The simple core and the complex periphery of natural language
A formal and a computational view

Petr SGALL
CKL, Charles University Prague
Malostranské nám. 25
118 00 Praha 1, Czech Rep.
sgall@ckl.mff.cuni.cz

Alena BÖHMOVÁ
CKL, Charles University Prague
Malostranské nám. 25
118 00 Praha 1, Czech Rep.
bohmova@ckl.mff.cuni.cz

Abstract

A complex procedure of syntactic annotation of a large text corpus may be helpful in checking a rich descriptive framework (the Praguian Functional Generative Description) that makes it possible to distinguish between the core of natural language, structured in a relatively simple way, and its large periphery with indistinct borderlines. Such a procedure underlies the Prague Dependency Treebank, within which about 20 000 Czech sentences from running texts have been analyzed in their underlying structure; for 2000 sentences also their Topic-Focus structures have been specified. We illustrate the wide range of the phenomena handled, i.e. the syntactic relations proper (arguments and adjuncts), coordination, topic-focus articulation, word order, deletion, positions of focusing particles, morphological categories such as number, tense, modality, their morphemic and analytical means of expression, and so on.

1 Introductory remarks

1.1 The aim

We want to point out how a complex procedure of syntactic annotation of a large text corpus may be helpful in checking a rich descriptive framework, which makes it possible to specify the frequently required distinction between the core of natural language, structured in a relatively simple way, and its large periphery with indistinct borderlines. On the background of the Praguian Functional Generative Description (see Sgall et al. 1986, Hajičová et al., 1998), this procedure underlies the **Prague Dependency Treebank (PDT)**, within which sentences from the Czech National Corpus, i.e. from running texts, are analyzed. Up to now, about 20 000 sentences have been annotated at the level of (underlying) syntax, out of which 2000 have been analyzed also in their Topic-Focus structures. Our illustrations should characterize the wide range of the phenomena handled, i.e. the syntactic relations proper (arguments and adjuncts), coordination, topic-focus articulation, word order, deletion, positions of focusing particles (operators), morphological categories such as number, tense, modality, their morphemic and analytical means of expression, and so on.

1.2 Formalization and linguistics as a cumulative science

The general approach we apply in looking for a formal description of natural language is based on our conviction that linguistics should not lose its character of a **cumulative science**; interruptions of its development may be reduced by systematic discussions between different theoretical approaches. Especially, the tradition of the **Prague school** of functional and structural linguistics is not to be forgotten, since it exhibits certain advantages; some of them are based on the fundamental distinction between unmarked (primary, i.e. prototypical) and **marked** (secondary) phenomena. This distinction makes it possible to handle the core of language structure as based on simple general principles, while its periphery consists of more or less marginal layers limited by contextual and other restrictions (highly different from one language to the other), and thus can be described only by specific rules.

Structural syntax in Europe has been based on **dependency** (the relation between head and modifier) since its beginnings, thus differing from descriptivist and Chomskyan trends, which work with constituency. The tradition of dependency-based syntax is much older, starting in the 1830s in Germany and elaborated then also in France and the Slavic countries, see the writings of L. Tesničre, V. Šmilauer and others. This approach is well suited for a high degree of modularity of language description, which perhaps underlies its frequent use in natural language processing. The theoretical potential of such a description may be clearly seen if one does not work only with some kind of surface syntax, but investigates the **underlying** structure, appropriate to serve as the input to semantic(-pragmatic) interpretation. From the viewpoint of linguistic **typology**, which primarily studies the relationships between underlying (syntactic) and surface (morphemic) representations (cf. Ramat 1985), the difference between an ending and a **function word** is directly relevant only for morphemics (cf. Skalička 1979; see Holenstein's 1975 evaluation of the Jakobsonian view of implication laws, which underlies this view). On the other hand, one of the main aspects of sentence structure can be found in the articulation of the sentence into its **Topic** and **Focus**, analyzed already in th 19th century by H. Weil, G. von der Gabelentz, P. Wegener, then by V. Mathesius and others, now see Hajičová et al. (1998), where also a formal treatment of its interpretation, based on the 'aboutness' relation, is discussed.

