Corpus Annotation as a Test of a Linguistic Theory

Eva Hajičová and Petr Sgall

Institute of Formal and Applied Linguistics, Charles University, Prague
Malostranské náměstí 25, Prague 1, 118 00, Czech Republic
{hajicova,sgall}@ufal.mff.cuni.cz

Abstract

In the present contribution we claim that corpus annotation serves, among other things, as an invaluable test for linguistic theories standing behind the annotation schemes, and as such represents an irreplaceable resource of linguistic information for the build-up of grammars (Sect. 1.). To support this claim we present four linguistic phenomena for the study and relevant description of which in grammar a deep layer of corpus annotation as introduced in the Prague Dependency Treebank has brought important observations, namely the information structure of the sentence (Sect. 2.), condition of projectivity and word order (Sect. 3.), types of dependency relations (Sect. 4.) and textual coreference (Sect. 5.).

1. Introductory remarks

1.1. Annotation of corpus

It has been already commonly accepted in computational and corpus linguistics that grammatical (or lexical-semantic, etc.) **annotation** does not 'spoil' a corpus, since the annotation is done 'in addition' to the raw corpus. Thus, on the contrary, annotation may and should bring an **additional value** to the corpus. Necessary conditions for this aim are:

- its scenario is carefully (i.e. systematically and consistently) designed, and
- it is based on a sound linguistic theory.

This view is corroborated by the existence of annotated corpora of various languages such as Penn Treebank (English), its successors as PropBank or Penn Discourse Treebank, further Tiger (German), Prague Dependency Treebank (Czech).

Corpus annotation is not a self-contained task: it serves, among other things, as

- a support for projects of natural language processing,
- an invaluable test for linguistic theories standing behind the annotation schemes,
- an irreplaceable resource of linguistic information for the build-up of grammars.

It is important to note that the annotation concerns not only the surface and morphemic shape of sentences, but also (and first of all) the underlying sentence structure, which ellucidates phenomena hidden on the surface although unavoidable for the representation of the meaning and functioning of the sentence, for modelling its comprehension and for studying its semantico-pragmatic interpretation.

1.2. The Prague Dependency Treebank (PDT)

The multi-layered annotation of the **Prague Dependency Treebank** (PDT, see e.g. (Hajič, 1998)) as carried out at Charles University in Prague is based on the framework of the Functional Generative Description (FGD), described in previous publications (see e.g. (Sgall et al., 1986)). The process of the annotation during the last decade and its results have allowed for an enrichment of this framework in several points.

PDT is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus (CNK), with a mark-up on three layers: (a) morphemic, (b) surface shape, and (c) underlying (tectogrammatical). Its current version (publicly available since summer 2005, http://ufal.mff.cuni.cz/pdt2.0), annotated on all three layers, contains 3168 documents (text segments mainly from journalistic style) comprising 49442 sentences and 33357 occurrences of word forms (including punctuation marks).

FGP distinguishes the levels of **morphemics** (with a morphemic representation of the sentence having the form of a string of more or less narrowly connected items, i.e. lexical, derivational and inflectional morphemes) and of **tectogrammatics**, or underlying syntactic structure.

The underlying sentence structure is represented in the annotations in the form of tectogrammatical tree structures (TGTSs), in which a dependency tree representing (one of) the (literal) meaning(s) of a sentence is combined with added information concerning coordination and apposition, if present. Only autosemantic words are represented as nodes of the tree, function words having indices of node labels as their counterparts on this level (among these, the functors represent the dependency relations, i.e. arguments and adjuncts, and the values of grammatemes represent morphological units such as tenses, numbers, modalities, and so on). New nodes (not present in the morphemic form of the sentence) are added to account for surface deletions. Each of the edges of the tree instantiates one type of dependency (more exactly, dependency can be understood as a set of binary relations, i.e. of arguments and adjuncts; certain technical adjustments have been necessary for including the relations of coordination, apposition and parenthesis). In the valency frame of the head word (contained in its lexical entry), it is specified which arguments and adjuncts are obligatory with this word. The annotation within PDT has confirmed that, in most cases, the annotators agree the assignment of the tree structure (i.e. in establishing the edges,

except for complex combinations of dependency with coordination, see (Hajičová et al., 2002)).
In PDT, a technical supplementary layer, namely the so-called **analytical** level has been added to the two theoretically substantiated levels. This makes it possible to work with analytical tree structures (ATSs) as trees including a specific node for every item present in the surface form of the sentence (not only function words, but also punctuation marks are represented here as nodes) and a linear ordering of the nodes corresponding to the surface word order.

