MTT 2007, Klagenfurt, May 21 – 24, 2007
Wiener Slawistischer Almanach, Sonderband 69, 2007

Dependency-based Sentence Synthesis Component for Czech

Jan Ptáček (1), Zdeněk Žabokrtský(2)

(1,2) Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Prague, Czech Republic
{ptacek,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We propose a complex rule-based system for generating Czech sentences out of tectogrammatical trees, as introduced in Functional Generative Description (FGD) and implemented in the Prague Dependency Treebank 2.0 (PDT 2.0). Linguistically relevant phenomena including valency, diathesis, condensation, agreement, word order, punctuation and vocalization have been studied and implemented in Perl using software tools shipped with PDT 2.0. Parallels between generation from the tectogrammatical layer in FGD and deep syntactic representation in Meaning-Text Theory are also briefly sketched.

Keywords

Natural Language Generation, Prague Dependency Treebank.

1 Introduction

Natural Language Generation (NLG) is a sub-domain of Computational Linguistics; its aim is studying and simulating the production of written (or spoken) discourse. Usually the discourse is generated from a more abstract, semantically oriented data structure. The most prominent application of NLG is probably transfer-based machine translation, but NLG is relevant also for dialog systems, systems for text summarizing, systems for generating technical documentation, etc.

In this paper, the NLG task is formulated as follows: given a Czech tectogrammatical tree – as introduced in Functional Generative Description (Sgall, 1967), (Sgall et al., 1986) and recently elaborated in more detail within the PDT 2.0[1] project – generate a Czech sentence the meaning of which corresponds to the content of the input tree. Note that in the context of PDT 2.0, synthesis of written sentences can be viewed as a process inverse to treebank annotation.

[1] http://ufal.mff.cuni.cz/pdt2.0/

Not surprisingly, the presented research is motivated by the idea of transfer-based machine translation with the usage of tectogrammatics as the highest abstract representation.

Outside the domain of Czech language, a NLG task is thoroughly explored. We mention here the text-generation systems based on the Meaning-Text Theory (Mel'čuk, 1988).

First we explore the data structure as expected on input to generation procedure. There is not a standard input defined for a general NLG problem. Regarding systems like LSF or AlethGen (Iordanskaja et al., 1992), (Coch, 1996) the generation is based on non-linguistic data stored in a database or obtained interactively. Unlike our generator, such systems are focused on a particular domain and deal with text and sentence planning. But the generation process can also start from a deep syntactic representation such as system RealPro (Lavoie & Rambow, 1997). We differ though in the definition of the deep syntactic representation.

Second criterion is the mechanism of grammar rules application. A graph rewriting approach suggested by Mel'čuk (Mel'čuk, 1988) dominanates here. Such approach treats grammar as a separable resource and needs a nontrivial framework (such as MATE (Bohnet & Wanner, 2001)) for its processing. Our grammar of Czech is 'hardwired'; written in the Perl programming language. It is modularized and uses pluggable resources as seen in Figure 2. Procedural design results in quick prototyping and also natural order of operations is highlighted.

2 PDT 2.0 in a Nutshell

In the Prague Dependency Treebank 2.0 annotation scenario, based on the theoretical framework of Praguian Functional Generative Description, three layers of annotation are added to Czech sentences (Jan Hajič et al., 2006):

Morphological layer (m-layer), on which each token in each sentence of the source texts is lemmatized and tagged with a positional POS-tag.[2]

Analytical layer (a-layer), on which a sentence is represented as a rooted ordered tree with labeled nodes and edges, corresponding to the surface-syntactic relations; each a-layer node corresponds to exactly one m-layer token.

