CLARE

A System for Classificatory Knowledge Representation

Hermann Helbig
Marion Schulz
Claus Herold

FernUniversität Hagen
58084 Hagen, Germany

Hermann.Helbig@FernUni-Hagen.de
Marion.Schulz@FernUni-Hagen.de
Claus.Herold@FernUni-Hagen.de

Tel: +49 2331 987 4520
Fax: +49 2331 987 375

Abstract

A sound framework for **cla**ssificatory knowledge **re**presentation (CLARE) is presented which is used for natural language understanding in connection with the system LINAS. In contrast to other approaches, a fundamental distinction is drawn between intensional and preextensional representations which are explicitly discerned as different layers in the knowledge base.

Furthermore, a global classification by different dimensions is provided comprising the following ones: generality, collectivity, quantification, referentiality and factuality. Whereas these dimensions are bipolar (dichotomic or semi-continuous transition between two distinct poles) another hierarchically structured classification is added: a taxonomy of sorts, poly–hierarchically organized features and special means for lexical subcategorization.

Contents

1 Introduction **3**

2 Layers of semantic interpretation **5**

3 Dimensions of Classification **7**

3.1 Generality 7

3.2 Collectivity 8

3.3 Referentiality 9

3.4 Factuality 10

3.5 Quantification 10

4 Levels of Classification **12**

4.1 Epistemic–Ontological Level 12

4.2 Socio–Cultural Level 14

4.3 Linguistic Subtlety Level 14

5 Conclusion **16**

List of Figures

1.1 The Classification System 4

3.1 Interrelations[1] in Counterfactuals 11

4.1 The Three Level Classification 13

Chapter 1

Introduction

Classificatory knowledge plays an important part in all scientific domains. In knowledge representation systems it provides the backbone for the whole representational apparatus. In many AI systems, this type of knowledge is expressed almost exclusively by a single hierarchy or poly–hierarchy of sorts. This restriction leads to a mixture of multiple classification criteria which blur important distinctions. In the LILOG[1] ontology for instance, some sorts are of fundamental importance, others are only motivated by a special application domain. Another example is DAHLGREN's[2] cross classification system where sorts like collective occur besides non stationary reflecting extensional aspects and temporal aspects respectively. In contrast to other approaches, we draw a fundamental distinction between intensional and preextensional representations which are explicitly represented as different layers in our framework for **cla**ssificatory knowledge **re**presentation (CLARE). Furthermore, there are five dimensions used for characterization of concepts. An overview is given in figure 1.1. Whereas these dimensions have a bipolar character (dichotomic or semi-continuous transition between two distinct poles) a second classification by three levels is hierarchically structured. These levels are especially important for describing lexical knowledge. The structure of the classificatory inventory parallels the nature of the information to be represented. It makes

[1] A major research effort by IBM Germany and a number of German universities, see [Herzog and Rollinger, 1991]

[2] [Dahlgren, 1988]

