PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 69–124

Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study

Jan Hajič, Eva Hajičová, Jiří Mírovský, Jarmila Panevová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

A case study based on experience in linguistic investigations using annotated monolingual and multilingual text corpora; the "cases" include a description of language phenomena belonging to different layers of the language system: morphology, surface and underlying syntax, and discourse. The analysis is based on a complex annotation of syntax, semantic functions, information structure and discourse relations of the Prague Dependency Treebank, a collection of annotated Czech texts. We want to demonstrate that annotation of corpus is not a self-contained goal: in order to be consistent, it should be based on some linguistic theory, and, at the same time, it should serve as a test bed for the given linguistic theory in particular and for linguistic research in general.[1]

1. Introduction

It is now quite easy to have access to large corpora for both written and spoken language. Corpora have become popular resources for computationally minded linguists and computer science experts developing applications in Natural Language Processing (NLP). Linguists typically look for various occurrences of specific words

[1] The present contribution is based in part on our previous summarizing study on annotation (Hajič et al., 2015), and also on studies concerning some particular linguistic phenomena quoted in the respective Sections below. We are grateful to our colleagues for providing us their material and expertise. Most importantly, we owe our thanks to Markéta Lopatková for her careful reading of the prefinal version of this paper and for her invaluable comments. The authors highly appreciate the comments and suggestions given by the two anonymous reviewers and have tried to take them into account when preparing the final version of the paper. All the responsibility, however, rests with the authors of the present paper.

 Corresponding author: hajicova@ufal.mff.cuni.cz
Cite as: Jan Hajič, Eva Hajičová, Jiří Mírovský, Jarmila Panevová. Linguistically Annotated Corpus as an Invaluable Resource for Advancements in Linguistic Research: A Case Study. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 69–124. doi: 10.1515/pralin-2016-0012.

or patterns, computational specialists construct language models and build taggers, parsers, and semantic labelers to be used in various applications.

It has also already been commonly accepted in computational and corpus linguistics that grammatical, lexical, or semantic, etc. annotation does not "spoil" a corpus, if the annotation is done in such a way that it does not remove substantial information about the raw corpus, such as spacing etc. (ideally, as stand-off annotation). On the contrary, annotation may and should bring an additional value to the corpus. Necessary conditions for this aim are:

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

This view is corroborated by the existence of annotated corpora of various languages: Penn Treebank (English; Marcus et al., 1993), its successors as PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers et al., 2004) or Penn Discourse Treebank (Prasad et al., 2008), Tiger (Brants et al., 2002) and Salsa (German; Burchardt et al., 2006), Prague Dependency Treebank (Czech; Hajič et al., 2006; Bejček et al., 2013), and many others.

The aim of our contribution is to demonstrate, on the basis of our experience with the annotated corpus of Czech, the so-called Prague Dependency Treebank (PDT), how annotation process and its results help to test a linguistic theory, to develop it further and also to compare it with other theories, that is, how such work may contribute to a better understanding of the language system.

We first present a brief account of PDT in its current state (Sect. 2), passing over to a layer-by-layer description of individual cases which may serve as examples of phenomena for the understanding of which the annotated treebank was instrumental (Sections 3 and 4). In Sect. 5 we add some statistical information on PDT data and on the tools available as well as some remarks on the annotation process as such. We sum up our observations in Sect. 6 highlighting first in which points the existing theoretical framework has been complemented and adding one particular aspect the study of which has been made possible by the consistent and systematic annotation.

2. The Prague Dependency Treebank in a nutshell

The Prague Dependency Treebank is an effort inspired by the PennTreebank; the work started as early as in the mid-nineties and the overall scheme was already published in 1997 (see Hajič et al., 1997 and Hajič, 1998). The basic idea was to build a corpus annotated not only with respect to the part-of-speech tags and some kind of (surface) sentence structure, but also capturing the syntactico-semantic, deep structure of sentences.

The annotation scheme of PDT is based on a solid, well-developed theory of an (integrated) language description, the so-called Functional Generative Description (FGD) (see, e.g., Sgall, 1967; Sgall et al., 1969; Sgall et al., 1986); at the time of the devel-

opment of the annotation scheme this theory had already been applied to an analysis of multifarious linguistic phenomena, mostly concentrated on Czech but also in comparison with English, Russian or some other (mainly Slavonic) languages. The principles of FGD were formulated as a follow-up to the functional approach of the Prague School and with due respect to the strict methodological requirements introduced to linguistics by N. Chomsky. The FGD framework was formulated as a generative description that was conceived of as a multi-level system proceeding from linguistic function (meaning) to linguistic form (expression), that is from the generation of a deep syntactico-semantic representation of the sentence through the surface syntactic, morphemic and phonemic levels down to the phonetic shape of the sentence. From the point of view of formal grammar, both syntactic levels were based on the relations of dependency rather than constituency. The main focus was laid on the account of the deep syntactic level, called "tectogrammatical" (the term borrowed from Putnam's (1961) seminal paper on phenogrammatics and tectogrammatics). On this level, the representation of the sentence has the form of a dependency tree, with the predicate of the main clause as its root; the edges of the tree represent the dependency relations between the governor and its dependents. Only the autosemantic (lexical) elements of the sentence attein the status of legitimate nodes in the tectogrammatical representation; functional words such as prepositions, auxiliary verbs and subordinate conjunctions are not represented by separate nodes and their contribution to the meaning of the sentence is captured within the complex labels of the legitimate nodes (see below on the characteristics of the tectogrammatical level in PDT). An important role in the derivation of sentences is played by the information on the valency properties of the governing nodes, which is included in the lexical entries: the valency values are encoded by the so-called functors, which are classified into arguments and adjuncts. It is assumed that each lexical entry in the lexicon is assigned a valency frame including all the obligatory and optional arguments appurtenant to the given entry; the frame also includes those adjuncts that are obligatory with the given entry; in accordance with the frame, the dependents of the given sentence element are established in the deep representation of the sentence and assigned an appropriate functor as a part of their complex label. The representation of the sentence on the tectogrammatical level also captures the information structure of the sentence (its topic–focus articulation) by means of the specification of individual nodes of the tree as contextually bound or non-bound and by the left-to-right order of the nodes. Coordination and apposition is not considered to be a dependency relation as they cannot be captured by the usual binary directional dependency relation. Coordinated sentence elements (or elements of an apposition) introduce a non-dependency, "horizontal" structure, possibly n-ary and/or nested, but still undirectional, where all elements have (in the standard dependency sense) a common governor (the only exception is formed by coordinated main predicates which naturally have no common governor). The coordinated (or appended) elements can also have common dependent(s). All the depen-

dency relations expressed in a sentence with coordination(s) and/or apposition(s) can be extracted by "multiplying" the common dependency relations concerned.

The design of the annotation scenario of PDT (see, e.g., Hajič, 1998; Böhmová et al., 2003; Hajič et al., 2006; Bejček et al., 2011; Bejček et al., 2013) follows the above conception of FGD in all of the fundamental points:

(i) it is conceived of as a multilevel scenario, including the underlying semantico-syntactic layer (tectogrammatical),
(ii) the scheme includes a dependency based account of syntactic structure on both (surface and deep) syntactic levels,
(iii) the scheme also includes the basic features of the information structure of the sentence (its topic–focus articulation) as a component part of the underlying syntax, and
(iv) from the very beginning, both the annotation process and its results have been envisaged, among other possible applications, as a good test of the underlying linguistic theory.

PDT consists of continuous Czech texts, mostly of the journalistic style (taken from the Czech National Corpus) analyzed on three levels of annotation (morphology, surface syntactic structure, and underlying syntactic structure). At present, the total number of documents annotated on all the three levels is 3,165, amounting to 49,431 sentences and 833,193 (occurrences of) word forms and punctuation marks (tokens). PDT, Version 1.0 (with the annotation of the first two levels) is available from the Linguistic Data Consortium, as is Version 2.0 (with the annotation of the third, underlying level). PDT Version 2.5 (with some additions) as well as the current PDT Version 3.0 are available from the LINDAT/CLARIN repository.[2]

The original annotation scheme has the following multilevel architecture:

(a) **morphological layer**: all tokens of the sentence get a lemma and a (disambiguated) morphological tag,
(b) **analytical layer**: a dependency tree capturing surface syntactic relations such as subject, object, adverbial; a (structural) tag reflecting these relations is attached to the nodes as one of the component parts of their labels,
(c) **tectogrammatical layer** capturing the underlying ("deep") syntactic relations: the dependency structure of a sentence on this layer is a tree consisting of nodes only for autonomous meaningful units (function words such as prepositions, subordinating conjunctions, auxiliary verbs etc. are not included as separate nodes in the structure, their contribution to the meaning of the sentence is cap-

[2] http://www.lindat.cz

tured by the complex labels of the autonomous units). Every node of the tectogrammatical representation is assigned a complex label consisting of:[3]

- the lexical value of the word (for verbs and certain nouns, with a reference to its sense captured in the corresponding valency lexicon entry),
- its '(morphological) grammatemes' (i.e. the values of morphological categories such as Feminine, Plural etc. with nouns, Preterite, etc. with verbs),
- its 'functors' (such as Actor, Patient, Addressee, Origin, Effect and different kinds of circumstantials (adjuncts), with a more subtle differentiation of syntactic relations by means of subfunctors, e.g. 'in', 'at', 'on', 'under', 'basic', 'than', etc.), and
- the topic–focus articulation (TFA) attribute containing the values for contextual boundness, on the basis of which the topic and the focus of the sentence can be determined. Pronominal coreference is also annotated.

In addition to the above-mentioned three annotation layers in PDT, there is also one non-annotation layer representing the "raw-text". In this layer, called the "word layer", the text is segmented into documents and paragraphs and individual tokens are recognized and associated with unique identifiers. Figure 1 displays the relations between the neighboring layers as annotated and represented in the data. Thus, for example, the Czech sentence *Můžete to vysvětlit například?* (lit.: "Can-you it explain on-example", E. translation: "Could you explain it with an example? ") contains a modal verb, a pronoun, a content verb, and a prepositional phrase (with a typo).

One methodological comment should be made. Though partitioned into layers, the annotation scheme of the Prague Dependency Treebank was built as a complex one: we have annotated all the language phenomena on the same collection of texts rather than to select only some phenomenon or phenomena of a particular layer without taking into account other phenomena of the same layer. At the same time, however, each layer of annotation is accessible separately, but with a possible explicitly annotated link to the other layers of annotation. The relations between the layers are in part captured in the associated valency lexicon for verbs and their arguments, along the lines suggested in (Hajič and Honetschläger, 2003; Hajič and Urešová, 2003).

In the process of the further development of PDT, additional information has been added to the original in the follow-up versions of PDT, such as the annotation of basic relations of textual coreference and of discourse relations in the Prague Discourse Treebank (PDiT), multiword expressions etc.

---

[3] In Fig. 1 there is only a very simplified tectogrammatical representation of the given sentence as the Figure is meant to illustrate the interlining of layers of annotation rather than to bring a full annotation of the sentence on each of the layers. On the tectogrammatical layer (t-layer), the modal verb *můžete* [can you] does not obtain a node of its own and the modal meaning is captured by an index attached to the lexical verb *vysvětlit* [explain], which is however not displayed in the Figure, and also the morphological categories are omitted. (The index 'inter' stands for interrogative mood, and, e.g., #Gen is a label of a node representing a "general" participant, ADDR standing for Addressee.)

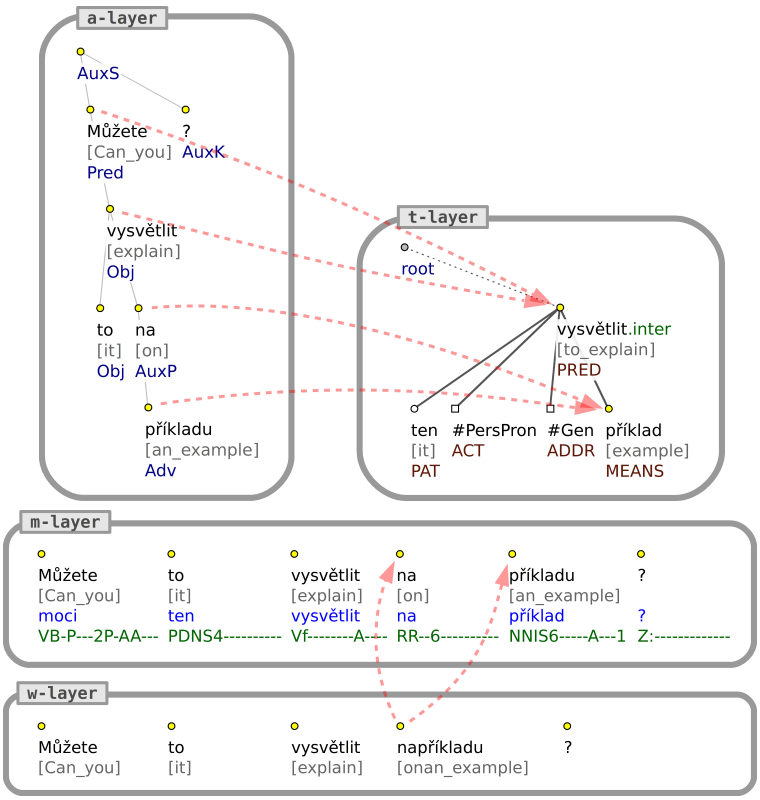

Figure 1. An example of the interlinking of the PDT layers in the Czech sentence "Můžete to vysvětlit na-příkladu?" [Lit.: Can-you it explain on-an-example?] [Can you explain it on-an example?]. The arrows represent non-1:1 relations among nodes on different layers of annotation; square nodes in the tree on the t-layer represent 'newly' generated nodes – nodes without a surface counterpart.

3. Case studies I: Morphology, surface and deep syntax

The contribution of corpus annotation for the theoretical description of language was greater than we expected at the beginning of the process. There are two phases in which this contribution comes out: In the first, preparatory decision-making phase, the proposed scheme was tested before a definite scenario with detailed instructions was approved for the build-up of the annotated corpus; at that point, the development of the annotation scenario itself reflected the state-of-the-art of the object of annotation. The tuning of the scenario (and the theoretical reflections there-off) was mainly

based on the annotators' feedback. The second phase started when the annotated corpus was ready for its exploitation for studies of theoretical issues at the end of the creation of the annotated corpus. The collection of data annotated according to the consistent scenario opened new horizons for the theoretical study of the particular phenomena on the basis of rich, real material not readily available before the time of corpus linguistics.

In the following subsections, we present an analysis of some grammatical issues based on the annotation process which has stimulated a modification of the theoretical framework of FGD or has made it necessary to supplement the existing handbooks of Czech grammar. For this purpose we have selected the following issues: Sect. 3.1 presents arguments for the necessity of an introduction of new morphological grammatemes constituting the category of diathesis, in Sect. 3.2 the peculiarities of counting objects occurring typically in pairs or groups and their morphological consequences are discussed, while in Sections 3.3 and 3.4 issues connected with valency are analyzed and an introduction of the "quasi-valency" class of modifiers is substantiated. Selected types of deletions are described in Sect. 3.5. In Sect. 3.6 variants of nominative subjects are analyzed. Sect. 4 is devoted to issues the analysis of which has brought modifications of the FGD approach to some particular aspects of the information structure of the sentence (Sect. 4.1) or of phenomena that concern the domain of discourse, which go beyond the domain of grammar and as such have been out of the scope of FGD interests (discourse relations in Sect. 4.2 and associative and coreference relations in Sect. 4.3).

