On Distance between Deep Syntax and Semantic Representation

Václav Novák
Institute of Formal and Applied Linguistics
Charles University
Praha, Czech Republic
novak@ufal.mff.cuni.cz

Abstract

We present a comparison of two formalisms for representing natural language utterances, namely deep syntactical *Tectogrammatical Layer* of Functional Generative Description (FGD) and a semantic formalism, *MultiNet*. We discuss the possible position of MultiNet in the FGD framework and present a preliminary mapping of representational means of these two formalisms.

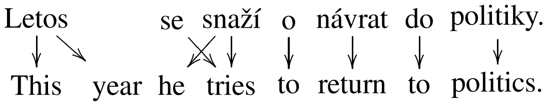
1 Introduction

The Prague Dependency Treebank 2.0 (PDT 2.0) described in Sgall et al. (2004) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5M words), and complex semantic (tectogrammatical) annotation (0.8M words); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level.

The theoretical basis of the treebank lies in the Functional Generative Description (FGD) of language system by Sgall et al. (1986).

PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current computational-linguistics research needs. The corpus itself is embedded into the latest annotation technology. Software tools for corpus search, annotation, and language analysis are included. Extensive documentation (in English) is provided as well.

An example of a tectogrammatical tree from PDT 2.0 is given in figure 1. Function words are removed, their function is preserved in node attributes (*grammatemes*), information structure is annotated in terms of topic-focus articulation, and every node receives detailed semantic label corresponding to its function in the utterance (e.g., *addressee*, *from_where*, *how_often*, . . .). The square node indicates an obligatory but missing valent. The tree represents the following sentence:



(1)

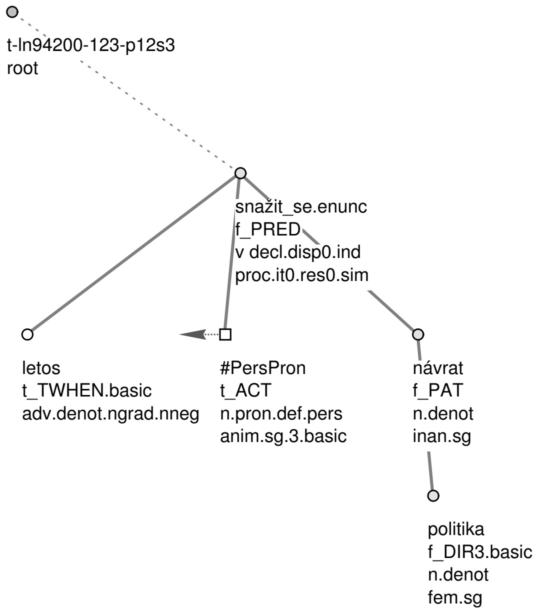


Figure 1: Tectogrammatical tree of sentence (1)

1.1 MultiNet

The representational means of Multilayered Extended Semantic Networks (MultiNet), which are

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 78–85,
Sydney, July 2006. ©2006 Association for Computational Linguistics

described in Helbig (2006), provide a universally applicable formalism for treatment of semantic phenomena of natural language. To this end, they offer distinct advantages over the use of the classical predicate calculus and its derivatives. The knowledge representation paradigm and semantic formalism MultiNet is used as a common backbone for all aspects of natural language processing (be they theoretical or practical ones). It is continually used for the development of intelligent information and communication systems and for natural language interfaces to the Internet. Within this framework, it is subject to permanent practical evaluation and further development.

The semantic representation of natural language expressions by means of MultiNet is mainly independent of the considered language. In contrast, the syntactic constructs used in different languages to describe the same content are obviously not identical. To bridge the gap between different languages we can employ the deep syntactico-semantic representation available in the FGD framework.

An example of a MultiNet structure is given in figure 2. The figure represents the following discourse:

Max gave his brother several apples.
This was a generous gift.
Four of them were rotten.
(2)

MultiNet is not explicitly model-theoretical and the extensional level is created only in those situations where the natural language expressions require it. It can be seen that the overall structure of the representation is not a tree unlike in Tectogrammatical representation (TR). The layer information is hidden except for the most important QUANT and CARD values. These attributes convey information that is important with respect to the content of the sentence. TR lacks attributes distinguishing intensional and extensional information and there are no relations like SUBM denoting relation between a set and its subset.

Note that the MultiNet representation crosses the sentence boundaries. First, the structure representing a sentence is created and then this structure is assimilated into the existing representation.