Having in mind such basic insights gained by classical linguistics, the Functional Generative Description has been elaborated as a formal framework in which the syntactic **tectogrammatical** representations (**TRs**) are viewed as the interface level of the language system and the layers of cognition (in which also the specification of reference, the inferencing based on contextual and other knowledge and a truth-conditional or other basis of semantics are relevant, cf. Sgall 1994). The TRs contrast with the morphemic („surface„) representations, i.e. strings of closely and loosely connected morphemes, directly expressed by phonemic strings.

2 The transparently patterned core

2.1 The core in a formal description:

2.1.1 Tectogrammatical dependency trees

In the unmarked case (in which no coordination constructions occur), the TR has the shape of a **dependency tree**, with its root labelled by the (underlying counterpart of the) verb, which occupies the position of PRED(icate) and displays in its **valency** frame the **functors**, characterizing types of its dependents, i.e. arguments and adjuncts (either of which can be obligatory or optional with the given head), such as PAT(ient or Objective), ACT(or), APP(urtenance, broader than "Possessive"), DIR(ectional)1. A formal specification of the TRs can be found in Plátek et al. (1984) and in Petkevič (1995). In the present paper we can only give some illustrations, having the computational treatment as our main aim.

A simplified TR of sentence (1) is given in Fig. 1; note that, in Czech, the opposition of the direct reference in the ACT and the relational character of the predicate noun (PAT, with the copula) is determined by the possible occurrence of Instrumental case in PAT, rather than by word order, which is the primary case in English. In our translation of (1), the secondary position of the intonation center (marked by capitals) expresses the Focus as preceding, in this specific case, the Topic, i.e. the contextually bound item, CB (cf. Section 2.1.2 below). The fact that a word such as *Miss* can occur as CB in an 'out of the blue' sentence is due to its appurtenance to a set of items assumed by the speaker to be at hand for the hearer/reader in the given situation (typically, this set of 'permanently established items' contains the indexicals, such as *me, you, here, now*, and words referring to entities well known in the given culture, e.g. *Europe, Shakespeare, mountains*).

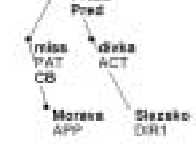

Figure 1

(1) *Miss Moravy je dívka ze Slezska.*
lit. Miss Moravia-Gen is girl from Silesia.

A girl from SILESIA is Miss Moravia.

Along with its lexical value and its functor, the label of a node contains a list of values of morphological **grammatemes**, which are not reproduced in Fig. 1. For Czech we work with grammatemes such as the attributes of **gender**, **number**, degree of **comparison**, **tense** (with a complex relationship of 'absolute' and 'relative' tense values, see Panevová's analysis in Sgall et al. 1986, Section 2.43), **aspect**, **iterativeness**, verbal **modality**, deontic modality (with the values DECL (declarative), DEB (debitive), HRT (hortative), VOL (volitive), POSS (possibilitive), PERM (permissive), FAC (facultative), mostly expressed by purely modal verbs), and sentence modality (for the head nodes of sentences and of certain types of clauses), with the values ENUNC(iative), EXCL(amative), DESID (desiderative), IMPER (imperative), INTER (interrogative).

2.1.2 Tectogrammatics and morphemics

If the relationships between tectogrammatical and morphemic representations are examined from a typological viewpoint, it can be claimed that those languages which do not exhibit a maximal preference of phenomena of a single type (such as Turkish with agglutination), the following means of expression are prototypically ("most naturally") used:

inflection applies in the core of syntax: inflectional endings, which often serve more than one function, express arguments (‚actants'), as well as the main kinds of adnominal adjuncts (APP, RSTR), and of finite verb forms,

analysis appears in the periphery of syntax (adverbal adjuncts, ‚circonstants' are expressed by prepositions and by subordinating conjunctions), and

agglutination (with derivational affixes) occurs in word formation.

The surface (morphemic) **word order** typically corresponds to the left-to-right order of the nodes in the TR, i.e. to the scale of **Communicative Dynamism** (CD, determined by ‚systemic ordering' in the Focus, see Sgall et al. 1995), in which Topic precedes Focus, or, more precisely, the **contextually bound** (CB) nodes precede their non-bound sister nodes and heads.