As the empirical material of our analysis is Czech, we accompany the Czech example sentences with their translations to English, in most cases both literal and free. When necessary, we add (simplified) glosses capturing the information on relevant morphological and syntactic features of the Czech forms.[4]

3.1. Diathesis[5]

The morphological meanings of verbs connected with the verbal voice were usually limited to the opposition active – passive. Our analysis of the Czech data has demonstrated that there are other constructions productive enough in Czech to be considered as members of the same category as active and passive. Due to their productivity and due to the consequences they have for the syntactic structure we proposed to assign these analytical verb forms a new morphological category (grammateme)[6] called

[4] It should be noted that in order to make the glosses easier to survey we accompany them only by those features that are necessary for the understanding of the point under discussion. We assume that the abbreviations in the glosses are self-explaining and correspond to the Leipzig glossing rules; if not, we add a commentary in the text or in a footnote.

[5] In many cases the analytical passive diathesis and simple resultatives seems to be ambiguous, but some formal criteria how to distinguish their diathesis values are given in Panevová and Ševčíková (2013).

[6] For the notion of grammateme, as applied in FGD, see Sect. 2 above.

"diathesis" with the values active, analytical passive, resultative diathesis (simple and possessive) and recipient passive.

Our classification slightly differs from the Czech traditional descriptions, in which these constructions are analyzed as a special verbal tense, with a certain analogy to perfect tenses in other languages (Mathesius, 1925) or as a special verbal category called "resultative state" ("výsledný stav" in Czech, see Hausenblas, 1963)[7]. Our analysis (Panevová et al., 2014) supports the idea about the position of this construction within the diathesis paradigm. These types of diathesis are expressed by different morphemic forms, they have different morphological meanings, and they influence the syntactic structure, which is different from their unmarked active counterparts. The active sentences (1) and (4) have the following counterparts differentiated by the value of diathesis: (2) and (3) for (1), (5) for (4),

(1) Dcera už připravila matce oběd.
daughter-NOM-sg already prepare-3-sg-PST mother-DAT-sg lunch-ACC-sg
[The daughter has already prepared lunch for her mother.]

(2) Oběd už je připraven.
lunch-NOM-sg-M-Sbj already be-AUX-3-sg-PRS prepare-PTCP-PASS-sg-M
[Lunch is already cooked.]

(3) Matka už má oběd připraven.
mother-NOM-sg-F-Sbj already have-AUX-3-sg-PRS lunch-ACC-sg-M prepare-PTCP-PASS-sg-M
[lit. Mother has her lunch already prepared.]

(4) Nakladatelství zaplatilo autorovi honorář včas.
Publishing_house-NOM-sg-Sbj pay-3-sg-PST author-DAT-sg-M fee-ACC-sg-M in_time
[The publishing house has paid the fees to the author on time.]

(5) Autor dostal od nakladatelství honorář zaplacen včas.
author-NOM-sg-M-Sbj receive-AUX-3-sg-PST-M from-PREP publishing_house-GEN-sg fee-ACC-sg-M pay-PTCP-PASS-ACC-sg-M in_time
[The author has received his fees from the publishing house in time.]

[7] A detailed analysis of resultative constructions in contemporary Czech from theoretical and empirical view is presented in Giger (2003).

In (2) and (3) the action of the preparation of lunch is presented from the point of view of the result, while actions in the active constructions are presented from the point of view of the Actor. In the simple resultative (ex. (2)) the result (*oběd* [lunch]) of the action (Patient of the verb) is shifted to the position of surface subject and the Actor is omitted. In the possessive resultative constructions (ex. (3)) two kinds of restructuring are possible: in (3) mother could be understood to be the actor of the lunch preparation, but the meaning that somebody else has prepared a lunch (for mother) is also possible. In (3) the verb *mít* [have] is used in possessive resultative as an auxiliary and this construction enters the verbal paradigm;[8] since there exist verbs for which this type of diathesis is not applicable, the feature *+res_poss* indicating the possible participation of the given verb in this kind of diathesis is included in the lexicon.

Example (5) represents a less frequent diathesis where the verb *dostat* [receive] is used as an auxiliary. Contrary to its unmarked counterpart (4), the Addressee of the action in (5) is shifted into the position of the surface subject; the Actor of the action could be optionally expressed (here by the prepositional phrase *od nakladatelství* [from the publishing house]).[9]

As a result of these observations and analysis the original set of morphological categories was rearranged and extended in the modified version of the theoretical framework of FGD and in PDT (Urešová, 2011a).

3.2. Number of nouns

In the category of number the Czech nouns enter a basic opposition: singular (sg) and plural (pl). However, this category exhibits some peculiarities, especially with nouns denoting pairs or typical groups (such as *boty* [shoes], *rukavice* [gloves], *sirky* [matches], *klíče* [keys]). With other nouns we use the class of basic numerals, see *jedna kniha* [one book-sg], *dvě knihy* [two books-pl], *tři knihy* [three books-pl], etc. For counting the objects denoted by pair and group nouns, the set numerals are obligatorily used instead of the basic numerals. Rich material provided by the PDT supported an introduction of a new morphological category called pair/group meaning. Thus, we work with two paradigmatic patterns of the meaning of number: the former is connected with counting single objects, the latter with counting pairs of them or the typical sets of them (e.g. *jedna bota, tři boty* [one-*basic numeral* shoe-sg, three-*basic numeral* shoes-pl], *jeden klíč, pět klíčů* [one-*basic numeral* key-sg, five-*basic numeral* keys-pl] vs. *jedny* [set numeral] *boty, troje* [set numeral] *boty* [one pair of shoes, three pairs of shoes]; *jedny* [set numeral] *klíče, patery* [set numeral] *klíče* [one set of keys, five sets of

[8] The grammaticalization of this category indicates that Czech belongs to the class of "habere" languages (see Clancy, 2010).

[9] The syntactic diathesis (deagentization, dispositional constructions and reciprocals) has been implemented in PDT 3.0 and was described from the theoretical point of view in Panevová et al. (2014).

keys]). The differences between Czech and English demonstrate that in Czech the pair and set meaning of the nouns is grammaticalized, since a special type of compatibility with numerals is required.

If nouns occurring typically in groups or sets occur in a plural form without a numeral the sentences are often ambiguous. In (6a) the regular plural form (unmarked as for pair/group meaning) of the noun *rukavice* [glove] is used. For (6b), (7a) and (7b) several interpretations are possible; their English translations reflect their preferred meanings chosen on the basis of world knowledge or a broader context. In (7b), e.g., the knowledge of the habits used in this office would help for disambiguation if the charwoman has a single key belonging to each office or if for any office a set of keys were needed.

(6a) Často něco ztrácím, teď mám doma několik levých
often something loose-1-sg-PRS just-now have-1-sg-PRS at-home several left
rukavic.[10]
glove-pl
[I usually lose my things, just now I have at home several left gloves.]

(6b) Musím si koupit nové rukavice.
need-1-sg-PRS REFL-DAT buy-INF new glove-sg-PAIR/GROUP
[I have to buy a new pair of gloves.]

(7a) Ztratila jsem klíče od
Loose-1-sg-PST be-AUX-1-sg-PRS key-sg-PAIR/GROUP from-PREP
domu.
house-GEN-sg
[I have lost my keys from my home.]

(7b) Uklízečka má klíče od všech
Charwoman-NOM-sg have-3-sg-PRS key-ACC-pl from-PREP all
pracoven.
office-GEN-pl
[The charwoman has keys from all of the offices.]

The introduction of the new morphological category pair/group meaning is based first of all on the requirement of economy of the description of these nouns in the lexicon: A single lexical entry is sufficient for the nouns referring either to a single (particular) object, or to a typical pair, or a typical set of these objects. The compatibility of the members of the opposition +pair/group vs. -pair/group meaning with a different class of numerals is also a strong argument in favour of the introduction of

[10] In order to explain the pair/group meaning as a new unit we use in the glosses for (6) and (7) the meanings of the number rather than their forms.

Noun lemma	# of plural forms	# of pl. forms with the pair/group meaning	Percentage
dvojče [twin]	5	5	100.0%
pouto [tie]	5	5	100.0%
ledvina [kidney]	7	7	100.0%
vlas [hair]	11	11	100.0%
kopačka [football shoe]	5	5	100.0%
ucho [ear]	9	9	100.0%
lyže [ski]	13	13	100.0%
schod [stair]	6	6	100.0%
ruka [hand, arm]	81	77	95.1%
prst [finger/toe]	10	9	90.0%
oko [eye]	89	80	89.9%
rameno [shoulder]	9	8	88.9%
rukavice [glove]	8	7	87.5%
kolej [rail]	16	14	87.5%
noha [foot, leg]	20	17	85.0%
kulisa [scene]	6	5	83.3%
koleno [knee]	5	4	80.0%
bota [shoe]	30	24	80.0%
klíč [key]	8	5	62.5%
zub [tooth]	14	8	57.1%
rodič [parent]	87	37	42.5%
křídlo [wing]	17	5	29.4%
doklad [document]	35	8	22.9%
cigareta [cigarette]	17	3	17.6%
lék [medicine]	16	2	12.5%
brambor [potato]	9	1	11.1%
těstovina [pasta]	7	0	0.0%
Total	**618**	**414**	**67.0%**

Table 1. Noun lemmas with five or more plural occurrences in the PDT 2.0

a special category assigned to forms used for the meanings of the noun number. The choice between the values proposed here was checked manually in the data of PDT 2.0 by two annotators; the plural forms of nouns suspected for their use typically in the pair/group meaning with the frequency equal and higher than 5 were selected and the task of the annotators was to make choice between three possibilities: “one pair/group”, “several pairs/groups”, “undecided between preceding two groups”.

Table 1 lists noun lemmas with five or more plural occurrences in the PDT 2.0 data arranged according to the percentage of occurrences assigned the pair/group meaning out of all plural occurrences of these nouns in the final annotation.

3.3. Valency in the lexicon and in the sentence

The theoretical framework for verbal valency was elaborated within FGD in the 1970's (see Panevová, 1974–75, 1977, 1994 and others) and it was based partially on Tesnière's approach, partially on Fillmore's case grammar. The lexicographical aspects as the other obligatory part of valency description was a challenge for building valency dictionaries; the FGD theory was applied in the VALLEX dictionary (Lopatková et al., 2008). The framework for verbal valency was based on the division of verbal modifications into the class of participants (actants, arguments) and free modifications (adjuncts, circumstantials). The modifications determined by the empirical tests as participants enter the valency frame (for the tests, see the publications quoted above). For the labeling of the 1st and 2nd participants a modified Tesnière's approach is applied: if the verb has one participant, it is the Actor; if it has two participants, they are labeled as Actor and as Patient. In labeling the 3rd and other participants their semantics is taken into account. Valency frame is defined as a set of modifications classified as valency slots of the lexical item. Every modification satisfying the criteria for the participants enter the valency frame of the respective verb: they fill either an obligatory position (*vyžadovat co-ACC* [to require sth], *věřit komu-DAT* [to believe sb], *vzpomínat na koho-Prep-ACC* [to remember sb/sth] or an optional position[11] (*koupit někomu-DAT* něco [to buy *sb/sth* to somebody], *požadovat* něco *od někoho-Prep-GEN* [to ask sb *for sth*], *překvapit* někoho *něčím-INS* [to surprise sb *by sth*]).

In addition, the valency frame also contains such adjuncts that were determined by the above mentioned test as obligatory with the particular lexical item (*směřovat někam* [to lead up somewhere], *trvat jak dlouho* [to last how long], *tvářit se nějak* [to look somehow]). According to one of the important theoretical principles of this valency theory, an occurrence of the same lemma with different valency signals the ambiguity of the given lemma. This principle caused some difficulties for annotators during the annotation procedure. To overcome these difficulties the valency dictionary PDT-VALLEX (Hajič et al., 2003; Urešová, 2011b,a) was built as an on-line tool helping the annotators to check the existing valency frames and/or to add a new valency frame.

Some other additions needed to account for the complexity of the valency theory were stimulated by practical problems within the process of annotation. One of them is connected with the types of omissions of valency members on the surface without an influence on grammaticality.[12]

[11] Optional positions are denoted by italics.

[12] Here we do not have in mind an omission of a valency member conditioned by the textual deletions occurring esp. in dialogues.

An omission of a valency member has different reasons:[13]

(i) The participant is marked in the valency dictionary as optional and as such can be omitted.
(ii) The participant is obligatory, but its lexical setting is generalized.

The notion of generalization is interpreted as a group of persons/objects/circumstances typical/usual for this position. In (8a), (8b) and (9a), (9b) the differences between (a) and (b) sentences are connected with a filled valency position and a generalized valency position expressed by a null, respectively, and the verbs concerned represent one class of verbs with the deletion of an obligatory participant under special conditions. In (8b) and (9b) the generalized participants with a null form on the surface are interpreted as: *this dog does not bite anybody, Paul is able to read everything/any text*, respectively. In the tectogrammatical (deep) representation the positions of Patient (in (8b)) and Effect (in (9b)) are filled by the lemma #Gen, and in (8a) and (9a) all positions prescribed for the verbs *kousat* [bite] and *číst* [read] by their respective valency frames are used. In general, this phenomenon is known and described in linguistics as "an intransitive usage of transitive verbs", but a full description of the morphosyntactic conditions allowing for an omission of the participant is not usually taken into account. Perfective aspect of the verbs concerned[14] excludes the omission of a valency member as demonstrated by ex. (9c). The generalization of the valency member is supported by the morphological categories of gnomic present tense (often connected with the ability mood) and imperfective aspect (as in ex. (9b)).

(8a) Tenhle pes hodné lidi nekouše.
this dog-NOM-sg good people-ACC-pl not_bite-3-sg-PRS-IPFV
[This dog does not bite good people.]

(8b) Tenhle pes-NOM-sg nekouše.
this dog not_bite-3-sg-PRS-IPFV
[This dog does not bite.]

(9a) Pavel čte všechny nové romány.
Paul read-3-sg-PRS-IPFV all new novel-ACC-pl
[Paul reads all new novels.]

[13] We also leave aside here the zero subject position which is typical for Czech as a pro-drop language, because the comparison of overt and missing subjects represents a separate empirically non-trivial problem which we discuss elsewhere. For a detailed, theoretically based as well as empirically tested typology of the so called null subject languages, see Camacho (2013).

[14] In FGD and in VALLEX the aspectual pairs of verbs are understood as morphological forms of the same lexical unit.

(9b) Pavel už dobře čte.
Paul already well read-3-sg-PRS-IPFV
[Paul already reads well.]

(9c) *Pavel už dobře přečte.
Paul already well read-3-sg-PFV
[Paul already reads well.][15]

Examples of the difficulties connected with the annotation procedure representing another subclass of verbs allowing for generalization (in this case of Addressee) are given in (10). In (10b) and (10c) the noun expected as the filler of the valency position of Addressee is "generalized"; generalization of the Addressee is acceptable be this verb in the perfective as well as in the imperfective aspect. The realizations (10a), (10b), (10c) correspond to the verbal frame of the lexical entry for *prodat/prodávat* [sell-PFV/ sell-IPFV]: ACT (NOM), PAT_{Gen} (ACC), $ADDR_{Gen}$ (DAT). In the deep structure of (10b) the position of ADDR has the lemma #Gen. The lower index *Gen* assigned to the participants in the valency frame used here to demonstrate that the possibility to generalize this valency slot must be treated in the dictionary. In ex. (10c) both PAT and ADDR can be generalized (see Fig. 2), because they satisfy the conditions prescribed for a possible deletion if the verb is used in the form of gnomic present and imperfective aspect.[16]

(10a) Jan prodal auto sousedovi.
John-NOM sell-3-sg-PST-PFV car-ACC-sg neighbour-DAT-sg
[John sold his car to his neighbour.]

