Function Words in Praguian Functional Generative Description

EVA HAJIČOVÁ, JARMILA PANEVOVÁ,
MARIE MIKULOVÁ, AND JAN HAJIČ

Charles University, Prague

Abstract

The present contribution is aimed at a complex description of the treatment of the so-called function words within the framework of a dependency-based Functional Generative Description as proposed in Prague by Petr Sgall and his team and its reflection in the Prague Dependency Treebank, an original annotated corpus of Czech. Both the framework and the treebank are based on a stratificational model of language, a part of which are two levels of dependency-based syntactic structure, one oriented towards the syntactic structure of the sentence on the surface layer called analytical and the other oriented towards the underlying, deep sentence structure called tectogrammatical. The dependency tree structure of the sentence on the analytical level contains all the words present in the sentence as separate nodes, while the dependency representation of a sentence on the tectogrammatical level conceived of as a linguistically structured meaning of the sentence contains only content words as its nodes. On the analytical level, a distinction is made between different classes of function words, the main boundary being between the function words that function within verbal complexes and which contain information about the morpho-syntactic properties of verbs (i.e., auxiliaries), and those being parts of nominal groups (prepositions) or connecting clauses (or, as the case may be, parts of clauses) into one whole (conjunctions). It is argued in this article that auxiliaries should be considered to be dependents on the verb that is their governor and to which they "belong," and that the prepositions and conjunctions, on the contrary, should be considered to be the heads of the nouns or clauses whose form they "govern" or "control." On the tectogrammatical level, the semantic contribution of the function words to the meaning of the sentence is reflected by informa-

Linguistic Analysis, 43 3-4
2024 Linguistic Analysis
8832 SW Dilworth Rd, Vashon, WA 98070

tion attached to the nodes of the tectogrammatical tree in the form of complex labels. Auxiliaries, prepositions and conjunctions are the most evident classes of function words, though there are some groups of words such as particles that are located on the borderline between function words and content words, to which we also pay attention in this study.

1. Introduction

The classification of words into subclasses is one of the topics that has been paid attention to by linguists of most different orientations and the criteria for such classification have been sought in morphology (subclasses known as word classes classified with regard to their participation in the declension/conjugation paradigms), in syntax (with regard to their function in the sentence as the term "function words" indicates), or with regard to their semantics as reflected in the opposition of autosemantic (content) vs. synsemantic (function) words. It is no wonder then that also formal descriptions of language such as dependency grammars have had to take these classification issues into consideration.

The Praguian Functional Generative Description (FGD in the sequel) we subscribe to is a dependency-based formal description of language that understands the language system as a multistratal system of levels the units of which stand in the function—form relation (Sgall 1967, Sgall et al. 1969, Sgall et al. 1986). In other words, a unit on a given level is understood to represent a form of (an) element(s) of (a) unit(s) of a next higher level that is/are its/their function(s). Besides the phonological and morphological levels, there are two syntactic levels in the system, one reflecting the structure of the sentence on the surface level (later called analytical), and the other reflecting the underlying syntactic structure of the sentence (called tectogrammatical). i.e., a structure understood as the linguistically structured meaning of the sentence. The representations of the sentence on these levels have the form of a dependency tree,[1] while the

[1] Formally, the representations are rooted (oriented) trees with labelled nodes. Each sentence is represented by a single tree (i.e. a continuous graph). Whenever labels are considered for edges, they are technically made part of the (complex) label of the dependent node. A complex label is a set of attribute-value pairs, organized in hierarchical groups (if necessary).

representations of the sentences on the lower levels have the form of a string (labelled sequence of tokens).

The main difference between the two syntactic levels—on top of the repertoire of the syntactic relations—lies in the fact that on the analytical level, every word of the surface shape of the sentence (including punctuation tokens) is represented by a node of its own and at the same time, no newly added nodes are allowed. On the tectogrammatical level only the autosemantic lexical items (content words) are represented by a separate node and the synsemantic lexical items (function words) lose their independent status; their contribution to the meaning of the sentence is captured within the (labels of the) nodes for the autosemantic word classes.

This implies (among other things) that in the case of surface deletions, some words present in the surface sentence shape are lacking their governor or their dependent. To keep the representation a continuous tree structure, words with a missing governor are placed in the position at which the missing (elided) word would have been and receive a special label to mark what may be called "a false dependency." In the tectogrammatical trees, the cases of surface deletion are resolved by inserting new nodes into an appropriate place in the dependency structure (Hajič et al. 2015, Hajičová et. al. 2015).