1.3. Objectives of the study

The present contribution concentrates on four linguistic phenomena for the study and relevant description of which in grammar a deep layer of corpus annotation has brought important observations:

- information structure (topic-focus articulation) of the sentence (Hypotheses A1, A2 in Sect. 2.),
- condition of projectivity and word order (Hypothesis B in Sect. 3.),
- types of dependency relations (Hypothesis C in Sect. 4.), and
- textual coreference (Hypothesis D in Sect. 5.).

2. Information structure of the sentence

2.1. Topic-Focus Articulation in PDT

Along with the dependency pattern, the tectogrammatical representations capture the **topic-focus articulation** (TFA), interpreted so that in a declarative sentence its F(ocus) is asserted to hold about its T(opic), or not to hold about T, in a negative sentence. Thus, in the prototypical case, F constitutes the scope of negation. Contextual Boundness (the linguistic counterpart of the cognitive opposition of given and new information) is seen as determining the dichotomy of T and F, in that a **contextually bound** (CB) item typically belongs to T, and a **non-bound** (NB) item belongs to F (in those marked cases in which the item does not depend directly on the main verb, being more deeply embedded, it is possible to find CB items in F and NB items in T, such as *my* and *nice*, respectively, in *This nice book belongs to my neighbor*). Further theoretical research connected with the possibility of a semantico-pragmatic interpretation of TFA by means of Partee's tripartite structures has indicated that **contrastive** and non-contrastive CB elements are to be distinguished both for Czech and for English (see (Hajičová et al., 1998), 151f; more details are given in (Hajičová and Sgall, 2004)). This enrichment of the descriptive framework is supported by a parallel tectogrammatical and prosodic annotation of a small spoken corpus of Czech, which has documented that the acoustical F0 characteristics of the sector of contrastive T are different from both those of the sector of non-contrastive T and from those of the F sector of the sentence (see (Veselá et al., 2003)).
In PDT, the attribute specifying TFA has three values:

t - contextually bound non-contrastive,
c - contextually bound contrastive,
f - contextually non-bound.

2.2. Bipartition of the sentence into Topic and Focus

To document the usefulness of corpus annotation for the study of TFA we present in this Section the results of our examination of two hypotheses.

Hypothesis A1
The division of the sentence into its T and F can be derived from the contextual boundness of the individual lexical items contained in the sentence.

If the preliminary definition of T and F (see (Sgall, 1981), also (Sgall et al., 1986), 216f) is "translated" into the PDT notation, i.e. using not only the values t and f, but also c of the TFA attribute, we get the following rules for the identification of the basic bipartition of the sentence in T and F:

(a) If the main verb has the TFA value f, it belongs to F. Else, it belongs to T.
(b) All the nodes immediately dependent on the main verb and carrying the TFA value t or c belong to T, together with all nodes depending on them, except the sentences in which the specific condition of rule (d) holds.
(c) All the nodes immediately dependent on the main verb and carrying the TFA value f belong to F together with all nodes depending on them.
(d) If the main verb carries the value t and all the nodes directly depending on the main verb also carry the value t, then follow the rightmost edge leading from the main verb down to the first node(s) on this path carrying the value f; this/these node(s) and all the nodes depending on it/them belong to F.

