



Deliverable 3.2.1

Tectogrammatical (Semantic) Representation of English and Czech

Authors: Jan Hajič
Institute of Formal and Applied Linguistics, Charles University, Prague
hajic@ufal.mff.cuni.cz

Silvie Cinková
Institute of Formal and Applied Linguistics, Charles University, Prague
cinkova@ufal.mff.cuni.cz

Authors: Other authors:
see the Appendices

Work-package: WP 3.2

Type: Technical Report w/2 Appendices

Distribution: Public (PU)

Status: Final

Date: 20.09.2007

Deliverable Coordinator: Jan Hajič (CU)

Reviewers: None

Area Coordinator: Björn Gambäck (SICS) and Jan Hajič (CU)

Project Coordinator: Yorick Wilks, University of Sheffield

EU Project Officer: Michel Brochard

ABSTRACT

This document describes the semantic (“tectogrammatical”) representation of Czech and English in its initial, yet annotation-proven version. It has three parts – an exposition (this document) to the scheme and its role in the Companions project, and then two Appendices. Appendix 1 (1471 pages) describes the technical details of the tectogrammatical representation for the Czech language, as applied to the Prague Dependency Treebank 2.0, and Appendix 2 (282 pages) describes the tectogrammatical representation of English, as it is currently being used in large-scale annotation of written English, with some preliminary annotation having been done on the first Companions dialog corpora.

TABLE OF CONTENT

1	Introduction	4
2	The Role of Semantics (“Tectogrammatics”) in the Project	4
2.1	The idea	4
2.2	Interfacing scheme	5
2.2.1	The “lower” interface	5
2.2.2	The “upper” interface	6
2.3	Generality	7
2.4	Breadth	7
2.5	Machine Learning capability	7
3	The Tectogrammatical Representation	8
3.1	Semantics in the Functional Generative Description	8
3.2	Tree structure	8
3.3	Functors and subfunctors	9
3.4	Valency	10
3.5	Anaphora and ellipsis resolution	10
3.6	Topic-focus articulation	10
4	Data and reference sources	11
4.1	Prague Dependency Treebank 2.0	11
4.2	Prague English Dependency Treebank	12
4.3	Prague Dependency Treebank of Spoken Czech	13
5	Tectogrammatical annotation of dialogs	14
5.1	From speech to dialog annotation	14
5.2	Preparation of the corpora	14
6	References	14

1 Introduction

The Deliverable 3.2.1 presents the semantic representation of English and Czech in the form of annotation manuals, which are used to build language resources for machine learning tasks such as text and speech analysis, as well as text and speech generation (described in more detail by the Deliverable 3.3). Rendering the transition between the (surface) linguistic features and the semantic content, the semantic representation is an indispensable interface between the dialog management module(s) and those performing the automatic speech recognition (ASR) and text-to-speech (TTS), which are going to be employed in the development of both the Senior and the Lifestyle (or any other scenario(s) of) Companion(s). This Deliverable comes in three parts – this short overview, and two appendices (Appendix 1 with the complete specification for Czech, and Appendix 2 with the current specification for English).

2 The Role of Semantics (“Tectogramatics”) in the Project

2.1 The idea

The novel ideas in the project proposal (and in the Description of Work as found in the Technical Annex) that concern language analysis and generation and that differentiate this project from previous dialog projects are:

- the very existence of the separate language component and novel interfacing to the rest of the system;
- generality (language, domain);
- breadth;
- machine learning capability.

Dialog systems usually concentrate on the core component – the dialog management module and its associated knowledge representation system with inferencing etc. The language part (even in systems that can deal with real speech) is often only rudimentary and ad hoc, meaning it is heavily language and domain dependent, and has to be redone for every new language and domain. We are not claiming that we can develop a completely domain-independent component (the less so a language independent one), but the use of the proposed semantic representation should at least go this direc-

tion and ease some of the “language burden” for those who are interfacing language with the core dialog management system.

It follows from the above what this semantic representation is *not*: it is not supposed to replace any of the knowledge representation systems possibly in use in the dialog management systems or in connection with it, nor is it supposed to be used directly for inferencing or dialog planning. These things are not only language independent (which the representation necessarily is not, at least not completely), and there are better and proven formalisms to deal with these issues.

On the other hand, individual language features that are part of the language description can be used by either the interfacing module(s) (e.g., for named entity resolution, predicate and argument identification, sentence mode resolution (question/command/statement/...) etc.). Details of the language features can be, of course, found in the actual representation description (see the Appendix 1 and Appendix 2 as part of this Deliverable).

