


# An Empirical Assessment of Semantic Interpretation

Martin Romacker & Udo Hahn

Text Understanding Lab,  Group,  
Freiburg University, Freiburg, D-79085, Germany

{mr,hahn}@coling.uni-freiburg.de

## Abstract

We introduce a framework for semantic interpretation in which dependency structures are mapped to conceptual representations based on a parsimonious set of interpretation schemata. Our focus is on the empirical evaluation of this approach to semantic interpretation, i.e., its quality in terms of recall and precision. Measurements are taken with respect to two real-world domains, *viz.* information technology test reports and medical finding reports.

## 1 Introduction

Semantic interpretation has been an actively investigated issue on the research agenda of the logic-based paradigm of NLP in the late eighties (e.g., Charniak and Goldman (1988), Moore (1989), Pereira and Pollack (1991)). With the emergence of empirical methodologies in the early nineties, attention has almost completely shifted away from this topic. Since then, semantic issues have mainly been dealt with under a lexical perspective, *viz.* in terms of the resolution of lexico-semantic ambiguities (e.g., Schütze (1998), Pedersen and Bruce (1998)) and the generation of lexical hierarchies from large text corpora (e.g., Li and Abe (1996), Hirakawa et al. (1996)) massively using statistical techniques.

The research on semantic interpretation that was conducted in the pre-empiricist age of NLP was mainly driven by an interest in logical formalisms as carriers for appropriate semantic representations of NL utterances. With this representational bias, computational matters — how can semantic representation structures be properly derived from parse trees for a *large* variety of linguistic phenomena? — became a secondary issue. In particular, this research lacked entirely quantitative data reflecting the accuracy of the proposed semantic interpretation mechanisms on real-world language data.

One might be tempted to argue that recent evaluation efforts within the field of information extraction (IE) systems (Chinchor et al., 1993) are going to remedy this shortcoming. Given, however, the fixed number of knowledge templates and the restricted types of entities, locations, and events they encode

as target information to be extracted, one readily realizes that such an evaluation framework provides, at best, a considerably biased, overly selective test environment for judging the understanding potential of text analysis systems which are not tuned for this special application.

On the other hand, the IE experiments clearly indicate the need for a quantitative assessment of the interpretative performance of natural language understanding systems. We will focus on this challenge and propose such a general evaluation framework. We first outline the model of semantic interpretation underlying our approach and then focus on its empirical assessment for two basic syntactic structures of the German language, *viz.* genitives and auxiliary constructions, in two domains.

## 2 The Basic Model for Semantic Interpretation

The problem of semantic interpretation can be described as the mapping from syntactic to semantic (or conceptual) representation structures. In our approach, the syntactic representation structures are given as dependency graphs (Hahn et al., 1994). Unlike constituency-based syntactic descriptions, dependency graphs consist of lexical nodes only, and these nodes are connected by vertices, each one of which is labeled by a particular dependency relation (cf. Figure 1).

For the purpose of semantic interpretation, dependency graphs can be decomposed into *semantically interpretable subgraphs*.<sup>1</sup> Basically, two types of semantically interpretable subgraphs can be distinguished. The first one consists of lexical nodes which are labeled by content words only (lexical instances of verbs, nouns, adjectives or adverbs) and which are *directly* linked by a *single* dependency relation of any type whatsoever. Such a subgraph is illustrated in Figure 1 by *Speicher – genatt – Computers*. The second type of subgraph is also delimited by labels of content words but, in addition, a series of  $n = 1 \dots 4$  intermediary lexical nodes may

<sup>1</sup>This notion and all subsequent criteria for interpretation are formally described in Romacker et al. (1999).

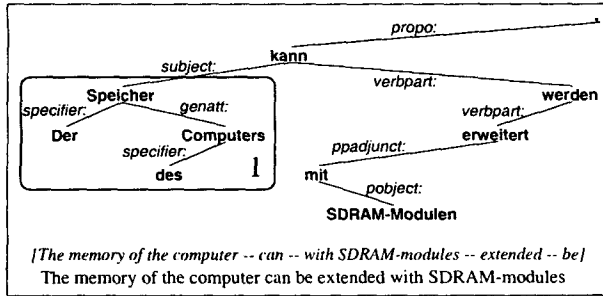


Figure 1: Dependency Graph for a Sample Sentence

appear between these content words, all of which are labeled by non-content words (such as auxiliary or modal verbs, prepositions). Hence, in contrast to direct linkage we speak here of *indirect* linkage between content words. Such a subgraph, with two intervening non-content words – the modal “*kann*” and the auxiliary “*werden*” –, is given in Figure 1 by *Speicher* – subject – *kann* – verbpart – *werden* – verbpart – *erweitert*. Another subgraph with just one intervening non-content word – the preposition “*mit*” – is illustrated by *erweitert* – ppadjunct – *mit* – pobject – *SDRAM-Modulen*. From these considerations follows that, e.g., the subgraph spanned by *Speicher* and *SDRAM-Modulen* does not form a semantically interpretable subgraph, since the content word *erweitert* intervenes on the linking path.