Tectogrammatical layer (t-layer), on which the sentence is represented as a deep-syntactic dependency tree structure (t-tree) built of nodes and edges. T-layer nodes represent auto-semantic words (including pronouns and numerals) while functional words such as prepositions, subordinating conjunctions and auxiliary verbs have no nodes of their own in the tree. Each tectogrammatical node is a complex data structure – it can be viewed as a set of attribute-value pairs, or even as a typed feature structure. Word forms occurring in the original surface expression are substituted with their t-lemmas. Only semantically indispensable morphological categories (called grammatemes) are stored in the nodes (such as

[2] Technically, there is also one more layer called w-layer (word layer) 'below' the mlayer; on this lowest layer the original raw text is only segmented into documents, paragraphs and tokens, and all these units are enriched with identifiers.

number for nouns, or degree of comparison for adjectives), but not the categories imposed by government (such as case for nouns) or agreement (congruent categories such as person for verbs or gender for adjectives). Each edge in the t-tree is labeled with a functor representing the deep-syntactic dependency relation.[3] Coreference and topic-focus articulations are annotated in t-trees as well. See (Mikulová et al., 2005) for a detailed description of the t-layer.

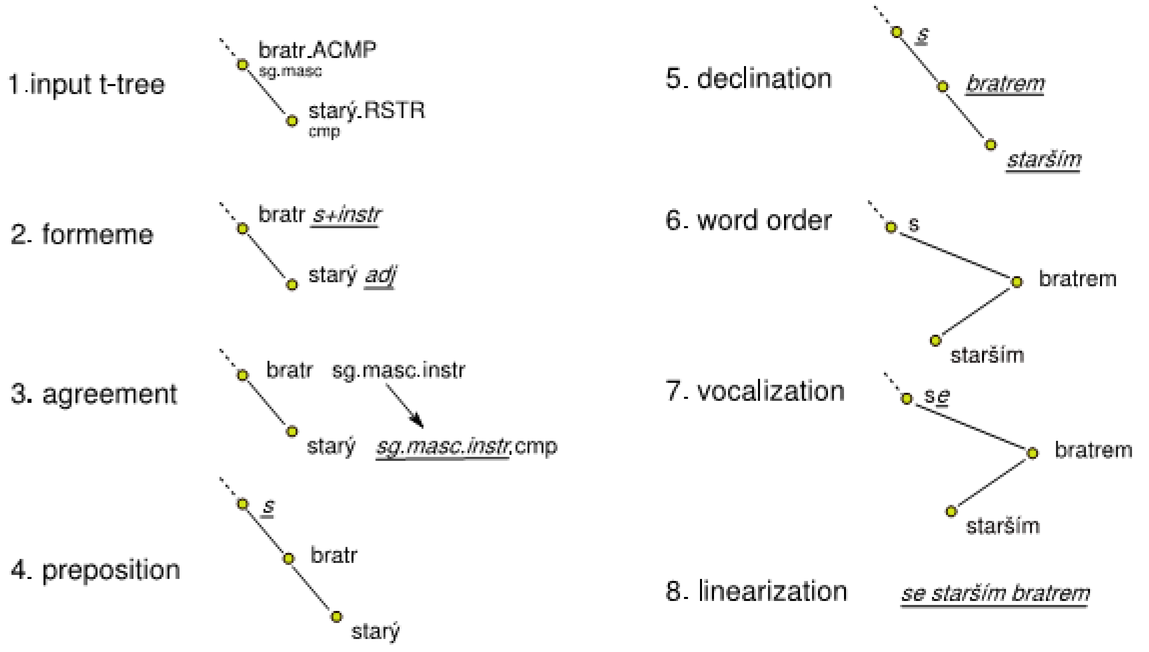

Figure 1: Illustration of the individual steps of the generating procedure when applied on a t-tree fragment corresponding to the expression *se starším bratrem* (lit. with older brother).

3 Synthesis Procedure

Unlike stochastic 'end-to-end' solutions, rule-based approach, which we adhere to in this paper, requires careful decomposition of the task (due to the very complex nature of the task, a monolithic implementation could hardly be maintainable). The decomposition was not trivial to find, because many linguistic phenomena are to be considered and some of them may interfere with others; the presented solution results from several months of experiments and a few re-implementations.

In our system, the input tectogrammatical tree is gradually changing – in each step, new node attributes and/or new nodes are added. Step by step, the structure becomes (in some aspects) more and more similar to a-layer tree. After the last step, the resulting sentence is obtained simply by concatenating word forms which are already filled in the individual nodes, the ordering of which is also already specified.

[3] Edge labels are in fact treated and visualized as attributes of dependent nodes.

