LEVELS OF REPRESENTATION IN NATURAL LANGUAGE BASED INFORMATION SYSTEMS AND THEIR RELATION TO THE METHODOLOGY OF COMPUTATIONAL LINGUISTICS

G. ZIFONUN, INSTITUT FUER DEUTSCHE SPRACHE,
D-6800 MANNHEIM, FEDERAL REPUBLIC GERMANY

Summary

In this paper the methodological basis of the 'computational linguistics approach' for representing the meaning of natural language sentences is investigated. Its adherance to principles of formal linguistics and formal philosophy of language like the 'separation of levels of syntactic and semantic analysis', and the "Fregean" principle may be contrasted with the 'artificial intelligence approach'. A "Montague" style method of mapping the syntax of natural language onto the syntax of the 'semantic language' used as the means of internal representation in the information system PLIDIS is presented. Rules for defining subsequent levels of representation like 'syntax-interpretative level', 'redundancy' free level' are given.

Introduction

The present paper presents ideas concerning a methodology of the 'semantics in computational linguistics' (COL-semantics).

There is the following hypothesis underlying:

In the field of COL-semantics algorithms and computer programs are developed which deliver structures of linguistic analysis and representation that can be compared with those of formal linguistic semantics and satisfy the adequacy criteria of certain linguistic theories. They therefore are suitable instruments for developing and testing such theories.

COL-semantics hence proceeds in a way different from the semantic processing as it is found in the framework of artificial intelligence (AI-semantics). AI-semantics is not so much linked to the semantics of formal linguistics or logic but rather to cognitive psychology, problem solving theory and the theory of knowledge representation which has been recently put forward within AI itself.[1] Between both branches of semantic processing of natural language that are realized in computer systems there therefore exists a difference in aims, theories and methods.

Starting from a brief sketch ot the aims and theories of both approaches one essential methodological principle of COL-semantics will be elaborated in the second chapter of the paper. In the third chapter COL-semantic methods will be exemplified by a concrete application, the production of semantic representations in an information system. Stress will not be laid on the question of w h a t a COL-semantic representation should look like but h o w levels of a semantic representation can be systematically related with natural language and with each other.

Aims and theoretical concepts of COL-semantics and AI-semantics

The difference of aims and methods can only be outlined here as far as it is relevant with respect to the methodological divergence which will be dealt with in detail: Aim of AI-semantics is the simulation of the human language understanding and/or language generating process that is to be understood as a manifestation of intelligent human problem solving behaviour. Aim of COL-semantics is the algorithmic generation of descriptive structures (of a generativ-semantic, interpretative, logico-semantic or other type) out of a given natural language input. Both purposes can be partial aims or intermediate steps within a larger project like 'simulation of dialogue behaviour', 'natural language information or question answering system'.

Thus the AI-approach leads to a theory where the object of explanation (or simulation) is "rational human behaviour"[2] or more specifically human language behaviour as a rational psychic process, whereas in the theory of linguistic semantics language is being objectified as a generated structure or a system which can be considered independently from the associated mental processes. In linguistic semantics and also in COL-semantics meta-linguistic notions which refer to language as a system like 'synonymy', 'equivalence' and (particularly in the formal linguistics based on logic)'truth' and 'entailment' are crucial; in AI-semantics however we have the 'behaviour' oriented concepts of 'inferencing','disambiguating', 'reasoning', 'planning' etc.[3]

A methodological principle of COL-semantics

A distinctive feature of linguistics, especially logico-linguistic theories, is the separation of different "expression" and "content" levels of analysis and representation and the speci-

fication of mapping rules between them (surface structure versus deep structure, syntactic structure versus semantic structure). In Montague grammar this differentiation between a well defined syntactic level and an also well defined semantic level of description is a methodologically necessary consequence of the "Fregean" principle. The Fregean principle states that the meaning of an expression can be determined on the basis of the meanings of its logically simple constituent expressions and the syntactic structure of the whole expression. This principle has been revived by Montague and has been realized in his theory of language in such a way that the syntactic and the semantic structure of a natural language expression are respectively represented as expressions of formal systems (syntax and meaning algebras) between which systems there exist well defined formal relationships (homomorphisms).