Our approach to semantic interpretation subscribes to the principles of locality and compositionality. It operates on discrete and well-defined units (subgraphs) of the parse tree, and the results of semantic interpretation are incrementally combined by fusing semantically interpretable subgraphs.

As semantic target language we have chosen the framework of KL-ONE-type description logics (DL) (Woods and Schmolze, 1992). Since these logics are characterized by a settheoretical semantics we stay on solid formal ground. Furthermore, we take advantage of the powerful inference engine of DL systems, the description classifier, which turns out to be essential for embedded reasoning during the semantic interpretation process. By equating the semantic representation language with the conceptual one, we follow arguments discussed by Allen (1993).

The basic idea for semantic interpretation is as follows: Each lexical surface form of a content word is associated with a set of concept identifiers representing its (different) lexical meanings. This way, lexical ambiguity is accounted for. These conceptual correlates are internal to the domain knowledge base, where they are described by a list of attributes or conceptual roles, and corresponding restrictions on permitted attribute values or role fillers are associated with them.

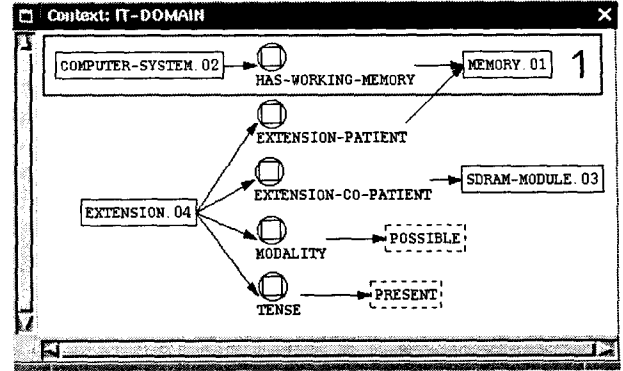


Figure 2: Concept Graph for a Sample Sentence

As an example, consider the description for the concept *COMPUTER-SYSTEM*. It may be characterized by a set of roles, such as *HAS-HARD-DISK* or *HAS-WORKING-MEMORY*, with corresponding restrictions on the concept types of potential role fillers. *HAS-WORKING-MEMORY*, e.g., sanctions only fillers of the concept type *MEMORY*. These conceptual constraints are used for semantic filtering, i.e., for the elimination of syntactically admissible dependency graphs which, nevertheless, do not have a valid semantic interpretation.

Semantic interpretation, in effect, boils down to finding appropriate conceptual relations in the domain knowledge that link the conceptual correlates of the two content words spanning the semantically interpretable subgraph, irrespective of whether a direct or an indirect linkage holds at the syntactic level. Accordingly, Figure 2 depicts the semantic/conceptual interpretation of the dependency structure given in Figure 1. Instances representing the concrete discourse entities and events in the sample sentence are visualized as solid rectangles containing a unique identifier (e.g., *COMPUTER-SYSTEM.02*). Labeled and directed edges indicate instance roles. Dashed rectangles characterize symbols used as makers for tense and modality.<sup>2</sup>

Note that in Figure 2 each tuple of content words which configures a minimal subgraph in Figure 1 has already received an interpretation in terms of a relation linking the conceptual correlates. For example, *Speicher* – genatt – *Computers* (cf. Figure 1, box 1) is mapped to *COMPUTER-SYSTEM.02 HAS-WORKING-MEMORY MEMORY.01* (cf. Figure 2, box 1). However, the search for a valid conceptual relation is not only limited to a simple one-link slot-filler structure. We rather may determine conceptual relation paths between conceptual correlates of lexical items, the length of which may be greater than 1.

<sup>2</sup>We currently do not further interpret the information contained in tense or modality markers.