(10b) Jan prodává auto.
John-NOM sell-3-sg-PRS-IPFV car-ACC-sg
[John is selling his car.]

(10c) Lucie prodává v supermarketu.
Lucy-NOM sell-3-sg-PRS-IPFV in-PREP supermarket-LOC-sg
[Lucy sells in a supermarket.]

The missing Patient and Addressee in (10c) are understood as goods usually sold in the supermarkets to the usual customers of the supermarket, respectively. The generalized members are again filled into the deep syntactic representation with the lexical label #Gen.

[15] Strictly speaking, no translation can be assigned to (9c) different from that for (9b) because in English there is no equivalent of the perfective form of the Czech verb.

[16] An alternative solution would be the introduction of a new lexical unit for *prodávat* [sell] with the meaning *být prodavačem* [to be a shop assistant].

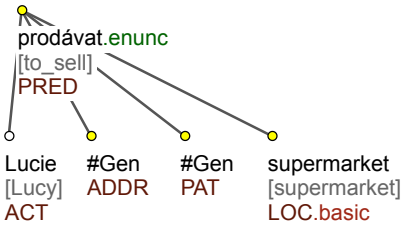

Figure 2. Sentence (10c): Lucie prodává v supermarketu.

The class of verbs with which the generalization of Patient is not limited to an imperfective form can be exemplified by (11), (12), though their valency frames contain an obligatory Patient.[17]

(11) Pokojská uklidila.
Chambermaid-NOM-sg clean-3-sg-PST-PFV
[The chambermaid has (already) cleaned.]

(12) Každé ráno ustelu a vyvětrám.
every morning make_a_bed-1-sg-PRS-PFV and ventilate-1-sg-PRS-PFV
[Every morning I make the bed and ventilate.]

Generalization is present also in the constructions with the possessive resultative, see (13) and (14), where the obligatory Patient in both sentences is generalized.

(13) Dnes máme vyprodáno.
today have-AUX-1-pl-PRS sell_out-PTCP-N-sg
[Today we are sold out.]

(14) Už mám zaplaceno.
already have-AUX-1-sg-PRS pay_for-PTCP-N-sg
[I have already paid my bill.]

The examples (8) through (14) illustrate that even though the theory of valency applied was formulated thoroughly, an extension of the theory is needed because of many empirical problems: the omissions (either connected with generalized valency positions or with other empirical issues) need an account of the restrictions on the morphological meanings available for the rules for deletion which influence the treatment of lexical entries as well as the syntactic structure.

[17] For a remark on the verb *vyvětrat* [ventilate], see (iii) below.

(iii) The omission of a verbal participant (usually the Patient) occurs also with verbs where the construction without the given valency member is close to the domain of phraseology or at least to the lexicalization of the verb, see (15), (16) and (17), where an explicit insertion of the Patient is either impossible or it does not bring novel information.

In variants (a) the Patient (*pivo* [beer], *prostor* [space] and *cigareta* [cigarette], respectively) is expressed overtly, in (b) PAT is missing. Ex. (15b) differs from the other two examples in the degree of lexicalization: the only reading of (15b) is *Janův otec je opilec* [John's father is a drunk]. In (16b) and (17b) the empty position for the valency member can be easily filled by a noun, which is semantically restricted excluding a free choice of a filler for the Patient.

(15a) Janův otec pije hodně pivo.
John's father-NOM-sg drink-3-sg-PRS very_much beer-ACC-sg
[John's father drinks beer very much.]

(15b) Janův otec hodně pije.
John's father-NOM-sg very_much drink-3-sg-PRS
[John's father drinks a lot.]

(16a) Po požáru vyvětrali všechny prostory.
After-PREP fire-LOC-sg ventilate-3-pl-PST all space-ACC-pl
[After the fire they have ventilated all spaces.]

(16b) V pokoji bude příjemněji, až
In-PREP room-LOC-sg be-3-sg-FUT pleasant-ADV-ALL after-CONJ
vyvětráš.
ventilate-2-sg-FUT-PFV
[It will be more pleasant in the room after you ventilate.]

(17a) Pohodlně se usadila a zapálila si
comfortably REFL-ACC sit_down-3-sg-PST-F and light-3-sg-PST-F REFL-DAT
cigaretu.
cigarette-ACC-sg
[She sat down comfortably and lighted a cigarette.]

(17b) Zapálila si a začala vyprávět své
light-3-sg-PST-F REFL-DAT and start-3-sg-PST-F relate-INF her
zážitky.
experience-ACC-pl
[lit. She lighted and started to relate her experiences.]

	Total	Generalized	
ACT(or)	87,118	6,910	7.9%
PAT(ient)	68,030	2,574	3.8%
ADDR(ressee)	10,150	3,640	35.9%
EFF(ect)	7,207	178	2.5%
ORIG(in)	847	30	0.4%

Table 2. Frequencies of participants and their generalization

Examples (15) through (17) point again to the necessity of cooperation between the lexical and the syntactic modules of the language description. Any type analyzed here needs a subtle treatment in the lexicon in order to offer a solid basis for sentence generation. The technical implementation of the new results reflecting the conditions for deletions of valency positions is in progress.

In Table 2, we present the frequency of particular participants depending on a verb as attested in PDT 3.0. Numbers in the first column correspond to all occurrences of the participant with a verbal head, in the second and third columns their generalized position is indicated.

3.4. Introduction of the notion of "quasi-valency"

During the extended studies of empirical data relevant for valency we have come across modifications that have properties typical for the class of participants ((i) through (iii)) as well as those typical for the class of free modifications ((iv) and (v)):

(i) they occur with a limited class of verbs
(ii) their morphological forms are given by their head
(iii) they cannot be repeated with a single verb occurrence
(iv) they have a specific semantics, contrary to the Actor, Patient and Effect (the semantics of which is usually heterogeneous)
(v) they are mostly optional

On the basis of these properties new functors were introduced: the modifiers Obstacle (OBST) and Mediator (MED) represent a more subtle division of the general modification of Means/Instrument.

(18a) Jan zakopl nohou o stůl
John-NOM stumble-3-sg-PST leg-INS-sg over-PREP table-ACC-sg
[John stumbled over the table *with his leg*.]

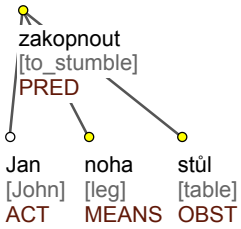

Figure 3. Sentence (18a): Jan zakopl nohou o stůl.

(18b) Matka se píchla nůžkami.
Mother-NOM-sg REFL-ACC prick-3-sg-PST scissors-INS-sg-PAIR/GROUP
[Mother pricked herself *with the scissors*.]

(18c) Šípková Růženka se píchla o trn.
Sleeping Beauty-NOM REFL-ACC prick-3-sg-PST by-PREP thorn-ACC-sg
[Sleeping Beauty pricked herself *by a thorn*.]

In (18a) *noha* [the leg] is a proper Means (Instrument), while the construction *o stůl* [over the table] is rather an Obstacle (see Fig. 3). Similar considerations concern the construction *o trn* [by a thorn] in (18c), which is also classified as an Obstacle. In (18b) *nůžky* [scissors] functions as an Instrument in the proper sense, its semantics implies the semantics of handling this instrument (which implies its movement). In (18b) a manipulation with scissors is supposed, while in (18a) and (18c) the referent of the noun stays fixed. The feature of an unconscious action is typical of (18a) and (18c), while in (18b) the action can be either conscious or unconscious.

Up to now, we have found only one Czech verb with an obligatory Obstacle (*zavadit* [to brush against]); otherwise with verbs listed in the dictionary as compatible with OBST this modification is optional.

Another semantic specification of the Instrument is expressed in Czech by the prepositional group *za + ACC*; we proposed to call it Mediator (see ex. (19)).

(19) Jan přivedl psa za obojek.
John-NOM bring-3-sg-PST dog-ACC-sg by-PREP collar-ACC-sg
[John brought the dog by its collar.]

In example (20), the ear of the boy is understood to be an object that mediates the contact between father's left hand and the boy. A supportive argument for the distinction to be made between the "classical" Instrument (*ruka* [hand]) and the Mediator

(*ucho* [ear]) is the fact that the Instrument and the Mediator *ucho* [ear] can co-occur in a single sentence.[18]

(20) Otec chytil kluka levou rukou za
Father-NOM-sg catch-3-sg-PRF boy-ACC-sg left-INS-sg hand-INS-sg by-PREP
ucho.
ear-ACC-sg
[Father has caught the boy's *ear* by his left *hand*.]

Because of the introduction of the class of quasi-valency modifiers into the formal framework of FGD the list of functors (semantic relations) originally used was checked and as the consequence of these observations a new list of valency members was provided: the modifiers of Intention (INTT) and Difference (DIFF) were shifted from the list of free modifiers into the list of quasi-valency members. For the modifier of Intention, see (21) and for the modifier of Difference, see (22):[19]

(21) Jan jel navštívit svou tetu.
John-NOM went-3-sg-PST visit-INF his-POSS aunt-ACC-sg
[lit. John left *to visit* his aunt.]

(22) Náš tým zvítězil o dvě branky.
our-POSS team-NOM-sg win-3-sg-PST by-PREP two-ACC goal-ACC-pl
[Our team won *by two goals*.]

3.5. Selected types of deletions

According to the annotation scenario for the surface layer of annotation in PDT only elements present on the surface are represented by separate nodes in the dependency tree. However, there are elements obviously missing for the complete meaning of the sentence. The following technical solution for such cases was proposed for the surface (analytical) layer of annotation: if the governor of some member of the sentence is not present, the syntactic function of this member receives the value ExD (with meaning "extra-dependency"). The nodes with this value are an excellent challenge for the studies of deletions (ellipsis) which must be reconstructed in the deep (tectogrammatical) structure.

In this section we present only selected types of grammatical deletions conditioned or even required by the grammatical system of language.[20] One special type of dele-

[18] See also Fillmore (1977), quoted from Fillmore (2003, p. 189): "A reason for feeling sure that two roles are distinct is that the same two nouns, preserving their case roles, can also occur together … in a single sentence."

[19] A detailed analysis and argumentation for these modifiers is given in Panevová et al. (2014).

[20] For a detailed discussion on the reconstruction of deletions, see Hajič et al. (2015) and Hajičová et al. (2015).

tions, namely the surface deletion of valency members, was analyzed in more details above in Sect. 3.3. Here we want to comment upon some complicated cases of deletions.

Comparison structures are a very well known problem for any language description aiming at a representation of the semantic (deep/underlying) structure. These considerations concern the comparison with the meaning of equivalence (introduced usually by the expression *jako* [as]; the subfunctor we use has the label 'basic') and the comparison with the meaning of difference (introduced usually by the conjunction *než* [than]; the subfunctor is called 'than').[21]

There are some comparison structures where the restoration of elements missing on the surface seems to be easy enough from the point of view of semantics (see (23a) and its restored version (23b), but most comparisons are more complicated, see (24) through (26):

(23a) Jan čte stejné knihy jako jeho
John-NOM read-3-sg-PRS same-ACC-pl book-ACC-pl as-CONJ his-POSS
kamarád.
friend-NOM-sg
[John reads the same books as his friend.]

(23b) Jan čte stejné knihy jako (čte)
John-NOM read-3-sg-PRS same-ACC-pl book-ACC-pl as-CONJ (read-3-sg-PRS)
(knihy) jeho kamarád.
(book-ACC-pl) his-POSS friend-NOM-sg
[John reads the same books as his friend (*reads books*).]

The introduction of the deleted elements into (24a) seems to be as easy as in (23b), however, for the expansion of "small clauses" expressing comparison illustrated by ex. (24b) such a solution is not sufficient: (24b) is not synonymous with (24a). More complicated expansion for (24b) is proposed and exemplified by (24c) as its deep structure counterpart.

(24a) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako jeho rodiče.
as-CONJ his-POSS parents-NOM-sg
[John lives in the country as comfortably as his parents.]

[21] More simple comparative structures expressed by secondary prepositions with nouns (such as *na rozdíl od* [in contrast to], *ve srovnání s* [in comparison with], *proti* [against], e.g. in *Ve srovnání s minulým rokem je letos úroda brambor vyšší* [Lit. In comparison with the last year the crop of potatoes is in this year higher] are left aside here.

(24b) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako u svých rodičů.
as-CONJ with-PREP his-POSS parents-GEN-sg
[John lives in the village comfortably as well as with his parents.]

(24c) Jan žije na vesnici stejně pohodlně
John-NOM live-3-sg-PRS in-PREP village-LOC-sg same-ADV comfortably-ADV
jako (Jan) (žít) (nějak) u svých
as-PREP (John-NOM) (live-3-sg-PRS) (some way-ADV) with/PREP his-POSS
rodičů.
parents-GEN-sg
[John lives in the village comfortably as well as he lives (somehow) with his parents.]

The compared members in (24a) and (24b) are not apparently of the same sort: the two modifications are collapsed in a single "small clause". This phenomenon contradicts the notation used in dependency based representations in FGD: the two functions (comparison and location) could not be assigned to the single node introduced by the comparison construction.

Though some extensions of the embedded predication (e.g. (24c), (26b)) do not sound natural, they represent only a theoretical construct required by the shortened surface shape (for details, see Panevová and Mikulová 2012). In the deep structure of (24c), the inserted node *žít* [to live] is labeled as CPR (comparison) and the node *rodiče* [parents] bears the functor LOC [location] depending on the restored node governing the comparison (*žít* [to live] in this case). While in (24a) the way of John's life in the country is compared with the identical way of his parents' life there, in (24b) John's life in the country is compared with the way of his (respective) life with his parents. John's way of life is presented as comfortable in the main clause, so his life with his parents may be assumed to be comfortable as well, however this assumption is not expressed explicitly. Therefore the adverbial specifying the way of life in the reconstructed representation is denoted by the underspecified artificial node *nějak* [in some way] rather than by a repetition of the lexical value *pohodlně* [comfortably]. In the tectogrammatical (deep) structure of (24c) the inserted node *žít/žije* [live] is labeled as comparison (CPR) and depends on the lexically identical predicate of the main clause, while *u rodičů* [with the parents] and *nějak* [in some way] are its children labeled as location (LOC) and manner (MANN), respectively.[22]

Examples (25) and (26) support the arguments presented for (24): (i) expansion of the surface shape of comparison structure is necessary, and (ii) fuzzy artificial lemmas

[22] For *nějak* [in some way] the artificial lemma #Some is used in PDT.

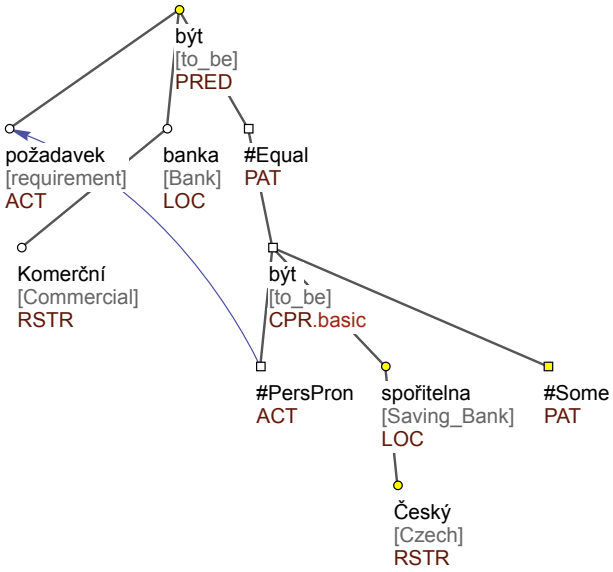

Figure 4. Sentence (25b): Požadavky u Komerční banky jsou jako u České spořitelny.

are introduced because of their lexical underspecification. Two types of comparison (identity in (25) and diference in (26)) are exemplified by the (25) and (26) as well.