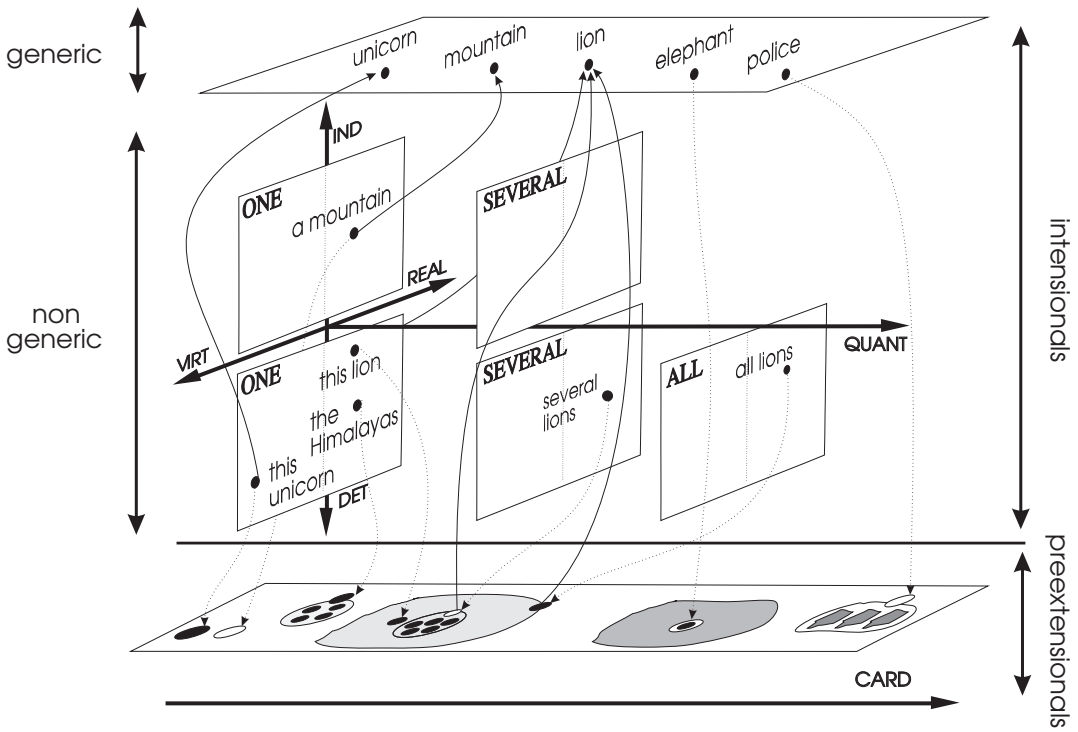

Figure 1.1: The Classification System

just the distinctions that allow to principally separate e.g. concepts such as *Caucasus* and *Prometheus*, *army* and *soldier* or *unicorn* and *eagle* in a sufficiently fine grained way. In natural language processing (NLP) systems, sorts, features and designators for the different dimensions of the knowledge base are important means to describe the lexical knowledge and the world knowledge as well.[3] In contrast with LANG's view[4], it is characteristic of our representation method that we use the same categories in describing both the lexical and the conceptual knowledge base. In this way, all classificatory concepts described in this paper can be found both in the lexicon COLEX[5] and in the multilayered extended semantic network MESNET[6] of the natural language access system LINAS[7] for bibliographic information retrieval of the FernUniversität Hagen.

[3] [Helbig *et al.*, 1994a]

[4] [Lang, 1992]

[5] [Helbig and Schulz, 1995]

[6] [Helbig and Herold, 1995]

[7] **Li**teraturrecherche in **na**türlicher **S**prache

Chapter 2

Layers of semantic interpretation

The importance of the distinction between intensional and extensional interpretations of concepts has been already discussed by CARNAP[1]. Nevertheless, the extensional aspect has hardly been properly accounted for in knowledge representation systems so far (first attempts in this direction: JANAS/SCHWIND[2] as well as ALLGAYER/REDDIG[3]). Knowledge representation is not able to deal with an extensional layer directly, however, because no system is powerful enough to model even parts of the real world. Also human beings reduce real world occurrences like sets or classes of objects by representatives.[4] That's why we call the layer corresponding to the extensional aspect preextensional. There are two main reasons which lead to the distinction between intensional and preextensional layer and to an explicit consideration of the extensional aspect in the knowledge representation itself. On the one hand, humans explicitly refer to extensional categories in natural language utterances. They speak about sets and cardinalities and these categories have also to be used in semantic knowledge representations to map the meaning of sentences like

[1][Carnap, 1947]

[2][Janas and Schwind, 1979]

[3][Allgayer and Reddig, 1990]

[4][Johnson-Laird, 1991]

(1) a. *Nearly all students ran to a boat.*

b. *Two of them wore yellow shoes.*

There are many cases where explicit reference to the preextensional layer is compelling as in example 1b where the underlying subset relation (connected with an explicitly stated cardinality, in this case 2) is a component of the preextensional layer. There are other linguistic indicators that make the extensional interpretation of concepts necessary, e.g. *the one* and *the others*, or determiners like *this one* and quantifiers like *every* or *each*.