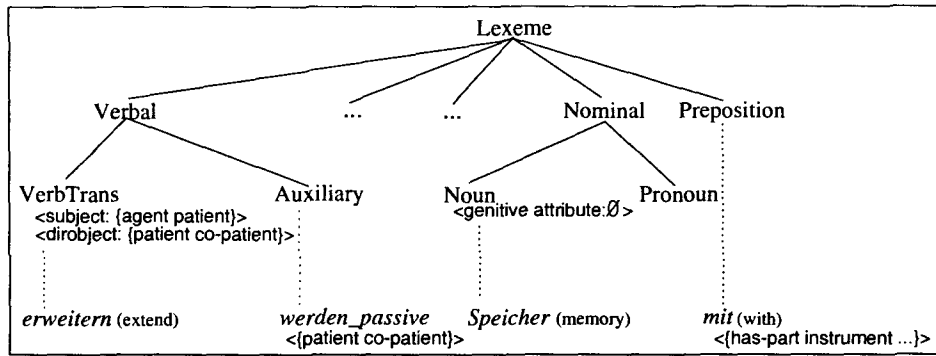


Figure 3: Fragment of the Lexeme Class Hierarchy

(Thus, the need for role composition in the DL language becomes evident.) The directed search in the concept graph of the domain knowledge requires sophisticated structural and topological constraints to be manageable at all. These constraints are encapsulated in a special path finding and path evaluation algorithm specified in Markert and Hahn (1997).

Besides these conceptual constraints holding in the domain knowledge, we further attempt to reduce the search space for finding relation paths by two kinds of syntactic criteria. First, the search may be constrained by the type of dependency relation holding between the content words of the currently considered semantically interpretable subgraph (direct linkage), or it may be constrained by the intervening lexical material, *viz.* the non-content words (indirect linkage). Each of these syntactic constraints has an immediate mapping to conceptual ones.

For some dependency configurations, however, no syntactic constraints may apply. Such a case of unconstrained semantic interpretation (e.g., for genitive attributes directly linked by the *genatt* relation) leads to an exhaustive directed search in the knowledge base in order to find all conceptually compatible role fillings among the two concepts involved.

Syntactic restrictions on semantic interpretation either come from lexeme classes or concrete lexemes. They are organized in terms of the lexeme class hierarchy superimposed on the fully lexicalized dependency grammar we use (Hahn et al., 1994). In the fragment depicted in Figure 3, the lexeme class of transitive verbs, VERBTRANS, requires that whenever a subject dependency relation is encountered, semantic interpretation is constrained to the conceptual roles AGENT or PATIENT and all their subrelations (such as EXTENSION-PATIENT). All other conceptual roles are excluded from the subsequent semantic interpretation. Exploiting the property inheritance mechanisms provided by the hierarchic organization of the lexicalized dependency grammar, all concrete lexemes subsumed by the lexeme class

VERBTRANS, like “*erweitern*” (*extend*), inherit the corresponding constraint. However, there are lexeme classes such as NOUN which do not render any constraints for dependency relations such as evidenced by *gen[itive] att[ribute]* (cf. Fig. 3).

It may even happen that such restrictions can only be attached to concrete lexemes in order to avoid overgeneralization. Fortunately, we observed that this only happened to be the case for closed-class, i.e., non-content words. Accordingly, in Figure 3 the preposition “*with*” is characterized by the constraint that only the conceptual roles HAS-PART, INSTRUMENT, etc. must be taken into consideration for semantic interpretation.

Since the constraints at the lexeme class or the lexeme level are hard-wired in the class hierarchy, we refer to the mapping of dependency relations (or idiosyncratic lexemes) to a set of conceptual relations (expanded to their transitive closure) as *static* interpretation. In contradistinction, the computation of relation paths for tuples of concepts during the sentence analysis process is called *dynamic* interpretation, since the latter process incorporates additional conceptual constraints on the fly.

The above-mentioned conventions allow the specification of high-level semantic interpretation schemata covering a large variety of different syntactic constructions by a single schema. For instance, each syntactic construction for which no conceptual constraints apply (e.g., the interpretation of genitives, most adjectives, etc.) receives its semantic interpretation by instantiating *the same* interpretation schema (Romacker et al., 1999). The power of this approach comes from the fact that these high-level schemata are instantiated in the course of the parsing process by exploiting the dense specifications of the inheritance hierarchies both at the grammar level (the lexeme class hierarchy), as well as the conceptual level (the concept and role hierarchies).