Note: More recently, the formulation of point (d) has been broadened, since it was found that in certain cases a NB node depends on a CB node that itself is subordinated to an NB node. The CB nodes to which a NB node is subordinated are called quasi-focus.
A tentative algorithm formulated in the mid-eighties has been implemented and tested on the whole of PDT; the results are reported by Kučová et al. (2005) and are summarized in the sequel. First let us present an example (see Fig. 1):

(1) *Nenadálou finanční krizi*
Lit.: (The) sudden financial crisis$_{Acc}$
podnikatelka řešila jiným způsobem.
(the) enterpreneur$_{Nom}$ solved by other means.
(context: The enterpreneur had to solve several problems before.)

An application of the above rules gives the following result:
Topic: *Nenadálou finanční krizi podnikatelka*
[the sudden financial crisis the enterpreneur]
Focus: *řešila jiným způsobem*
[solved by other means]
The implementation of the algorithm has led to a differentiation of five basic types of F:

(i) F consisting of the predicate and its subtrees,

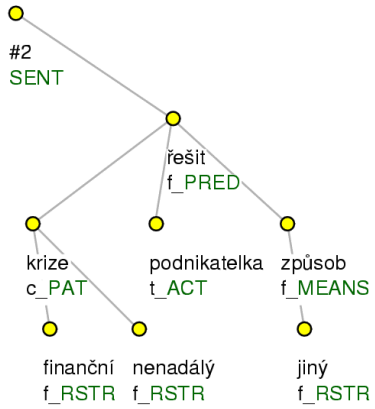

Figure 1: The preferred TGTS of sentence (1).

(ii) F consisting of the right-attached subtrees to a t-marked predicate
(iii) Quasi-focus with the t-marked main predicate
(iv) Quasi-focus with the f-marked main predicate
(v) F interrupted by a c-marked node

The frequency of these types as identified by the implementation of the algorithm to the TFA-annotated sentences in PDT is indicated in Table 1.

The results achieved by the implementation of the algorithm demonstrate that in Czech the boundary between T and F can be determined in principle on the basis of the consideration of the status of the main predicate and its direct dependents. In other words, the hypothesis has been significantly supported that in Czech the boundary between T and F is signalized by the position of the verb in the prototypical case; the boundary between T and F has been found immediately before the verb in 95% of the cases. It has also been confirmed that the TFA annotation leads to satisfactory results even with rather complicated "real" sentences in the corpus. To evaluate the results achieved by the implementation of the procedure identifying T and F on the basis of the contextual boundness of individual nodes of the underlying structure, a subcorpus of 10000 sentences has been annotated manually by three annotators in parallel as for the T/F bipartion. This will make it possible to check whether the procedure returns the results expected by the theory and also to evaluate the annotators agreement (see (Zikánová, 2006a)).

Even though some of the observations indicated above may – and should – lead to a certain modification of the annotation procedure, we do hope that the material gathered and analyzed in this way may be further used for the study of several aspects of the discourse patterning such as the linking of sentences in a text, a study of reference assignment based on the hierarchy of activation of elements the stock of knowledge the speaker assumes that the hearer(s) share(s) with him, and several other aspects of discourse structure.

2.3. Canonical order in Focus

Along with the ordering corresponding to the dependency relations, we work with a left-to-right linear ordering of the nodes in the TGTSs that may be interpreted as corresponding to the communicative dynamism, introduced by J. Firbas, i.e. as proceeding from T proper (the least dynamic, leftmost item) to F proper (most dynamic, rightmost). It has been assumed that within F this ordering prototypically is fixed, which can be formulated as the following hypothesis.

Hypothesis A2
In the focus part of the sentence the complementations of the verb (be they arguments or adjuncts) follow a certain canonical order in the tree, i.e. in the underlying representations, the so-called systemic ordering (not necessarily the same for all languages). In Czech, also the surface word order in F corresponds to systemic ordering in the prototypical case.

For the main dependency relations (functors in the sequel) in Czech, the following order is typical: Actor - Time:*since-when* - Time:*when* - Time: *how-long* - Time:*till-when* - Cause - Respect - Aim - Manner - Place - Means - Dir:*from-where* - Dir:*through-where* - Addressee - Origin - Patient - Dir:*to-where* - Effect. In English most of the adjuncts follow Addressee and Patient (see (Sgall et al., 1986)).