(25a) Požadavky u Komerční banky jsou jako
requirement-NOM-pl in-PREP Commercial Bank-GEN-sg be-3-pl-PRS as-CONJ
u České spořitelny.
in-PREP Czech Saving Bank-GEN-sg
[lit. The requirements in Commercial Bank are as in Czech Saving Bank.]

(25b) Požadavky u Komerční banky jsou (stejné)
requirement-NOM-pl in-PREP Commercial Bank-GEN-sg be-3-pl-PRS (same)
jako (jsou požadavky) u
as-CONJ (be-3-pl-PRS requirement-NOM-pl) in-PREP
České spořitelny (nějaké-#Some)
Czech Saving Bank-GEN-sg (some-#Some)
[lit. The requirements in Commercial Bank are (the same) as (are the requirements) in Czech Saving Bank.]

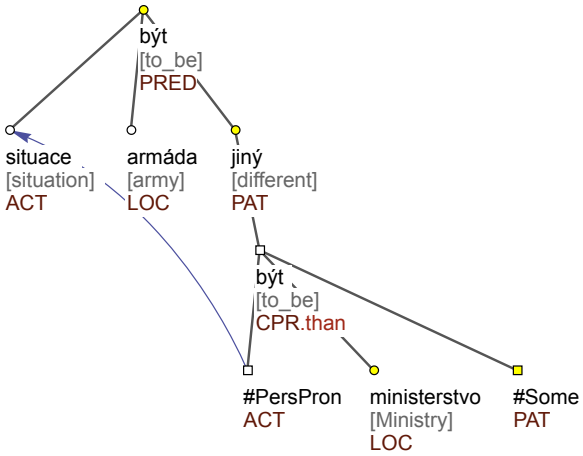

Figure 5. Sentence (26b): Situace v armádě je jiná než na ministerstvu.

(26a) Situace v armádě je jiná
situation-NOM-sg-F in-PREP army-LOC-sg be-3-sg-PRS different-NOM-sg-F
než na ministerstvu.
than-CONJ at-PREP Ministry-LOC-sg
[lit. The situation in the army is different than at the Ministry.]

(26b) Situace v armádě je jiná
situation-NOM-sg-F in-PREP army-LOC-sg be-3-sg-PRS different-NOM-sg-F
než (je situace) na ministerstvu
than-CONJ (be-3-sg-PRS situation-NOM-sg-F) at-PREP Ministry-LOC-sg
(nějaká-#Some)
(some-#Some)
[lit. The situation in the army is different than (the situation) at the Ministry is (some).]

Also the analysis of other types of adverbials points to the possibility to restore in the deep structure representation a whole embedded predication (e. g. adverbial phrases introduced by the expressions *kromě* [except for, besides], *místo* [instead of]). In the surface structure there are again two types of modifications, see (27a), where the adverbial of direction is embedded into the adverbial of substitution sharing the same predicate on the surface. As we have mentioned above when analyzing complex comparisons, the FGD framework does not allow for an assignment of more than a single

function to a single sentence member. The extension of this predication on the underlying level is illustrated by (27b):

(27a) Místo do Prahy přijel Jan do
instead-of-PREP at-PREP Prague-GEN arrive-3-sg-PST John-NOM at-PREP
Vídně.
Vienna-GEN
[Instead of arriving at Prague, John arrived at Vienna.]

(27b) Místo toho, aby přijel do
instead-of-PREP that-PRON AUX-3-sg-COND arrive-3-sg-PST at-PREP
Prahy, přijel Jan do Vídně.
Prague-GEN arrive-3-sg-PST John-NOM at-PREP Vienna-GEN
[Instead of arriving at Prague, John arrived at Vienna.]

From the point of view of their surface form, these deletions are not as transparent as e.g. dropped subject or comparison constructions, but the difficulties the annotators had during the annotation procedure stimulated a more detailed analysis sketched briefly above and presented in detail in Panevová et al. (2014).

3.6. Non-nominative subjects

The non-nominative subjects are the topic of many typologically oriented studies (see e. g. Bhaskararao and Subbarao, 2004). The fact that in some languages the subject is expressed by the dative, genitive and other forms is well known. Such marginal forms are present in Czech as well, where prototypical subjects have the form of nominative. The traditional term "dative subject" is applicable for the Czech examples as (28) and (29), in the deep structure of which the dative is understood as the Actor; a similar structure is assigned to the sentences where nominative is present, but it is not understood as an Actor, see (30), due to the semantic parallel structure with different formal exponents (see (31)).

This solution corresponds the theory of valency used in FGD: Any verb has in its valency frame in the lexicon a slot for the Actor (1st actant according to Tesnière, 1959). Actor is prototypicaly expressed by Nominative, however there are two types of exceptions: either the verb has an unprototypical patterning of its valency complementations (see ex. (28) – (31)), or the participant of Actor is stylistically or semantically modified (see ex. (32) – (35); semantic modifications of Actor are represented by the subfunctors of the Actor.

(28) Je mu smutno.
be-3-sg-PRS he-DAT-M-sg-Sbj sad-ADV
[He is sad.]

(29) V Praze se rodičům líbí.
in Prague REFL-ACC parents-DAT-Sbj like-3-sg-PRS-ACT
[My parents like Prague.]

(30) Bolí mě hlava.
ache-3-sg-PRS-ACT I-ACC head-NOM-Sbj
[I have a headache.]

(31) Bolí mě v krku.
ache-3-sg-PRS-ACT I-ACC-Sbj in-PREP throat-LOC-sg
[I have a sore throat.]

The genitive subject occurs in Czech sentences as a stylistic variant of the nominative, see (32) and (33), where the negative genitive and partitive genitive, respectively, are used. The genitive forms, however, carry some additional semantic information with respect to the unmarked nominative form, but in contemporary Czech they are accepted as a little bit archaic, therefore they are rare in the PDT. The semantic contribution to the unmarked nominative forms is expressed by the introduction of "subfunctors" (rendering semantic variations of the nominative subject/Actor); in addition new semantic shades of the construction (formally rendered by the value of subfunctors) are expressed by some prepositions, see (34) displayed in Fig. 6 and (35).

(32) Z vyhlazovacích táborů nebylo úniku.
from-PREP extermination camp-GEN-pl not_be-3-sg-PST-N escape-GEN-sg-Sbj
[From the extermination camps there was no escape.]

(33) Přibývá podnikatelů, kteří nemají kancelář a
increase enterpreneur-GEN-pl-Sbj who not_have-3-pl-PRS office-ACC-sg and
podnikají doma.
do_business-3-pl-PRS home
[The number of entrepreneurs who have no office and who do business from their homes increases.]

(34) Své expozice bude mít okolo 60
their exposition-ACC-pl be-3-sg-FUT have-INF approximately-PREP 60
stavebních firem.
building firm-GEN-pl-Sbj
[Approximately 60 building firms will have their own expositions.]

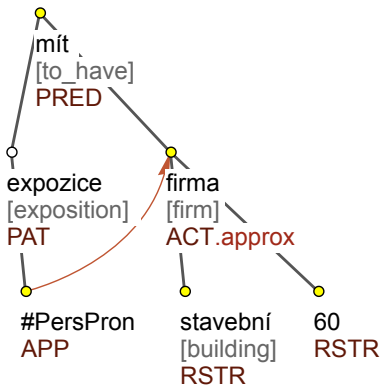

Figure 6. Sentence (34): Své expozice bude mít okolo 60 stavebních firem.

(35) Za každou stranu přišlo po pěti
for-PREP every party-ACC-sg come-3-sg-PST-N by-PREP five-LOC
delegátech.
deputy-LOC-pl-Sbj
[Every party was represented by five deputies.]

An approximate amount of firms is expressed in (34) by the preposition *okolo/kolem, na* + accusative [around], the distributive meaning is expressed in (35) by the prepositional case *po* + Locative [by]. The subfunctors *approximity, distributivity* corresponding to these meanings were introduced into the list of subclassified meanings of the main syntactic functors (Actor in this case).

In Sections 3.1 to 3.6 we presented examples of grammatical phenomena which either were not yet described explicitly or were not described at all. Some of these issues are known, but their consequences for a consistent description have not yet been fully considered. Some of these results are reflected in the annotation guidelines and all of them enriched the theoretical description of Czech grammar.

4. Case studies II: Information structure of the sentence, discourse relations and coreference

4.1. Topic–focus annotation in the Czech corpus

In the theoretical account of topic–focus articulation (TFA in the sequel, see e.g. Sgall, 1967; Sgall et al., 1973, 1980; Hajičová et al., 1998) within the framework of the Functional Generative Description, the dichotomy of topic and focus – which divides the sentence into what the sentence is about (its topic) and what it says about the topic

(its focus) – is understood as based on the primary notion of contextual boundness. The TFA information/feature is an integral part of the representation of sentences on the underlying (tectogrammatical) sentence structure, since TFA is semantically relevant.[23] Every node of the tectogrammatical dependency tree carries – in addition to other characteristics such as the tectogrammatical lemma, the type of dependency or (morphological) grammatemes – an index of contextual boundness: a node can be either contextually bound or non-bound. This feature, however, does not necessarily mean that the entity is known from a previous context or new but rather how the sentence is structured by the speaker as for the information structure. Thus, for example, in the second sentence of the discourse segment *When we walked around the town, we met Paul and his wife. I immediately recognized HIM, but not HER* (capitals denoting the intonation center, if the sentences are pronounced), both Paul and his wife are mentioned in the previous context, but the sentence is structured as if they are a piece of non-identifiable information, i.e, marked as contextually non-bound. Contrary to that, the above segment from the information structure point of view can also be structured in a different way, which, in the surface form of the sentence in English, would involve a different placement of the intonation center: *When we walked around the town, we met Paul and his wife. I immediately RECOGNIZED him*. In this segment, both Paul and his wife are also introduced in the first sentence, but in the second sentence, it is the event of recognizing which is structured as bringing 'new' (non-identifiable) information, while Paul – being referred to by a non-stressed pronoun – is taken as contextually bound (identifiable). In Czech, the two situations would be expressed by a different word order and different forms of the pronoun corresponding to English *him*, namely *jeho* vs. *ho*: *Hned jsem poznal JEHO* versus *Hned jsem ho POZNAL.*

The annotation of PDT follows the theoretical description rather closely: to each node of the dependency tree on the tectogrammatical layer of PDT a special attribute of TFA is assigned which may obtain one of the three values: *t* for a non-contrastive contextually bound node, *c* for a contrastive contextually bound node and *f* for a contextually non-bound node. [24]

The left-to-right dimension of a tectogrammatical tree serves as the basis for the specification of the scale of communicative dynamism: communicative dynamism is specified as the deep word order, with the dynamically lowest element standing in the

[23] The semantic relevance of TFA has been documented in the writings quoted above and elsewhere by such sentences differing only in their TFA structure as *I work on my dissertation on SUNDAYS.* vs. *On Sundays, I work on my DISSERTATION.*, or *English is spoken in the SHETLANDS.* vs. *In the Shetlands ENGLISH is spoken.*, or *Dogs must be CARRIED.* vs. *DOGS must be carried.*, etc. The difference in TFA is expressed, in the surface structure, either by word order (as in Czech), or by a different position of the intonation centre denoted here by capitals (this holds both for Czech and for English) or even by some specific sentence structure (e.g., cleft constructions in English).

[24] There are 206,537 tectogrammatical nodes annotated as contextually bound, out of them 30,312 are contrastively contextually bound. Further, 354,841 nodes are contextually non-bound and for 38,493 nodes (and for 2,849 technical roots), contextual boundness is not annotated (e.g., for coordinating nodes).

leftmost position and the most dynamic element (the focus proper of the sentence) as the rightmost element of the dependency tree.

4.1.1. The identification of the boundary between topic and focus

For the identification of the dichotomy of topic and focus (which is supposed to be very important especially for the specification of the scope of negation) on the basis of the information on contextual boundness for each node, a rather strong hypothesis was formulated, namely that the topic-focus distinction can be made depending on the status of the main verb (i.e. the root) of the sentence and its immediate dependents: Basically, 1. if the verb is contextually bound (*t*, *c*) then the verb and all the contextually bound nodes depending immediately on the verb and all nodes subordinated to these nodes constitute the topic, the rest of the sentence belonging to its focus; 2. if the verb is contextually non-bound (*f*), then the verb and all the non-bound nodes immediately depending on it and all nodes subordinated to these nodes constitute the focus, the rest of the sentence belonging to its topic; 3. if both the main verb and all nodes immediately depending on the main verb are contextually bound, then follow the rightmost edge leading from the main verb to the first node(s) on this path that are contextually non-bound; this/these node(s) and all the nodes subordinated to it/them belong to focus (see the definition of topic and focus by Sgall, 1979, see also Sgall et al., 1986, 216f).

To test this hypothesis on the PDT data, we have proceeded in three steps:

(i) a minor modification and implementation of the algorithm so that it can be applied to the data of the whole PDT,
(ii) manual parallel annotation of the control raw data as for the topic and focus of the individual sentences,
(iii) comparison of the values obtainded from the manual annotation with the automatically assigned Topic–Focus bipartition and evaluation of the results.

The results of the implementation of the modified algorithm indicate that a clear division of the sentence into topic and focus according to the hypothesized rules has been achieved in 94.28% of sentences to which the procedure has been applied; 4.41% of sentences contained the so-called proxy focus (itself a part of topic but a part that has the focus subordinated to it).[25] The real problem of the algorithm then rests with

[25] More exactly, proxy focus is a node A such that A is contextually bound, A differs from the main verb and the focus of the sentence is subordinated to A. The introduction of the notion of proxy focus was invoked to handle cases where the focus of the sentence is so deeply embedded that it does not include the verb or any of its immediate dependents (see Hajičová et al., 1998). Thus in *I met the teacher of CHEMISTRY* as an answer to *Which teacher did you meet yesterday?* the focus *chemistry* depends on a head (*teacher*) that has a specific status, it is a proxy focus: it is contextually bound and thus does not belong to the focus; however, it is the only part of the upper subtree of the sentence that lies on the path from the root of the tree (the verb) to the focus.

the case of ambiguous partition (1.14%) and cases where no focus was recognized (0.11%) as the assumption of the TFA theory is that all sentences should contain focus (though there may be topicless sentences, e.g., those that bring hot news: *KENNEDY was assassinated*!) but this is a very small part of the data analyzed.