Thus, in (1), the **functors** PAT, APP and ACT are expressed by endings (zero, *-y* and *-a*, respectively), together with number and gender; also the verb ending has several functions (3rd Pers. Sing., Pres., Indic.), DIR1 is rendered by a preposition, and the word order corresponds to CD.

2.2 The core in automatic corpus annotation

2.2.1 Four stages of annotation of the PDT

The annotation procedure consists of the following steps (see Hajič 1998): morphemic analysis, morphemic tagging, syntactic tagging on the intermediate 'analytical level' and tectogrammatical level tagging, which leads to full underlying representations of the sentences from the corpus. The latter are handled in the shape of **tectogrammatical tree structures (TGTSs)**, which differ from the theoretically postulated TRs in that they contain specific nodes for coordinating conjunctions, instead of displaying more than two dimensions. Note that the left-to-right order of coordinated nodes in the TGTSs does not reflect CD.

The first three steps have been automatized to a high degree, using (i) the **morphemic** analyzer (Hajič 2002), which yields all possible values of the word forms present in the outer form of the sentence, (ii) a morphemic **tagger** (Hladká 2000), which chooses one of the values, (iii) the **'analytical** level', which has been developed as a technical device that has no immediate theoretical significance, but constitutes the first stage of syntactic annotations, bridging the gap between the morphemic string and the TGTS. In the analytical tree structure (ATS), every word of the sentence, including the punctuation marks, is represented by a single node. The ATSs produced by Collins' dependency parser (Collins et al. 1999), which yields the ATSs, are manually edited; instructions for the editing have been formulated (Bémová et al. 1997) and approximately 100000 sentences have been annotated at the analytical level. The final step is annotation on the underlying level – the **tectogrammatical** representation, in which only the autosemantic words correspond to nodes of the dependency tree. Auxiliary words and punctuation marks are captured as the grammatemes of the nodes. The relations between the nodes are marked with a more finegrained set of functors. The procedure of transition from ATS to TGTS is partly automatized, and the result of the automatic procedures is manually finalised by humans.

To illustrate the intermediate steps of the procedure, let us present some of the morphemic tags of the word forms of sentence (1), as identified by the morphemic tagger, with N and V indicating Noun and Verb, respectively, NOM and GEN standing for the cases Nominative and Genitive, and INDIC for the mood of Indicative:

```
        Miss.N.FEM.SG.NOM
        Moravy.N.FEM.SG.GEN
        je.V.3RD.SG.INDIC.PRES
        dívkaN.FEM.SG.NOM  ze.PREP
        Slezska.N.NEUT.SG.GEN
```

The ATS of (1) is presented in Figure 2:

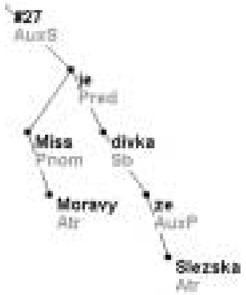

Figure 2

The values of functors are identified automatically in the following three cases. Value 'ACT' (actor) is assigned to every node indicated in the ATS as the subject of an active verb. If there is a subject and an 'object' depending on a passive verb, these two nodes are assigned functor 'PAT' and 'ACT', respectively. The head verbs of the sentences are assigned the functor 'PRED' (predicate). Additions to this part of the procedure are being prepared, so that further typical cases can be specified, esp. a node indicated in the ATS as an 'object' and expressed morphemically by the Dative case may get the functor ADDR(essee), adverbs may be lexically classified as corresponding to functors such as MEANS, MANN(er), EXT(ent), and so on.

2.2.2 Function words

To characterize how function **words** are treated automatically as grammatical morphemes in the TRs, let us note that every **preposition** node is deleted and its lexical value is stored in the attribute fw (function word) of the noun. The preposition will be used for the future (at least partly automatized) determination of the value of the syntactic grammateme of the noun (i.e. an index of its functor). Thus, if a preposition group depends on a verb, then e.g. *na* 'at, on', and *v* 'in' can be distinguished as different indices accompanying functor LOC, *do* 'into' and *k* 'to' can yield indices of DIR1, and so on.