In contrast to CLASSIC (Brachman et al., 1991) and other KL-ONE networks, MultiNet contains a predefined final set of relation types, encapsulation of concepts, and attribute layers concerning cardinality of objects mentioned in discourse.

In Section 2, we describe our motivation for extending the annotation in FGD to an even deeper level. Section 3 lists the MultiNet structural counterparts of tectogrammatical means. We discuss the related work in Section 4. Section 5 deals with various evaluation techniques and we conclude in Section 6.

2 FGD layers

PDT 2.0 contains three layers of information about the text (as described in Hajič (1998)):

Morphosyntactic Tagging. This layer represents the text in the original linear word order with a tag assigned unambiguously to each word form occurence, much like the Brown corpus does.

Syntactic Dependency Annotation. It contains the (unambiguous) dependency representation of every sentence, with features describing the morphosyntactic properties, the syntactic function, and the lexical unit itself. All words from the sentence appear in its representation.

Tectogrammatical Representation (TR). At this level of description, we annotate every (autosemantic non-auxiliary) lexical unit with its tectogrammatical function, position in the scale of the communicative dynamism and its grammatemes (similar to the morphosyntactic tag, but only for categories which cannot be derived from the word's function, like number for nouns, but not its case).

There are several reasons why TR may not be sufficient in a question answering system or MT:

1. The syntactic functors Actor and Patient disallow creating inference rules for cognitive roles like *Affected object* or *State carrier*. For example, the axiom stating that an affected object is changed by the event ($(v$ AFF $o) \rightarrow (v$ SUBS `change.2.1`$)$) can not be used in the TR framework.

2. There is no information about sorts of concepts represented by TR nodes. Sorts (the upper conceptual ontology) are an important source of constraints for MultiNet relations. Every relation has its signature which in turn

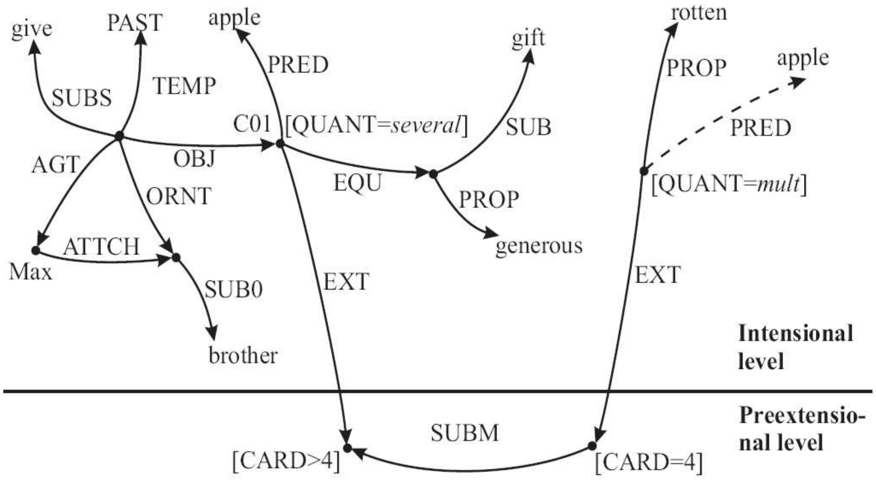


Figure 2: MultiNet representation of example discourse (2)

reduces ambiguity in the process of text analysis and inferencing.

3. Lexemes of TR have no hierarchy which limits especially the search for an answer in a question answering system. In TR there is no counterpart of SUB, SUBR, and SUBS MultiNet relations which connect subordinate concepts to superordinate ones and individual object representatves to corresponding generic concepts.

4. In TR, each sentence is isolated from the rest of the text, except for coreference arrows heading to preceding sentences. This, in effect, disallows inferences combining knowledge from multiple sentences in one inference rule.

5. Nodes in TR always correspond to a word or a group of words in the surface form of sentence or to a deleted obligatory valency of another node. There are no means for representing knowledge generated during the inference process, if the knowledge doesn't have a form of TR. For example, consider axiom of temporal precedence transitivity (3):

$$(a \text{ ANTE } b) \wedge (b \text{ ANTE } c) \rightarrow (a \text{ ANTE } c) \quad (3)$$

In TR, we can not add an edge denoting $(a \text{ ANTE } c)$. We would have to include a proposition like "a precedes c" as a whole new clause.

For all these reasons we need to extend our text annotation to a form suitable to more advanced tasks. It is shown in Helbig (2006) that MultiNet is capable to solve all the above mentioned issues.