Let's also mention that on both levels, a complete node ordering is defined. On the analytical level, the linear node ordering corresponds to the word order of the tokens in the sentence (which is easy to achieve since each token is represented by a single unique node); this implies that the analytical trees can be non-projective in principle (Havelka 2005, 2007; Zeman 1998, 2004). On the other hand, the node order on the tectogrammatical level is given by the so-called communicative dynamism (related to information structure; Sgall and Hajičová 1987; Veselá et al. 2004; Hajičová et al. 2004), and by definition tectogrammatical tree structures are always projective.[2]

The theoretical multilevel approach of FGD is reflected in the conception of the Prague Dependency Treebank (PDT) that was being built in Prague beginning in the nineties of the 20th century

[2] While there are many formal definitions of projectivity (not all of them equivalent, see e.g., Marcus (1965)), the basic idea is as follows: a dependency tree of a sentence is projective, if all yields of all subtrees are continuous (i.e., without an intervening word that is not represented by some node in that subtree); a yield of a subtree is the set of all words that are represented in the subtree, arranged linearly using their original positions in the sentence.

(Hajič 1998; Hajič et al. 2017) originally for Czech (its annotated data serving i.a. for a complex description of Czech syntax; Panevová et al. 2014). It was gradually enlarged both by a build-up of a parallel English-Czech treebank or treebank of spoken Czech and by the range of annotated relations (discourse, coreference, multiword expressions, etc.). One can thus speak about a PDT-family of treebanks.[3] For the purpose of our discussion of the nature and position of function words presented in this article, we use the PDT data-style as the reference point taking into consideration both the analytical (analytical tree structures, ATS) and the tectogrammatical (tectogrammatical tree structures, TGTS) levels.

In the main part of our contribution, we pay attention to the following two fundamental issues; as our discussion will indicate, they are in a mutual relationship:

(i) What is the position of the function words within ATS structure with regard to the content words? (Sect. 3)

(ii) How is the contribution of the function word to the "meaning" of the content word to which the given fucntion word is related represented within TGTS? (Sect. 4)

2. Syntactic levels in FGD and PDT

The fundamental principles of the FGD are deeply rooted in the original structural and functional tenets of the Prague School, especially in regard not only to the language forms but also to their functions. This has led among other things to the fact that out of the two syntactic levels within the FGD multistratal system, it is the tectogrammatical level that is the focus of attention both for the theoretical model as well as for its reflection in the PDT annotation. The criteria applied in the design of the tectogrammatical level are of a syntactico-semantic nature: the dependency structure of the TGTS of sentences contains "full-value" lexical items, i.e., content words. The information carried by the function words is reflected in the attributes of the TGTS nodes; this concerns both the morphosyntactic

[3] The latest consolidated release of the existing PDT-corpora of Czech data is The Prague Dependency Treebank—Consolidated 1.0 (Hajič et al. 2020a, 2020b).

features (such as temporality, modality with verbs) reflected in the so-called grammatemes,[4] or syntactico-semantic values of the relations between nodes (at the edges of the tree) reflected in the values of the so-called functors and subfunctors.

The analytical syntactic level plays a subsidiary role. It serves as a bridge between the string-like surface shape of the sentence on the morphological level to the tree-like syntactico-semantic dependency shape of TGTS viewed as a linguistically structured meaning of the sentence. In ATS, on the one hand, the lexical items present in the surface shape of the sentence preserve their status as separate nodes, on the other hand, their position in the dependency tree is given mostly by syntactic considerations of their status in the language system.

The difference between the ATS and TGTS representation is illustrated in Fig. 1, which displays ATS and TGTS of the sentence in (1). In general, the labels of the nodes in ATS starting with Aux denote function words, Pred stands for the Predicate and Sb, Obj, Adv, Atr stand for the basic analytical syntactic functions Subject, Object, Adverbial and Attribute, respectively. In TGTS, the labels such as ACT, ADDR, PAT, DIR3 and RSTR stand for the dependency relations (functors) of Actor, Addressee, Patient, Direction-where-to and Restriction, respectively. We can observe that function words such as prepositions, conjunctions, auxiliary verbs, i.e., *to be* (present in ATS) do not have their own node in TGTS. Their contribution to the meaning of a sentence is described by grammateme attributes and functor and subfunctor attributes, see (Fig. 1).

(1) Řekl *jsem* Tomovi, že *bych* nastoupil *na* nabízené místo.
Told *am* Tom *that would* accept *on* offered position
'I told Tom that I would accept the offered position.'

The example trees throughout this paper are in a form in which they are canonically visualized using the tools available for browsing and editing the PDT. We do not explain all the attributes and values that are used in the example trees, only those that are necessary to illustrate the phenomenon described (a simple explanation of all the labels used in the trees can be found in the list of abbreviations

[4] The morphosyntactic features (which we reflect in the grammateme attributes) are carried not only by function words but also by other formal means such as the inflection of nouns and verbs. This applies to morphosyntactic features such as the number of nouns, degree of adjectives, etc.

and labels at the end of this article). Furthermore, as opposed to the complete tectogrammatical annotation, the TGTS visualization of the example trees is simplified (for example, coreference or topic-focus articulation (information structure) is not shown). Examples of ATS are complete, reflecting also the surface word order. For the detailed description, see the annotation manuals for analytical level (Hajič et al. 1999) and for tectogrammatical level (Mikulová et al. 2006).