The validity of the hypothesis has been tested with a series of psycholinguistic experiments (with speakers of Czech, German and English); however, PDT offers a richer and more consistent material. Checking the hypothesis in PDT, we apply (a) the specification of F according to the rules mentioned above in Sect. 2.2, (b) the assumed order according to the scale of systemic ordering, and (c) the surface word order. In the TGTSs, the functors referring to the values of the dependency relation (valency slot) provide the information on the type of the complementation and the TFA annotation provides the information what is the focus part of the sentence (as judged by the annotators in the broader context by the assignment of one of the three values of the TFA attribute). These two pieces of information can be then used to check the order of the complementations in the actual sentence (preserved for the time being in the TGTS). The work is in progress and the final results will be reported in the report by (Zikánová, 2006b).

3. Condition of projectivity and word order

One of the issues frequently discussed in linguistic literature on the relation between syntactic structure and word order is the strongly restrictive condition of **projectivity**, which says that if a node *a* depends on *b* and there is a node *c* between *a* and *b* in the linear ordering, *c* is subordinated to *b* (where *subordinated* means an irreflexive transitive closure of *depends*). The more restricted the formal syntactic description is, the more valuable are the observations based on it; in this sense the condition of projectivity might well serve its purpose. However, there are seemingly many non-projective constructions in the surface shapes of the sentences. The task then is to attempt to classify the constructions, in which the condition of projectivity is not met in the analytical trees (ATSs, preserving the surface word order), and to attempt at a description meeting the condition as far as the core of the language system is concerned, but accounting by some simple well-defined means also for the cases of non-projectivity of analytical trees.

Type of F	No. of trees	Rel. frequency
F consisting of the predicate and its subtrees	46588	85.70
F consisting of the right-attached subtrees to a t-marked predicate	4664	8.58
Quasi-focus with the t-marked main predicate	1415	2.60
Quasi-focus with the f-marked main predicate	986	1.81
F interrupted by a c-marked node	30	0.06
Trees with which the identification of T and F was not unambiguous	617	1.14
Trees in which no F was identified	60	0.11
TOTAL	54360	100.00

Table 1: The frequency of the types of F as identified by the implementation of the algorithm to the TFA-annotated sentences in PDT.

If, in a theoretical description, we work without the analytical level, the relation between the linear ordering of the nodes of tectogrammatical structures and the morphemic word order is specified as a transition from projective trees to strings of morphemes, in which the condition of projectivity cannot be applied; cf. the examples of movement rules in (Sgall, 1997). We may formulate this assumption as the following hypothesis.

Hypothesis B
The TGTSs are projective; the marked cases in which the surface word order is not in accordance with projectivity, can be specified by movement rules.

As mentioned above, one of the important features of TGTSs consists in the fact that they do not contain nodes for function words; from this it follows that in the numerous cases in which the "non-projectivity" of surface word order concerns auxiliary verbs or conjunctions, etc., the projectivity of TGTSs is not at stake.
For an illustration of this point, see the highly simplified ATS and TGTS for sentence (2) in Fig. 3 and 3, respectively.

(2) *Pro podnikatele by tu mohl být ráj.*
Lit.: For enterpreneurs $_{Cond}$ here cold be paradise
(content: For enterpreneurs there could be a paradise here.)

The PDT with its multi-layered scenario provides an extremely precious material for the classification of non-projectivities in the surface shape of the sentence and for an examination of the reasons of them, as documented by the doctoral thesis of Zeman (2004) and the paper by Hajičová et al. (2004). This material also serves well for the purpose of checking Hypothesis B.
Zeman's data contain 73 088 sentences annotated on the analytical layer, which comprise 1 255 590 occurrences of words. The condition of projectivity is broken by 23 691 pairs of words (1,9 %). The number of sentences in which the condition is broken is 16 920 (23,2 % of all sentences). As was demonstrated in (Hajičová et al., 2004), these cases can be divided into three groups:

(i) leftpreposing of items exhibiting specific grammatical properties (e.g. of reflexive and interrogative words, or of items depending on infinitives or on comparative constructions) or which belong to closed lexical groups, esp. idioms (37 %);

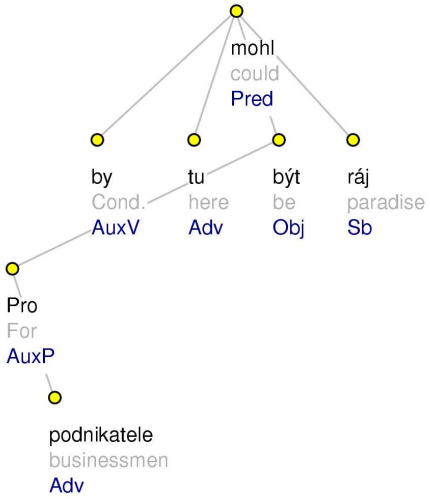

Figure 2: An ATS of sentence (2).

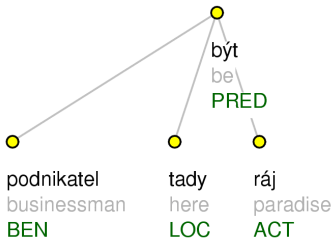

Figure 3: A TGTS of sentence (2).

(ii) syntagms divided into parts of T and of F (6 %);
(iii) analytic verb forms and other constructions with function words (57 %).

Group (iii) concerns only the technical layer of ATSs, rather than TGTSs, and group (i) can be relatively easily delimited. Thus, only the smallest group (ii) is problematic; it has been found that in most of such cases the preposed item has the value c, i.e. is a contrastive CB element. In some cases it is a dependent item, e.g. in *Společnou máme tuto odpovědnost* (lit.: Common we have this responsibility), in others it is a head noun, e.g. in *Záruky nemám žádné* (Guar-

anties I have none). Along with these cases, there are also collocations without contrast, such as *Měl plné kapsy peněz* (He had full pockets of money).

4. Types of dependency relations

In FGP, the level of **tectogrammatics**, or underlying syntactic structure, is based on **dependency** syntax, i.e. on the valency of the verb and of other words (with the relations of coordination and of apposition understood as specific orderings, the combinations of which with dependency are narrowly restricted, so that a linearized sentence representation with parentheses and indices is possible). In the domain of **valency**, i.e. of the dependency relations, arguments (inner participants, i.e. Actor, Patient, Addressee, Origin and Effect) are distinguished from adjuncts (free complementations, such as Locative, several directional and temporal complementations, Manner, Cause, Condition, etc.). The arguments prototypically are obligatory with the individual head words, each of them occurs as dependent on head words from a limited group, the morphemic shape of an argument is specific for the head word and semantically an argument is vague, in certain cases being further blurred by the "shifting" of the relationships to the corresponding cognitive roles, cf. e.g. *He left the town* (Patient corresponding to the role that typically is represented by Origin), *They addressed us* (Addressee - Patient). On the other hand, adjuncts are optional in the non-marked case, they may accompany any word from the given class, they are semantically homogeneous (without the "shifting") and they they are expressed by different prepositions and cases (relevant for such semantic opposition as that between *in, on, under*, etc. with Locative, Positive or Negative with Benefactive, and so on). These theoretical considerations can be formulated as Hypothesis C.

Hypothesis C
There are two types of valency slots, arguments and adjuncts, distinguished by operational criteria.

The corpus annotation and the work on a valency dictionary related to the annotation indicates that this hypothesis is too strong and that a third type of relation should be distinguished, i.e. complementations sharing certain properties with the arguments and other with the adjuncts, such as e.g. Obstacle (*He stumbled over the table*), Difference (*We won by two goals*) or Mediator (*They pulled the dog by its collar*); see (Lopatková and Panevová, 2005) where the term 'quasi-valency' is proposed for this type of valency slots.