Helbig (1986) describes a procedure for automatic translation of natural language utterances into MultiNet structures used in WOCADI tool for German. WOCADI uses no theoretical intermediate structures and relies heavily on semantically annotated dictionary (HagenLex, see Hartrumpf et al. (2003)).

In our approach, we want to take advantage of existing tools for conversions between layers in FGD. By combining several simpler procedures for translation between adjacent layers, we can improve the robustness of the whole procedure and the modularity of the software tools. Moreover, the process is divided to logical steps corresponding to theoretically sound and well defined structures. On the other hand, such a multistage processing is susceptible to accumulation of errors made by individual components.

3 Structural Similarities

3.1 Nodes and Concepts

If we look at examples of TR and MultiNet structures, at first sight we can see that the nodes of TR mostly correspond to concepts in MultiNet. However, there is a major difference: TR does not include the concept encapsulation. The encapsulation in MultiNet serves for distinguishing definitional knowledge from assertional knowledge about given node, e.g., in the sentence "The old man is sleeping", the connection to *old* will be in the definitional part of *man*, while the connection to the state *is sleeping* belongs to the assertional part of the concept representing the *man*. In TR, these differences in content are represented by differences in Topic-Focus Articulation (TFA) of corresponding words.

There are also TR nodes that correspond to no MultiNet concept (typically, the node representing the verb "be") and TR nodes corresponding to a whole subnetwork, e.g., *Fred* in the sentence "Fred is going home.", where the TR node representing *Fred* corresponds to the subnetwork[1] in figure 3.

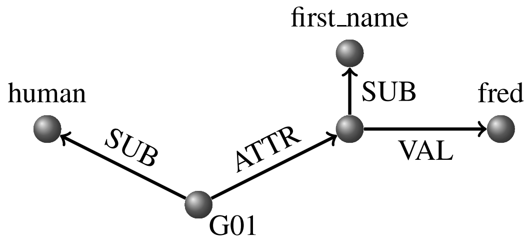


Figure 3: The MultiNet subnetwork corresponding to TR node representing *Fred*

3.2 Edges, relations and functions

An edge of TR between nodes that have their conceptual counterparts in MultiNet always corresponds to one or more relations and possibly also some functions. In general, it can be said that MultiNet representation of a text contains significantly more connections (either as relations, or as functions) than TR, and some of them correspond to TR edges.

3.3 Functors and types of relations and functions

There are 67 functor types in TR (see Hajičová et al. (2000) for description), which correspond to 94 relation types and 19 function types in MultiNet (Helbig, 2006). The mapping of TR functions to MultiNet is given in table 1:

TR functor	MultiNet counterpart
ACMP	ASSOC
ACT	AFF, AGT, BENF, CSTR, EXP, MEXP, SCAR
ADDR	ORNT
ADVS	SUBST, OPPOS
AIM	PURP
APP	ASSOC, ATTCH
APPS	EQU, NAME
ATT	MODL
AUTH	AGT, ORIG
BEN	BENF
CAUS	CAUS, JUST
CNCS	CONC
CM	*ITMS, MODL
COMPL	PROP except for sentential complements
COND	COND
CONFR	OPPOS
CONJ	*IMTS-I, *TUPL
CONTRA	OPPOS
CONTRD	CONC
CPR	*COMP
CRIT	METH, JUST, CIRC, CONF
CSQ	CAUS, JUST, GOAL
DIFF	*MODP, *OP
DIR1	ORIGL, ORIG
DIR2	VIA
DIR3	DIRCL, ELMT
DISJ	*ALTN2, *VEL2
EFF	MCONT, PROP, RSLT
EXT	QMOD
HER	AVRT
ID	NAME
INTT	PURP
LOC	LOC, LEXT
MANN	MANNR, METH
MAT	ORIGM
MEANS	MODE, INSTR
MOD	MODL
OPER	*OP, TEMP
ORIG	AVRT, INIT, ORIGM, ORIGL, ORIG
PARTL	MODL
PAT	AFF, ATTR, BENF, ELMT, GOAL, OBJ, PARS, PROP, SSPE, VAL
PREC	REAS, OPPOS
REAS	CAUS, GOAL
REG	CONF
RESL	CAUS, GOAL
RESTR	*DIFF
RHEM	MODL
RSTR	PROP, ATTR
SUBS	SUBST

continued . . .

continued . . .

[1] In fact the concept representing the man is the concept G01, i.e. only one vertex. However, the whole network corresponds to the TR node representing *Fred*.