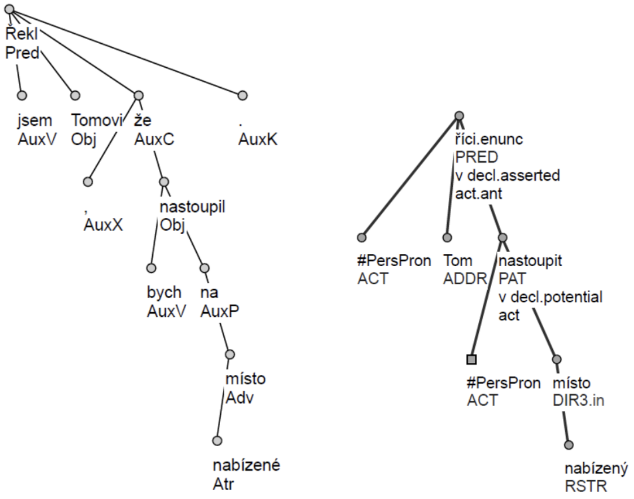

Fig. 1. ATS and TGTS of the sentence
(1) *Řekl jsem Tomovi, že bych nastoupil na nabízené místo.*

3. Function words on the analytical level

3.1 General considerations

When considering the position of the function words in ATS, a decision has to be made whether the function words should have the position of heads, and if yes, then under which condition(s). Tesnière in his fundamental treatment of dependency (Tesnière 1959, quoted from the English translation 2015, ch. 22, 23) introduces the notion of a nucleus, defined as a set that joins together the structural node (as a structural center) and its semantic functions (semantic center); the presence of both the two centers is obligatory but need not be fulfilled by a single word: each can be fulfilled by a separate word. In such a case, one deals with a "dissociated nucleus"; Tesnière gives as an example a periphrastic tense verb. In their most recent paper on Universal Dependencies,[5] the authors (de Marneffe et al. 2021) refer to the notion of a dissociated nucleus and choose the lexical word (Tesnière's "structural center") as the head and the function words[6] (comparing them to Tesnière's semantic functions) as the dependents.[7] In the treatment of function words with regard to content words on the analytical level of FGD, a distinction is made between those function words that further specify the meaning of the content words, with typical examples such as markers of tense and mood with verbs, and those that further specify the syntactic role of a node. In the former case, function words are considered to be dependents of the content word to which they belong as well as to the meaning which they further specify; this is mainly the case of complex verb forms. In the latter case, the function word influences the semantics of the relation between two content words; in a certain sense it may be understood as a "label" of the edge between them that exists in TGTS. According to this view, the node for the function word is placed between the two content words and as such is in the position of the head of the second node. This situation occurs mostly in the case of prepositional phrases and with subordinate sentence structures.

3.2 Categories of function words and their position in the analytical dependency tree

3.2.1 Verb groups

From the point of view of their complexity, the predicate of a sentence can be expressed by a single word or by a multiword expression that functions as a single predicate. The simple forms are used, for example, for the present tense (2).

(2) *Píšu* článek.
write paper
'I am writing a paper.'

[5] For a general overview of Universal Dependencies, see the seminal paper by Nivre et al. (2016).

[6] A critical analysis of such an approach from the point of view of syntactic consistency can be found in Osborne and Gerdes (2019).

[7] De Marneffe et al. (2021) admit that Universal Dependency representations are thus "midway" if compared to such multilevel systems as FGD and Mel'čuk's Meaning-Text Theory (Mel'čuk 1988) or Bresnan's Lexical Functional Grammar (Bresnan et al. 2016); however, this issue is beyond the scope of our present discussion (see also de Marneffe et al this volume).

With a certain simplification, three different groups of predicates composed of multiple words can be distinguished:

(i) Complex verb forms. The verb forms of past tense (3), future tense of imperfective verbs (4), present and past conditional (5) and passive infinitive are complex verb forms (traditionally called analytical or periphrastic forms). They consist of a content (lexical) verb connected with one, two, or three occurrence(s) of the auxiliary *to be* (labelled by AuxV in ATS).

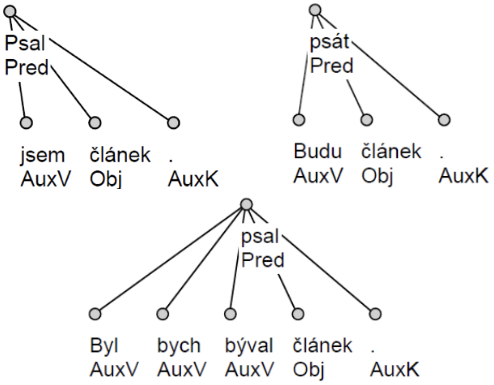

Fig. 2. ATS of the sentences:

(3) *Psal jsem článek.*
(4) *Budu psát článek.*
(5) *Byl bych býval psal článek.*

(3) *Psal* *jsem* článek. (Fig. 2)
wrote-Sg-M *am*-Sg-1st paper
'I *wrote* a paper.'

(4) *Budu* *psát* článek. (Fig. 2)
will-Sg-1st *to-write* paper
'I *will write* a paper.'

