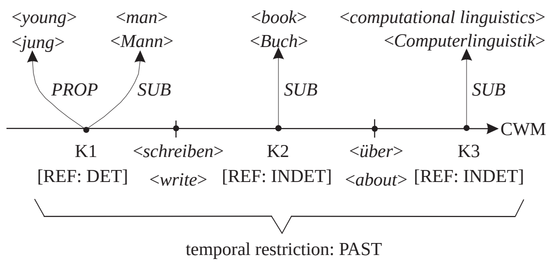

Figure 7: The CWM structure on the first level

(the activation of the word–class agents and the changing of the WCAM states are described in Section 3), three elementary semantic constituents have been created which have their correspondence in *mental kernels* built during human language understanding. In Figure 7, which shows the structure of the CWM at the end of the first level, these kernels are labelled K1 to K3.[8]

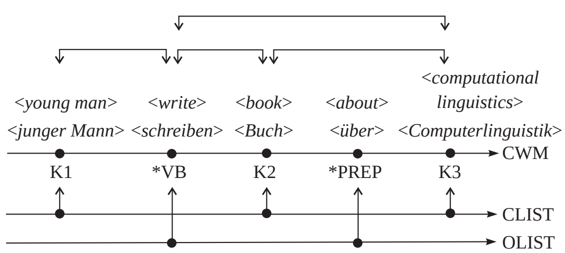

Figure 8: OList and CList on the second level

5.2 COMPLEX CONSTITUENTS

The task of the WCAM on the second level consists in constructing complex constituents by connecting the elemenary constituents determined on the first level. A difficult question concerns whether the elementary semantic constituents should be subordinated to a preceding elementary constituent or to the verb.

The elementary constituents K1 to K3 are marked by CL (closed) because they are permitted to satisfy the valencies of other constituents of the sentence. The verb *schreiben* and the preposition *über* are marked by OP (open) because these words open valencies (cf. Section 3). The constituents marked by CL are stored in the CList and the words marked by OP in the OList.

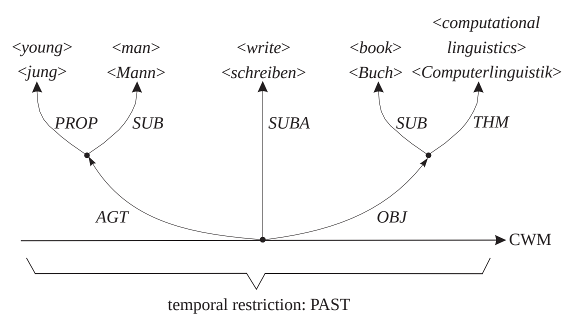

Figure 9: The structure of the CWM after creating an elementary proposition on the third level

Given that the elements of the CList and the OList are arranged in the order of the current sentence, a useful technique that takes advantage of the structure of the two lists is to analyse the sentence in reverse order. In our example, the kernel K3 may be subordinated to the kernel K2 or to the verb (see Figure 8). The lexical entries of the noun *Buch* and the verb *schreiben* shown in Figures 3 and 4 allow both subordinations. Following the disambiguation principle of *Right Association* (see Section 4.1),

[8]Determinate vs. indeterminate reference of concepts is marked by DET and INDET, respectively. This contrastive pair forms one of the semantic dimensions mentioned in Section 4.3.

K3 will be attached to the preceding kernel K2. Thus, a complex kernel is built up by connecting the kernels K2 and K3.

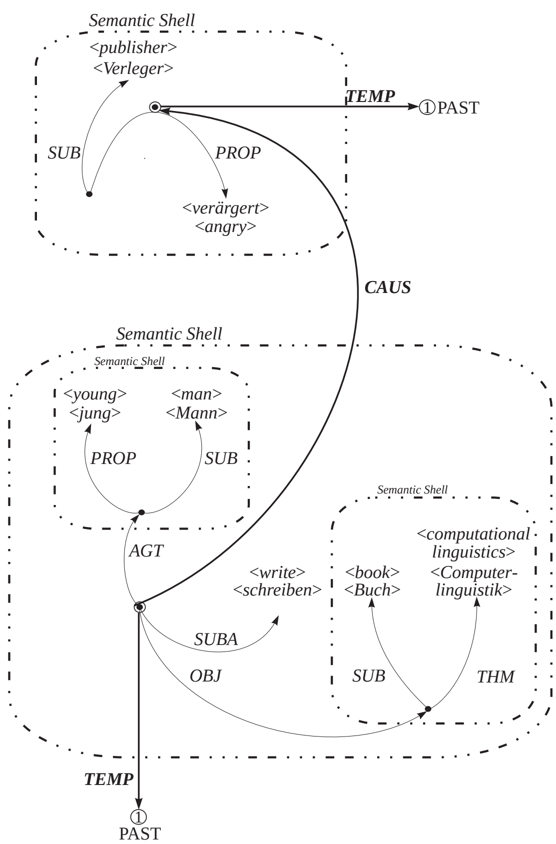

Figure 10: The structure of the CWM after constructing a complex proposition on the fourth level

5.3 ELEMENTARY PROPOSITIONS

On the third level, the kernels determined on the levels before are subordinated to the verb in order to saturate its open valencies. The valencies are processed in a descending order. First the obligatory valencies of the verb and then the optional valencies are satisfied. Finally the adverbial qualifications are subordinated to the verb. The subordination of the kernels is done by making use of the subcategorization feature SELECT for the obligatory and optional valencies and the compatibility feature COMPAT for the adverbial qualifications (cf. Figure 4).