However, in order to validate the hypothesis it is necessary to compare the results achieved by the automatic identification of topic and focus with the judgements of Czech speakers (step (ii) above). For the control annotation, PDT documents comprising a total of 11,000 sentences have been analyzed manually, most of them in three parallel annotations (about 10,000 sentences), and about 600 sentences in six parallel annotations (a detailed description of the project is given in Zikánová et al., 2007; we present here a brief summary of the methodology used and the results). The annotators were mostly high school students, having some (common sense) basic idea of the dichotomy of topic and focus (as "the aboutness relation") but were not familiar with the theoretical framework TFA is based on. They worked with the raw texts (i.e. without any annotation) and were instructed to mark – according to their understanding – every single word in the sentence as belonging either to topic or to focus; they were supposed to take nominal groups as an integrated element and they were also told that they may assign all the elements of the sentences to the focus. At the same time, they were supposed to mark which part of the sentence they understand as topic and which part as focus. In subordinated clauses and in coordinated constructions they were asked to mark each clause separately. One of the important subtasks of this project was to follow annotators' agreement/disagreement. The disagreement in the assignments of the two parts of the sentence as a whole was rather high and indicates that the intuitions concerning the division of the sentence into its topic and focus parts may dramatically differ. However, it is interesting to note that the annotators' agreement in the assignments of individual words in the sentences to topic or to focus was much higher (about 75% in both the three and six parallel analyses compared to 36% of the assignments of the topic and the focus as a whole) than the assignments of the topic–focus boundary.

The work on the step (iii) is still in progress. It is a matter of course that in that step, the variability of manual solutions must be taken into considerations; the annotators were asked to assign a single, most plausible TFA annotation, different annotators for the same text may have chosen a different interpretation. We are aware of the fact that while we get only a single, unambiguous result from the automatic procedure, more ways of interpretation could be possible. This mostly occurs with the assignment of the verb: actually, it is the assignment of the verb to topic or to focus, in which the annotators differed most frequently.[26]

[26] See the discussion in K. Rysová et al. (2015a). It should be added that no machine learning methods for TFA assignment have been considered so far.

4.1.2. Systemic ordering as the order of elements in the focus

The empirical study of Czech texts has led to the assumption (Sgall et al., 1980, p. 69) that the ordering of the elements in the focus part of the sentence is primarily given by the type of the complementation of the verb. This assumption resulted in a rather strong hypothesis called systemic ordering of the elements in the focus of the sentence. The hypothesis was empirically tested pairwise (i.e., successively for two of the complementation types) and it was also supported by several psycholinguistic experiments (Sgall et al., 1980, p. 72ff; Preinhaelterová, 1997). The following ordering has been established for Czech:

Actor – Temporal (when – since when – till when – how long) – Location (where) – Manner – Extent – Measure – Means – Addressee – From where – Patient – To where – Effect – Condition – Aim – Cause

Even at the time of the formulation of the hypothesis, several accompanying assumptions were taken into account:

(i) It was assumed that systemic ordering is a universal phenomenon and that at least in most European languages the order of the principle verb complementations (such as Actor – Addressee – Patient) is the same, which was also attested by experiments for English and German; at the same time it was clear that languages may differ in the (underlying) order of the particular elements.
(ii) It was understood that there are several factors that may influence the underlying order in focus such as the rhythmical factor (short complementation before the longer one), or the lexical meaning of some verbs which may be associated more closely with a certain type of complementation (e.g., the verb *pay* in construction with Patient: *pay the debts*); such a construction may have a character of a phraseological expression (*to pave the way*, *to make claims*, etc.).
(iii) In the original formulation no difference was made between sentential and non-sentential structures expressing the given complementation. This difference certainly influences the ordering and has to be taken into account.
(iv) The question has remained open as for the character of the ordering: does each complementation have a separate position in the scale or is it the case that more than a single type of complementation occupy a given position on this scale?
(v) It was clear from the very beginning that the hypothesis of systemic ordering is very strong and that in spite of the fact that it was based on the examination of hundreds of examples, further investigation based on a much broader material is needed, which may lead to a more precise specification or modification(s), as is the case with all empirical statements.

The material of the Prague Dependency Treebank opened the possibility to validate the hypothesis. After the first attempts made by Zikánová (2006), a deeper and a more complex analysis is presented by K. Rysová (2014a), who arrives at several interesting

and important observations summarized in the sequel. 1. First of all, she confirms that the sentential character of a complementation is a very important factor in that there is a tendency of a contextually non-bound element expressed by a clause to follow the non-sentential element (which is apparently connected with the 'weight' of the element mentioned above in point (iii)). 2. She also points out the influence of the form of the complementation: the assumed order Manner – Patient is more frequent if the complementation of Manner is expressed by an adverb and the complementation of Patient by a nominal group.[27] 3. When examining the position of the Actor on the scale, a substantial number of counterexamples of the original hypothesis (with the position of Actor at the beginning of the scale) concern cases for which the outer form of the Actor plays an important role: in sentences with the verb *být* (to be) in structures of the type *je nutné* (PAT) *přiznat* (ACT) (*it is necessary to acknowledge*), where Actor is expressed by infinitive, Patient precedes Actor, while the hypothesized order Actor – Patient is attested to if both complementations are expressed by nominal groups.

Rysová's analysis (using the PDT material with the manual annotation) is based on examples where there are two complementations in the focus of the sentence; her analysis confirms that there is a considerable tendency that in such pairs one ordering prevails over the other, which, as a matter of fact, was the starting point of the postulation of the systemic ordering hypothesis. However, with some pairs, such as Patient and Means, there was a balance between the frequency of the two possible orders, which may indicate that for some particular complementations more than a single complementation occupy one position on the scale (see point (iv) above). She also mentions the possibility that the order might be influenced by the valency characteristics of the verbs, namely by the difference in the optional/obligatory character of the given complementations: she assumes that there is a tendency that obligatory complementations seem to follow the optional ones, but she admits that this tendency is not a very influential word order factor.

Rysová observes that in some cases the decisions of the annotators are not the only possible ones and that this fact has to be taken into consideration when drawing conclusions. This observation is confirmed also by the data on the annotators' agreement/disagreement, see also Veselá et al. (2004) or Zikánová (2008) and below in Section 5.

4.1.3. Rhematizers (focusing particles, focalizers)

A specific function of certain particles from the point of view of a bipartitioning of the sentence was noted first by Firbas (1957) in connection with his observation of a specific rhematizing function of the adverb *even.* It should also be mentioned at this point

[27] As one example for all, let us mention a combination of a node with the functor MANN and a node with functor PAT, both contextually non-bound and directly depending on a node with a verbal semantic part of speech. There are 1,111 such cases, in 933 out of them, MANN precedes PAT in the surface order (in agreement with the systemic ordering), in 174 cases MANN follows PAT

that a semantic impact of the position of several kinds of adverbials and quantifiers was substantiated already by Sgall (1967), who exemplifies the semantic relevance of topic/focus articulation on the English quantifier *mostly*. Sgall's argumentation is followed by Koktová (1999, but also in her previous papers), who distinguishes a specific class of adverbials called attitudinal.

The same class of words was studied later in the context of formal semantics by Rooth (1985) in relation to the prosodic prominence of the words that followed them; he called this class 'focalizers'.

Both terms – rhematizer and focalizer – refer to the apparent function of these particles, namely as being 'associated' with the focus of the sentence; the position of the focalizer (and the accompanying placement of the intonation center) indicates which reading of the sentence is being chosen from the set of alternatives. However, the assumption of such an exclusive function of these particles has been found to be too simplistic, an analogy with a semantic analysis of negation was claimed to be a more adequate approach (Hajičová, 1995). A distinction has been made between 'the (global) focus' of the sentence and 'the focus' of the focalizer (specified as the part of the sentence that follows the focalizer) by Hajičová et al. (1998). Comparing the analysis of the semantic scope of negation and the analysis of the function of focalizers, it is necessary to also consider the possibility of a secondary interpretation of the position of the focalizers. This issue was demonstrated in examples such as *JOHN criticized even Mother Teresa as a tool of the capitalists*. This sentence may occur in a context illustrated by the question *Who criticized even MOTHER TERESA as a tool of the capitalists?* The predicate of the indicative sentence *criticized even Mother Teresa as a tool of the capitalists* is repeated from the question; the only part of this sentence that stands in the focus is *John* (with a paraphrase 'the person who criticized even Mother Teresa as a tool of capitalists was John'). Such an understanding would compare well with the sometimes indicated recursivity of topic/focus articulation.

Based on the observations on the scope of focalizers as reflected in PDT and a similarly based annotation of English in the so-called Prague English Dependency Treebank (see Cinková et al., 2009), some complicated (and intricate) cases have been singled out, concerning first of all the occurrence of focalizers with a restricted freedom of position, with a distant placement of focalizers and their possible postpositions, and the semantic scope of focalizers. The function and the diversity of expressions originally called rhematizers has been studied in detail by Štěpánková (2013).

It is interesting to notice that contrary to the general characteristics of Czech as a language with a relatively "free" word order (i.e. without grammatical word-order restrictions), in the placement of the focalizer *only* English is more flexible than Czech is: this particle can be placed either immediately before the element it is 'associated with' or between the subject and the verb in English.

In Czech, a backward scope of focalizers is not that frequent as in English, but it is also possible. For example, the intonation center in the sentence quoted here from the Prague Czech–English Dependency Treebank as (36), if pronounced, would

be placed on the word *inflation* (as indicated here by capitalization); the postposited focalizer *only* having its scope to the left. In the Czech translation of this sentence, the focalizer *jen* (only) has to be placed in front of the focused element. It is interesting to note that there was a single example of a backward scope of a rhematizer in the whole of the PDT.

(36) Scénář 1, známý jako „konstantní zmrazení dolaru", nahrazuje Pentagonu výdaje jen kvůli INFLACI.
[Scenario 1, known as the "Constant Dollar Freeze", reimburses the Pentagon for INFLATION only.]

Štěpánková's comprehensive and detailed analysis (Štěpánková, 2013) based on the PDT material demonstrates that the class of focalizers is larger than originally (and usually) assumed; properties similar to those of 'prototypical' focalizers *only*, *even*, *also* are evident also with *alone*, *as well*, *at least*, *especially*, *either*, *exactly*, *in addition*, *in particular*, *just*, *merely*, *let alone*, *likewise*, *so much as*, *solely*, *still/much less*, *purely*, and several others (prototypical Czech rhematizers are *pouze*, *jen*, *jenom*, *zejména*, *zvláště*, *především*, *obzvlášť*, *hlavně*, *jedině*, *například*, *toliko*, *ne*, *ano*, *výhradně*, *výlučně*). Even more importantly, her material provides evidence that according to the context in which they are used, these elements are ambiguous and may obtain functions other than that of a focalizer. Table 3 quoted from Štěpánková's dissertation (Štěpánková, 2013) based on the Czech data from PDT illustrates the ambiguity of a rhematizer obtaining also a function that is classified as a free modification (adverbial modifier).

Expressions that function in some contexts as rhematizers may also obtain – in other contexts – an attitudinal function, especially in cases when the given expression relates to the whole sentence irrespective of the position in which it occurs in the surface shape of the sentence, see the difference between (37) and (38). In (37), the expression *třeba* functions as an adverbial of attitude (ATT) (translated to E. *maybe*), in (38) the same expression obtains the function of a rhematizer (translated to E. *for instance*).

(37) Třeba.ATT Honza se tam bude nudit.
[Maybe Honza will feel bored.]

(38) Třeba.RHEM HONZA se tam bude nudit.
[For instance HONZA will feel bored.]

Examples of such ambiguous expressions in Czech are *to*, *leda*, *též*, *rovněž*, *také*, *taktéž*, *zároveň*, *prakticky*, *spíše*, *třeba* (in English a similar homonymy concerns expressions such as *only*, *at best*, *also*, *at the same time*, *practically*, *rather*, *maybe*, …).

One specific issue connected with the analysis of constructions with rhematizers is the scope of rhematizers. Since the scope is relevant for the meaning of the sentence, it must be possible to derive it on the basis of tectogrammatical representations. One possibility is to represent the scope of rhematizers on the basis of the indication of the

Expression	Used in the function of a rhematizer	Function of an adverbial	Used in the function of an adverbial
nejvýše nanejvýš	*I would have given him at most a home prison.*	EXT – specification of a numeral	*It cost at most one hundred crowns.*
už již	*Already KOMENSKÝ spoke about it.*	TWHEN – meaning "now"	*The time has already come to go to bed.*
zrovna právě teprve	*Exactly THIS I have told him.*	TWHEN – meaning "now", or EXT – "exactly"	*He has just left the car. Invite just one hundred people.*
až	*It looked too bad.*	EXT – meaning "up to", "almost"	*The meeting will be attended by up to 100 people.*
zase	*I am bad and Jim for his part well.*	TWHEN	*I will come again.*
přímo	*He was quite amazing.*	DIR2 – meaning "directly" MANN	*The road went directly to the village. Tell me downright.*
zvlášť	*Take care especially of the kids.*	MANN	*We will pay separately.*
hned	*He took right away three apples.*	TWHEN	*I will come back immediately.*
naopak	*George on the contrary ran away.*	MANN – meaning: in an opposite way, contrary to	*He did everything contrary to what they TOLD him.*

Table 3. Ambiguity of Czech rhematizers obtaining also a function of a free modification

topic–focus articulation, namely on the contextual boundness of individual nodes of the tree and the boundary between topic and focus. The rhematizer that signals the focus of the sentence has in its scope all the contextually non-bound items that follow it in the surface shape of the sentence; the scope of the rhematizer signaling the contrastive topic is basically the first element with the value of contrastive contextually bound element (together with its dependents) that follow it. If the rhematizer is the

only contextually non-bound element of the sentence, it is assumed to have a backward scope. However, these basic assumptions have not yet been put under a detailed scrutiny and wait for their validation on the PDT material.

To sum up, the analysis based on the PDT material has confirmed that there is a special class of particles that have a specific position in the TFA of the sentence and that these particles have some common features with negation. It has also been demonstrated that these particles called in linguistic literature rhematizers, focalizers or focussing particles need not be restricted to a position indicating the focus (rheme) of the sentence, rather, they can also occur in the topic of the sentence; also, there can be more than a single rhematizer in the sentence. In the theoretical description, these observations lead to the conclusion that it is necessary to distinguish between the focus of the whole sentence and the focus of a focalizer. Finally, we have observed that the scope of a focalizer has important consequences for the semantic interpretation of the sentence.

4.1.4. Contrastive study of TFA based on a parallel corpus

The existence of parallel corpora equipped with basically the same scheme of annotation offers an invaluable material for contrastive linguistic studies and thus for a re-evaluation of existing hypotheses. Let us quote here one of the numerous examples based on the comparison of a particular phenomenon in Czech and English.

A similarly based annotation as in PDT, though not covering all the features captured by the Czech corpus, exists for English in the so-called Prague Czech–English Dependency Treebank 2.0 (PCEDT; Hajič et al., 2011, see also K. Rysová et al., 2015b)[28] comprising an annotation of Czech and English parallel texts (almost 50 thousand sentences for each part) along the lines of PDT. This material has allowed for a more detailed contrastive analysis of tectogrammatical (underlying syntactic) sentence structures also concerning the topic–focus structure of Czech and English sentences. As an example, we present here the results of a case study concerning the use of the indefinite article with the subject of an English sentence.

Basically, in both languages a common strategy in communication is to proceed from retrievable, identifiable information to an unretrievable one. This strategy can be documented for Czech by the fact that in PDT, there is only a small portion of cases in which a contextually bound item in the topic of the sentence does not provide a coreferential link (i.e., it does not serve as an anaphor; it should be noted that the coreference annotation in PDT captures so far only relations of a nominal group to an antecedent, see below). As for English, a good indicator of such a rare situation is the appearance of the indefinite article in the subject position of sentences, if one assumes the unmarked position of the intonation center at the end of the sentence. Such cases

[28] http://ufal.mff.cuni.cz/pcedt2.0/en/index.html

are rather rare and can be explained by an interaction of other factors as documented on the material from the Czech–English corpus in Hajičová et al. (2011).