When this concept is transferred to the operationalizing of linguistic analysis in a computer system it will be excluded to conceive the mapping from natural language into semantic representation as a simple integrated pass, where in the course of parsing a sentence the valid semantic interpretation is assigned to each occurring item or group of items and where the possibilities of inference and association with stored background knowledge are 'locally' realized without ever generating a full syntactic analysis. Saving an explicit level of syntactic representation seems to be compatible with the Fregean principle only under the condition that the algorithm incorporates a grammar (in the technical sense of a consistent set of generating or accepting syntactic rules), but for reasons of optimization directly associates or applies semantic 'values' or 'rules' in processing the corresponding syntactic 'nodes' or 'rules'[4], or even allows a semantic control of rule selection without leaving the parsing mode. This condition however is mostly not maintained in AI parsing approaches where the one step processing is understood as a cognitively adequate analogue of human linguistic information processing and where even the terminal and non terminal symbols of the "grammar" are interpreted as semantic categories.[5]

Syntactic and semantic representation in an information system

The way of processing natural language according to the principles of COL-semantics shall be demonstrated by the linguistic component of a natural language information system. The description is oriented at the application area and the structure of the system PLIDIS (information system for controlling industrial water pollution, developed at the Institut fuer deutsche Sprache, Mannheim).[6] Giving only the over all structure of the system we have the following processings and levels:

morphological analysis of natural language input → syntactic analysis (level of syntactic representation) → transduction into formal representation language (level of semantic representation) → interpretation (evaluation) against the database → answer generation

The formal representation language is the language KS an extended first order predicate calculus, where the features going beyond predicate calculus are many sorted domain of individuals, lambda-abstraction and extended term building.[7] In the following two aspects of the semantic representation will be treated:

- the mapping between syntactically analyzed natural language expressions and their KS counterparts will be investigated
- a differentiation between three levels of semantic representation will be accounted for: (level 1) syntax-interpretative level, (level 2) canonical level, (level 3) database-related level,

All three levels follow the same syntax, i.e. the syntax of KS and have the same compositional model theoretic semantics; they differ in their non logical constant symbols.

Mapping natural language into the semantic representation language KS

In analogy with Montague's "theory of translation" in "Universal Grammar" we assume that the syntactic structures of natural language (NL, here German) and the semantic language (here KS) are similar, i.e. there exists a translation function f, such that the following holds:

(1.1.) Given the categories of a categorial grammar of NL, f is mapping from these categories on the syntactic categories of KS. I.e. If α, $\beta_1, \ldots, \beta_n$ are basic categories of German, then $f(\alpha)$, $f(\beta_1), \ldots, f(\beta_n)$ are syntactic categories of KS. If $\alpha/\beta_1/\ldots/\beta_n$ is a derived category (functor category) of NL, then $f(\alpha)/f(\beta_1)/\ldots/f(\beta_n)$ is a derived category of KS.

(1.2.) If a is an expression of category δ in NL (aδ), then $f(a)$ is an expression of category $f(\delta)$ in KS $(f(a)_{f(\delta)})$.

(1.3.) The concatenation of an expression of the derived category $\alpha/\beta_1/\ldots/\beta_n$ with expressions of category $\beta_1, \ldots, \beta_n$ resulting in an expression of category α

$\alpha/\beta_1/.../\beta_n \frown \beta_1 \frown ... \frown \beta_n \rightarrow \alpha$

is rendered in KS by the construction of a list

$[f(\alpha/\beta_1/.../\beta_n)\ f(\beta_1)\ ...\ f(\beta_n)]$

with the category $f(\alpha)$ (concatenation and list construction are defined for categories instead of expressions in order to improve readability).

Thus the 'transduction grammar' NL-KS is the triple

$$< G_{NL}, G_{KS}, f >$$

We now specify a minimal categorial grammar of German G_{NL}. A particular of G_{NL} is the analysis of verbs as m-ary predicates, i.e. in the categorial framework, as functions from m NP into S[8] and the analogue treatment of nouns as functor categories[9] taking their attributes as arguments.