(5) *Byl* *bych* *býval*[8] *psal* článek. (Fig. 2)
was-Sg-M *would*-Sg-1st *was*-Sg-M *wrote*-Sg-M paper
'I *would have written* the paper (if I had had the time).'

[8] In the past conditional, the auxiliary *býval* is optional.

(ii) Reflexives. Reflexive forms *se*/*si* are multifunctional and as such may belong to different POS (parts of speech). They actually stand on the boundary between the two classes: pronouns and function words. As parts of a predicate, the forms *se*/*si* are function words of two types. First, they function as a part of the lexical meaning of particular verbs. Such verbs are included in dictionaries as a special lexical entry (e.g., *bát se* 'to be afraid,' *dívat se* 'to look at,' *šířit se* 'spread' (6); Panevová and Karlík 2016). In this case, reflexive forms *se*/*si* are labelled by AuxT in ATS. Second, the forms *se*/*si* serve as an indication of (grammatical) diathesis, namely deagentive (7), reciprocal (8), and dispositional. In these cases, *se*/*si* gets the label AuxR.

(6) Virus *se* rychle *šíří*. (Fig. 3)
virus *Refl* quickly *spreads*.
'Virus *spreads* quickly.'

(7) *Píše se* článek. (Fig. 3)
writes Refl paper
'The paper is (already) being written.'

(8) Účastníci *si vyměňují* vizitky.
Participants *Refl exchange* business-cards.
'Participants exchange their business cards (among each other).'

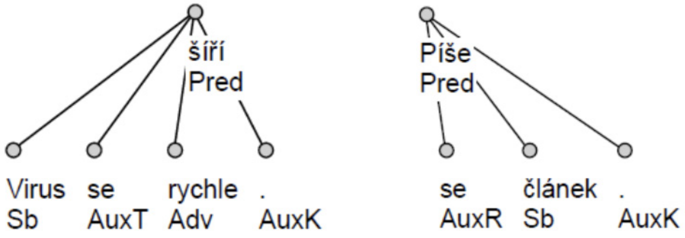

Fig. 3. ATS of the sentences
(6) *Virus se rychle šíří.*
(7) *Píše se článek.*

Apart from being a part of the predicate, the forms *se*/*si* function as a pronoun corefering with its antecedent (usually the subject of the sentence). In this case, the forms *se*/*si* are labelled by their respective syntactic function (mostly as an Object, (9)).

(9) Anna *se* *umyla*.
Anne *self* *washed*.
‘Anne washed herself.’

(iii) Modal and other compound predicates. A modal predicate consists of a modal verb, which expresses the modal meaning of the predicate, and the infinitive of a content verb, carrying the main lexical meaning of the expression as a whole (10). The large group of compound predicates further includes light verb constructions, copula verbs, idioms, etc.[9] They consist not only of verbal forms, but also of words belonging to other POS (especially nouns, but also adjectives, etc.; cf. (11) with the copula verb *to be*[10]).

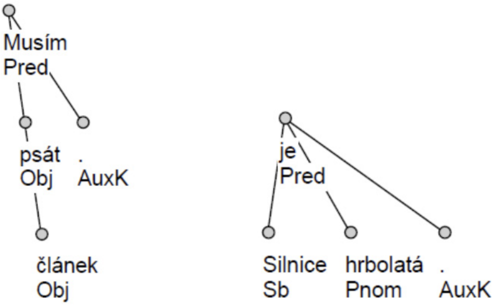

Fig. 4. ATS of the sentences
(10) *Musím psát článek.*
(11) *Silnice je hrbolatá.*

(10) *Musím* *psát* článek. (Fig. 4)
must *to-write* paper
‘I must write a paper.’

(11) Silnice *je* hrbolatá. (Fig. 4)
road *is* bumpy
‘The road is bumpy.’

[9] Light verbs (considered as lexically weak verbs with stress on the nominal part of the compound) and copula verbs seem to be good candidates for function words; there are, however, syntactic arguments against this approach; see 4.2.1.

[10] The verb *být* ‘to be’ has many functions. It is used as an auxiliary with complex verb forms (type (i); examples (3), (4), (5)); as a copula it connects the subject with the nominal part of the predicate (type (iii), (11)); or it has an existential meaning (e.g., *Moderní počítače jsou všude na světě*. ‘Modern computers are all over the world.’).

The three types of predicates expressed by multiword verb groups as described above have different structure in ATS. For each predicate (sub)type (i)-(iii), a decision had to be made which of the words is the head and which dependency relation will be assigned to the dependents. While having the content verb as the root would be the simplest rule, and a simple conversion would then exist to the shape of TGTS (Sect. 4), morphosyntactic considerations have prevailed, since the analytical level has been conceived as a syntactic counterpart of the tectogrammatical one. In the case of predicates, agreement has been chosen as the most important criterion with the goal of having a single, direct dependency edge between the two nodes being "in agreement." In Czech, agreement in number, gender, and person exists primarily between the subject and the predicate. Within the verb group, agreement features carried by subject can be expressed by the content verb as well as by the auxiliaries, in various combinations depending on mood, modality, tense, etc. In such cases, rules have been set to choose which word will be the head of the predicate group; typically, it is the word with the strongest agreement with the subject, with additional consistency considerations (see points (a)—(c) below) which might sometimes prevail (such as imperfective future tense, (4)).