Also every subordinating **conjunction** node is deleted. Its lexical value is stored in the attribute fw of the head verb of the subordinate clause.

A **modal** verb proper, which typically expresses one of the marked values of DEONTMOD, is merged with the auto-semantic verb depending on it in the ATS. The latter verb becomes the head of the subtree and the attribute of DEONTMOD of this verb is assigned its value according to the lexical value of the modal verb (see Table 1). The modal verb node is deleted. The analysis of sentence modality is based on the final punctuation and on the presence of certain words or verb forms in the sentence.

Modal verb	**Engl. Transl.**	**Attribute `deontmod` assigned**
Chtít	Want	VOL
Muset	Must	DEB
Moci, dát_se	May, can	POSS
Smět	be allowed	PERM
Umět, dovést	Can	FAC
Mít	Should	HRT

Table 1: Modal verbs

3 The vast and complex periphery

3.1 Markedness and restricted phenomena

There is a difference between (i) those marked items that belong to the core of the language systems (e.g. plural, preterite, future, the non-DECL modalities present in Table 1 above), and (ii) such marked phenomena which characterize non-prototypical subdomains either of tectogrammatics itself, or of its relationships to morphemics.

Two characteristic examples of marked aspects of **tectogrammatics** are:

(a) coordination - a syntactic relation of another type than dependency; their manysided possible combinations, cf. e.g. sentence (2), require TRs to constitute more-dimensional networks, which, however, allow for a univocal linearized

notation, in which every dependent item is enclosed in a pair of parentheses (with its functor assigned to the parenthesis oriented towards its head), cf. (3); the equivalence of TRs to such linear representations documents that the core of language is patterned in a way not remote from proposition calculus, which is significant for discussions concerning both language acquisition and an approach to linguistic description able to reflect the interactive nature of language (cf. Schnelle 1991) without requiring an excessively complex innate mechanism;
(b) focusing particles in their different positions (cf. the discussion of examples (4) and (5) below).

(2) Mary and Jim, who are our friends, live in Boston.
(3) ((Mary Jim)$_{\text{CONJ}}$ ($_{\text{DESCR}}$ (who)$_{\text{ACT}}$ be ($_{\text{PAT}}$ (we)$_{\text{APP}}$ friends)))$_{\text{ACT}}$ live ($_{\text{LOC}}$ Boston)

Marked phenomena in **morphemics** include especially:
(a) irregular morphemic paradigms (with synonymy of different sets of case endings, of personal endings, etc., and with ambiguity of many endings),
(b) function words, the tectogrammatical counterparts of which we mentioned in Section 2.2.2,
(c) deviations of the "surface" word-order from CD, cf. Section 2.1.2 above and the discussion of examples (4) - (6) in Section 3.2.2 below.

3.2 Towards an automatic parser

3.2.1 A semi-automatic way from the ATS to tectogrammatics

The semi-automatic procedure includes
(a) abolition of ATS nodes corresponding to function words and to most punctuation marks, with an indication of their functions by indices at the autosemantic units the function words belong to; cf. Section 2.2.1 on the exception concerning special nodes for coordinating conjunctions;
(b) assigning every lexical unit one of the more than 40 **functors** and a set of morphological **grammatemes** (marking the values of tense, modalities, number, etc.), as well as syntactic grammatemes (values such as 'in, on, under, among');
(c) addition of nodes for items **deleted** in the surface shapes of the input sentences;
(d) indication of the position of every node in the **topic-focus** articulation, with the scale of CD being represented as the left-to-right order of the nodes.
A considerable part of this procedure is handcrafted up to now, although along with the automatic treatment of large sets of prototypical phenomena, mentioned in Section 2.2, another set of automatic steps has been prepared, which completes some of the manual operations in cases in which it has not been difficult to formulate general rules. Thus, e.g., the build-up of the lexicon with entries including several kinds of grammatical data, especially the valency frames, automatic assignment of functors, or assignment - on the basis of the values of the existing attributes of coreference - of degrees of activation in the 'stock of shared knowledge,' as far as derivable from the use of nouns and pronouns in subsequent utterances, are under consideration. The word derivation is also planned to be enriched, as up to now only the most productive classes of distinct POSs are handled on the basis of the lemmas of the source words.