We currently supply up to ten semantic interpretation schemata for declaratives, relatives, and pas-

sives at the clause level, complement subcategorization via PPs, auxiliaries, all tenses at the VP level, pre- and and postnominal modifiers at the NP level, and anaphoric expressions. We currently do not account for control verbs (work in progress), coordination and quantification.

### 3 The Evaluation of Semantic Interpretation

In this section, we want to discuss, for two particular types of German language phenomena, the adequacy of our approach in the light of concrete language data taken from the two corpora we work with. This part of the enterprise, the empirical assessment of semantic interpretation, is almost entirely neglected in the literature (for two notable exceptions, cf. Bonnema et al. (1997) and Bean et al. (1998)).

Though similarities exist (*viz.* dealing with the performance of NLP systems in terms of their ability to generate semantic/conceptual structures), the semantic interpretation (SI) task has to be clearly distinguished from the information extraction (IE) task and its standard evaluation settings (Chinchor et al., 1993). In the IE task, a small *subset* of the templates from the entire domain is selected into which information from the texts are mapped. Also, the design of these templates focus on particularly *interesting* facets (roles, in our terminology), so that an IE system does not have to deal with the full range of qualifications that might occur — even relating to relevant, selected concepts. Note that in any case, *a priori* relevance decisions limit the range of *a posteriori* fact retrieval.

The SI task, however, is far less restricted. We here evaluate the adequacy of the conceptual representation structures relating, in principle (only restricted, of course, by the limits of the knowledge acquisition devices), to the *entire* domain of discourse, with *all* qualifications mentioned in a text. Whether these are relevant or not for a particular application has to be determined by subsequent data/knowledge cleansing. In this sense, semantic interpretation might deliver the raw data for transformation into appropriate IE target structures. Only because of feasibility reasons, the designers of IE systems equate IE with SI. The cross-linking of IE and SI tasks, however, bears the risk of having to determine, in advance, what will be relevant or not for later retrieval processes, assumptions which are likely to be flawed by the dynamics of domains and the unpredictability of the full range of interests of prospective users.

#### 3.1 Methodological Issues

Our methodology to deal with the evaluation of semantic interpretation is based on a triple division of test conditions. The first category relates to checks

whether so-called *static* constraints, effected by the mapping from a single dependency relation to one or more conceptual relations, are valid (cf. Figure 3 for restrictions of this type). Second, one may investigate the appropriateness of the results from the search of the domain knowledge base, i.e., whether a relation between two concepts can be determined at all, and, if so, whether that relation (or role chain) is adequate. The conceptual constraints which come into play at this stage of processing are here referred to as *dynamic* constraint propagation, since they are to be computed on the fly, while judging the validity of the role chain in question.<sup>3</sup> Third, interactions between the above-mentioned static constraints and dynamic constraint propagation may occur. This is the case for the interpretation of auxiliaries or prepositions, where intervening lexical material and associated constraints have to be accounted for simultaneously.

In our evaluation study, we investigated the effects of category II and category III phenomena by considering genitives and modal as well as auxiliary verbs, respectively. The knowledge background is constituted by a domain ontology that is divided into an upper generic part (containing about 1,500 concepts and relations) and domain-specific extensions. We here report on the two specialized domains we deal with — a hardware-biased information technology (IT) domain model and an ontology covering parts of anatomical medicine (MED). Each of these two domain models adds roughly about 1,400 concepts and relations to the upper model. Corresponding lexeme entries in the lexicon provide linkages to the entire ontology. In order to avoid error chaining, we always assume a correct parse to be delivered for the semantic interpretation process.

We took a random selection of 54 texts (comprising 18,500 words) from the two text corpora, *viz.* IT test reports and MEDical finding reports. For evaluation purposes (cf. Table 1), we concentrated on the interpretation of genitives (as an instance of direct linkage; GEN) and on the interpretation of periphrastic verbal complexes, i.e., passive, temporal and modal constructions (as instances of indirect linkage; MODAUX).

The choice of these two grammatical patterns allows us to ignore the problems caused by syntactic ambiguity, since in our data no structural am-

<sup>3</sup>Note that computations at the domain knowledge level which go beyond mere type checking are usually located outside the scope the semantic considerations. This is due to the fact that encyclopedic knowledge and its repercussions on the understanding process are typically not considered part of the semantic interpretation task proper. While this may be true from a strict linguistic point of view, from the computational perspective of NLP this position cannot seriously be maintained. Even more so, when semantic and conceptual representations are collapsed.

biguities occurred. If one were to investigate the combined effects of syntactic ambiguity and semantic interpretation the evaluation scenario had to be changed. Methodologically, the first step were to explore the precision of a semantic interpretation task without structural ambiguities (as we do) and then, in the next step, incorporate the treatment of syntactic ambiguities (e.g., by semantic filtering devices, cf. Bonnema et al. (1997)).