With regard to the auxiliary verb *to be* in complex verb forms (type (i) above), the auxiliary is always captured as a dependent on the content verb ((3), (4), (5) above, (12) below). There are at least three arguments in favor of the placement of the content word as the head in the case of complex verb forms and for the treatment of auxiliaries as dependents on the content verb:

(a) the morphological information is included in the single auxiliary (e.g., in the future tense of imperfective verb (4)) or it is spread over the individual words that form the predicate (e.g., in the past tense (3), the agreement in person is between the subject and the auxiliary, while agreement in gender is with the content verb and agreement in number is with both; in the future tense of passive forms (12), the auxiliary *to be* is used in the appropriate form of the person and number, the agreement with gender and number is expressed within the past passive participle of the content verb). Sometimes it is duplicated (e.g., number in past tense (3) or gender and number in conditionals (5)).

(12) Snad *bude* *pozvána*.
Perhaps *will*-3rd-Sg *invited*-F-Sg
'Perhaps she'll be invited.'

(b) in the case of multiple auxiliaries, it would not be clear which of them competes for the position of the head of the whole complex verb group (cf. (5) with the complex forms consisting of three auxiliaries).
(c) there are specific cases where there is no auxiliary *to be*; compare examples for the past tense (3), where the auxiliary *to be* (*jsem* 'I am') for the 1st person is expressed, and (13) where there is no auxiliary *to be* for the 3rd person.

(13) *Psal* článek.
wrote-Sg-M paper
'He wrote a paper.'

For reflexive forms *se/si* (type (ii) above), the reflexive particle, regardless of its type and function within the verb group (AuxT or AuxR), is positioned as dependent on the root of the verb group, which is the content verb (in most cases; (6), (7)) or a modal or light verb (in case of the combination of types (ii) and (iii)). For modals, copulas, and light verbs (type (iii) above), according to the agreement principle, the finite verb form is made the head. The dependent part is assigned the function Obj in the case of modal predicate (10) and light verb constructions, and Pnom in the case where the head is a copula verb (11).

3.2.2 Prepositions

In the prepositional group, the preposition is the parent of the noun governed by it. In Czech, the preposition determines the case of the noun mostly unambiguously; with some prepositions the morphological case might have one of two values (e.g., *na* 'on' with Accusative (*dej to na stůl* 'put it on the table') and Locative (*visí to na stěně* 'it hangs on the wall')), and even one of three values (e.g., *za* 'behind/under' with Instrumental (*stojí za stromem* 'he stands behind the tree'), Accusative (*dal to za stůl* 'he put it behind the table') and with Genitive (*žil za komunismu* 'he lived under the communist régime')). The force of government required by the preposition is a strong argument in these cases. In ATS, a preposition is always labelled as AuxP. Such a standard representation is applied also to prepositions of a foreign origin (14).

There exist also multiword prepositions, called secondary prepositions (in traditional Czech grammars) due to the fact that their nucleus is constituted by a noun that stands on the boundary between content elements and function words. As this is a phenomenon concerning also other multiword expressions such as conjunctions, we devote a special section to it below (Sect. 3.2.6).

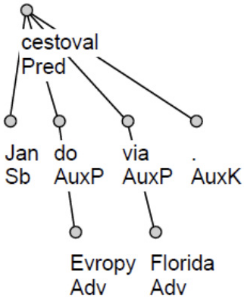

Fig. 5. ATS of the sentence
(14) *Jan cestoval do Evropy via Florida.*

(14) Jan cestoval do Evropy *via* Florida. (Fig. 5)
John travelled *to* Europe *via* Florida.
'John travelled to Europe via Florida.'

3.2.3 Subordinate conjunctions

The class of function words having traditionally the POS label of *conjunction* plays an important role in connecting subordinate clauses with their governing clause. In ATS, the subordinate conjunctions (labelled AuxC) are dependents of the predicate of their governing clause and they themselves govern the predicate of the dependent subordinate clause. This clause may obtain any syntactic function, such as Adverbial (15), Subject (16) or Object (17).

(15) Turisté šli na výlet, *přestože* pršelo. (Fig. 6)
'Tourists went for an excursion, *though* it was raining.'

(16) Není správné*, že* ho rodičům nepředstaví.
'It is not correct *that* she does not introduce him to her parents.'

(17) Oznámil přátelům, *že* se bude ženit.
‘He announced to friends *that* he will get married.’

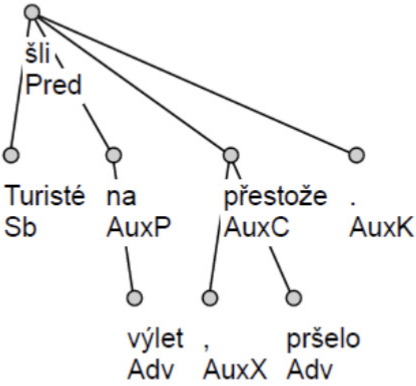

Fig. 6. ATS of the sentence
(15) *Turisté šli na výlet, přestože pršelo*.