We started from the hypothesis that one of the possibilities how to "topicalize" Patient (Object) in English is passivization. In PCEDT, with the total number of 49,208 sentences (and, for comparison, with the total number of 54,304 predicates – roughly: clauses) there were 194 cases of an occurrence of a nominal group with an indefinite article in the function of a subject of a passive construction. These cases were compared with their Czech counterparts and can be classified into four groups as follows:

(a) Most frequent constructions contain a General Actor, not expressed in the surface (see Sect. 3.3 above)

These sentences are translated into Czech with the subject (expressing the Patient) at the end of the sentence (in Focus!); in English, the postposition of the subject into the final position is not possible due to the grammatically fixed English word-order, see (39) with the assumed position of intonation centre denoted by capitals:

(39) (Preceding context: Soviet companies would face fewer obstacles for exports and could even invest their hard currency abroad. Foreigners would receive greater incentives to invest in the U.S.S.R.) Alongside the current non-convertible ruble, a second CURRENCY would be introduced that could be freely exchanged for dollars and other Western currencies.
[Czech equivalent: Zároveň se současným nekonvertibilním rublem bude zavedena druhá MĚNA, která by mohla být volně směnitelná za dolary a další západní měny.]

(b) The indefinite article is used with the meaning "one of the", see (40):

(40) A seat on the Chicago Board of Trade was sold for $ 390,000, unchanged from the previous sale Oct. 13.
[Czech equivalent: Členství (meaning: membership, e.g., the status of a member) v Chicagské obchodní radě bylo prodáno za 390 000 dolarů, což je nezměněná cena od posledního prodeje 13. října.]

(c) Interesting though few cases involve a contrast in the topic part, see (41), with the assumed intonation center (in focus) on the year 1984 and a contrastive accent (in topic) on *faster*:

(41) (Preceding context: The "Designated Order Turnaround" System was launched by the New York Stock Exchange in March 1976, to offer automatic, high-speed order processing.) A faster version, the SuperDot, was launched in 1984.
[Czech translation (in the indicated context, with the same intonation contour): Rychlejší verze SuperDot byla spuštěna v roce 1984.]

4.2. Annotation of discourse relations

The annotation of the textogrammatical layer of PDT also serves as the starting point of the annotation of discourse relations and the basic relations of textual coreference. Though we do not consider these relations to belong to the underlying layer of language description as understood in the theoretical framework of Functional Generative Description, however, technically, the annotation of these phenomena is based on the tectogrammatical layer of PDT. As claimed in Mírovský et al. (2012), Nedoluzhko and Mírovský (2013), and Jínová et al. (2012), such an approach has its advantages: the annotators (and, eventually, an automatic preprocessing procedure) can take into account the information relevant for discourse relations that is already present in the underlying representation of the sentence (e.g., the dependency relation between the governing clause and its dependent clauses in case of the relation of cause and admission); in addition, the tectogrammatical representations contain a "reconstruction" of the deleted items in the surface structure (see Sect. 3.5 above), which is very important for the identification of coreference relations, but also relevant for the establishment of certain discourse relations.

The annotation of discourse relations in PDT 3.0 (present also in the Prague Discourse Treebank, PDiT, see Poláková et al., 2013) is based on the annotation scenario applied to the annotation of English texts in the Pennsylvania Discourse Treebank (Prasad et al., 2008). In the process of annotation, the annotators identify so-called connectives and for each of the connective they look for its so-called arguments, i.e., pieces of the text that are connected by some kind of discourse relation indicated by the connective. In this approach, it is assumed that there should be always two arguments connected by one connective.[29]

Fig. 7 exhibits the annotation of a discourse relation between the sentences: *Slovenská elita byla zklamána politickou volbou Slovenska.* [The Slovak elite were disappointed by the political choice of Slovakia.] and *Proto většina kvalitních odborníků zůstala v Praze*. [Therefore, most of the good specialists stayed in Prague.]. A discourse relation between the trees is marked with a thick curved arrow; the type of the relation (reason) is displayed next to the tectogrammatical functor of the starting node. The connective assigned to the relation (*proto* [therefore]) is also displayed at the starting node, as well as the range of the arguments entering the relation (range: 0 -> 0, indicating that in this case, only the two mentioned trees (clauses) enter the relation).

As indicated above, discourse annotation in PDT 3.0 is focused on an analysis of discourse connectives, the text units (or arguments) they connect and on the semantic relation expressed between these two units. A discourse connective is defined as a predicate of a binary relation – it takes two text spans (mainly clauses or sentences) as its arguments. It connects these units and signals to a semantic relation

[29] It should be noted that while the annotation of the discourse relations in the Pennsylvania Discourse Treebank was carried out on running texts, in case of PDiT the disscourse relations are annotated on the tree structures (of the PDT tectogrammatical layer).

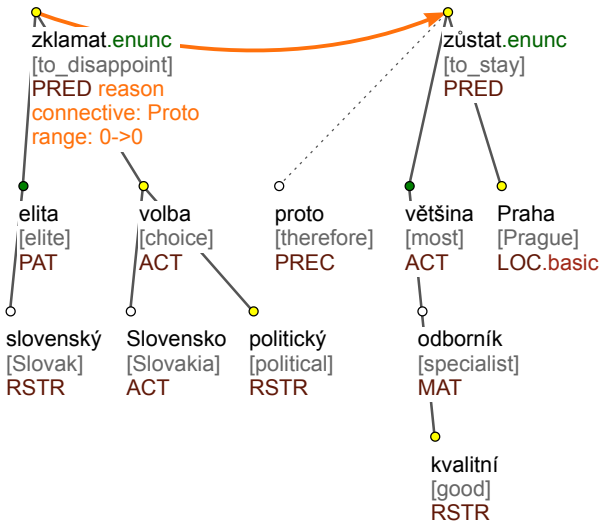

Figure 7. Annotation of a discourse relation between the sentences: Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze. [The Slovak elite were disappointed by the political choice of Slovakia. Therefore, most of the good specialists stayed in Prague.]

between them at the same time. Discourse connectives are morphologically inflexible and they never act as grammatical constituents of a sentence. Like modality markers, they are "above" or "outside" the proposition. They are represented by coordinating conjunctions (e.g. *a* [and], *ale* [but]), some subordinating conjunctions (e.g. *protože* [because], *pokud* [if], *zatímco* [while]), some particles (e.g. *také* [also], *jenom* [only]) and sentence adverbials (e.g., *potom* [afterwards]), and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like *jinými slovy* [in other words] or *naproti tomu* [on the contrary]. The annotation is focused only on discourse relations indicated by overtly present (explicit) discourse connectives – the relations not indicated by a discourse connective were not annotated in the first stage of the project.[30]

The taxonomy of discourse relations in PDT 3.0 is based on the taxonomy used in the Penn Discourse Treebank[31] but it is modified by taking into account the theory of the Functional Generative Description and the tradition of the Czech studies (e.g., the

[30] There are 18,161 discourse relations annotated in the data, out of them 5,538 relations are inter-sentential, 12,623 relations are intra-sentential.

[31] See The Penn Discourse TreeBank 2.0 Annotation Manual (Prasad et al., 2007).

addition of the relation of gradation and explication). The taxonomy contains four basic groups of relations: temporal, contingency, contrast, and expansion.

Within these main groups several subgroups are being distinguished, namely synchronous and asynchronous with temporal relations, reason – result, condition, explication, and purpose in the contingency group; confrontation, opposition, concession, correction, and gradation in the group of contrast, and conjunction, exemplification, specification, equivalence ,and generalization in the group of relations of semantic extension.

In addition to the discourse relations proper, some other types of information have been included in our annotation scheme as well as the appurtenance of the text into the so-called genres (Poláková et al., 2014). This complex annotation makes it possible to search in the annotated corpus for the combination of the deep syntactic structure, information structure, coreference, and genre information.

The process of manual checking of the consistency of annotation that was carried out after the whole treebank was annotated has led not only to a necessary unification of the understanding of some relations but also to interesting observations concerning the complexity of some relations, or to an analysis of multiword connectives, of multiple coordination, etc.

The analysis of annotated data (see Table 4) helped us observe which types of relations are more frequently expressed within the frame of a single sentence and which hold rather between complexes of sentences (divided by final punctuation marks). Up to now, this distribution could be only approximated on the basis of language intuition.

The largest proportion of occurrences within a single (complex) sentence is documented for the relation of purpose, condition, and disjunctive alternative. These relations only rarely occur between two independent sentences. On the basis of these calculations, a preliminary hypothesis can be formulated that the semantic content expressed by the arguments of the above relations are more closely bound together than with the other relations. Also, the relatively high position of the conjunction relation is surprising as one would expect a more balanced distribution, perhaps similar to that found with opposition.

In the course of the annotation, it came out that some connective means connect implicatures or deletions hidden in the text rather than arguments expressed explicitly in the text. To capture these relations, a category called “pragmatic” relations has been introduced, see (42), where the second sentence containing a connective *však* [however] does not express an opposition to the fact that several orders are in progress but an opposition to the unexpressed implication that to have many orders means a large income for the firm.

Group of relation	Type of relation	Intra-sentential	Inter-sentential
Contingency	Purpose	100%	0%
	Condition	99%	1%
	Pragmatic condition	93%	7%
	Reason–result	61%	39%
	Explication	43%	57%
	False reason–result	31%	69%
Contrast	Correction	73%	27%
	Concession	70%	30%
	Confrontation	53%	47%
	Gradation	52%	48%
	Pragmatic opposition	46%	54%
	Opposition	43%	57%
	Restrictive opposition	37%	63%
Expansion	Disjunctive alternative	95%	5%
	Specification	82%	18%
	Conjunction	81%	19%
	Conjunctive alternative	79%	21%
	Equivalence	40%	60%
	Exemplification	19%	81%
	Generalisation	9%	91%
Temporal	Synchronous	77%	23%
	Asynchronous	70%	30%

Table 4. Ratio of types of discourse relations occurring intra-sententially and inter-sententially

(42) Podle vedoucího výroby Miloše Přiklopila má Seba rozpracovanou celou řadu zakázek. Zákazníci však vyvíjejí velký tlak na snižování cen tkanin.
[According to the production manager M.P. several orders are in process in SEBA. The customers, however, make a big pressure on the lowering of the price of the material.]

It will be a matter of future research to see, which relations are more frequently indicated by explicit connectives and which can be easily implied implicitly. Such research may bring interesting results when based on parallel corpora; the fact that we also have at our disposal a parallel Czech–English treebank makes such research possible.

Another research topic relates to the fact that an analysis of discourse relations cannot be restricted to a study based on a rather narrowly defined class of connective devices. Similarly as in the Penn Discourse Treebank (see Prasad et al., 2008), in the current stage of discourse annotation we have focused on the so-called alternative lexicalizations (AltLex, or secondary connectives, see M. Rysová, 2012), that is expressions connecting discourse arguments but not belonging to a narrow class of connectors; the structure of these expressions ranges from a single word to a whole sentence. The first attempt to single out these secondary connectives resulted in a list of 1,201 occurrences of relations signaled by them in PDT. Contrary to the annotation phase that worked only with primary connectives that related clausal arguments, secondary connectives may relate also nominalizations (in 310 cases out of the 1,201 relations rendered by these secondary connectives, e.g., *He was absent because he was ill*. vs. *The reason for his absence was his illness*.). Similarly as is the case of primary connectives, also secondary connectives are not typically component parts of the arguments, they, as it were, stand outside them. However, the situation is different with some verbs of saying (such as *to add, to complement, to continue*) where the verb behaves as a secondary connective but also represents the second argument: its meaning includes also the information that somebody said, wrote, etc., something before (see M. Rysová, 2014b).

4.3. Annotation of coreferential and associative relations

In addition to discourse relations, the annotation scheme of PDT has been enriched by the annotation of coreferential and associative relations. As for coreference, we distinguish between grammatical and textual coreference. Grammatical coreference is restricted to certain syntactic relations within a sentence and the antecedent can be determined in principle on the basis of grammatical rules of the given language. Table 5 shows basic types of grammatical coreference distinguished in PDT.

As for textual coreference, it can be expressed not only by grammatical or lexical means (such as pronominalization, grammatical agreement, repetition of lexical units, use of synonyms, paraphrasing, use of hyponyms, or hypernyms within lexical cohesion) but it can also follow from the context and pragmatics; in contrast to grammatical coreference, textual coreference often goes beyond sentence boundaries.[32]

Two types of textual coreference between nominal groups can be distinguished, namely that specific and generic reference, see (43) as an example of the former type and (44) as an example of the latter type.

(43) **Marie** a Jan spolu odjeli do Izraele, ale **Marie** se musela vrátit kvůli nemoci.
[**Mary** and John left together for Israel, but **Mary** had to return because of illness.]

[32] Textual coreference can be found in 84,306 cases, grammatical coreference in 20,624 cases. There are 30,470 examples of bridging anaphora.

Type of relation	Example
Coreference of reflexive pronouns	*Dcera se musela dlouho přesvědčovat, aby pokračovala v tréninku.* [in the reading of: *The daughter* had to persuade *herself* to continue in training.]
Coreference of relative means (*který*, *jenž*, *což* etc.)	*Za informační dálnici se považuje světová telekomunikační síť, po níž lze přenášet zvuk, data i obraz a která tak otevírá přístup k množství informatických služeb.* [The information motorway is such a world wide telecommunication network, which … and which …]
Relation of "control" present with a specific group of verbs, e.g. *začít* [begin to], *dovolit* [allow to], *chtít* [want to], *dokázat* [prove] to etc.)	*Vedení sekce plánuje vyklidit knihovnu.* [The management of the department plans to empty the library.] (the unexpressed subject of the infinitive *to empty* is in a coreference relation to the Actor of the main clause: *the management*)
Coreference with a complement with so-called "double dependency"	*Honza zastihl Hanku běhat kolem rybníka.* [Honza found Hana *to run* round the pool.] (coreference of the unexpressed subject of the verb to run with the Patient of the verb *found* – Hana)

Table 5. Types of grammatical coreference in PDT

(44) **Psi** štěkají. To je způsob, jak [**oni**] vyjadřují své emoce.
[**Dogs** are barking. This is the way [**they**] express their emotions.]

The border line between these two types is not always clearcut and the interpretation may be rather subjective, see (45), where the expression *hospoda* [restaurant] may have either a specific reference (the concrete enterprise) or a generic one (restaurant as a type of enterprise).

(45) Začal jsem provozováním **hospody**, která byla mnohokrát vykradena. [… 2 sentences follow …] **Hospoda** byla jen startem, polem k podnikání s masem a masnými výrobky.
[I started with opening a **restaurant**, which was many times visited by thieves. [… 2 sentences follow…] The **restaurant** was just a start, an opportunity to deal with meat and meat products …]

We are fully aware that coreference relations may exist not only between nominal groups but also between verbs which denote events. For the time being, however,

our scheme captures only cases where a verb appears as an antecedent of a nominal group. This phenomenon is referred to in literature as a textual deixis.

Side by side with coreference, several other textual relations contribute to the cohesion of text and help the addressee to understand a certain expression as referring to a known entity even if the two entities are not in a strict coreferenctial relation. We call such relations associative anaphora (Nedoluzhko, 2011); in English oriented studies, they are called *bridging anaphora/relation*, *indirect anaphora*, *associative anaphora* etc.