Basic categories of NL

S category of sentences
O-N category of "saturated" common noun phrases
NP category of noun phrases (singular terms)
NPR category of proper nouns
(If M_{NP} is the set of noun phrases, M_{NPR} the set of proper nouns
$M_{NPR} \subset M_{NP}$
holds.)

derived categories of NL

S/NP/.../NP (m times) category of m-ary verbs

O-N/NP/.../NP (n times) category of common noun phrases taking n attributes

NP/NP category of prepositions
NP/O-N category of articles (determiners)

syntactic rules (expansion of (1.3.), NL-part)

(1) NP/NP $\frown$ NP $\rightarrow$ NP
(2) NP/O-N $\frown$ O-N $\longrightarrow$ NP
(3) O-N/NP/.../NP (n times) $\frown$ $NP_1 \frown ... NP_n \rightarrow$ O-N
(4) S/NP/.../NP (m times) $\frown$ $NP_1 \frown ... \frown NP_m \longrightarrow$ S

application of f to the basic categories:

f(S) = FORMEL
f(O-N) = LAMBDAABSTRAKT
f(NP) = TERM
f(NPR) = KONSTANTE, with $M_{KONSTANTE} \subset M_{TERM}$

to the derived categories:

f(S/NP/.../NP) = f(S)/f(NP)/.../f(NP) = FORMEL/TERM/.../TERM (m times)
for short: PRAED stel m

f(O-N/NP/.../NP) = f(O-N)/f(NP)/.../f(NP) = LAMBDAABSTRAKT/TERM/.../TERM (n times) where LAMBDAABSTRAKT itself i a functor category in KS:

LAMBDAABSTRAKT = FORMEL/TERM

f(NP/NP) = f(NP)/f(NP) = TERM/TERM
f(NP/O-N) = f(NP)/f(O-N) = TERM/LAMBDAABSTRAKT, for short: QUANT

syntactic rules of KS (expansion of (1.3.) KS part)

(1-KS) [TERM/TERM TERM] $\rightarrow$ TERM
(2-KS) [TERM/LAMBDAABSTRAKT LAMBDAABSTRAKT] $\rightarrow$ TERM
for short:
[QUANT LAMBDAABSTRAKT] $\rightarrow$ TERM
(3-KS) [LAMBDAABSTRAKT/TERM/.../TERM (n times)
$TERM_1$ ... $TERM_n$] $\rightarrow$ LAMBDAABSTRAKT
where an expression
$a_{LAMBDAABSTRAKT} = a_{FORMEL/TERM}$
is written as
[LAMBDA x a x].

In a Lambdaabstrakt

[LAMBDA x [a_1 b_1 ... b_n]x]

a_1 has the function of a n+1-ary predicate (PRAED), seen from the viewpoint of predicate calculus, such that we can rewrite

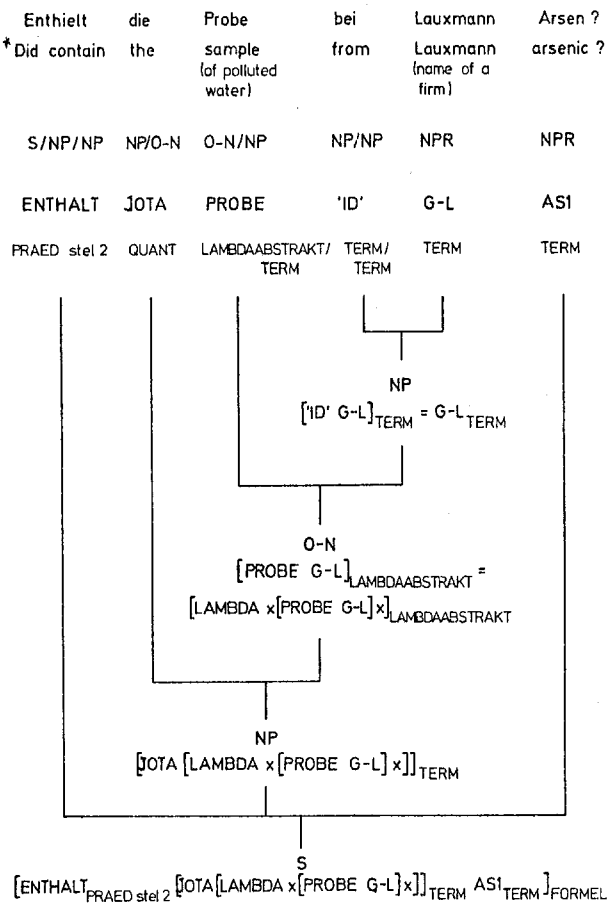
[LAMBDA x [a_1 b_1 ... b_n]x] as
[LAMBDA x [a_1 b_1 ... b_n x]].