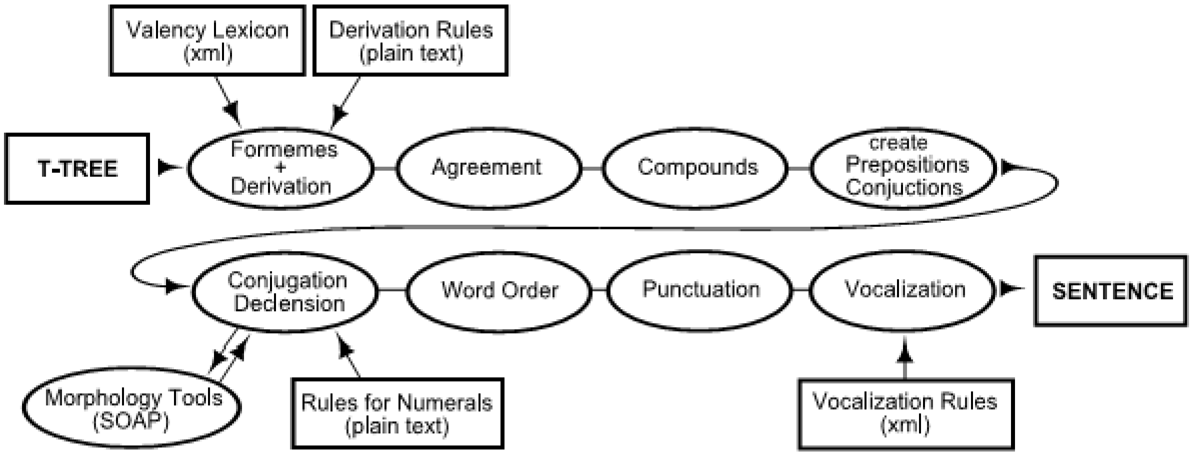

Figure 2: Data-flow diagram representing the process of sentence synthesis.

A simplified data-flow diagram corresponding to the generating procedure is displayed in Figure 2. All the main phases of the generating procedure will be outlined in the following subsections, some of them are illustrated on an artificial t-tree fragment in Figure 1, or on an authentic sentence from PDT 2.0 in Figures 3 and 4. The procedure has been implemented in Perl within the tred/btred[4] tree processing environment developed by Petr Pajas.

3.1 Formeme Selection, Diatheses, Derivations

In this phase, the input tree is traversed in the depth-first fashion, and so called *formeme* is specified for each node. Under this term we understand a set of constraints on how the given node can be expressed on the surface (i.e., what morphosyntactic form is used). Possible values are for instance simple case *gen* (genitive), prepositional case *pod+7* (preposition *pod* and instrumental), *v-inf* (non-finite verb), *že+v-fin* (subordinating clause introduced with subordinating conjunction *že*), *adj* (syntactic adjective), etc.

Several types of information are used when deriving the value of the new *formeme* attribute. At first, the valency lexicon[5] is consulted: if the governing node of the current node has a nonempty valency frame, and the valency frame specifies constraints on the surface form for the functor of the current node, then these constraints imply the set of possible formemes. In case of verbs, it is also necessary to specify which diathesis should be used (active, passive, reflexive passive, etc.; depending on the type of diathesis, the valency frame from the lexicon undergoes certain transformations). If the governing node does not have a valency frame, then the formeme default for the functor of the current node (and subfunctor, which specifies the

[4] http://ufal.mff.cuni.cz/˜pajas/tred/

[5] There is the valency lexicon PDT-VALLEX ((Hajič et al., 2003)) associated with PDT 2.0. On the t-layer of the annotated data, all semantic verbs and some semantic nouns and adjectives are equipped with a reference to a nonempty valency frame in PDT-VALLEX, which was used in the given sentence.

type of the dependency relations in more detail) is used. For instance, the default formeme for the functor ACMP (accompaniment) and subfunctor basic is *s+7* (with), whereas for ACMP.wout it is *bez+2* (without).