TR functor	MultiNet counterpart
TFHL	DUR
TFRWH	TEMP
THL	DUR
THO	QUANT layer
TOWH	SUBST, TEMP
TPAR	TEMP, DUR
TSIN	STRT
TTILL	FIN
TWHEN	TEMP

Table 1: Mapping of TR functors to MultiNet

There are also TR functors with no appropriate MultiNet counterpart: CPHR, DENOM, DPHR, FPHR, GRAD, INTF, PAR, PRED and VOCAT

Table 2 shows the mapping from MultiNet relations to TR functors:

MultiNet	TR counterpart
Relations:	
AFF	PAT, DIR1
AGT	ACT
ANTE	TWHEN
ARG1/2/3	ACT, PAT, . . .
ASSOC	ACMP, APP
ATTCH	APP
ATTR	RSTR
AVRT	ORIG, ADDR, DIR1
BENF	BEN
CAUS	CAUS, RESL, REAS, GOAL
CIRC	CRIT
CONC	CNCS
COND	COND
CONF	REG, CRIT
CSTR	ACT
CTXT	REG
DIRCL	DIR3
DUR	TFHL, PAR, THL
ELMT	DIR3, DIR1
EXP	ACT
FIN	TTILL
GOAL	see RSLT, DIRCL and PURP
IMPL	CAUS
INIT	ORIG
INSTR	MEANS
JUST	CAUS
LEXT	LOC
LOC	LOC
MANNR	MANN
MCONT	PAT, EFF
MERO	see PARS, ORIGM, *ELMT, *SUBM and TEMP
METH	MANN, CRIT
MEXP	ACT
MODE	see INSTR, METH and MANNR
MODL	MOD, ATT, PARTL, RHEM
NAME	ID, APPS
OBJ	PAT
OPPOS	CONTRA
ORIG	ORIG, DIR1, AUTH
ORIGL	DIR1
ORIGM	ORIG
ORNT	ADDR
PROP	COMPL, RSTR
PROPR	COMPL, RSTR
PURP	AIM
QMOD	RSTR
REAS	see CAUS, JUST and IMPL
RPRS	LOC, MANN
RSLT	PAT, EFF
SCAR	ACT
SITU	see CIRC and CTXT
SOURC	see INIT, ORIG, ORIGL, ORIGM and AVRT
SSPE	PAT
STRT	TSIN
SUBST	SUBS
SUPPL	PAT
TEMP	TWHEN
VAL	RSTR, PAT
VIA	DIR2
Functions:	
*ALTN1	CONJ
*ALTN1	DISJ
*COMP	CPR, grammateme DEGCMP
*DIFF	RESTR
*INTSC	CONJ
*ITMS	CONJ
*MODP	MANN
*MODQ	RHEM
*MODS	MANNR
*NON	grammateme NEGATION
*ORD	grammateme NUMERTYPE
*PMOD	RSTR
*QUANT	MAT, RSTR

continued . . .

MultiNet	TR counterpart
∗SUPL	grammateme DEGCMP
∗TUPL	CONJ
∗UNION	CONJ
∗VEL1	CONJ
∗VEL2	DISJ

Table 2: Mapping of MultiNet relations to TR

There are also MultiNet relations and functions with no counterpart in TR (stars at the beginning denote a function): ANLG, ANTO, CHEA, CHPA, CHPE, CHPS, CHSA CHSP, CNVRS, COMPL, CONTR, CORR, DISTG, DPND, EQU, EXT, HSIT, MAJ, MIN, PARS, POSS, PRED0, PRED, PREDR, PREDS, SETOF, SUB, SYNO, VALR, *FLPJ and *OP.

From the tables 1 and 2, we can conclude that although the mapping is not one to one, the preprocessing of the input text to TR highly reduces the problem of the appropriate text to MultiNet transformation. However, it is not clear how to solve the remaining ambiguity.

3.4 Grammatemes and layer information

TR has at its disposal 15 grammatemes, which can be conceived as node attributes. Note that not all grammatemes are applicable to all nodes. The grammatemes in TR roughly correspond to layer information in MultiNet, but also to specific MultiNet relations.

1. NUMBER. This TR grammateme is transformed to QUANT, CARD, and ETYPE attributes in MultiNet.

2. GENDER. This syntactical information is not transformed to the semantic representation with the exception of occurences where the grammateme distinguishes the gender of an animal or a person and where MultiNet uses SUB relation with appropriate concepts.