(4-KS) [FORMEL/TERM/.../TERM (m times)
$TERM_1$ ... $TERM_m$] $\rightarrow$ FORMEL
for short:
[PRAED stel m $TERM_1$ ... $TERM_m$] $\rightarrow$ FORMEL

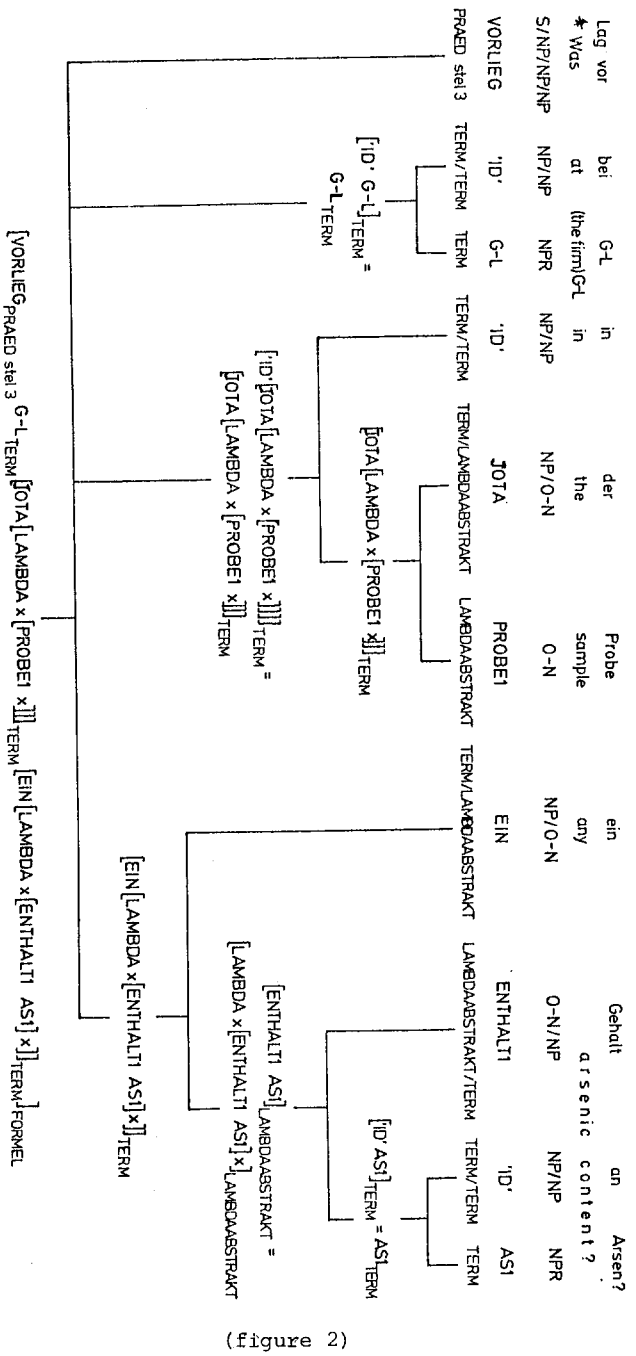
By applying the function f we have got a grammar G_{KS} for our semantic language KS in an inductive way. We now give the following lexical correspondence rules for some non logical expressions of NL, taken from the application area of PLIDIS.