On the other hand, in the first spontaneous understanding people do not care to elaborate all extensional aspects of a semantic interpretation. For instance, they often do not necessarily discern between distributive and collective reading. For instance, no one would really care to figure out when hearing sentence 1a whether *nearly all* means one hundred or only fourteen or whether there is one boat for all or one for each of these students. So, analogously, the representation on the intensional layer will contain information e.g. in form of linguistic quantifiers like NEARLY ALL and TWO. Only on the preextensional layer and only if needed, e.g. *14* as an explicit cardinality for students and *14* as cardinality for *boat* assuming the distributive reading has to be stated (if there is any information at all to spell out these exact cardinalities). Generally, in a first and spontaneous understanding of a natural language sentence a condensed structure (the so–called intension of the sentence) will be generated, containing all information possibly needed for a further extensional interpretation. The latter is carried out only if it becomes necessary.

On the base of this fundamental distinction CLARE opens a natural way to deal with the semantic representation of collective nouns as well as indeterminate and determinate reference. Also referential identity as in FREGE's famous example can be captured: Whereas *morning star* and *evening star* have different conceptual associations, i.e. they are intensionally different, they are extensionally identical, the star Venus, and they have the same preextensional representation. Of course, there is an intensional representation for the concept *Venus* (comprising for instance the concept of the goddess bearing the same name in its meaning field) just as for the other two concepts.

Chapter 3

Dimensions of Classification

Additionally to the two basic layers of semantic interpretation explained above, we introduce different dimensions which run across the sorts and which are more related to the representation of world knowledge than of lexical knowledge. In the following, we propose five dimensions with their fundamentals.

3.1 Generality

We oppose individual and generic as a general dichotomy expressed by special layers. This distinction is relevant for all sorts (cf. section 4) and therefore has to be given another status, which we call a classificatory dimension. Consider the following examples:

(2) *elephants vs. elephants that cross the Alps with Hannibal* (both are discrete)

(3) *sickness vs. Peter's sickness* (abstracted states)

(4) *at stations vs. at the main station of New York* (locations)

(5) *at night vs. yesterday night* (times)

Generic concepts are of importance because they allow the attribution of properties that hold for whole classes of individuals (e.g. *Bears* are mammals). These properties can be inherited by instances of the class, be it parametrized individuals (e.g. All tourists saw *a bear* last night.) or determinate instances

(*Balu the bear*) each subordinated to the generic concept. Generic concepts serve also as a basis for drawing inferences.

Furthermore, the distinction between individual and generic concepts allows to capture certain selectional restrictions. E.g. related to the role of an affected object (AFF) of the verb *to die out* certain generic concepts are possible role fillers.

(6) a. *The whale is about to die out.*

b. **This whale named Timmy is about to die out.*

The admissibility of individual concepts (see 6b. and sentence 7) depends on the collectivity dimension, treated in the next section.

3.2 Collectivity

The difference between collective and non–collective nouns is marked by means of a separate dimension.

Intensionally, a collective term like *Caucasus* is considered as an individual concept denoting a set of mountains in the preextensional layer containing e.g. the representative of the mountain Elbrus as an element. This contrasts with non–collective terms like *Prometheus* which are also considered as individual concepts but which denote a single element on the preextensional layer.

Therefore collective entities fulfill the selectional restrictions of the verb *to die out* equally well as generic entities. For illustration see the following sentence:

(7) *The Adams family will never die out.*

Usually, noun phrases in plural like *20 men* denote sets while in rare cases morphologic plural forms (as for instance *scissors* or Russian *kleščі – pliers*) designate non–collectives. Also this distinction can be expressed by the collectivity dimension. Generally, collectives are possible candidates for agent role fillers in constructions like

(8) *The team met at 9 a.m. for breakfast.*

(9) a. *The 20 men met at 9 a.m. for breakfast.*
b. **The man met at 9 a.m. for breakfast.*

The role of sets in the representation will be dealt with in section 3.5.