5. Textual coreference

Another domain we analyze is that of **textual coreference** (differing from grammatical coreference rendered by grammatical means - reflexive and relative pronouns, control relations induced by verbs or nouns of control). Although it goes beyond the frame of grammar, textual coreference is reflected in a certain extent in the annotation of sentences in PDT; the antecedents of demonstrative, personal and possessive pronouns are identified, as well as those of the zero form of the 3rd Person subject pronoun (cf. C. Fillmore's "silent" anaphors). Anaphorical links of different kinds are distinguished: (a) to a particular node, (b) to the governing node of a (sub)tree (including clauses and sentences), (c) to a text segment, (d) deixis, exophora. Along with coreference, also bridging (associative) anaphora is being studied. The assumption that CB items typically are coreferential has been tested on a small subcorpus of PDT (80 text segments), also the coreference of nouns is being studied; see (Hajičová et al., 2006).

It is an important question what enables the addressee to identify the reference of referring expressions in discourse. The following hypothesis is studied in the Prague group.

Hypothesis D
A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge (information) shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation of referents.

Heuristic rules aiming at the specification of the changing degrees of activation have been presented by Hajičová (1993) and in earlier writings quoted there. These rules cover the basic layer of the course of activation changes, taking into account the positions of potential antecedents in TFA, as well as coreferential and anaphoric links. Also a procedure yielding a visualization of the development of activation in the form of a schematic graph has been prepared, see (Hajičová et al., 2006). Texts are analyzed as for discourse segmentation and for the types of corefering expressions. This will make it possible to check Hypothesis D on a richer material. Already the first steps in this analysis confirm that this orientation of discourse studies lead to interesting observation and that Hypothesis D can be understood as plausible.

Conclusions

We wanted to document on certain selected grammatical and discourse phenomena that systematic and consistent corpus annotation on the level of (underlying) syntax constitutes a useful means for testing a linguistic descriptive framework. PDT offers up-to-now absent possibilities of such testing, i.e. of enriching processes of construction and enrichment of grammars.

Acknowledgement

The research reported on in this contribution has been supported by the grant of the Czech Ministry of Education MSM0021620838 and by the project of Information Society No. 1ET201120505.

6. References

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Prague: Karolinum.

Eva Hajičová and Petr Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In *Information structure – Theoretical and empirical aspects*, pages 1–13. Berlin-New York: Walter de Gruyter.

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

Eva Hajičová, Petr Pajas, and Kateřina Veselá. 2002. Corpus annotation on the tectogrammatical layer: Summarizing of the first stages of evaluation. *The Prague Bulletin of Mathematical Linguistics*, (77):5–18.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, (81):5–22.

Eva Hajičová, Barbora Hladká, and Lucie Kučová. 2006. From a sentence to discourse patterns: Annotated corpus as a test bed. (in press).

Eva Hajičová. 1993. *Issues of Sentence Structure and Discourse Patterns*. Prague: Charles University.

Lucie Kučová, Kateřina Veselá, Eva Hajičová, and Jiří Havelka. 2005. Topic-focus articulation and anaphoric relations: A corpus based probe. *The Prague Bulletin of Mathematical Linguistics*, (84):5–12.

Markéta Lopatková and Jarmila Panevová. 2005. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, pages 83–92. Veda Bratislava, Slovakia.

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

Petr Sgall. 1981. Towards a definition of focus and topic. In *Prague Studies in Mathematical Linguistics 7*, pages 173–198.

Petr Sgall. 1997. On the usefulness of movement rules. In *Actes du 16e Congrès International des Linguistes*. Oxford: Elsevier Sciences.

Kateřina Veselá, Nino Peterek, and Eva Hajičová. 2003. Topic-Focus articulation in PDT: Prosodic characteristics of contrastive topic. *The Prague Bulletin of Mathematical Linguistics*, (79-80):5–12.

Daniel Zeman. 2004. *Parsing with a Statistical Dependency Model*. Ph.D. thesis, Charles University in Prague, Faculty of Mathematics and Physics.

Šárka Zikánová. 2006a. Identification of Topic and Focus in Czech: Comparative evaluation on PDT. (in prep.).

Šárka Zikánová. 2006b. What do the data in PDT say about systemic ordering in Czech? (in prep.).