Similarly to prepositions, subordinate conjunctions may also consist of more than a single word; for the treatment of such cases see Sect. 3.2.6.

Let us add that there is another type of subordinate clause that is attached to the governing clause by pronouns and adverbs such as *kdo* ‘who,’ *který* ‘who/which,’ *kdy* ‘when,’ *kam* ‘where’ (so called *wh*-sentences) rather than by a function word. This pronoun or adverb has its own function within the dependent clause in which it occurs; as such, it gets a label of the respective analytical function (see (18) and (19)).

(18) Jan navštívil kamaráda, *kterého* dlouho neviděl. (Fig. 7)
‘Jan visited a friend *whom* he had not seen for a long time.’

(19) Anna nevěděla, *kdy* pes zmizel. (Fig. 7)
‘Anne did not know, *when* the dog disappeared.’

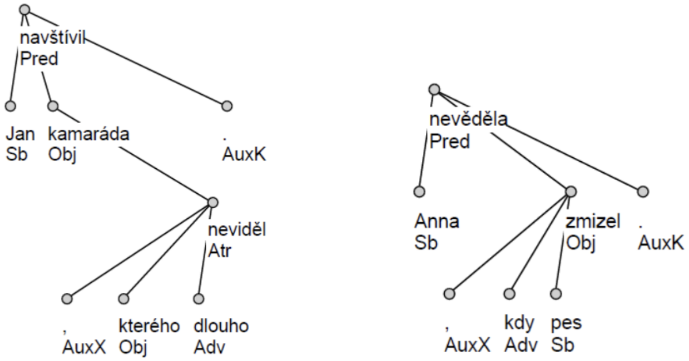

Fig. 7. ATS of the sentences
(18) *Jan navštívil kamaráda, kterého dlouho neviděl.*
(19) *Anna nevěděla, kdy pes zmizel.*

3.2.4 Coordinate conjunctions and conjunctions for apposition

Coordination is a complicated phenomenon for a description in any dependency (or for any other) formalism. Coordination may arise among members (constituents) of a sentence ((20), (21)), as well as among clauses or whole sentences (22); the parts of a coordinate structure are connected either by coordinate conjunctions ((20), (22)), or they are connected by a punctuation mark (21). Coordinate groups do not imply only two member structures (20), but the set of coordinated members could be theoretically non-finite.

In ATS, a coordinate conjunction (or just a punctuation mark) is captured as the head of the coordinated structure and gets the label Coord as a marker of the syntactic relation of coordination. The coordinated members are formally represented as dependents of the coordinate conjunction. The members of the coordinate group (having the *Coord*-labelled node as their parent) each carry the suffix _Co to mark their membership within the coordinate structure. The type of coordination (conjunction, adversative, exclusive, etc.) is (implicitly) reflected by the lexical part of the node Coord (*a* 'and,' *ale* 'but,' *nebo* 'or'). Cf. Fig. 8.

(20) Bratr *a* sestra opustili rodiče. (Fig. 8)
'Brother *and* sister left the parents.'

(21) Na podzim dozrávají jablka, hrušky, švestky.
'In Autumn apples, pears, plums ripen.'

(22) Otec sedí, čte noviny *a* poslouchá hudbu. (Fig. 8)
'Father is sitting, reading the newspaper *and* listening to music.'

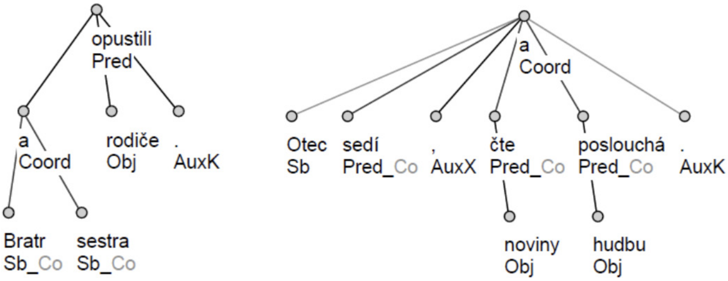

Fig. 8. ATS of the sentences
(20) *Bratr a sestra opustili rodiče.*
(22) *Otec sedí, čte noviny a poslouchá hudbu.*

In case members of the coordination share a dependent, the shared dependent is a child of the head of the coordinate structure. Its label does not carry the _Co suffix to signal its shared dependency relation to its sister nodes (cf. (22) demonstrating a coordination of three clauses with a shared subject).

The syntactic relation of apposition, as a phenomenon close to coordination structures, is represented in a similar way as coordination. The members of the apposition formally depend on the head of the apposition structure (for which we use the apposition conjunction) and they usually express the same function marked with the suffix _Ap. The head of the apposition structure is labelled Apos (23).

(23) Chová chobotnatce, *neboli* slony. (Fig. 9)
'He breeds proboscideans, *i.e.* elephants.