Several guidelines were defined for the evaluation procedure. A major issue dealt with the correctness of a semantic interpretation. In cases *with* interpretation, we considered a semantic interpretation to be a *correct* one, if the conceptual relation between the two concepts involved was considered adequate by introspection (otherwise, *incorrect*). This qualification is not as subjective as it may sound, since we applied really strict conditions adjusted to the fine-grained domain knowledge.<sup>4</sup> Interpretations were considered to be correct in those cases which contained exactly one relation, as well as cases of semantical/conceptual ambiguities (up to three readings, the most), presumed the relation set contained the correct one.<sup>5</sup> A special case of incorrectness, called *nil*, occurred when no relation path could be determined though the two concepts under scrutiny were contained in the domain knowledge base and an interpretation should have been computed.

We further categorized the cases where the system failed to produce an interpretation due to at least one concept specification missing (with respect to the two linked content words in a semantically interpretable subgraph). In all those cases *without* interpretation, insufficient coverage of the *upper model* was contrasted with that of the two *domain models* in focus, MED and IT, and with cases in which concepts referred to *other domains*, e.g., fashion or food. Ontological subareas that could neither be assigned to the upper model nor to particular domains were denoted by phrases referring to *time* (e.g., “the beginning of the year”), *space* (e.g.,

<sup>4</sup>The majority of cases were easy to judge. For instance, “the infiltration of the stroma” resulted in a correct reading – STROMA being the PATIENT of the INFILTRATION event –, as well as in an incorrect one – being the AGENT of the INFILTRATION. Among the incorrect semantic interpretations we also categorized, e.g., the interpretation of the expression “the prices of the manufacturers” as a conceptual linkage from PRICE via PRICE-OF to PRODUCT via HAS-MANUFACTURER to MANUFACTURER (this type of role chaining can be considered an intriguing example of the embedded reasoning performed by the description logic inference engine), since it did not account for the interpretation that MANUFACTURERS fix PRICES as part of their marketing strategies. After all, correct interpretations always boiled down to entirely evident cases, e.g., HARD-DISK PART-OF COMPUTER.

<sup>5</sup>At the level of semantic interpretation, the notion of semantic ambiguity relates to the fact that the search algorithm for valid conceptual relation paths retrieves more than a single relation (chain).

“the surface of the storage medium”), and *abstract* notions (e.g., “the acceptance of IT technology”). Finally, we further distinguished *evaluative* expressions (e.g., “the advantages of plasma display”) from *figurative* language, including idiomatic expressions (e.g., “the heart of the notebook”).

At first glance, the choice of genitives may appear somewhat trivial. From a syntactic point of view, genitives are directly linked and, indeed, constitute an easy case to deal with at the dependency level. From a conceptual perspective, however, they provide a real challenge. Since *no static* constraints are involved in the interpretation of genitives (cf. Figure 3, lexeme class NOUN) and, hence, no prescriptions of (dis)allowed conceptual relations are made, an unconstrained search (apart from connectivity conditions imposed on the emerging role chains) of the domain knowledge base is started. Hence, the main burden rests on the *dynamic* constraint processing part of semantic interpretation, i.e., the path finding procedure muddling through the complete domain knowledge base in order to select the adequate conceptual reading(s). Therefore, genitives make a strong case for test category II mentioned above.

Dependency graphs involving modal verbs or auxiliaries are certainly more complex at the syntactic level, since the corresponding semantically interpretable subgraphs may be composed of up to six lexical nodes. However, all intervening non-content-word nodes accumulate constraints for the search of a valid relation for semantic interpretations and, hence, allows us to test category III phenomena. The search space is usually pruned, since only those relations that are sanctioned by the intervening nodes have to be taken into consideration.

### 3.2 Evaluation Data

We considered a total of almost 250 genitives in all these texts, from which about 59%/33% (MED/IT) received an interpretation.<sup>6</sup> Out of the total loss due to incomplete conceptual coverage, 56%/58% (23 of 41 genitives/57 of 98 genitives) can be attributed to insufficient coverage of the domain models. Only the remaining 44%/42% are due to the residual factors listed in Table 1.