It should be noted that the formeme constraints depend also on the possible word-forming derivations applicable on the current node. For instance, the functor APP (appurtenance) can be typically expressed by formemes *gen* (genitive) and *adj* (possessive adjective), but in some cases only the former one is possible (some Czech nouns do not form derived possessive adjectives).

3.2 Propagating Values of Congruent Categories

In Czech, which is a highly inflectional language, several types of dependencies are manifested by agreement of morphological categories (agreement in gender, number, and case between a noun and its adjectival attribute, agreement in number, gender, and person between a finite verb and its subject, agreement in number and gender between relative pronoun in a relative clause and the governor of the relative clause, etc.). As already mentioned, the original tectogrammatical tree contains only those morphological categories which are semantically indispensable. After the formeme selection phase, value of case should be also known for all nouns. In this phase, oriented agreement arcs (corresponding to the individual types of agreement) are conceived between nodes within the tree, and the values of morphological categories are iteratively spread along these arcs until the unification process is completed.

3.3 Expanding Complex Verb Forms

Only now, when person, number, and gender of finite verbs are known, it is possible to expand complex verb forms where necessary. New nodes corresponding to reflexive particles (e.g., in the case of reflexiva tantum), to auxiliary verbs (e.g., in the case of complex future tense), or to modal verbs (if deontic modality of the verb is specified) are attached below the original autosemantic verb.

3.4 Adding Prepositions and Subordinating Conjunctions

In this phase, new nodes corresponding to prepositions and subordinating conjunctions are added into the tree. Their lemmas are already implied by the value of node formemes.

3.5 Determining Inflected Word Forms

After the agreement step, all information necessary for choosing the appropriate inflected form of the lemma of the given node should be available in the node. To perform the inflection, we employ morphological tools (generator and analyzer) developed by Hajič (Hajič, 2004). The generator tool expects a lemma and a positional tag (as specified in (Hana et al., 2002)) on the input, and returns the inflected word form. Thus the task of this phase is effectively reduced to composing the positional morphological tag; the inflection itself is performed by the morphological generator.

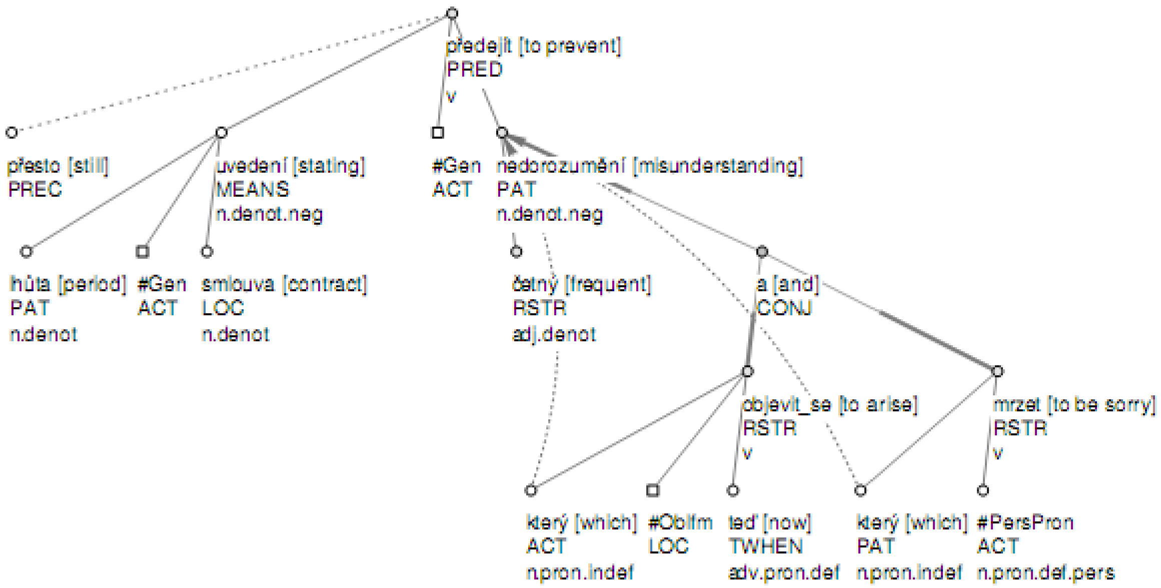

Figure 3: (Simplified) PDT 2.0 t-tree corresponding to the sentence '*Přesto uvedením lhůty ve smlouvě by se bylo předešlo četným nedorozuměním, která se nyní objevila a která nás mrzí.*' (But still, stating the period in the contract would prevent frequent misunderstandings which have now arisen and which we are sorry about.)