NL word	NL category	KS translation	KS category
Probe ("sample of sewage water")	(a) O-N/NP	PROBE	LAMBDAABSTRAKT/TERM
	(b) O-N	PROBE1	LAMBDAABSTRAKT
enthalten	S/NP/NP	ENTHALT	PRAED stel 2
vorliegen	S/NP/NP/NP	VORLIEG	PRAED stel 3
der, die, das	NP/O-N	JOTA	QUANT
ein	NP/O-N	EIN	QUANT
bei	NP/NP	'ID' (identity: ['ID' a_{TERM}] = a_{TERM}	TERM/TERM
an	NP/NP	'ID'	TERM/TERM
in	NP/NP	'ID'	TERM/TERM
Arsen	NPR	AS1	KONSTANTE
Lauxmann	NPR	G-L	KONSTANTE
Gehalt	O-N/NP	ENTHALT1	LAMBDAABSTRAKT/TERM

With the given syntactic and lexical rules we can generate the following level 1 representations of two natural language sentences:



(figure 1)



(figure 2)

Meaning postulates for generating canonical representations

Both sentences have received different representations on level 1, they are nevertheless synonymous at least as far as the context of information seeking is concerned.

An important principle in COL-semantics is the notion of structural (not lexical) synonymy. The following intuitively valid synonymy postulates (meaning postulates) can be formulated.

(1) A NL noun phrase containing n ($n \geq o$) attributes (category O-N/NP/.../NP)

n times

is synonymous with an NP containing n+1 attributes, where the n+1st attribute is an unspecified "place holder" attribute, under the precondition that the central noun of the NP systematically admits[10] n+1 attributes:

eine Probe	is synonymous with	*eine Probe bei einem Betrieb*
('a sample of sewage water')		('a sample of an industrial plant')

The application of this principle may be iterated.

(2) There are verb classes the elements of which have no descriptive meaning ("non-content verbs"), in German the so called "Funktionsverben", the copula *sein* and others). In such cases the NP as object or subject of the verb is the content bearer or 'principal' NP, e.e. it becomes the predicate of the proposition. Such a sentence is synonymous with a corresponding sentence containing a content verb equivalent in meaning to the content bearing NP. For example:

Arsengehalt liegt in der Probe vor. ('There exists an arsenic content in the sample.')	is synonymous with	*Die Probe enthält Arsen.* ('The sample contains arsenic.')

In such a non-content verb proposition a noun phrase with a place holder attribute can also function as a "second order" principal NP, i.e. its unspecified attribute can be replaced by a "filler" NP, occurring as argument of the non-content verb:

Arsengehalt liegt bei Lauxmann in der Probe vor. is synonymous with *Die Probe bei Lauxmann enthält Arsen.*

Both postulates shall be applied for transducing the level 1 representations of NL sentences into level 2 representations. We first give a definition of 'principal term', i.e. the KS construction corresponding to a 'principal NP'.

(Def.) A principal term in a formula containing as PRAED the translation of a non content verb is a term that is capable, according to its semantic and syntactic structure, to embed other argument terms or the translation of the non content verb as its arguments.

The operationalized version of the two principles is now after having shifted them onto the KS level:

(1: maximality principle)When a NL-expression has n analysis ($n \geq 2$) in level 1 which only differ in the number of arguments, then the level 2 representation consists of the 'maximal' level 1 expression, i.e. the expression containing the largest number of arguments. Any failing arguments are to be substituted by (existentially bound) variables.

(2: transformation principle)
(2.1.) When the PRAED of a formula is the translation of a non-content verb, at least one of its arguments must be a principal term.
(2.2.) A formula containing the translation of a non content verb must be transformed into an expression which contains the PRAED of a principal term as predicate iff there is an unambiguous mapping of the arguments of the translation of the non-content verb

a) into arguments of a principal term

or

b) into a principal term

such that a well-formed formula of level 2 is obtained.