In our sample, the number of syntactic constructions containing modal verbs or auxiliaries amount to 292 examples. Compared to genitives, we obtained a more favorable recall for both domains: 66% for MED and 40% for IT. As for genitives, lacking interpretations, in the majority of cases, can be attributed to insufficient conceptual coverage. For the IT domain, however, a dramatic increase in the number of missing concepts is due to gaps in the upper model (78 or 63%) indicating that a large number of

<sup>6</sup>Confidence intervals at a 95% reliability level are given in brackets in Table 1.

	MED-GEN	IT-GEN	MED-MODAU	IT-MODAU
# texts	29	25	29	25
# words	4,300	14,200	4,300	14,200
recall	57%	31%	66%	40%
precision	97%	94%	95%	85%
# occurrences ...	100	147	58	234
... <b>with interpretation</b>	59 (59%)	49 (33%)	40 (69%)	111 (47%)
[confidence intervals]	[48%-67%]	[24%-41%]	[56%-81%]	[40%-53%]
..... correct (single reading)	53 (53%)	28 (19%)	38 (66%)	88 (38%)
..... correct (multiple readings)	4 (4%)	18 (12%)	0 (0%)	6 (3%)
..... incorrect	0	3	0	14
..... nil	2	0	2	3
... <b>without interpretation</b>	41 (41%)	98 (67%)	18 (31%)	123 (53%)
..... domain model (MED/IT)	23 (23%)	57 (39%)	11 (19%)	42 (34%)
..... upper model	3	23	5	78
..... other domains	0	4	0	0
..... time	0	15	0	1
..... space	7	8	0	5
..... abstracta, generics	11	12	0	16
..... evaluative expressions	0	8	0	3
..... figurative language	1	17	2	24
..... miscellaneous	0	1	0	3

Table 1: Empirical Results for the Semantic Interpretation of Genitives (GEN) and Modal Verbs and Auxiliaries (MODAU) in the IT and MED domains

essential concepts for verbs were not modeled. Also, figurative speech plays a more important role in IT with 24 occurrences. Both observations mirror the fact that IT reports are linguistically far less constrained and are rhetorically more advanced than their MED counterparts.

Another interesting observation which is not made explicit in Table 1 concerns the distribution of modal verbs and auxiliaries. In MED, we encountered 57 passives and just one modal verb and no temporal auxiliaries, i.e., our data are in line with prevailing findings about the basic patterns of medical sublanguage (Dunham, 1986). For the IT domain, corresponding occurrences were far less biased, *viz.* 80 passives, 131 modal verbs, and 23 temporal auxiliaries. Finally, for the two domains 25 samples contained both modal verbs and auxiliaries, thus forming semantically interpretable subgraphs with four word nodes.

One might be tempted to formulate a null hypothesis concerning the detrimental impact of the length of semantically interpretable subgraphs (i.e., the number of intervening lexical nodes carrying non-content words) on the quality of semantic interpretation. In order to assess the role of the length of the path in a dependency graph, we separately investigated the results for these subclasses of combined verbal complexes. From the entire four-node set (cf. Table 2) with 25 occurrences (3 for MED and 22 for IT), 16 received an interpretation (3 for MED, 13 for IT). While we neglect the MED data due to the small absolute numbers, the IT domain revealed

	MED 4-nodes	IT 4-nodes
recall	–	59%
precision	–	85%
# occurrences ...	3	22
... <b>with interpretation</b>	3	13
..... correct	3	11

Table 2: Interpretation Results for Semantically Interpretable Graphs Consisting of Four Nodes

59% recall and 85% precision. If we compare this to the overall figures for recall (40%) and precision (85%), the data might indicate a gain in recall for longer subgraphs, while precision keeps stable.

The results we have worked out are just a first step into a larger series of broader and deeper evaluation efforts. The concrete values we present, sobering as they may be for recall (57%/31% for genitives and 66%/40% for modal verbs and auxiliaries), encouraging, however, for precision (97%/94% for genitives and 95%/85% for modal verbs and auxiliaries), can only be interpreted relative to other data still lacking on a broader scale.

As with any such evaluation, idiosyncrasies of the coverage of the knowledge bases are inevitably tied with the results and, thus, put limits on too far-reaching generalizations. However, our data reflect the intention to submit a knowledge-intensive text understander to a realistic, i.e., conceptually unconstrained and therefore “unfriendly” test environment.