3.6 Special Treatment of Definite Numerals

Definite numerals in Czech (and thus also in PDT 2.0 t-trees) show many irregularities (compared to the rest of the language system), that is why it seems advantageous to generate their forms separately. Generation of definite numerals is discussed in (Ptáček, 2005).

3.7 Reconstructing Word Order

Ordering of nodes in the annotated t-tree is used to express information structure of the sentences, and does not directly mirror the ordering in the surface shape of the sentence. The word order of the output sentence is reconstructed using simple syntactic rules (e.g., adjectival attribute goes in front of the governing noun) and topic-focus articulation. Special treatment is required for clitics: they should be located in the 'second' position in the clause (Wackernagel position); if there are more clitics in the same clause, simple rules for specifying their relative ordering are used (for instance, the clitic *by* always precede short reflexive pronouns).

3.8 Adding Punctuation Marks

In this phase, missing punctuation marks are added to the tree, especially (i) the terminal punctuation (derived from the sentmod grammateme), (ii) punctuations delimiting boundaries of clauses, of parenthetical constructions, and of direct speeches, (iii) and punctuations in multiple coordinations (commas in expressions of the form *A, B, C and D*).

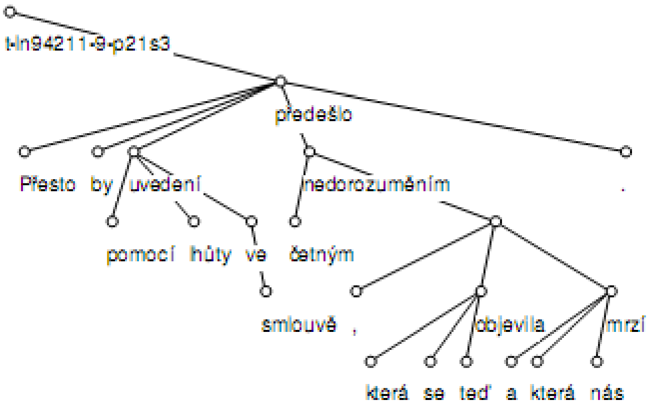

Figure 4: One of the intermediate phases during the processing of the t-tree from Figure 3. Almost all processing steps are performed (see the added nodes with functional words and punctuation marks, the inflected word forms, properly placed clitics etc). After performing the last step – concatenation of the word forms into one string – the following synthesized sentence is obtained: *Přesto uvedením lhůty ve smlouvě by se bylo předešlo četným nedorozuměním, která se nyní objevila a která nás mrzí.*

Besides adding punctuation marks, the first letter of the first token in the sentence is also capitalized in this phase.

3.9 Vocalizing Prepositions

Vocalization is a phonological phenomenon: the vowel *-e* or *-u* is attached to a preposition if the pronunciation of the prepositional group would be difficult without the vowel (e.g., *ve výklenku* instead of **v výklenku*). We have adopted vocalization rules precisely formulated in (Petkevič, 1995) (technically, we converted them into the form of an XML file, which is loaded by the vocalization module).

3.10 Linearization

At this moment, the resulting structure has roughly the shape of surface-syntactic tree (one inflected word form or punctuation mark per node, see Figure 4). The last thing to do is to merge the tokens into the final sentence string, which is a trivial task complicated only by the question of placement of spaces around quotation marks and other special symbols.

4 Final Remarks

In this paper we have presented our approach to generating Czech sentences from tectogrammatical trees. More information about the system (including some implementation details and evaluation of the generator performance by measuring BLEU-score distance between the original sentences in the PDT 2.0 and the generated sentences) is given in (Ptáček, 2005).