2.2 Interfacing scheme

Having a separate language component that handles semantics raises a question how to interface it to the rest of the system.

2.2.1 The “lower” interface

As “lower”, we describe the interface between the speech (audio) component and the language system. Generally, the interface is very simply – it is on the text level both for input to the language module (from the ASR), as well for its out (from the NLG to the TTS module), although some information might be different in either direction.

For the “front-end”, i.e., analysis of the (user’s) spoken input, the interface is the transcribed text using the standard conventions (single case, no punctuation, etc.). It is the responsibility of the language component to deal with issues which differ in the transcription from “usual” text input (see also the extension of the representation for spoken input in Chapter 5 of this document).

For TTS, we expect to pass on the full textual representation as generated by the NLG module, including punctuation, capitalization etc., so that the TTS module can take advantage of it. We are also ready to pass the information about emotion or specific prosody, if passed onto the generation as a whole from the dialog management system.

2.2.2 The “upper” interface

This is the crucial new and relatively complicated part of the system. By “upper” interface we understand the internal interface between the semantic component and the dialog management module(s) – again, in both directions (analysis (user input) / generation (output of the system presented *to* the user)).

The “interfacing” or “bridging” interface will be necessary to be delivered together with the language analysis and generation proper. We as the authors of this Deliverable and of the representation itself will work closely with other partners to identify exactly what the needs of the dialog representation are (e.g., the named entity recognition module) and will work closely together while developing these “bridging” modules. They should benefit from the language “preprocessing” (i.e., from the availability of the semantic representation as output by the language analyzer, and from the fact that the language generator does not have to be told the language specifics when creating the system’s response).

This interface (and software) depends largely on the dialog management formalism used. For analysis, it is typically crucial to identify predicates, their arguments and their features (such as semantic number, deontic modality, tense, aspect etc.), recognize named entities, and identify sentence modality. For all of these except named entities the semantic representation is very close to the usual predicate expressions used in dialog systems, and only “technical” conversion is assumed; for the named entity resolution task, it can benefit again from language features normally not available to named entity recognizers that work off plain text. Machine learning can be employed in this stage, too.

For generation, the situation is similar, even though the generation module will probably take advantage of the fact that some system responses are quite fixed (such as “Yes” or other short, clearly defined reactions), and a simple conversion of these pre-defined responses to the languages in question (English, Czech, ...) is all that is needed. However, for the rest of the responses, if they are “real” sentences (or even partial ones with ellipsis etc.) with values of “variables” (such as person names, events, etc.) to be filled in, one can benefit from the fact that only the dependency tree with some basic semantic information must be passed on to the natural language generation module which then produces text for the TTS to take over. It is obvious, however, that even though these modules exist (Deliverable 3.3., to be followed by English NLG soon), they will have to be adapted during the remaining course of the project

to at least pass through or take advantage of the discourse-related features that are specific to dialogues (e.g., proper situation ellipsis generation, emotion and stress to be passed over to the TTS etc.).

2.3 Generality

Generality is the core issue, in the sense that the formal representation is reusable for other domains and languages, while the concrete instantiation of certain labels might be language dependent (but still domain independent). For example, the determination information crucial in English might not be present in Czech or other languages which do not use determiners (yet it is minimized if it can be derived from context). On the other hand, we expect the differences to be small; in the core semantic label set, only a few labels are not shared across Czech and English in our representation (see Chapter 7 of the Czech specification – Deliverable 3.2.1, Appendix 1, vs. Chapter 6 of the English specification – Deliverable 3.2.1, Appendix 2).

2.4 Breadth

The representation also aims to be as broad as possible, in order to achieve the claimed domain independence. The goal is not to have to rewrite the language component every time the domain is changed. In the Companions project, the two scenarios (Senior Companion and the Lifestyle Companion) will demonstrate that, and we are now ready to change to other domain if necessary without a change in the formal specification.

2.5 Machine Learning capability

Machine learning is a novel approach to dialog systems as well. We will use the developed specification to annotate data in such quantities so that automatic tools can be developed by machine learning to perform both language analysis and language generation. Our initial generation module(s), which are obviously simpler than their analysis counterparts, will not be using machine learning so extensively, but we will gradually add to them once more data becomes available (the Czech one has already been delivered – Deliverable 3.3., the English one is being worked on). The existence of the annotated data, shared with the annotation of dialog acts, speech transcription, speech standardization, emotion etc. done by other partners within the project will also allow using global machine learning approach on the top level.