Judged from the figures of our recall data, there is no doubt, whatsoever, that conceptual coverage of the domain constitutes *the* bottleneck for any knowledge-based approach to NLP.<sup>7</sup> Sublanguage differences are also mirrored systematically in these data, since medical texts adhere more closely to well-established concept taxonomies and writing standards than magazine articles in the IT domain.

## 4 Related Work

After a period of active research within the logic-based paradigm (e.g., Charniak and Goldman (1988), Moore (1989), Pereira and Pollack (1991)), work on semantic interpretation has almost ceased with the emergence of the empiricist movement in NLP (cf. Bos et al. (1996) for one of the more recent studies dealing with logic-based semantic interpretation in the framework of the VERBMOBIL project). Only few methodological proposals for semantic computations were made since then (e.g., higher-order colored unification as a mechanism to avoid over-generation inherent to unconstrained higher-order unification (Gardent and Kohlhase, 1996)). An issue which has lately received more focused attention are ways to cope with the tremendous complexity of semantic interpretations in the light of an exploding number of (scope) ambiguities. Within the underspecification framework of semantic representations, e.g., Dörre (1997) proposes a polynomial algorithm which constructs packed semantic representations directly from parse forests.

All the previously mentioned studies (with the exception of the experimental setup in Dörre (1997)), however, lack an empirical foundation of their various claims. Though the MUC evaluation rounds (Chinchor et al., 1993) yield the flavor of an empirical assessment of semantic structures, their scope is far too limited to count as an adequate evaluation platform for semantic interpretation. Nirenburg et al. (1996) already criticize the 'black-box' architecture underlying MUC-style evaluations, which precludes to draw serious conclusions from the shortcomings of MUC-style systems as far as single linguistic modules are concerned. More generally, in this paper the rationale underlying *size* (of the lexicons, knowledge or rule bases) as the major assessment category is questioned. Rather dimensions relating to the *depth* and *breadth* of the knowledge sources involved in complex system behavior should be taken more seriously into consideration. This is exactly what we intended to provide in this paper.

As far as evaluation studies are concerned dealing with the assessment of semantic interpretations, few

<sup>7</sup>At least for the medical domain, we are currently actively pursuing research on the semiautomatic creation of large-scale ontologies from weak knowledge sources (medical terminologies); cf. Schulz and Hahn (2000).


have been carried out, some of which under severe restrictions. For instance, Bean et al. (1998) narrow semantic interpretation down to a very limited range of spatial relations in anatomy, while Gomez et al. (1997) bias the result by preselecting only those phrases that were already covered by their domain models, thus optimizing for precision while shunting aside recall considerations.

A recent study by Bonnema et al. (1997) comes closest to a serious confrontation with a wide range of real-world data (Dutch dialogues on a train travel domain). This study proceeds from a corpus of annotated parse trees to which are assigned type-logical formulae which express the corresponding semantic interpretation. The goal of this work is to compute the most probable semantic interpretation for a given parse tree. Accuracy (i.e., precision) is rather high and ranges between 89,2%–92,3% depending on the training size and depth of the parse tree. Our accuracy criterion is weaker (the intended meaning must be included in the set of all readings), which might explain the slightly higher rates we achieve for precision. However, this study does not distinguish between different syntactic constructions that undergo semantic interpretation, nor does it consider the level of *conceptual* interpretation (we focus on) as distinguished from the level of *semantic* interpretation to which Bonnema et al. refer.

## 5 Conclusions

The evaluation of the quality and adequacy of semantic interpretation data is still in its infancy. Our approach which confronts semantic interpretation devices with a random sample of textual real-world data, without intentionally constraining the selection of these language data, is a real challenge for the proposed methodology and it is unique in its experimental rigor.

However, our work is just a step in the right direction rather than giving a complete picture or allowing final conclusions. Two reasons may be given for the lack of such experiments. First, interest in the deeper conceptual aspects of text interpretation has ceased in the past years, with almost all efforts devoted to robust and shallow syntactic processing of large data sets. This also results in a lack of sophisticated semantic and conceptual specifications, in particular, for larger text analysis systems. Second, providing a gold standard for semantic interpretation is, in itself, an incredibly underconstrained and time-consuming process for which almost no resources have been allocated in the NLP community up to now.

**Acknowledgements.** We want to thank the members of the  group for close cooperation. Martin Romacker is supported by a grant from DFG (Ha 2097/5-1).