The case frame of the verb *schreiben* requires an AGT and at least one of the two roles OBJ or DAT. In our example, the AGT–role and the OBJ–role are expressed by the elementary kernel *<junger Mann>* and the complex kernel *<Buch über Computerlinguistik>*, respectively. At this state, the semantic representation of the given sentence is stored in the CWM (cf. Figure 9).

In addition, the case frame of *schreiben* allows for an optional thematic role (THM) which may be satisfied by the kernel K3. This saturation is blocked, however, for the reasons mentioned above.

5.4 COMPLEX PROPOSITIONS

On the fourth level, the elementary proposition which has been determined in the preceding level may be connected to another elementary proposition. This process of the WCAM will be illustrated by analysing another variation of the example introduced before:

- *Der Verleger war verärgert, weil der junge Mann ein Buch über Computerlinguistik geschrieben hatte.*
- ∘ *The publisher was angry because the young man had written a book about computational linguistics.*

Now a clause of reason being added to the example, two elementary propositions can be determined by the WCAM on the third level and, in each case, the nodes and relations of the propositions are combined in separate semantic shells (cf. Figure 10).

On the fourth level, first the temporal information will be added to the SN and then the main task of the WCAM consists in analysing the subordinating conjunction *weil*. In order to guarantee the interpretation of the conjunction, the COMPLETE–act of the word–class agent *CONJ is activated and, as a result of this process, the two propositions are combined by the corresponding semantic relation *CAUS*.

6 CONCLUSIONS

In this paper we summarized the fundamentals of our approach to word agent based NLP. We further discussed the syntactic–semantic analysis of the WCAM and the different levels of language processing. In addition, we introduced the system components which are directly connected to the WCAM.

The word based approach presented in this paper can be characterized briefly by the following essential properties:

- the word and its combinatory potential plays the central part in natural language understanding;
- each word–class is connected to a corresponding word–class agent representing its grammatical function;
- integration of syntactic and semantic interpretation;
- the word–class agents are divided into two different acts corresponding to the opening and satisfying of valencies;
- the word–class agents are triggered and inspected by the word–class agent machine;
- the language processing is organized in four main levels: elementary and complex constituents, elementary and complex propositions;
- the disambiguation processes are supported by a group of strategies and, in addition, by special information in the lexicon;
- the result of the analysis is represented by means of a multi–layered extended semantic network.

Our approach is practically applied in a natural language understanding system and the components described in the previous sections are implemented and successfully used in the bibliographic information retrieval system LINAS at the University of Hagen.

REFERENCES

[Allen 87] Allen, J.: Natural Language Understanding. Menlo Park, California: Benjamin/Cummings 1987

[Boves, Nijholt 94] Boves, L., Nijholt, A. (eds.): Speech and Language Engeneering, Proceedings of the eighth Twente Workshop on Language Technology, Enschede, 1 and 2 December 1994, Universiteit Twente, Faculteit Informatica 1994

[Bröker et al. 93] Bröker, N., Hahn, U., Schacht, S.: Ein Plädoyer für Performanzgrammatiken. In: [DGfS–CL 93], pages 6–11

[Bröker et al. 94] Bröker, N., Hahn, U., Schacht, S.: Concurrent Lexicalized Dependency Parsing: The ParseTalk Model. In: COLING 94 – Proceedings of the 15th International Conference on Computational Linguistics, August 5–9, 1994, volume I, Kyoto, Japan 1994, pages 379–385,

[DGfS–CL 93] Deklarative und prozedurale Aspekte der Sprachverarbeitung, Tagungsband der 4. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft, Universität Hamburg 1993

[Eimermacher 86] Eimermacher, M.: Wortorientiertes Parsing mit erweiterter Chart–Repräsentation. In: Rollinger, C.–R., Horn, W. (eds.): GWAI–86 und 2. Österreichische Artificial–Intelligence–Tagung, 22.–26. September 1986, Ottenstein/Niederösterreich. Berlin etc.: Springer–Verlag 1986, pages 131–142

[Eimermacher 88] Eimermacher, M.: Wortorientiertes Parsen. Dissertation. Technische Universität Berlin, Fachbereich Informatik 1988

[Ford et al. 82] Ford, M., Bresnan, J., Kaplan, R. M.: A competence–based theory of syntactic closure. In: Bresnan, J. W. (ed.): The Mental Representation of Grammatical Relations. Cambridge, MA: MIT Press 1982, pages 727–796