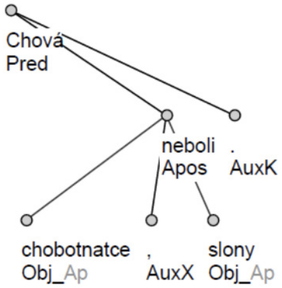

Fig. 9. ATS of the sentence
(23) *Chová chobotnatce, neboli slony.*

3.2.5 Borderline cases of function words

In addition to the relatively well-defined categories of function words within the nominal and verbal groups, there are several other cases that lack such a straightforward specification. They share the lexico-semantic characteristics of the content words, but at the same time their function in the sentence is somewhat auxiliary and difficult to specify in syntactic terms.

A primary example is the set of particles, the role of which is to emphasize some part of the sentence or the sentence as a whole; in the annotation, they obtain the label AuxZ. The list of particles belonging to this group is an open list. Most of them belong to the class of the so-called focalizers,[11] i.e., a fairly limited set of words foregrounding that part of the sentence that is in their semantic "scope." Several of these words are homonymous, i.e., the words may obtain, given their context, different (lexical) meanings and as such may be assigned different POS (Štěpánková 2014). For example, *až* may function as an adverb ('as far as'), a subordinating conjunction ('when') or a focalizer ('as far as'); *jen* may be classified as an adverb ('merely'), a particle (*jen se opovaž* 'just try'), or a focalizer ('only'); *i* may appear in the function of a coordinate conjunction ('and'), a focalizer ('only') or as a part of complex subordinate conjunctions (*i když* 'even when').

As mentioned above, focalizers by definition foreground a word or a part of the sentence that is in their scope and it is then natural that they are considered to be dependents of the words which they modify ((24) and (25)).

[11] Also called focussing adjuncts, rhematizers, focussing adverbials, emphasizing particles, focus sensitive particles, etc.

(24) Tom poznal *jenom* Marii. (Fig. 10)
‘Tom recognized *only* Mary.’

(25) Tom *jenom* tancoval. (Fig. 10)
‘Tom *only* danced.’

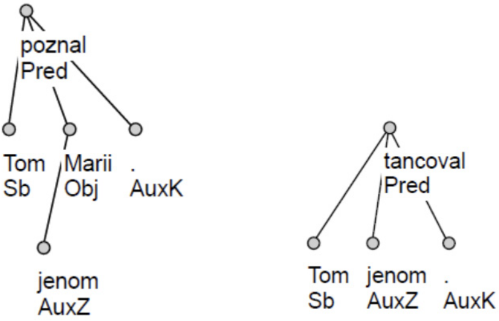

Fig. 10. ATS of the sentences
(24) *Tom poznal jenom Marii.*
(25) *Tom jenom tancoval.*

The situation gets rather complicated when the scope of a focalizer covers more than a single word. In case the focalizer modifies a whole nominal group or a whole subordinate clause, it is placed as a dependent on the head of the respective subtree; (26)-(28).

(26) *Jen* příliší optimisté vytrvali. (Fig. 11)
‘*Only* extreme optimists persisted.’

(27) Zprávy vyšly *aspoň* v první verzi.
‘The news was published *at least* in the first version.’

(28) Opustil dům, *právě* když začalo pršet. (Fig. 11)
‘He left the house *jus*t when it began to rain.’

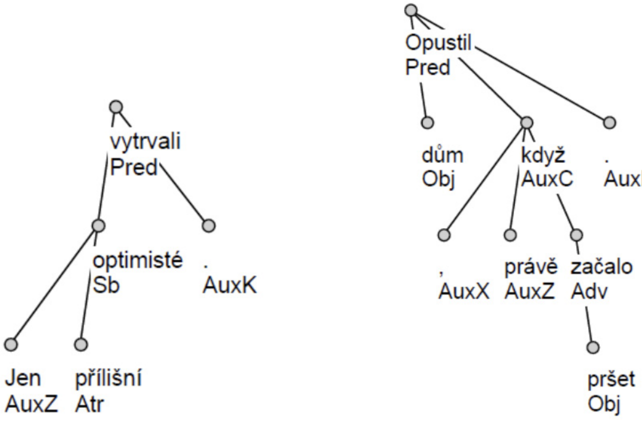

Fig. 11. ATS of the sentences
(26) *Jen přílišní optimisté vytrvali*.
(28) *Opustil dům, právě když začalo pršet*.

However, if there are, for example, multiple nominal groups in the scope of the focalizer that do not form a single subtree (cf. (64) in Sect 4.2.5), its scope cannot be adequately captured in ATS and it is delegated to the tectogrammatical level (as described in Sect 4.2.5).

3.2.6 Multiword function words

Prepositions (described in Sect. 3.2.2), subordinate (Sect. 3.2.3) and coordinate (Sect. 3.2.4) conjunctions, focalizers, and other particles (Sect. 3.2.5) may also have the form of a multiword expression. In such cases, the last word of the multiword function word is considered the governor and thus gets the label of the class of function words it belongs to and the remaining parts depend directly on it. These parts are assigned the label AuxY to indicate that they are parts of a multiword expression. Cf. (29) with a multiword preposition (*bez ohledu na* 'without regard to'), (30) with a coordinate conjunction (*buď-nebo* 'either-or') and (31) with a multiword focalizer (*přece jen* 'in the end').