Finally, we would like to note that the task of generating sentences from t-trees is in our opinion very similar to generating sentences from DSyntR (Deep-Syntactic Representation) as defined in Meaning-Text Theory. Most of the consequences of the common features could have been seen in the previous section. However, in the following paragraphs, we try to make them explicit using the list of resemblances between t-trees and DSyntR enumerated in (Žabokrtský, 2005).

(1) "*The skeleton of both representations is formed by dependency tree (unordered in MTT, ordered according to information structure in FGD).*" – In other words, lexicalization and hierarchization of a message (and each sentence in particular) is more or less specified already in the input of the generating procedure (unlike the case of generation e.g., from SemR).

(2) "*Only semantically full lexemes (autosemantic words) do have nodes of their own (semantically empty lexemes/synsemantic words, such as prepositions, subordinating conjunctions, auxiliary verbs, etc. are introduced only in the surface-syntactic structure).*" – This implies two things: both in the generation from t-layer and DSyntR, full lexemes must undergo inflection and functional words have to be added.

(3) "*Each lexeme is associated with appropriate semantically full grammemes (grammatemes in FGD terminology); grammemes imposed only by government and agreement are excluded.*" – During generation, values of grammemes have to be distributed along the agreement links also to the places, where they are not semantically indispensable, but are manifested by inflection.

(4) "*Each dependency tree is accompanied with (non-tree) grammatical coreferential relations, together forming dag (directed acyclic graph).*" – To generate a grammatical sentence, coreferential links cannot be ignored: they are important e.g., for detection of reflexive pronoun or agreement of relative pronouns.

We also believe that the two approaches could mutually enrich each other: for example, it would be very useful to adopt the notion of lexical functions for FGD, especially if a similar notion is de facto used in PDT for relating e.g., deadjectival adverbs with their primary adjectives, or possessive adjectives with their primary nouns.

Acknowledgements

The research has been funded by projects 1ET101120503, 1ET201120505 and MSM0021620838.

Bibliography

Bohnet, B.,Wanner, L.: On Using a Parallel Graph Rewriting Grammar Formalism in Generation. *Proceedings of the 8th European Natural Language Generation Workshop at the Annual Meeting of the Association for Computational Linguistics*, Toulouse (2001)

Coch, J.: Overview of AlethGen. *Demonstrations and Posters of the Eighth International Natural Language Generation Workshop* (1996) 25–28

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová-Řezníčková, V., Pajas, P.: PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, Vaxjo University Press (2003) 57–68

Hajič, J.: *Disambiguation of Rich Inflection – Computational Morphology of Czech. Charles University* – The Karolinum Press, Prague (2004)

Hana, J., Hanová, H., Hajič, J., Vidová-Hladká, B., Jeřábek, E.: *Manual for Morphological Annotation*. Technical Report TR-2002-14 (2002)

Hajič et al.: *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, CAT LDC2006T01, ISBN 1-58563-370-4 (2006)

Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B., Polguere, A.: Generation of extended bilingual statistical reports. *Proceedings of the 15th International Conference on Computation Linguistics* (1992) 1019–1023

Lavoie, B., Rambow, O.: RealPro – a fast, portable sentence realizer. *Proceedings of the Conference on Applied Natural Language Processing* (1997)

Mel'čuk, I.: *Dependency Syntax: Theory and Practice.* State University of New York Press (1988)

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., Kučová, L.: *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka*. Technical Report TR-2005-28, ÚFAL MFF UK (2005)

Petkevič, V., ed.: Vocalization of Prepositions. In: *Linguistic Problems of Czech*. (1995) 147–157

Ptáček, J.: *Generování vět z tektogramatických stromů Pražského závislostního korpusu.* Master's thesis, MFF, Charles University, Prague (2005)

Sgall, P.: *Generativní popis jazyka a česká deklinace.* Academia (1967)

Sgall, P., Hajičová, E., Panevová, J.: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* D. Reidel Publishing Company, Dordrecht (1986)

Žabokrtský, Z.,: Resemblances between Meaning-Text Theory and Functional Generative Description. In Jurij D. Apresjan, L.L.I., ed.: *Proceedings of the 2nd International Conference of Meaning-Text Theory*, Moscow, Russia, June 23-25, Slavic Culture Languages Publishers House (2005) 549–557