3.3 Referentiality

The referentiality dimension expresses the nature of a reference into extension. Our definition of determinate reference is similar to the one given by SOMMERS[1] for genuine reference, with the exception that the entity referred to need not exist in reality but maybe in a possible world (The utterance of a referring expression must contain a descriptive expression that unequivocally applies to the referred object and presents this object demonstratively). Consider the following examples:

(10) *The winner cried tears of joy.*

(11) a. speaker 1: *Why are these guys running so fast?*
b. speaker 2: *The winner will get $ 500 !*
c. speaker 1: *Ah. – And why do the others run?* [2]

In sentence 10, the linguistically definite 'the winner' refers to a determinate entity already known whereas in sentence b of example 11, 'the winner' (whoever it will be) denotes also a determinate entity but in a future world which is only intensionally characterized. In contrast to the definition as determinate with respect to time, we understand 'the winner' in sentence b as a cognitively indeterminate entity because its reference is not known at the time of utterance. Speaker 1 (sentence c) wrongly assumes it to be determinately referring to an extensionally given object which is indicated by the phrase 'the others', thereby producing a comic effect.

For indeterminate reference, a further subdistinction can be made between indeterminate reference playing the part of constants on the one hand (*He saw a man.*) and of variables on the other (*Every boy saw a man.*): the first

[1] [Sommers, 1982]

[2] example taken from [Allgayer and Reddig, 1990]

does not depend on other entities, and therefore the entities introduced by constantly indeterminate reference can be used for further reference also in separate sentences. The second depends on other entities, represented by a special dependency–relation DPND in the intensional layer. In the following example, *a dog* normally is considered to denote a different individual for each of the postmen in question. In this reading a further reference is blocked.

(12) *Every postman has been attacked by a dog. *It was a Rottweiler.*[3]

3.4 Factuality

The dimension of factuality expresses the distinction between *real* and *virtual* concepts, i.e. whether a counterpart of the concept in the real world exists (*elephant*) or not (e.g. *unicorn*)[4]. More important for practical applications, the factuality dimension is used in the domain of states of affairs to express whether or not a truth value can be assigned to a state or event. Especially in counterfactuals, this distinction becomes crucial for representation as it can be seen in sentences like the following (cf. figure 3.1):

(13) *If Hannibal had not taken elephants with him, he would not have become so famous.*

3.5 Quantification

The quantification dimension has a different character than the others. It is quantitative whereas the rest is intrinsically qualitative. Touched upon in section 2 on page 6 the quantification dimension is used on the intensional layer to model the quantification in natural language. There are definite numbers like FOUR or FIFTYFIVE as well as fuzzy quantifiers, e.g. NEARLY ALL or A FEW.

[3] If the reference is to be considered as grammatically correct than another reading is compelling.

[4] which at best has its counterpart in a fairy world

[5] COND – Conditional Relation, CAUS – Causal Relation, ASSOC – Associative Relation

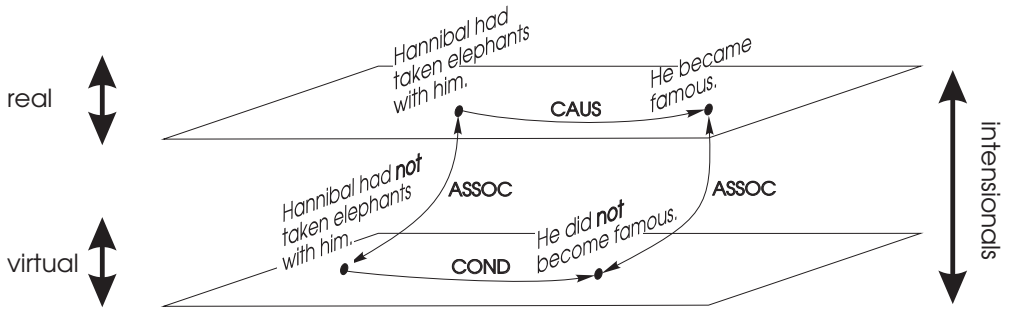

Figure 3.1: Interrelations[5] in Counterfactuals

This dimension has no single pair of poles but a semi-continuous transition between opposite ends of a scale from ONE to ALL.