Of course, there is always a risk that the resulting tools will not perform as expected, or will not be able to perform at least as well as those developed “by hand” for the specific domain. However, this is the risk as in all research projects; we will do our best to deliver also the necessary software (for both Czech and English) to do well, maintaining the main goals (generality, breath, and the possibility to use ML) intact.

3 The Tectogrammatical Representation

3.1 Semantics in the Functional Generative Description

The semantic representation is based on the Functional Generative Description (FGD), a formal natural-language description framework, developed in Prague since the 1960’s [Sgall, Hajičová and Panevová 1986]. As FGD always has combined the structuralist linguistic tradition with current trends within computational linguistics, it is easy to implement in treebank annotation.

FGD stratifies the language in several levels, with the most important ones being the *morphological*, the *analytical* (surface-syntax) and the *tectogrammatical* (underlying-syntax) levels. The tectogrammatical (semantic) level is the topmost and most abstract level within FGD. The essential features of the tectogrammatical representation (henceforth TR), i.e. syntactic dependencies, semantic labeling, valency (predicate-argument structure), ellipsis resolution and coreference, along with topic-focus articulation annotation, reflect *the linguistic meaning* of each sentence. TR describes syntactic as well as semantic relations among autosemantic words within a sentence, with coreference (pronominal anaphora) markup exceeding the sentence boundaries. This section gives a brief survey of the essential principles of TR. For complete reference see the attached annotation manuals (Appendices 1 and 2).

3.2 Tree structure

TR (as it is implemented in treebank annotation) displays each sentence as a rooted dependency tree with nodes and edges. Each tree (i.e. each sentence) has a technical root node with sentence ID. The effective root of the sentence, though, is its governing predicate. (Fig. 1 displays the TR of a sentence from the NAP corpus – it comes from within the Senior Companion scenario).

The dependency relations between autosemantic nodes are represented as edges. Besides, non-dependency edges are employed with certain node pairs, e.g. with paratactic structure root nodes.

Each autosemantic word has its own node, which contains a set of attribute values. The attribute values are used to store technical parameters (e.g. node-ID, position in the word order) as well as linguistic markup. The essential attributes are *t-lemma* (mostly the basic form of the word) and *functor* (a label describing the semantic relation of the given node to its parent node). Function words (auxiliary verbs, articles, prepositions etc.) do not have their own nodes on TR. Instead, they are stored as references to ID's of the corresponding nodes on the surface-syntax level (a-level) within the relevant nodes. E.g. the prepositional phrase *in the cars* in the sentence “*There was no air conditioning in the cars.*” would be represented as one node with the t-lemma *car* and the functor LOC (location). It would contain references to the ID's of the a-level nodes *in* and *the*. It would be governed by the node with the t-lemma *be* (which, as the main predicate, would have the functor PRED).

With some POS the t-lemma does not correspond to the dictionary-codified basic form of the given word, having a *t-lemma substitute* instead. This is typically the case of pronouns, whose class and (if relevant) person, number etc. is defined by the so-called *grammatemes* – additional attribute values capturing lexical and semantic derivation. Other t-lemma substitutes belong to artificially generated nodes that do not have any correspondence in the surface syntax. Artificially generated nodes are employed e.g. in verb control.

3.3 Functors and subfunctors

The semantic relation between a node and its parent node is defined by a semantic label, in FGD called *functor*. There are approximately 70 functors in FGD. Five functors (*Actor*, *Patient*, *Addressee*, *Origo* and *Effect*) are assigned to obligatory arguments of verbs, nouns, adjectives and certain types of adverbs (see 2.4). The functors APPS, CONJ, DISJ etc. mark the roots of paratactic constructions like appositions (e.g. *Mark, i.e. my brother*) and various semantic types of coordination (e.g. *Mark and Jane*, *Mark or Jane*). Yet most functors classify free modifications. There are e.g. temporal functors (TWHEN, THL – “how long”, THO – “how often”), locative and directional functors (LOC, DIR1-3), functors for causal relations (CAUS, AIM etc.), functors expressing manner (MANN, EXT, MEANS etc.), functors for rhematizers and sentence adverbials (RHEM, MOD – “modality”, ATT – “attitude” etc.), functors for multi-word lexical units and foreign expressions, adnominal functors (restrictive/descriptive attribute and appurtenance), and the predicative complement functor.