We now state that PROBE and ENTHALT are 'maximal' expressions and PROBE1 and ENTHALT1 must be mapped into them respectively and that further holds:

VORLIEG is the translation of the non-content verb *vorliegen*
PROBE is the PRAED of a second order principal term with respect to a 'plant' argument
ENTHALT is the PRAED of a principal term with respect to a 'sample' argument

Then the two examples of level 1 are mapped into a single representation on level 2:

[ENTHALT[JOTA[LAMBDA x[PROBE G-L x]]]AS1]

The reduction of synonymous structures in the canonical level of representation meets the criteria of economy as they are necessary in a computer system.[11] As we have tried to show, however,it can be based upon general linguistic principles and need not be imputed to the field of "world semantics". On the other side admitting paraphrases as natural language input (as our examples are) improves the systems "cooperativeness" towards the user. In PLIDIS special aspects of the world model are accounted for in the level 3 representations which mirror the relational structure of the data model to some extent. We can not go into the details of the relationship between level 2 and level 3 for reasons of space.

Comparison with other approaches

Language processing systems that are oriented at Montague grammar or model theoretic semantics are being developed among others by Friedman et al., Sondheimer and the PHLIQA1 group. A theoretical discussion of the relationship between model theoretic semantics and AI-semantics can be found in Gunji and Sondheimer, cf. also Hobbs and Rosenschein, St. Bien and Wilks (with a contrary view). The methodological ideas presented here are most closely related with the approach of multi-level semantics pursued in PHLIQA1. But unlike the PHLIQA1 approach we regard the level(s) of linguistic representation not only under the more formal aspect of syntax interpretation but, as the last chapters show, we also take into account aspects of semantics of natural language word classes and structural synonymy.

Notes

1 There are certainly important interactions with empirial semantic work done in the last 10 years, so Ortony and Wilks stress the pervasive influence of Fillmore. Like any other systematic distinction the one between formal linguistic semantics and AI-semantics is somewhat simplifying: Within AI there are semantic approaches which are more or less oriented at formal logic, so the one of McCarthy, Creary or Nash-Webber and Reiter and others. As typical AI-semantic approaches we regard the ones of Schank and his colleagues, Wilks or Charniak (cf. for instance the articles in Charniak and Wilks).

2 Hayes, 9

3 Slightly exaggerating this tendency is formulated by Schank in Schank et al.):"Researchers in NPL (natural language processing in AI) have become less and less concerned with language issues per se. We are more interested in inferencing and memory models for example." (p. 1008)

4 Such systems are presented for instance in Riesbeck, Norman and Rumelhart, and even more programmatically in Schank et al., DeJong. Also in systems conceived as data base interfaces like LIFER (Hendrix) and PLANES (Waltz) "semantic" grammars are used. A theoretical discussion on the role of syntax can be found in Schank et al.

5 i.e. one has to check, whether in systems containing only "part grammars" or working with a syntactic "pre-processing" the syntactic rules which were effectively used, can be combined resulting in a coherent and consistent grammar. Questions of syntactic-semantic and purely semantic grammars underlying parsers are also discussed from a theoretical point of view in Wahlster.

6 The system PLIDIS is described in Kolvenbach, Lötscher and Lutz.

7 The language KS ("Konstruktsprache") is described in Zifonun.

8 Cresswell gives an analogous categorial description for verbs. Like in this minimal grammar in applying the rule of concatenation phenomena of word order are neglected.

9 Keenan and Faltz introduce the category of "function noun" (in our framework O-N/NP)

10 The vague condition of "systematically admitting" is made concrete in PLIDIS by prescribing a semantic "sort" for each argument of a predicate.

11 This reduction is done in PLIDIS with the help of meaning postulates which are interpreted by a theorem prover.

References

Bronnenberg, W.J.H.J./Bunt, H.C./Landsbergen, S.P.J./Scha, R.J.H./Schoenmakers, W.J./van Utteren, E.P.C.:"The Question Answering System PHLIQA1", in: L. Bolc (ed.) "Natural Language Question Answering Systems" (Natural Communication with Computers), Macmillan, London 1980, 217-305.

Charniak, E./Wilks, Y. (eds.): "Computational Semantics", ²ed. North Holland, Amsterdam 1976.