The counterpart on the preextensional layer is the cardinality which is always a concrete number. This distinction is necessary as it is not always possible and also not necessary to determine the exact cardinality that corresponds to linguistic quantifiers like *dozens of* or *several thousand*.[6]

(14) *Hannibal tooks* dozens *(in fact, 37) of elephants and* several thousand *(in fact, 8000) horseman with him.*

The case of collectives with an exact quantifier (*five baseball teams*) is even worse: not much is said about the cardinality of the set of players involved. Nevertheless, the number of men belonging to these teams could be easily computed on the base of background knowledge (number of players belonging to a baseball team).

[6]The definition of truth conditions of constructions containing linguistic quantifiers or explicit cardinalities of sets is quite another task compared with the task of knowledge representation. For this purpose techniques like those of Generalized Quantification Theory are used ([Barwise and Cooper, 1981]).

Chapter 4

Levels of Classification

A second, structured classification which relates mainly to the intensional representation is added: a) a taxonomy of sorts, b) poly–hierarchically organized features and c) special means for lexical subcategorization. The levels which are also used for efficient analysis of natural language sentences have been introduced for pragmatic reasons. In principle the information described by them could be expressed by the general knowledge representation devices, i.e. in our knowledge representation system by semantic relations. This would lead, however, to the dilemma that analysis had to deal with the whole knowledge base instead of the lexical knowledge alone which is merely a small part of it. The ambitious Cyc project[1] pursues just this aim of always taking the complete set of conceptual interrelations into account. The feasibility of Cyc has yet to be proven for practical applications.

Sorts and features are indispensable for definition purposes – they determine the domains and value restrictions of relations and functions – and for effective checking of selection restrictions (congruency tests). The structure of the three levels including a small set of examples is shown in figure 4.1.

4.1 Epistemic–Ontological Level

A taxonomy of disjunctive sorts can and should only provide a rather global classification into categories of primary importance. These categories have

[1] [Lenat and Guha, 1990]

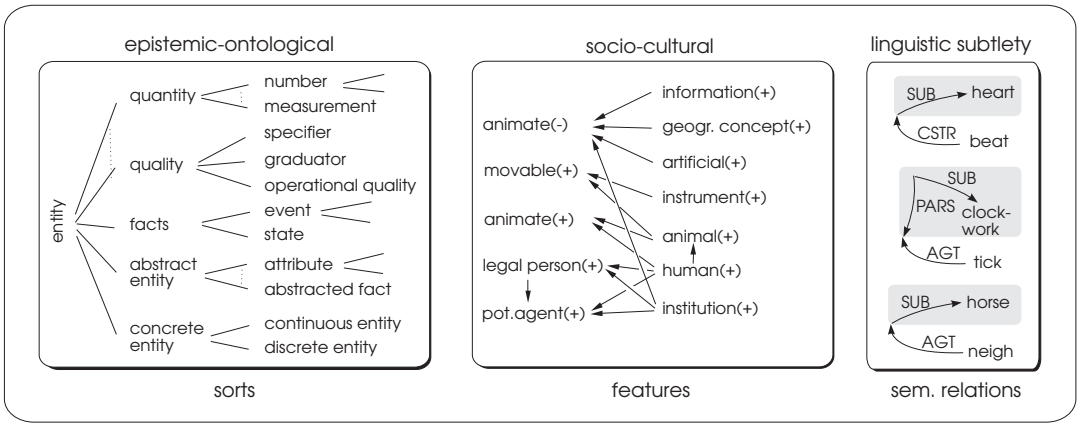

Figure 4.1: The Three Level Classification

much in common with categories philosophers discern.
We call them epistemic–ontological because they represent taxonomic knowledge about the world. They form the first level. Examples are quantity, quality or abstract vs. concrete entity. Sorts are mainly used to define the algebraic structure of the knowledge representation system.