These relations can be classified in several ways, according to the purpose of their description. In our scheme, we concentrate on the linguistic meaning of anaphora and therefore our classification is rather a detailed one. At the same time, however, we have tried to define the types rather strictly, in order to keep the consistency of the annotation. In PDT 3.0, the following types of associative relations are distinguished:

(a) Relation between a whole and a part (*a house – its roof*)
(b) Relation between a set and its subset or member (*a class – several pupils – a pupil*)
(c) Relation between an object and a function defined on that object (*a coach – a team*)
(d) Relation of a pragmatic and semantic contrast (*last year – this year – next year*).
(e) Non-coreferential anaphoric relation, in case of an explicit anaphoric reference to an non-coreferential antecedent (often accompanied by expressions *such as*, *the same*, *similar*, etc.)

In addition to these types, we also distinguish some other specific relations, such as family relations (*father – son*), place – inhabitant (*Prague – Praguians*), author – piece of work (*Rodin – Thinker*), a thing – its owner, event – argument (*enterprise – enterpreneur*), an object and a typical instrument (*copybook* – pen).

Contrary to the domains exemplified in Sections 3.1 through 3.6 above and in 4.1, in the analysis of which we could build upon our well-established theory, and in Sect. 4.2, in which we could modify or complement an existing scenario proposed by another team working on a similar project (namely the Penn Discourse Treebank), we have not found any consistent, uniform and well-developed scheme that would suit our purpose to integrate both the aspects – discourse relations and coreference in broad sense – into the overall system of PDT. In this sense, any step or proposal of a taxonomy of coreferential (and associative) relations within PDT was in itself a contribution to the development of a suitable and consistent approach to the description of these aspects of text coherence resulting in a basic annotation scenario for the phenomena concerned.

5. Some corpus statistics: Inter-annotator agreement

The strength of an annotated corpus lies not only in the quality of the underlying linguistic theory and in its contribution to this theory but also in three other aspects:

– the quality of the annotation process
– the size of the annotated data
– the quality of a search tool for the corpus

The quality of the annotation process can be measured by agreement between the annotations of the same data performed by two or more annotators. As annotation is an expensive process, usually the data are annotated only by one annotator and only a small part of the data is annotated in parallel by two annotators, just for the purpose of measuring the inter-annotator agreement. In this Section, we report on inter-annotator measurements in PDT or other corpora of the Prague dependency family.

The size of the annotated data is also very important, as a small corpus might not offer enough material for a sufficient analysis of scarce phenomena. We have included some figures concerning the size of the data and the frequency of some of the phenomena in the sections above at places for which these figures were relevant.[33] The size and complexity of a corpus are also in a close relation to the possibility to retrieve relevant examples from the data, which is a task for the search tool. For PDT (and other corpora using the same data format), a powerful and user-friendly querying system exists called PML Tree Query (PML-TQ; Pajas and Štěpánek, 2009).

Since the first years of annotation of PDT, the inter-annotator agreement has been measured for many individual annotation tasks. The measurements and the analysis of the disagreements help detect errors in the annotations, improve the annotation guidelines, and find phenomena difficult from the annotation point of view. We present numbers measured on PDT or on the Czech part of Prague Czech-English Dependency Treebank (PCEDT), which uses the same annotation scenario and annotates a similar type of data (journalistic texts).

For classification tasks (tasks where the places to be annotated are given, i.e., identifying such places is not a part of the annotator's decision) we use simple agreement ratio, i.e. percentage of the places where the annotators assigned the same value; sometimes we also mention Cohen's κ (Cohen, 1960), a measure that shows how much better the inter-annotator agreement is compared with the agreement by chance. For more complex tasks, where the identification of the place to be annotated is also a part of the annotator's decision, we use F1-measure, which is the harmonic mean of precision and recall.[34]

On the **morphological layer**, disambiguation of the automatic morphological analysis was done in parallel by pairs of annotators on the whole PDT data. The inter-annotator agreement on the assignment of the correct **morphological tag** to words

[33] If not stated otherwise, the numbers reported come from 9/10 of the whole PDT data, as the last tenth of the data is designated to serve as evaluation test data and as such should not be observed or used in any way other than testing. In these 9/10 of PDT (used as train and development test data), there are 43,955 sentences in 2,849 documents.

[34] http://en.wikipedia.org/wiki/F1_score

with an ambiguous morphological analysis was 95% (Bémová et al., 1999); if the unambiguous words are also counted, the agreement is 97% (Hajič, 2005). Note that in Czech, there are approx. 4.7 thousand different morphological tags.[35]

For the **analytical layer** in PDT, as far as we know, no measurements of the inter-annotator agreement have been published.

On the **tectogrammatical layer**, there are many annotation tasks. The measurements were performed during the annotation of PDT (the numbers for PDT on the tectogrammatical layer, unless specified otherwise, come from Hajičová et al., 2002) and the Czech part of PCEDT (numbers come from Mikulová and Štěpánek, 2010).

(i) The agreement on **linking the tectogrammatical nodes to** their counterparts from **the analytical layer** in PCEDT was 96% for the lexical counterparts and 93.5% for the auxiliary nodes.

(ii) The agreement on assigning **sentence modality** for 268 complex cases of coordinated clauses in PDT (ver. 3.0) was 93.7% with Cohen's κ 89% (Ševčíková and Mírovský, 2012).

(iii) The agreement on establishing the correct **dependency** between pairs of nodes (i.e. the establishment of dependency links together with the determination which member of the pair is the governor) was 91% (64 differences in 720 dependency relations) in PDT, and 88% in PCEDT.

(iv) The agreement on assigning the correct type to the dependency relation (the tectogrammatical **functor**) was 84% (112 differences in 720 relations) in PDT, and 85.5% in PCEDT.

(v) The agreement on assigning the correct value to individual nodes in the annotation of **topic-focus articulation** (i.e. the assignment of the values 'contextually bound' or 'contextually non-bound' within the TFA attribute; 'correct' here means 'as judged by the author of the manual', i.e. the agreement is measured pairwise between each annotator and the arbiter) was approx. 82% (81%, 82%, 76%, and 89% for different annotators) (Veselá et al., 2004).

(vi) In the task of marking **multiword expressions** in the data (which was done on top of the tectogrammatical layer for PDT 2.5), the authors used their own version of weighted Cohen's κ (with adjusted upper agreement bound) and report the agreement above chance of 64.4% (Bejček and Straňák, 2010).

The mismatches between annotators were carefully studied. A comparison of the agreement figures given in (iii) and (iv) indicates that annotators were more confident of their judgements when building the dependency structure rather than when labeling the nodes by functors. This observation indicates that it was not difficult to decide which node is the governor and which is the dependent. Discrepancies between an-

[35] For comparison with other projects, let us mention the inter-annotator measurement during the annotation of the German corpus NEGRA, as reported by Brants (2000). Their agreement in the part-of-speech annotation was 98.57%. However, the size of their part-of-speech tagset was only 54 tags.

notators were found in the decisions on the type of dependency relation, i.e. on the labels for valency members as well as for these of free modifications. This fact demonstrates that the boundaries between some pairs of functors are rather fuzzy, or perhaps they were not defined in an exhaustive way. The functor MEANS (Instrument) and EFF (Effect) were often interchanged as well as the functor BEN (Beneficient) and ADDR (Addressee), though the former member of the pair belongs to the class of free modifications and the latter to the class of valency members. These mismatches are connected with a more or less effective application of the criteria for obligatory positions in the valency frame of the corresponding items. However, there are only few mismatches which are systematic, most of discrepancies are subjective/individual.

Among the **phenomena crossing the sentence boundary**, we have measured the inter-annotator agreement in PDT for the extended (nominal) textual coreference, bridging anaphora and discourse relations. To evaluate the inter-annotator agreement in these annotations, we used several measures:

(i) The connective-based F1-measure (Mírovský et al., 2010) was used for measuring the agreement on the recognition of a **discourse relation**, the agreement was 83%.
(ii) The chain-based F1-measure was used for measuring the agreement on the recognition of a **textual coreference** or a **bridging anaphora**, the agreement was 72% and 46%, respectively.
(iii) A simple ratio and Cohen's κ were used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation, the agreement was 77% (Cohen's κ 71%) for **discourse**, 90% (Cohen's κ 73%) for **textual coreference**, and 92% (Cohen's κ 89%) for **bridging anaphora** (Poláková et al., 2013).[36]

The numbers of the inter-annotator agreement for the phenomena crossing the sentence boundary reveal some simple observations: it is quite clear that recognizing the presence of a textual coreference relation is easier than that of a bridging relation. For both textual coreference and bridging anaphora, it is more difficult to find the existence of a relation rather than to select its type – once the presence of the relation is agreed upon, the annotators are able to assign its type with high accuracy. For discourse relations, on the contrary, an assignment of the type of a relation seems to be more difficult than recognition of its presence.

As mentioned above, the nature of the tasks required to apply for the different annotation tasks different measures for the inter-annotator agreement. Although the numbers expressing different measures of evaluation are not – strictly speaking – directly comparable (especially Cohen's κ cannot be compared with other measures),

[36] For comparison, the simple ratio agreement on types of discourse relations in Czech (77%) is the closest measure to that of measuring the inter-annotator agreement used on subsenses (second level in their sense hierarchy) in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008).

they confirm the general idea that the deeper we go in the abstraction of the language description, the more difficult it is to achieve high values of the inter-annotator agreement.

Measuring the inter-annotator agreement and studying discrepancies between annotators repeatedly proved to be an indispensable part of the annotation process of PDT and other corpora. Not only is it necessary for ensuring a high quality annotation (for reasons mentioned above) but it may even reveal shortcomings in the underlying linguistic theory. It is the only way to establish and enumerate the difficulty of a given annotation task and to set a higher boundary for the accuracy we can expect from automatic methods of annotation.

6. Summary and outlook

6.1. Contributions of the annotation to the theory

In the present paper, we have presented several selected case studies based on the Prague Dependency Treebank Version 3.0 that are supposed to document the importance of corpus annotation at different linguistic layers for a verification of established linguistic hypotheses and for their eventual modifications, or, as the case may be, for making the linguistic description of some phenomena more precise.

The basic ideas of the theoretical Framework of the FGD were formulated before the large language resources were available and as such, they were applied in the design of the original annotation scenario of PDT. During the process of annotation of the raw texts the hypotheses formulated on the basis of the theory were tested and by testing them the accuracy of the theory itself was furthermore accessed, and the gaps within the list of morphological meanings, syntactic and semantic units have been identified. These gaps, however, should not be understood as errors in the original proposal since many of the phenomena concerned had not been noticed before by any reference grammar of Czech.[37] In the present contribution several of these improvements have been discussed at the end of each Section: the necessity of the two levels of syntax (surface and deep/underlying levels, called tectogrammatics) is supported by the introduction of the category of diathesis (see 3.1), by the new grammateme pair/group number (see 3.2) and by the restoration of elements missing on the surface structure and required by the deep representation (see 3.5). Also a new class of valency members (called quasivalency) was introduced (see 3.4). While in the classical version of the FGD the issues of lexicon were not in the focus of our attention, the introduction of new categories (functors, subfunctors, grammatemes) opened new aspects of the interplay between grammar and lexicon which were analyzed in particular case studies above and became a source of extension of the theoretical framework.

[37] The notion of "reference grammar" is not commonly used in Czech linguistics but the Mluvnice češtiny [Czech Grammar] (Komárek et al., 1986, Daneš et al., 1987) is supposed to be a standard source of references, and, as for Czech syntax, the monograph by Šmilauer (1947) is most frequently used in this sense as well.

In the domain of information structure, the annotated data helped us to develop in more detail the hypotheses concerning the deep word order (so-called systemic ordering) as documented in 4.1.2 and to achieve a more detailed analysis of the special class of particles called in linguistic literature rhematizers, focussing particles, or focalizers. The analysis of rich corpus material has indicated that the position of these particles need not be restricted to the focus of the sentence (as the term previously used for them may suggest) but that they may also occur in the topic; this observation has led to the introduction of the notion of contrastive topic and to the distinction between the focus of the sentence as a whole (global focus) and the local focus of the focalizer.

While Part I of the case studies (Section 3) contains analyses of phenomena that belong to grammar, Part II covers a domain that traditionally might be relegated to the domain of pragmatics. However, as the arguments presented in numerous writings on topic–focus articulation quoted in Section 4.1 and supporting the semantic relevance of this phenomenon, a description of the information structure of the sentence is an indispensable part of any functionally conceived grammatical theory. On the other hand, coreference relations (with the exception of grammatical coreference) and discourse relations do go beyond the sentence structure and therefore they were not analyzed in detail in the theoretical framework the PDT annotation is based on. In this sense, the analysis presented in Sections 4.2 and 4.3 brings observations that have not yet been included in a systematic way in any description of Czech.

An irreplaceable role in the process of recognition and implementation of the improvements to the theory is played by the human annotators themselves; though the manual annotation is rather expensive, its effect is doubtless: the annotators have to work consistently applying the existing guidelines and they supply many observations that uncover linguistic details hitherto not registered. The usefulness, abundance and originality of these observations is best documented by the publication of the modern scientific syntax of Czech based on PDT (Panevová et al., 2014).

6.2. Outlook

It is indisputable, however, that some particular phenomena require further analysis. Many empirical problems are connected with coordinated constructions. The studies of elliptic coordinations are planned for the detection of the formal criteria for the possibility of restoration their underlying representation in contrast to the pragmatic conditions for their application belonging to the domain of text structure and discourse relations. Another domain of further work relates to the reflection of the results achieved in our analysis in the dictionary build-up. Selected data extracted from PDT 3.0 will be incorporated into the valency dictionary: e.g. completion of the list of words with the ability of control and the proposal of the description of the interplay between morphological meanings of verbs and the realization of their valency frames in the sentence.

In the particular case of the Prague Dependency Treebank, there is one feature that distinguishes it from annotation schemes worked out for other languages, namely the fact that annotation on all layers together with the annotation of discourse relations, coreference, and associative relations is applied to the same collection of full texts (and partly also on parallel English texts). This makes it possible to look for an interplay of these layers and to try and use the complex annotation for some particular projects. For instance, we have started a research in the interplay of syntactic structure, information structure, and coreference relations based on the notion of the activation hierarchy of elements of the stock of knowledge as proposed by Hajičová and Vrbová (1982) and elaborated further e.g., in Hajičová (1993, 2003, 2012) and Hajičová and Vidová-Hladká (2008). The underlying hypothesis for our analysis of discourse structure was formulated as follows: 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 shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation (salience) of referents. The following three basic heuristics (a) through (c) based on the position of the items in question in the topic or in the focus of the sentence, on the means of expression (noun, pronoun) and on the previous state of activation have been formulated to determine the degrees of salience of the elements of the stock of shared knowledge:

(a) In the flow of communication, a discourse referent enters the discourse, in the prototypical case, first as contextually non-bound, thus getting a high degree of salience. A further occurrence of the referent is contextually bound, the item still has a relatively high degree of salience, but lower than an element referred to in the focus (as contextually non-bound) in the given sentence.
(b) If an item is not referred to in the given sentence, the degree of salience is lowered; the fading is slower with a referent that had in the previous co-text occurred as contextually bound; this heuristics is based on the assumption that a contextually bound item has been 'standing in the foreground' for some time (as a rule, it was introduced in the focus, then used as contextually bound, maybe even several times) and thus its salience is reinforced; it disappears from the set of the highly activated elements of the stock of shared knowledge in a slower pace that an item which has been introduced in the focus but then dropped out, not rementioned. If the referent has faded too far away it has to be re-introduced in the focus of the sentence.
(c) If the difference in the degree of salience of two or more items is very small, then the identification of reference can be done only on the basis of inferencing.