3.2.2 Present state of PDT annotations

To illustrate the present state of PDT annotations, we present examples which contain several marked cases. The treatment of the sentence (4) characterizes how coordination (with the binary conjunction *sice – ale*, similar to E. 'though' ... 'however'), deletion (of the second occurrence of the coordinated verb), and differences between the underlying and the morphemic (surface) word order are reflected in the TGTSs (cf. Section 2.1.2 for the notion of CD; note that an adjective primarily is more dynamic than its head noun, although the 'neutralized' morphemic order is A N), as well as a focusing particle in its primary position (at the beginning of Focus), with the functor RHEM(atizer); TWHEN, LOC, MANN are abbreviations for functors; the index CO marks the coordinated items (their sister node having no CO, such as the word *případ* 'case' in (4), represents an item depending on the coordinated construction as a whole):

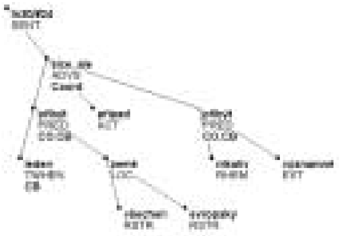

Figure 3

(4) *V lednu sice přibylo případů ve všech evropských zemích, nikoliv ale významně.*
Lit.: In January though raised cases in all European countries, not however significantly.
In January the number of cases raised in all European countries, but not significantly.

The next example shows the treatment of passive and of a case of inverted word order, with Topic following Focus in the morphemic string, which contrasts with the scale of CD in the TGTS. In the corpus, (5) is the first sentence of a newspaper text with the headline *Chřipky je prý zatím jen minimálně* 'Flu is supposed to have spread only minimally up to now', so that the reference to flu in (5) can be understood as contextually bound, belonging to Topic. The focusing particle *pouze* 'only' depends here on a CB noun (*člověk* 'man', the plural form of which is *lidé* 'people', Genitive *lidí*), which is connected with the fact that Focus is embedded more deeply than would correspond to the prototypical case, i.e. the head verb and all of its direct dependents are CB; a noun in a position similar to that of *člověk* has been called 'proxy focus' in Hajičová et al. (1998).

(5) *Pouze u devatenácti lidí v České republice byla letos v zimě prokázána chřipka.*
Lit. Only with nineteen people in Czech Republic was this-year in winter attested flu.
This winter, flu was attested in the Czech Republic only with nineteen people.

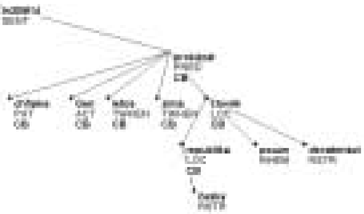

Figure 4

In the slightly simplified sentence (6) we can observe the handling of a subordinated (relative) clause, and also a case of morphemic word order differing from the scale of CD to such a degree that the condition of projectivity (similar to that of the continuity of constituents) appears not to be met by the outer form of the sentence, although it is assumed to hold in the TGTS. This appears to be conditioned, in the given case, by the two verbs constituting a specific cluster.

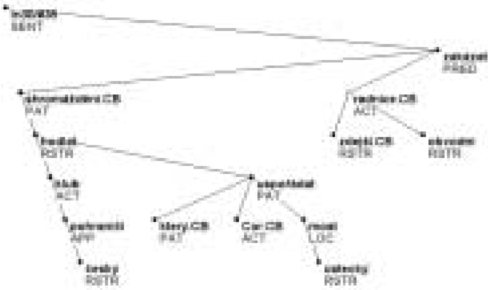

Figure 5

(6) *Permanentní shromáždění, které hodlá uspořádat Klub českého pohraničí na ústeckém mostě, zdejší obvodní radnice zakázala.*
Lit. Permanent assembly.Accus., which intends to-organize Club of-Czech borderland on Ústí bridge, local district Council prohibited.
The permanent assembly, which the Club of Czech Borderland intends to organize on the Ústí bridge, was prohibited by the local district Council.