Creary, L.G.: "Propositional Attitudes: Fregean Representation and Simulative Reasoning", Proc. 6th IJCAI Tokyo 1979, 176-182.

Cresswell, M.J.: "Logics and Languages", Methuen, London 1973.

DeJong, G.: "Predictional Substantiation: Two Processes that Comprise Understanding", Proc. 6th IJCAI Tokyo 1979, 217-222.

Friedman, J./Moran, D.B./Warren, D.S.: "Explicit Finite Intensional Models for PTQ and An Interpretation System for Montague Grammar", American Journal of Computational Linguistics Microfiche 74, 1978, 3-96.

Gunji, T./Sondheimer, N.: "The Mutual Relevance of Model-Theoretic Semantics and Artificial Intelligence", unpubl. paper, Department of Computer and Information Science The Ohio State University, February 1979.

Hayes, P.: "On the Difference between Psychology and Artificial Intelligence", AISB quarterly 34 July 1979, 8-9.

Hendrix, G.G.: "LIFER: A Natural Language Interface Facility", Tech. Note 135, AI Center Stanford Research Inst., Menlo Park, California 1976.

Hobbs, J.R./Rosenschein, S.J.: "Making Computational Sense of Montague's Intensional Logic", Artificial Intelligence 9, 1978, 287-306.

Keenan, E./Faltz, L.M.: "Logical Types for Natural Language", UCLA Occasional Papers in Linguistics 3, Fall 1978.

Kolvenbach, M./Lötscher, A./Lutz, H. (eds.): "Künstliche Intelligenz und natürliche Sprache. Sprachverstehen und Problemlösen mit dem Computer", Narr, Tübingen 1979.

McCarthy, J.: "First Order Theories of Individual Concepts and Propositions", in:D. Michie (ed.) Machine Intelligence 9, Edinburgh 1979.

Montague, R.: "Formal Philosophy", ed. by R. Thomason, Yale University Press, New Haven and London 1974.

Nash-Webber, B./Reiter, R.: "Anaphora and Logical Form: On Formal Representations for Natural Language", Proc.5th IJCAI Cambridge Mass. 1977, 121-131.

Norman, D.A./Rumelhart, D.E. (eds.): "Explorations in Cognition", Freeman, San Francisco 1975.

Ortony, A./Wilks, Y.: "Cognitive Science versus Artificial Intelligence", AISB quarterly 34, April 1979, 20-22.

Riesbeck, C.K.: "Conceptual Analysis", in: R.C. Schank (ed.) "Conceptual Information Processing", North Holland, Amsterdam 1975, 83-156.

Schank, R.C./Lebowitz, M./Birnbaum, L.: "Parsing directly into Knowledge Structures", Proc. 6th IJCAI Tokyo 1979, 772-777.

Schank, R.C. et al.: "Panel on Natural Language Processing",Proc. 5th IJCAI Cambridge Mass. 1977, 1007-1013.

Sondheimer, N.K./Gunji, T.: "Applying Model-Theoretic Semantics to Natural Language Understanding: Representation and Question Answering", Proc. 7th COLING Bergen 1978.

St. Bien, J.: "Computational Explication of Intensionality", Preprints 6th COLING Ottawa 1976.

Wahlster, W.: "ATN und semantisch-pragmatische Analysesteuerung, in: T. Christaller/D. Metzing (ed.) "Augmented Transition Network Grammatiken", vol I, Einhorn, Berlin 1979, 167-185.

Waltz, D.L.: "An English Language Question Answering System for a Large Relational Database", CACM 21.7, July 1978.

Wilks, Y.: "Philosophy of Language", in: E. Charniak, Y. Wilks (eds.) Computational Semantics, ²ed. North Holland, Amsterdam 1976, 205-233.

Zifonun, G.: "Formale Repräsentation natürlichsprachlicher Äußerungen", in: Kolvenbach, Lötscher, Lutz (eds.), Künstliche Intelligenz und natürliche Sprache. Sprachverstehen und Problemlösen mit dem Computer, Narr, Tübingen 1979, 93-134.