These three basic heuristics served as a basis for our formulation of several rules for the assignment of the degrees of salience, which have been applied to numerous text segments to check how the determination of these degrees may help reference resolution. Thanks to the richly annotated corpus of PDT, we basically have at our disposal

all of the information we need for an application of our rules for activation assignment: the underlying sentence representation with restored (superficial) deletions as well as with part-of-speech information, the Topic-Focus assignment (via the TFA attribute with values contextually-bound and contextually non-bound) and coreferential chains for nominal and pronominal realization of referential expressions. The activation algorithm has already been implemented and applied to (selected but full) documents, the 'activation' diagrams have been visualized and the task now is to test the hypotheses our approach is based on and the possibilities the approach offers for text analysis and generation on a larger portion of the PDT collection.

Acknowledgements

The authors gratefully acknowledge the support from the Grant Agency of the Czech Republic (project No. P406/12/0658) and from the Ministry of Education, Youth and Sports (projects LH14011 and LM2015071). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Bibliography

Bejček, Eduard and Pavel Straňák. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1–2):7–21, 2010.

Bejček, Eduard, Jarmila Panevová, Jan Popelka, Lenka Smejkalová, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, Josef Toman, Zdeněk Žabokrtský, and Jan Hajič. Prague Dependency Treebank 2.5. Data/software, 2011.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0. Data/software, 2013.

Bémová, Alevtina, Jan Hajič, Barbora Vidová Hladká, and Jarmila Panevová. Morphological and Syntactic Tagging of the Prague Dependency Treebank. In *Journées ATALA – Corpus annotés pour la syntaxe; ATALA Workshop – Treebanks*, pages 21–29, Paris, 1999. Université Paris.

Bhaskararao, Peri and Karumuri Venkata Subbarao. *Non-nominative subjects*, volume 1. John Benjamins Publishing, 2004.

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In *Treebanks: Building and Using Syntactically Annotated Corpora*, chapter 7, pages 103–128. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In Hinrichs, E. and K. Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, 2002.

Brants, Thorsten. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, 2000. European Language Resources Association.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 969–974, 2006.

Camacho, J. A. *Null Subjects*. Cambridge University Press, 2013.

Cinková, Silvie, Josef Toman, Jan Hajič, Kristýna Čermáková, Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů, and Zdeněk Žabokrtský. Tectogrammatical Annotation of the Wall Street Journal. *The Prague Bulletin of Mathematical Linguistics*, (92):85–104, 2009.

Clancy, Steven J. *The chain of being and having in Slavic*, volume 122. John Benjamins Publishing, 2010.

Cohen, Jacob. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Daneš, František, Helena Běličová, Mirek Čejka, Emil Dvořák, Miroslav Grepl, Karel Hausenblas, Zdeněk Hlavsa, Jana Hoffmannová, Josef Hrbáček, Jan Chloupek, Petr Karlík, Eva Macháčková, Olga Müllerová, Bohumil Palek, Jiří Nekvapil, Jiří Novotný, Petr Piťha, Hana Prouzová, Milena Rulfová, Blažena Rulíková, Otakar Šoltys, Ludmila Uhlířová, and Stanislav Žaža. *Mluvnice češtiny. 3. Skladba [Grammar of Czech. 3. Syntax]*. Academia, Prague, 1987.

Fillmore, Charles J. The Case for Case Reopened. *Syntax and Semantics*, 8(1977):59–82, 1977.

Fillmore, Charles J. *Form and Meaning in Language. Volume 1. Papers on Semantic Roles*. CSLI Publications, Stanford University Press, 2003.

Firbas, Jan. K otázce nezákladových podmětů v současné angličtině. Příspěvek k teorii aktuálního členění větného. *Časopis pro moderní filologii*, 39:22–42; 165–173, 1957. (An abbreviated and modified English version of this contribution was published as Non-thematic subjects in Contemporary English, TLP 2, Prague: Academia, 239–256.).

Giger, Markus. *Resultativa im modernen Tschechischen: unter Berücksichtigung der Sprachgeschichte und der übrigen slavischen Sprachen*, volume 69. Peter Lang, Bern – Berlin – Bruxelles – Frankfurt a.M. – New York – Oxford – Wien, 2003.

Hajič, Jan. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. Eva Hajičová)*. Karolinum, Charles University Press, Prague, 1998.

Hajič, Jan. Complex Corpus Annotation: The Prague Dependency Treebank. In Šimková, Mária, editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 54–73. Veda, Bratislava, 2005.

Hajič, Jan and Václav Honetschläger. Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *The Prague Bulletin of Mathematical Linguistics*, (79–80):61–86, 2003.

Hajič, Jan and Zdeňka Urešová. Linguistic Annotation: from Links to Cross-Layer Lexicons. In Nivre, Joakim and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 69–80, Vaxjo, Sweden, 2003. Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim and Erhard Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden, 2003. Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. Prague Dependency Treebank 2.0. Data/software, 2006.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech–English Dependency Treebank 2.0, 2011.

Hajič, Jan, Eva Hajičová, Marie Mikulová, Jiří Mírovský, Jarmila Panevová, and Daniel Zeman. Deletions and node reconstructions in a dependency-based mutlilevel annotation scheme. In Gelbukh, Alexander, editor, *16th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pages 17–31, Berlin / Heidelberg, 2015. Springer.

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

Hajičová, Eva. Aspects of discourse structure. In Vertan, Christina, editor, *Natural language processing between linguistic inquiry and system engineering*, pages 47–54, Iasi, 2003. Editura Universitatii Alexandru Ioan Cuza.

Hajičová, Eva. On scalarity in information structure. *Linguistica Pragensia*, XXII(2):60–78, 2012.

Hajičová, Eva and Barbora Vidová-Hladká. What Does Sentence Annotation Say about Discourse? In *18th International Congress of Linguists, Abstracts*, pages 125–126, Seoul, Korea, 2008. The Linguistic Society of Korea.

Hajičová, Eva, Petr Pajas, and Kateřina Veselá. Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations. *The Prague Bulletin of Mathematical Linguistics*, 77:5–18, 2002.

Hajičová, Eva, Jiří Mírovský, and Katja Brankatschk. A Contrastive Look at Information Structure: A Corpus Probe. In *Proceedings of the 6th Congres de la Societe Linguistique Slave*, pages 47–51, Aix-en-Provence, 2011. Univ. de Provence.

Hajičová, Eva, Marie Mikulová, and Jarmila Panevová. Reconstruction of Deletions in a Dependency-based Description of Czech: Selected Issues. In Hajičová, Eva and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 131–140, Uppsala, Sweden, 2015. Uppsala University.

Hajičová, Eva and Jarka Vrbová. On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding. In Horecký, Ján, editor, *Proceedings of the 9th Conference on Computational Linguistics*, pages 107–113, Prague, 1982. Academia.

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

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alevtina Bémová. A manual for analytic layer tagging of the Prague Dependency Treebank. Technical Report TR-1997-03, 1997.

Hajič, Jan, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. Prague Dependency Treebank. To be published in Handbook on Linguistic Annotation, eds. N. Ide and J. Pustejovsky. Berlin / Heidelberg: Springer, 2015.

Hajičová, Eva. Postavení rematizátorů v aktuálním členění věty [Position of Rhematizers in the Topic–Focus Articulation]. *Slovo a slovesnost*, 56(4):241–251, 1995.

Hausenblas, Karel. Slovesná kategorie výsledného stavu v dnešní češtině. *Naše řeč*, 46:13–28, 1963.

Jínová, Pavlína, Jiří Mírovský, and Lucie Poláková. Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In Hajičová, Eva, Lucie Poláková, and Jiří Mírovský, editors, *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, pages 43–58, Bombay, 2012.

Kingsbury, Paul and Martha Palmer. From TreeBank to PropBank. In *Proceedings of LREC 2002*, pages 1989–1993, Las Palmas, Canary Islands, Spain, 2002.

Koktová, Eva. *Word-order based grammar*, volume 121. Walter de Gruyter, 1999.

Komárek, Miroslav, Jan Petr, Jan Kořenský, Anna Jirsová, Naďa Svozilová, Karel Hausenblas, Jan Balhar, Emil Dvořák, Milena Rulfová, Zdeňka Hrušková, Jarmila Panevová, Eva Buráňová, Libuše Kroupová, and Oldřich Uličný. *Mluvnice češtiny. 2. Tvarosloví [Grammar of Czech. 2. Morphology]*. Academia, Prague, 1986.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves [Valency Dictionary of Czech Verbs]*. Nakladatelství Karolinum, Praha, 2008.

Marcus, Mitchell, Beatrice Santorini, and Marcinkiewicz Mary Ann. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.

Mathesius, Vilém. Slovesné časy typu perfektního v hovorové češtině [Verbal tenses of the perfective type in colloquial Czech]. *Naše řeč*, 9(7):200–202, 1925.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC 2004*, pages 803–806, Lisbon, Portugal, 2004.

Mikulová, Marie and Jan Štěpánek. Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1836–1839, Valletta, 2010. European Language Resources Association.

Mírovský, Jiří, Lucie Mladová, and Šárka Zikánová. Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In Huang, Chu-Ren and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 775–781, Beijing, 2010. Tsinghua University Press.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Does Tectogrammatics Help the Annotation of Discourse? In Kay, Martin and Christian Boitet, editors, *Proceedings of the 24th International Conference on Computational Linguistics*, pages 853–862, Bombay, 2012.

Nedoluzhko, Anna. *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu) [Textual coreference and associative anaphora: The conception of annotation of Czech data in the Prague Dependency Treebank]*. Charles University in Prague, Institute of Formal and Applied Linguistics, Prague, 2011.

Nedoluzhko, Anna and Jiří Mírovský. How Dependency Trees and Tectogrammatics Help Annotating Coreference and Bridging Relations in Prague Dependency Treebank. In Hajičová, Eva, Kim Gerdes, and Leo Wanner, editors, *Proceedings of the Second International Conference on Dependency Linguistics, Depling 2013*, pages 244–251, Praha, Czechia, 2013. Univerzita Karlova v Praze, Matfyzpress.

Pajas, Petr and Jan Štěpánek. System for Querying Syntactically Annotated Corpora. In Lee, Gary and Sabine Schulte im Walde, editors, *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, 2009. Association for Computational Linguistics.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description, Parts I, II. *The Prague Bulletin of Mathematical Linguistics*, 22, 23:3–40, 17–52, 1974–75.

Panevová, Jarmila. Verbal Frames Revisited. *The Prague Bulletin of Mathematical Linguistics*, 28: 55–72, 1977.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Ph. L., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. Benjamins Publ. House, Amsterdam-Philadelphia, 1994.

Panevová, Jarmila and Magda Ševčíková. The Role of Grammatical Constraints in Lexical Component in Functional Generative Description. In Apresjan, Valentina, Boris Iomdin, and Ekaterina Ageeva, editors, *Proceedings of the 6th International Conference on Meaning-Text Theory*, pages 134–143, Praha, Czechia, 2013. Univerzita Karlova v Praze.

Panevová, Jarmila, Eva Hajičová, Václava Kettnerová, Markéta Lopatková, Marie Mikulová, and Magda Ševčíková. *Mluvnice současné češtiny. 2 [Grammar of Modern Czech. 2]*. Karolinum, Prague, 2014.

Poláková, Lucie, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, 2013. Asian Federation of Natural Language Processing.

Poláková, Lucie, Pavlína Jínová, and Jiří Mírovský. Genres in the Prague Discourse Treebank. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1320–1326, Reykjavik, 2014. European Language Resources Association.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01, Philadelphia, 2007.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, 2008. European Language Resources Association.

Preinhaelterová, Ludmila. Systemic Ordering of Complementations in English as Tested with Native Speakers of British English. *Linguistica Pragensia*, 7(97):12–25, 1997.

Putnam, Hilary. Some Issues in the Theory of Grammar. In Jakobson, Roman, editor, *The Structure of Language and Its Mathematical Aspects, Proceedings of Symposia in Applied Mathematics*, pages 25–42, Providence, 1961. American Mathematical Society.

Rooth, Mats. *Association with Focus*. PhD thesis, GLSA, Dept. of Linguistics, University of Massachusetts, Amherst, 1985.

Rysová, Kateřina. *O slovosledu z komunikačního pohledu [On Word Order from the Communicative Point of View]*. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, 2014a.

Rysová, Kateřina, Jiří Mírovský, and Eva Hajičová. On an apparent freedom of Czech word order. A case study. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 93–105, Warszawa, Poland, 2015a. IPIPAN.

Rysová, Kateřina, Magdaléna Rysová, and Eva Hajičová. Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description. Technical Report TR 2015-59, Prague, Czechia, 2015b.

Rysová, Magdaléna. Alternative Lexicalizations of Discourse Connectives in Czech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2800–2807, Istanbul, 2012.

Rysová, Magdaléna. Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 930–935, Reykjavik, 2014b.

Ševčíková, Magda and Jiří Mírovský. Sentence Modality Assignment in the Prague Dependency Treebank. In Sojka, Petr, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 15th International Conference, TSD 2012*, pages 56–63, Berlin/Heidelberg, 2012. Springer.

Sgall, Petr. Functional Sentence Perspective in a Generative Description of Language. *Prague Studies in Mathematical Linguistics*, 2:203–225, 1967.

Sgall, Petr. Towards a Definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics*, 31:3–25, 1979.

Sgall, Petr, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová. *A Functional Approach to Syntax in Generative Description of Language*. American Elsevier Publishing Company, New York, 1969.

Sgall, Petr, Eva Hajičová, and Eva Benešová. *Topic, Focus and Generative Semantics*. Scriptor, Kronberg/Taunus, 1973.

Sgall, Petr, Eva Hajičová, and Eva Buráňová. *Aktuální členění věty v češtině [Topic–Focus Articulation in Czech]*. Academia, Prague, 1980.

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

Šmilauer, Vladimír. *Novočeská skladba [Syntax of Modern Czech]*. Ing. Mikuta, Prague, Czechia, 1947.

Štěpánková, Barbora. *K funkci výrazů částicové povahy ve výstavbě textu, zejména k jejich roli v aktuálním členění. [On the function of particles in the structure of text, especially on their role in topic-focus articulation]*. PhD thesis, Charles University, Prague, 2013.

Tesnière, Lucien. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, Paris, 1959.

Urešová, Zdeňka. *Valence sloves v Pražském závislostním korpusu [Valency of Verbs in the Prague Dependency Treebank]*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011a.

Urešová, Zdeňka. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex) [Valency Dictionary of the Prague Dependency Treebank (PDT-Vallex)]*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011b.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. Annotators' Agreement: The Case of Topic–Focus Articulation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2191–2194, Lisbon, 2004.

Zikánová, Šárka. What do the Data in Prague Dependency Treebank Say about Systemic Ordering in Czech? *The Prague Bulletin of Mathematical Linguistics*, 86:39–46, 2006.

Zikánová, Šárka. Problematické syntaktické struktury: k rozborům aktuálního členění v Pražském závislostním korpusu [Probematic syntactic structures: on topic-focus articulation analysis in the Prague Dependency Treebank]. In Polách, Vladimír, editor, *Svět za slovy a jejich tvary, svět za spojením slov*, pages 233–240. Univerzita Palackého, Olomouc, 2008.

Zikánová, Šárka, Miroslav Týnovský, and Jiří Havelka. Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. *The Prague Bulletin of Mathematical Linguistics*, 87:61–70, 2007.

Address for correspondence:
Eva Hajičová
hajicova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Charles University
Malostranské nám. 25
118 00 Prague 1
Czech Republic