In contradistinction to DAHLGREN[2], we strongly oppose to introducing a sort like collective, especially as a counterpart for individual. Rather, collective contrasts with non collective and individual with generic and these are characterized as belonging to different dimensions. For instance, *this crowd* picks out a collective individual – collective because it denotes a set as a whole and individual because it refers to a particular crowd. *A crowd*, on the other hand, can be used to denote the generic concept 'crowd' as exemplified in the following sentence: *It is easy to disappear in a crowd.* This analogously holds for non collective concepts like *Jim T.* and *a man.* There are also collective terms for times, e.g. *vacation* which is a pluraletantum in German (*Ferien*) or abstract entities like *police* in individual as well as in generic form.

So, instead of using sorts, we distinguish collective and non collective as well as individual and generic concepts by different dimensions of CLARE. These layers cannot be expressed in terms of sorts because they run across the sorts and they are not relevant for the definition of semantic relations and functions.

[2][Dahlgren, 1988]

4.2 Socio–Cultural Level

The second level provides a number of features that allow a finer characterization of first level sorts and it is more closely connected with the interrelations between human beings and other parts of the world, i.e. with human activities in the broadest sense. According to the character of this level, dealing with the world from a subjective human perspective, it can be seen as socio–cultural in a very wide sense. It must be emphasized that the denomination of the different levels gives no strong criteria for the division between sorts and features but only a certain orientation. The features comprise e.g. the classical animate – inanimate distinction as well as the one between movable and non-movable objects. If we tried to use these features in order to define disjunctive categories, this would at best produce rather artificial sorts. In contrast to the epistemologic–ontological level, there is no single hierarchy but a poly–hierarchy. For example, being human consistently involves being animate and natural (non artificial). Analogously, any institution is a legal entity and at the same time a potential agent.
An important distinction between sorts (like abstract or concrete) and features (like movable, non-movable) consists in the fact that the former are always categorically assigned to concepts while the latter generally characterize typical properties of entities which can be overridden in certain contexts, i.e. they have to be understood as defaults. Both levels make use of a fixed inventory of primitives – sorts or features respectively – and both provide a rather global classification.

4.3 Linguistic Subtlety Level

Sufficiently fine–grained specifications, which are necessary to characterize the combinatory potential of single lexemes and which are mainly used for lexical subcategorization, are provided on the third level. Parts of semantic networks achieve even the most specific characterization. Therefore, we call this level of categorization the linguistic subtlety level. For example, the verb *to neigh* subcategorizes for an agent which generally is a horse or *to tick* is classified

as an activity that usually relates to concrete objects having a clockwork as a part and the like. Through addition of the linguistic subtlety classification, a continuous transition from rather rough ontological specification to fine grained lexical subcategorization is achieved. It is not our aim to separate linguistically and non linguistically motivated levels in general. However, the association of linguistic motivation contained in all levels is perfectly clear on the linguistic subtlety level.

Especially in natural language processing and semantic representation of the meaning of texts and dialogues, sorts and features are used to characterize the nodes of semantic networks. Features are mainly used to specify the subcategorization of the semantic representatives of whole classes of verbs and nouns to rule semantic congruencies. The main motivation for the application of semantic relations and functions at the linguistic subtlety level is to provide an adequate framework for the development of a lexicon strongly interrelated with the general knowledge representation paradigm.

When a natural language system (NLS) starts from scratch, the lexical subcategorization is representative for rudimentary world knowledge connected to specific lexemes, e.g. the lexical entry *to neigh* comprises the restriction that only horses are able to do this. In a NLS with fully developed knowledge base, this level can be realized in the lexicon by exclusively using pointers to the nodes of the semantic network.

Chapter 5

Conclusion

The classification system presented above has proven to be very useful in two respects. It is sufficiently fine grained to represent most of the interesting classificatory aspects of meaning representation of natural language expressions and to explain conceptual differences. At the same time it is compact enough for efficient processing. Furthermore, this approach is inspired from cognitive psychology and can be used as a component of a cognitive model as well.