Many functors have additional semantically motivated attribute values (subfunctors). E.g. the functor EXT (“extent”) must be further specified by adding the subfunctors *basic*, *approx*, *more* and *less*.

3.4 Valency

In FGD, verbs, nouns, adjectives and some types of adverbs have their valency (predicate-argument structure) patterns (see [Panevová 1974, 1975, 1980]). They are listed in a lexicon of valency frames, which is interlinked with the data. Each occurrence of a POS with valency is related to the appropriate valency frame in the lexicon via a reference link.

3.5 Anaphora and ellipsis resolution

Pronominal anaphora is annotated within the coreference annotation. Anaphoric pronouns are linked to their text antecedents/postcedents. When an anaphoric pronoun refers to an entire text segment it is marked without delimiting the segment antecedent. Even exophoric anaphora is identified, and hence exophoric references can be easily related to named entities outside the text. Besides, the coreference markup includes grammatical coreference, i.e. verb control etc. Coreference annotation should provide substantial “help” to any name entity resolution software (which is an important part of any dialog system, the more so in the Senior Companion scenario) - current NER systems are typically unable to handle the assignment of pronouns to named entities.

Ellipses of two types are reconstructed. TR distinguishes between textual and grammatical ellipsis. When a node or a subtree is omitted in the sentence but is explicitly mentioned elsewhere in the text, the given node/subtree is copied. Grammatical ellipses are e.g. missing predicates that cannot be inferred from the context (but only from common knowledge), as in *Careful!* (= *be careful!*). Nodes of grammatical ellipsis do not have any corresponding nodes on the a-level, and therefore they get t-lemma substitutes.

3.6 Topic-focus articulation

On the tectogrammatical level, also the topic-focus articulation (TFA) is annotated. We consider TFA to be a phenomenon of the underlying structure of the sentence – two surface realizations of a sentence with differing TFA correspond to two different tectogrammatical trees. The TFA annotation comprises two phenomena:

- contextual boundness;
- communicative dynamism.

Contextual boundness is represented by the values of the attribute *tfa* for each node of the tectogrammatical tree. Contextual boundness is a property of an expression (be it expressed or absent in the surface structure of the sentence) which determines whether the speaker (author) uses the expression as given (for the recipient), i.e. uniquely determined by the context.

Communicative dynamism is represented by the underlying order of nodes. It is a property of an expression that reflects its relative degree of importance in comparison with other expressions in the sentence attributed to it by the speaker; we consider contextually non-bound expressions to be more dynamic than expressions contextually bound (be they non-contrastive or contrastive).

Annotated trees therefore contain two types of information – on the one hand the value of contextual boundness of a node and its relative ordering with respect to its sister nodes reflects its function within the topic-focus articulation of the sentence, on the other hand the set of all the TFA values in the tree and the relative ordering of subtrees reflects the overall functional perspective of the sentence, and thus enables to distinguish in the sentence the complex categories of topic and focus (however, these are not annotated explicitly).

4 Data and reference sources

Since the onset of powerful computer technologies the theory of FGD has been continuously verified and refined by being implemented in treebank annotation. There are two types of FGD-based language resources: corpus data and valency lexicons interlinked with the data. These resources will be used not only for verifying the annotation scheme (by evaluating annotator's agreement (kappa or similar) rates, etc.), but mainly for training the analysis and (parts of the) generation software that will actually carry out the analysis and generation, as described in the Section 2.5 above.

4.1 Prague Dependency Treebank 2.0

The first FGD-based corpus to build was the Prague Dependency Treebank, a parsed corpus of Czech texts, taken mainly from newspapers. Its second version [Hajič et al. 2006] comprises approx. 800 000 tokens manually annotated on the tectogrammatical

section into the FGD-like shape. The Czech counterpart comprises translations of PTB-WSJ [Mitchell et al. 1994]. The manual tectogrammatical annotation of the English data was launched in the fall of 2006. The current annotation focuses the following issues:

- correct tree structure, including mainly:
 - a) rules for coordination, apposition, parenthesis
 - b) some specific constructions like comparison, restriction, consecutive clauses with quantifiers etc.
 - c) determination of function words
- assigning and completing valency frames in verbs
- correct semantic labels (functors) in nodes
- correct t-lemmas
- correct links to a-layer

The following issues have been left aside for the moment:

- coreference (will be dealt with shortly, however, for its importance to the named entity recognition task)
- topic-focus articulation
- subfunctors, grammatemes.