3. PERSON. This verbal grammateme is reflected in cognitive roles connected to the event or state and is semantically superfluous.

4. POLITENESS has no structural counterpart in MultiNet. It can be represented in the conceptual hierarchy of SUB relation.

5. NUMERTYPE distinguishing e.g. "three" from "third" and "one third" is transformed to corresponding number and also to the manner this number is connected to the network.

6. INDEFTYPE corresponds to QUANT and VARIA layer attributes.

7. NEGATION is transformed to both FACT layer attribute and *NON function combined with modality relation.

8. DEGCMP corresponds to *COMP and *SUPL functions.

9. VERBMOD: *imp* value is represented by MODL relation to imperative, *cdn* value is ambiguous not only with respect to facticity of the condition but also with regard to other criteria distinguishing CAUS, IMPL, JUST and COND relatinos which can all result in a sentence with *cdn* verb. Also the FACT layer attribute of several concepts is affected by this value.

10. DEONTMOD corresponds to MODL relation.

11. DISPMOD is semantically superfluous.

12. ASPECT has no direct counterpart in MultiNet. It can be represented by the interplay of temporal specification and RSLT relation connecting an action to its result.

13. TENSE is represented by relations ANTE, TEMP, DUR, STRT, and FIN.

14. RESULTATIVE has no direct counterpart and must be expressed using the RSLT relation.

15. ITERATIVENESS should be represented by a combination of DUR and TEMP relations where some of temporal concepts have QUANT layer information set to *several*.

3.5 TFA, quantifiers, and encapsulation

In TR, the information structure of every utterance is annotated in terms of Topic-Focus Articulation (TFA):

1. Every autosemantic word is marked `c`, `t`, or `f` for contrastive topic, topic, or focus, respectively. The values can distinguish which part of the sentence belongs to topic and which part to focus.

2. There is an ordering of all nodes according to communicative dynamism (CD). Nodes with lower values of CD belong to topic and nodes

with greater values to focus. In this way, the degree of "aboutness" is distinguished even inside topic and focus of sentences.

MultiNet, on the other hand, doesn't contain any representational means devoted directly to representation of information structure. Nevertheless, the differences in the content of sentences differing only in TFA can be represented in MultiNet by other means. The TFA differences can be reflected in these categories:

- Relations connecting the topic of sentence with the remaining concepts in the sentence are usually a part of definitional knowledge about the concepts in the topic, while the relations going to the focus belong to the assertional part of knowledge about the concepts in focus. In other words, TFA can be reflected in different values of K_TYPE attribute.

- TFA has an effect on the identification of presuppositions (Peregrin, 1995a) and allegations (Hajičová, 1984). In case of presupposition, we need to know about them in the process of assimilation of new information into the existing network in order to detect presupposition failures. In case of allegation, there is a difference in FACT attribute of the allegation.

- The TFA has an influence on the scope of quantifiers (Peregrin, 1995b; Hajičová et al., 1998). This information is fully transformed into the quantifier scopes in MultiNet.

4 Related Work

There are various approaches trying to analyze text to a semantic representation. Some of them use layered approach and others use only a single tool to directly produce the target structure. For German, there is the above mentioned WOCADI parser to MultiNet, for English, there is a Discourse Representation Theory (DRT) analyzer (Bos, 2005), and for Czech there is a Transparent Intensional Logic analyzer (Horák, 2001).

The layered approaches: DeepThought project (Callmeier et al., 2004) can combine output of various tools into one representation. It would be even possible to incorporate TR and MultiNet into this framework. Meaning-Text Theory (Bolshakov and Gelbukh, 2000) uses an approach similar to Functional Generative Description (Žabokrtský, 2005) but it also has no layer corresponding to MultiNet.

There were attempts to analyze the semantics of TR, namely in question answering system TIBAQ (Jirků and Hajič, 1982), which used TR directly as the semantic representation, and Kruijff-Korbayová (1998), who tried to transform the TFA information in TR into the DRT framework.

5 Evaluation

It is a still open question how to evaluate systems for semantic representation. Basically, three approaches are used in similar projects:

First, the **coverage** of the system may serve as a basis for evaluation. This criterion is used in several systems (Bos, 2005; Horák, 2001; Callmeier et al., 2004). However, this criterion is far from ideal, because it's not applicable to robust systems and can not tell anything about the quality of resulting representation.

Second, the **consistency** of the semantic representation serves as an evaluation criterion in Bos (2005). It is a desired state to have a consistent representation of texts, but there is no guarantee that a consistent semantic representation is in any sense also a good one.