The classification system CLARE provides the basis for the expressional means of the multilayered extended semantic network MESNET described in [Helbig and Herold, 1995]. The classificatory knowledge also plays an important part in checking the selection restrictions of words during syntactic-semantic analysis and generation and in testing for semantic congruencies during natural language processing. This problem was only touched upon in this paper. Details are presented in connection with the word-class controlled analysis WCAM of the system LINAS in [Helbig and Mertens, 1994]. The role of classificatory knowledge for disambiguation processes is described in [Helbig *et al.*, 1994b]. In future work, we'll have to examine more intensely whether the whole classification which is appropriate for all object–like entities is also sufficient for the classification of states of affairs and abstract entities.

Bibliography

[Allgayer and Reddig, 1990] J. Allgayer and C. Reddig. What KL–ONE lookalikes need to cope with natural language – scope and aspect of plural noun phrases. In K. Bläsius, U. Hedstück, and C.-R. Rollinger, editors, *Sorts and Types in Artificial Intelligence*, pages 240–285. Springer, Berlin, Germany, 1990.

[Barwise and Cooper, 1981] J. Barwise and R. Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219, 1981.

[Carnap, 1947] R. Carnap. *Meaning and Necessity.* Chicago, 1947.

[Dahlgren, 1988] K. Dahlgren. *Naive Semantics for Natural Language Understanding.* Natural Language Processing and Machine Translation. Kluwer Academic Publishers, Boston, Dordrecht, London, 1988.

[Helbig and Herold, 1995] H. Helbig and C. Herold. MESNET – A Multilayered Extended Semantic Network. Informatik Berichte 185, FernUniversität Hagen, Hagen, Germany, 1995.

[Helbig and Mertens, 1994] H. Helbig and A. Mertens. Word Agent Based Natural Language Processing. In A. Nijholt and L. Boves, editors, *Proceedings of the 8th Twente Workshop on Language Technology – Speech and Language Engineering*, Enschede, 1994. Universiteit Twente, Fakulteit Informatica.

[Helbig and Schulz, 1995] H. Helbig and M. Schulz. COLEX – Ein Computerlexikon für die automatische Sprachverarbeitung. Informatik Berichte (forthcoming), FernUniversität Hagen, Hagen, Germany, 1995.

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

[Helbig *et al.*, 1994b] H. Helbig, A. Mertens, and M. Schulz. Die Rolle des Lexikons bei der Disambiguierung. In H. Trost, editor, *Tagungsband KONVENS'94 Verarbeitung natürlicher Sprache*, pages 151–160, Berlin, 1994.

[Herzog and Rollinger, 1991] O. Herzog and C.-R. Rollinger, editors. *Text Understanding in LILOG.* Springer, Berlin, Heidelberg, New York, 1991.

[Janas and Schwind, 1979] J. M. Janas and C. B. Schwind. Extensional Semantic Networks. In N. V. Findler, editor, *Associative Networks – Representation and Use of Knowledge by Computers*, pages 267 – 305. Academic Press, New York San Francisco London, 1979.

[Johnson-Laird, 1991] P. N. Johnson-Laird. *Deduction.* Laurence Erlbaum, Hillsdale, 1991.

[Lang, 1992] E. Lang. Linguistische vs. konzeptuelle Aspekte der LILOG Ontologie – Anfragen von außen. In G. Klose, E. Lang, and Th. Pirlein, editors, *Ontologie und Axiomatik der Wissensbasis von LILOG*, pages 23–45. Springer, Berlin, Heidelberg, 1992.

[Lenat and Guha, 1990] D. B. Lenat and R. V. Guha. *Building large knowledge–based systems: representation and inference in the Cyc project.* Addison–Wesley, Reading (Mass.), 1990.

[Sommers, 1982] F. Sommers. *The Logic of Natural Language.* Clarendon Press, 1982.