## References

- James F. Allen. 1993. Natural language, knowledge representation, and logical form. In M. Bates and R. M. Weischedel, editors, *Challenges in Natural Language Processing*, pages 146–175. Cambridge: Cambridge University Press.
- Carol A. Bean, Thomas C. Rindflesch, and Charles A. Sneiderman. 1998. Automatic semantic interpretation of anatomic spatial relationships in clinical text. In *Proceedings of the 1998 AMIA Annual Fall Symposium.*, pages 897–901. Orlando, Florida, November 7–11, 1998.
- Remko Bonnema, Rens Bod, and Remko Scha. 1997. A DOP model for semantic interpretation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the ACL*, pages 159–167. Madrid, Spain, July 7–12, 1997.
- Johan Bos, Björn Gambäck, Christian Lieske, Yoshiki Mori, Manfred Pinkal, and Karsten Worm. 1996. Compositional semantics in VERBMOBIL. In *COLING'96 – Proceedings of the 16th International Conference on Computational Linguistics*, pages 131–136. Copenhagen, Denmark, August 5–9, 1996.
- Eugene Charniak and Robert Goldman. 1988. A logic for semantic interpretation. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 87–94. Buffalo, New York, U.S.A., 7–10 June 1988.
- Nancy Chinchor, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: an analysis of the third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–447.
- Jochen Dörre. 1997. Efficient construction of underspecified semantics under massive ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the ACL*, pages 386–393. Madrid, Spain, July 7–12, 1997.
- George Dunham. 1986. The role of syntax in the sublanguage of medical diagnostic statements. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 175–194. Hillsdale, NJ & London: Lawrence Erlbaum.
- Claire Gardent and Michael Kohlhase. 1996. Higher-order coloured unification and natural language semantics. In *ACL'96 – Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 1–9. Santa Cruz, California, U.S.A., 24–27 June 1996.
- Fernando Gomez, Carlos Segami, and Richard Hull. 1997. Determining prepositional attachment, prepositional meaning, verb meaning and thematic roles. *Computational Intell.*, 13(1):1–31.
- Udo Hahn, Susanne Schacht, and Norbert Bröker. 1994. Concurrent, object-oriented natural language parsing: the PARSETALK model. *International Journal of Human-Computer Studies*, 41(1/2):179–222.
- Hideki Hirakawa, Zhonghui Xu, and Kenneth Haase. 1996. Inherited feature-based similarity measure based on large semantic hierarchy and large text corpus. In *COLING'96 – Proceedings of the 16th International Conference on Computational Linguistics*, pages 508–513. Copenhagen, Denmark, August 5–9, 1996.
- Hang Li and Naoki Abe. 1996. Clustering words with the MDL principle. In *COLING'96 – Proceedings of the 16th International Conference on Computational Linguistics*, pages 4–9. Copenhagen, Denmark, August 5–9, 1996.
- Katja Markert and Udo Hahn. 1997. On the interaction of metonymies and anaphora. In *IJCAI'97 – Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 1010–1015. Nagoya, Japan, August 23–29, 1997.
- Robert C. Moore. 1989. Unification-based semantic interpretation. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 33–41. Vancouver, B.C., Canada, 26–29 June 1989.
- Sergei Nirenburg, Kavi Mahesh, and Stephen Beale. 1996. Measuring semantic coverage. In *COLING'96 – Proceedings of the 16th International Conference on Computational Linguistics*, pages 83–88. Copenhagen, Denmark, August 5–9, 1996.
- Ted Pedersen and Rebecca Bruce. 1998. Knowledge lean word-sense disambiguation. In *AAAI'98 – Proceedings of the 15th National Conference on Artificial Intelligence*, pages 800–805. Madison, Wisconsin, July 26–30, 1998.
- Fernando C.N. Pereira and Martha E. Pollack. 1991. Incremental interpretation. *Artificial Intelligence*, 50(1):37–82.
- Martin Romacker, Katja Markert, and Udo Hahn. 1999. Lean semantic interpretation. In *IJCAI'99 – Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 868–875. Stockholm, Sweden, July 31 – August 6, 1999.
- Stefan Schulz and Udo Hahn. 2000. Knowledge engineering by large-scale knowledge reuse: experience from the medical domain. In *Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning*. Breckenridge, CO, USA, April 12–15, 2000.
- Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.
- William A. Woods and James G. Schmolze. 1992. The KL-ONE family. *Computers & Mathematics with Applications*, 23(2/5):133–177.