Third, the **performance in an application** (e.g., question answering system) is another criterion used for evaluating a semantic representation (Hartrumpf, 2005). A problem in this kind of evaluation is that we can not separate the evaluation of the formalism itself from the evaluation of the automatic processing tools. This problem becomes even bigger in a multilayered approach like FGD or MTT, where the overall performance depends on all participating transducers as well as on the quality of the theoretical description. However, from the user point of view, this is so far the most reliable form of semantic representation evaluation.

6 Conclusion

We have presented an outline of a procedure that enables us to transform syntactical (tectogrammatical) structures into a fully equipped knowledge representation framework. We have compared the structural properties of TR and MultiNet and found both similarities and differences suggesting which parts of such a task are more difficult and which are rather technical. The comparison shows that for applications requiring understanding of texts (e.g., question answering system) it is desirable to further analyze TR into another layer of knowledge representation.

Acknowledgement

This work was supported by Czech Academy of Science grant 1ET201120505 and by Czech Ministry of Education, Youth and Sports project LC536. The views expressed are not necessarily endorsed by the sponsors. We also thank anonymous reviewers for improvements in the final version.

References

Igor Bolshakov and Alexander Gelbukh. 2000. The Meaning-Text Model: Thirty Years After. *International Forum on Information and Documentation*, 1:10–16.

Johan Bos. 2005. Towards Wide-Coverage Semantic Interpretation. In *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, pages 42–53.

Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, Lori Alperin Resnick, and Alex Borgida. 1991. Living with CLASSIC: When and How to Use a KL-ONE-like Language. In John Sowa, editor, *Principles of Semantic Networks: Explorations in the representation of knowledge*, pages 401–456. Morgan-Kaufmann, San Mateo, California.

Ulrich Callmeier, Andreas Eisele, Ulrich Schäfer, and Melanie Siegel. 2004. The DeepThought Core Architecture Framework. In *Proceedings of LREC*, May.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum, Charles University Press, Prague, Czech Republic.

Eva Hajičová, Jarmila Panevová, and Petr Sgall. 2000. A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. Technical Report TR-2000-09, ÚFAL MFF UK, Prague, Czech Republic. in Czech.

Eva Hajičová, Petr Sgall, and Barbara Partee. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht.

Eva Hajičová. 1984. Presupposition and Allegation Revisited. *Journal of Pragmatics*, 8:155–167.

Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. 2003. The Semantically Based Computer Lexicon HaGenLex – Structure and Technological Environment. *Traitement automatique des langues*, 44(2):81–105.

Sven Hartrumpf. 2005. University of hagen at qa@clef 2005: Extending knowledge and deepening linguistic processing for question answering. In Carol Peters, editor, *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*, Wien, Österreich. Centromedia.

Hermann Helbig. 1986. Syntactic-Semantic Analysis of Natural Language by a New Word-Class Controlled Functional Analysis. *Computers and Artificial Inteligence*, 5(1):53–59.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer-Verlag, Berlin Heidelberg.

Aleš Horák. 2001. *The Normal Translation Algorithm in Transparent Intensional Logic for Czech*. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic.

Petr Jirků and Jan Hajič. 1982. Inferencing and search for an answer in TIBAQ. In *Proceedings of the 9th conference on Computational linguistics – Volume 2*, pages 139–141, Prague, Czechoslovakia.

Ivana Kruijff-Korbayová. 1998. *The Dynamic Potential of Topic and Focus: A Praguian Approach to Discourse Representation Theory*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Jaroslav Peregrin. 1995a. Topic, Focus and the Logic of Language. In *Sprachtheoretische Grundlagen für die Computerlinguistik (Proceedings of the Goettingen Focus Workshop, 17. DGfS)*, Heidelberg. IBM Deutschland.

Jaroslav Peregrin. 1995b. Topic-Focus Articulation as Generalized Quantification. In P. Bosch and R. van der Sandt, editors, *Proceedings of "Focus and natural language processing"*, pages 49–57, Heidelberg. IBM Deutschland.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing company, Dodrecht, Boston, London.

Petr Sgall, Jarmila Panevová, and Eva Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In A. Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, Massachusetts, USA. Association for Computational Linguistics.

Zdeněk Žabokrtský. 2005. Resemblances between Meaning-Text Theory and Functional Generative Description. In *Proceedings of the 2nd International Conference of Meaning-Text Theory*, pages 549–557.