[Frazier, Rayner 82] Frazier, L., Rayner, K.: Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. In: Cognitive Psychology, 14, pages 178–210

[Görz 92] Görz, G. (ed.): KONVENS 92 – 1. Konferenz Verarbeitung natürlicher Sprache, Nürnberg, 7.–9. Oktober 1992. Berlin etc.: Springer–Verlag 1992

[Helbig 86] Helbig, H.: Syntactic–semantic analysis of natural language by a new word–class controlled functional analysis (WCFA). Bratislava: Computers and AI 5 (1986) 1, pages 53–59

[Helbig, Mertens 94] Helbig, H., Mertens, A.: Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung. Teil I – Überblick über das Gesamtsystem. Informatik Berichte 158. FernUniversität Hagen 1994

[Helbig 94] Helbig, H.: Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung. Teil II – Die vier Verarbeitungsstufen. Informatik Berichte 159. FernUniversität Hagen 1994

[Helbig et al. 94a] Helbig, H., Mertens, A., Schulz, M.: Die Rolle des Lexikons bei der Disambiguierung. In: [Trost 94], pages 151–160

[Helbig et al. 94b] Helbig, H., Herold, C., Schulz, M.: A Three Level Classification of Entities for Knowledge Representation Systems. Informatik Berichte 162. FernUniversität Hagen 1994

[Hemforth et al. 92] Hemforth, B., Konieczny L., Scheepers, C., Strube, G.: SOUL–Processing: Semantik–orientierte Prinzipien menschlicher Sprachverarbeitung. In: [Görz 92], pages 198–208

[Hemforth et al. 94] Hemforth, B., Konieczny L., Scheepers, C.: Principle–based or Probabalistic Approaches to Human Parsing: How Universal is the Human Language Processor? In: [Trost 94], pages 161–170

[Hindle, Rooth 93] Hindle, D., Rooth, M.: Structural ambiguity and lexical relations. In: Computational Linguistics, 19, 1993, pages 103–120

[Hirst 87] Hirst, G.: Semantic interpretation and the resolution of ambiguity. Cambridge University Press: London etc. 1987

[Schmitz, Quantz 93] Schmitz, B., Quantz, J. J.: Defaults in machine translation. KIT–Report 106, Technische Universität Berlin 1993

[Sikkel, Nijholt 93] Sikkel, K., Nijholt, A. (eds.): Natural Language Parsing – Methods and Formalisms, Proceedings of the sixth Twente Workshop on Language Technology, Enschede, 16 and 17 December 1993, Universiteit Twente, Faculteit Informatica 1993

[Small 81] Small, S.: Viewing word expert parsing as linguistic theory. Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI–81), 24–28 August 1981, volume I, University of British Columbia, Vancouver 1981, pages 70–76

[Small 87] Small, S. L.: A Distributed Word–Based Approach to Parsing. In: Bolc, L. (ed.): Natural Language Parsing Systems. Berlin etc.: Springer–Verlag 1987, pages 161–201

[Stürmer 93] Stürmer, T.: Semantic–Oriented Chart Parsing with Defaults. In: [Sikkel, Nijholt 93], pages 99–106

[Trost 94] Trost, H. (ed.): KONVENS '94 – Verarbeitung natürlicher Sprache, Wien, 28.–30. September 1994. Berlin: Springer–Verlag 1994

FernUniversität Hagen
Praktische Informatik VII/Artificial Intelligence
D–58084 Hagen
Germany

The following reports of the Praktische Informatik VII/Artificial Intelligence can be obtained by anonymous FTP from `ftp.fernuni-hagen.de` (132.176.114.150). After `login` with user name `anonymous` change to the directory `/fachb/inf/pri7`. If you should have any problems please contact `christoph.doppelbauer@fernuni-hagen.de`.

IB 158 Hermann Helbig, Andreas Mertens:
Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung – Teil I: Überblick über das Gesamtsystem. Informatik Berichte 158 (4/1994), Fachbereich Informatik, FernUniversität Hagen, 1994

IB 159 Hermann Helbig:
Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung – Teil II: Die vier Verarbeitungsstufen. Informatik Berichte 159 (4/1994), Fachbereich Informatik, FernUniversität Hagen, 1994

IB 162 Hermann Helbig, Claus Herold, Marion Schulz:
A Three Level Classification of Entities for Knowledge Representation Systems. Informatik Berichte 162 (8/1994), Fachbereich Informatik, FernUniversität Hagen, 1994

IB 168 Hermann Helbig, Andreas Mertens, Marion Schulz:
Disambiguierung mit Wortklassenagenten. Informatik Berichte 168 (12/1994), Fachbereich Informatik, FernUniversität Hagen, 1994

IB 169 Hermann Helbig, Andreas Mertens:
The Word Class Agent Machine. Informatik Berichte 169 (1/1995), Fachbereich Informatik, FernUniversität Hagen, 1995