4 Conclusions

Since, as it is broadly acknowledged, discontinuous constituents or cases with word order different from CD are non-prototypical, it may be assumed that the core of the language system displays a basic patterning that, although having the form of more-dimensional nets, meets strong requirements on the relationships between the different dimensions. This is documented by the existence of a one-to-one linearization, see ex. (3) in Section 3.1. It may then be important to analyze the theoretical advantages such a decriptive framework may bring, if its relevance for the issues of language acquisition is taken into account. It might be found out that the core of language is substantially patterned in ways which do not significantly surpass the structure of proposition calculus, i.e. come closer to common human mental capacities than what is assumed by theories working with a complex innate mechanism specific for the language faculty.

It is inappropriate to attempt at a specification of both the core and the periphery of language at once. The periphery comprises most different details, differing not only from one language to the other, but also between dialects, generations, styles, etc., and thus reflected also in the small steps typical for language development. A much more realistic approach is to aim at a description of the core of the language system by general principles or rules (in which languages differ just in the repertoire of attributes and of their values, not in the basis of the structural pattern), and to capture the most different non-prototypical phenomena, i.e. phenomena restricted by contextual conditions, just by rules of a more specific nature. The latter kind of rules concerns esp. the relationship between underlying sentence structure and the means of its expression, phenogrammatics, i.e. morphemics.

Syntactic analysis and parsing can then in principle be perspicuous, although the non-prototypical phenomena with different degrees of specificity (from large classes down to individual exceptions) are not easy to handle in full detail. Moreover, it is necessary to consider inferencing (based on a semantic classification of words) and statistical procedures (or their combinations with structural ones), which are needed for the big task of disambiguation of linguistic forms.

Acknowledgements

The research reported in this paper was carried out in the frame of Project MSMT LN00A063.

References

Bémová A., Buráňová E., Hajič J., Kárník J., Pajas P., Panevová J., Štěpánek J., Urešová Z. (1997) *Anotace na analytické rovině - příručka pro anotátory [Annotations on the analytical level – Manual for annotators]*, Technical Report #4, LJD UFAL MFF UK, Prague, Czech Republic

Collins M., Hajič J., Brill E., Ramshaw L., Tillmann C. (1999) *A Statistical Parser of Czech*. In: Proceedings of ACL 1999, pp. 505-512, College Park, Maryland

Hajič J. (1998) *Building a syntactically annotated corpus: The Prague Dependency Treebank*. In: Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. by E. Hajičová). Prague: Karolinum, pp. 106-132.

Hajič J. (2002) *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Prague: Karolinum.

Hajičová E., Partee B. H. and P. Sgall (1998) *Topic-focus articulation, tripartite structures, and semantic content*. Dordrecht:Kluwer.

Hladká B. (2000) *Czech Language Tagging*. PhD Thesis, Chrles University, Prague.

Holenstein E. (1975) *Roman Jakobsons phänomenologischer Strukturalismus*. Frankfurt am Main.

Petkevič V.(1995) *A new formal specification of underlying structures*. Theoretical Linguistics 21:7-61.

Plátek M., Sgall J. and Sgall P. (1984) *A dependency base for a linguistic description*. In Sgall P., ed.: Contributions to functional syntax, semantics and language comprehension. Amsterdam:Benjamins - Prague:Academia,1984, 63-97.

Ramat P. (1985) *Typologie linguistique*. Paris.

Sgall P. (1994) *Meaning, reference and discourse patterns*. In: Ph. Luelsdorff (ed.). The Prague School of Structural and Functional Linguistics. Amsterdam/Philadelphia: J. Benjamins, 277-309.

Sgall P., Hajičová E. and Panevová J. (2000) *Manual for tectogrammatical annotation.* Technical Report #7, UFAL MFF UK, Prague

Sgall P., Hajičová E. and Panevová J. (1986) *The meaning of the sentence in its semantic and pragmatic aspects*. Ed. by J. Mey. Dordrecht:Reidel - Prague:Academia.

Sgall P., Pfeiffer O., Dressler W. U. and Půček M. (1995) *Experimental research on Systemic Ordering*. Theoretical Linguistics 21:197-239.

Schnelle H. (1991) *Natur der Sprache*. Berlin: W. de Gruyter.

Skalička V. (1979) *Typologische Studien*. Braunschweig:Vieweg.