The valency lexicon for the English TR annotation EngVallex [Cinková 2006] comprises only verbs at the moment. It was built by a semiautomatic conversion of the PropBank Lexicon [Palmer et al. 2004]. Manual corrections are going on in parallel with the data annotation. Currently, about 150,000 words are annotated for English (and linked to the English valency lexicon), the goal being (a) the full WSJ corpus (about 1 million words) and (b) dialog sample annotation for the analysis software adaptation to the dialog domain(s).

4.3 Prague Dependency Treebank of Spoken Czech

Since recently, TR has not been applied to speech in terms of actual annotation. The manual annotation of spoken data was launched last year. The data have been taken

from the Czech part of the MALACH corpus [Byrne et al. 2004], which comprises testimonies by holocaust survivors. The TR annotation was built upon a manual transcription synchronized with the source audio file. A manual speech-reconstruction annotation was added as a complement to the morphological level.

5 Tectogrammatical annotation of dialogs

5.1 From speech to dialog annotation

The Companions project has given rise to adapting the TR to dialogs. The project “Prague Dependency Trebank of Spoken Czech” has already proven that TR is capable of capturing speech. The experimental annotation of a part of the NAP corpus (approx. 120 sentences) has shown that no substantial alterations to the current TR annotation scheme are needed. All dialog-typical features like ellipses and fragmentary answers, turn switching and overlapping talk pose no problems. The dialog-specific features include expanding anaphora annotation, references to entities outside the text and additional linguistic links between questions and answers as previously used in the TIBAQ question answering system (*wh-path*) [Hajičová (ed.) 1995], which will be applied depending on which semantic features of speech acts the dialog-structure annotation provided by USFD (or any such annotation/formalism that will be used in the final scheme of the Companions) is going to provide (see the Deliverable 3.1.1). This will be used then for the interfacing to the dialog manager (as described in the section 2.2.2 above).

5.2 Preparation of the corpora

The TR annotation of dialogs could have been performed on two corpora so far: the AAA corpus and/or the NAP corpus. For the immediate future, the exact form of audio transcripts to be provided by AAA and NAP has already been defined and agreed upon among AAA, NAP, CU and ZCU. USFD is even considering the possibility of providing CU with speech reconstruction annotation (to be built upon the transcripts) carried out by native speakers (hired students). If this turns out to be impossible, CU will hire native English speakers in Prague.

6 References

[Byrne et al. 2004] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D.W. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Wei-

- Jing Zhu, *Automatic recognition of spontaneous speech for access to multilingual oral history archives*, IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing, 12(4), pp. 420-435, July 2004.
- [Cinková 2006] Silvie Cinková, *From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description*, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). 2006.
- [Hajič et al. 2003] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, Petr Pajas, *PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation*, in J. Nivre, E. Hinrichs (eds.): Proceedings of The Second Workshop on Treebanks and Linguistic Theories, Vaxjo University Press, Vaxjo, Sweden, 2003
- [Hajič et al. 2006] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, *Prague Dependency Treebank 2.0*, Linguistic Data Consortium, 2006, LDC Catalog No. LDC2006T01, 1-58563-370-4
- [Hajičová (ed) 1995] Eva Hajičová (ed.), *Text-And-Inference-Based Approach to Question Answering*, Prague, 1995
- [Mitchell et al. 1993] P. M. Mitchell et al., *Building a Large Annotated Corpus of English: The Penn Treebank*, Computational Linguistics, 1993
- [Palmer et al. 2004] Martha Palmer et al., *Proposition Bank I*, LDC2004T14, ISBN: 1-58563-304-6, Sep 01 2004
- [Panevová 1974] Jarmila Panevová, *On verbal frames in Functional generative description I*, Prague Bulletin of Mathematical Linguistics, 22, pp. 3-40, MFF UK, Prague, Czech Republic, 1974
- [Panevová 1975] Jarmila Panevová, *On verbal frames in functional generative description I*, Prague Bulletin of Mathematical Linguistics, 23, pp. 17-52, MFF UK, Prague, Czech Republic, 1975
- [Panevová 1980] Jarmila Panevová, *Formy a funkce ve stavbě české věty*, Prague:Academia, 1980
- [Sgall, Hajičová and Panevová 1986] Petr Sgall, Eva Hajičová, Jarmila Panevová, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Dordrecht:Reidel Publishing Company and Prague:Academia, 1986