(29) Odjeli *bez* *ohledu* *na* počasí. (Fig. 12)
left *without* *regard* *to* weather
'They left without regard to the weather.'

(30)	*Buď*	odejdi,	*nebo*	mlč! (Fig. 12)
	either	leave	*or*	shut-up

‘Either leave or shut up!’

(31)	*Přece*	*jen*	to	uhnilo. (Fig. 12)
	in-the-end	*even*	it	rotted-away

‘It has rotted away, in the end.’

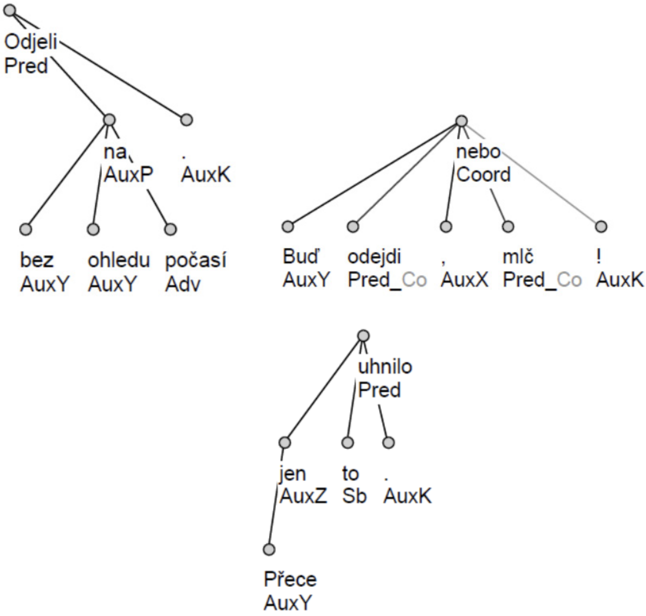

Fig. 12. ATS of the sentences
(29) *Odjeli bez ohledu na počasí.*
(30) *Buď odejdi, nebo mlč!*
(31) *Přece jen to uhnilo.*

3.2.7 Punctuation marks

The basic principle of ATS, namely to represent every word by a node of its own, applies not only to words (written in alphanumeric characters), but to all tokens, i.e., also to punctuation marks (and, for that matter, to all graphical symbols found within the sentence to be annotated). Although punctuation marks are not commonly understood as function words as they are not words in the true sense of the word, in some cases they play a role in the sentence comparable to function words.

In ATS, we distinguish three types of punctuation marks:

(i) terminal symbol of the sentence labelled AuxK, which is a child of the root of the tree.

(ii) comma; with the exception of the case in which a comma represents the head of a coordination or apposition (Sect. 3.2.4), its label is AuxX and it depends on the root of the subtree in which it appears or which is introduced or surrounded by comma(s).

(iii) other punctuation marks; with the exception of the case in which the mark is the head of a coordination or apposition (Sect. 3.2.4), its label is AuxG and it depends on the root of the subtree in which it appears or which is introduced or surrounded by it. Cf. (32) in which all types of punctuation marks are included.

(32) Zasadili: brambory, cibuli (máslovou dýni). (Fig. 13)
planted: potatoes, onion (butter pumpkin).
'They planted: potatoes, onions (butter pumpkin).'

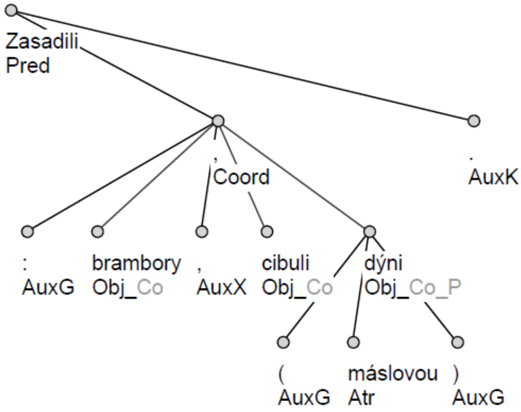

Fig. 13. ATS of the sentence
(32) *Zasadili: brambory, cibuli (máslovou dýni).*

4. Function words and tectogrammatical level

4.1 General remarks

As already mentioned in Sect. 2 above and illustrated by Fig. 1, the FGD approach is conceived of as a multilevel system, working with two syntactic levels, one corresponding to the surface structure of the sentence (called analytical), and the other representing the underlying (deep) syntactic sentence structure (called tectogrammatical). The tectogrammatical level represents a linguistically structured meaning of the sentence. These two levels reflect the relation of function and form as understood by the FGD. Such a multilevel approach allows for an appropriate reflection of the dual character of function words: on the one hand, as forms, they have their place in the surface shape of the sentence (as represented in ATS), on the other hand they contribute to the meaning of the sentence only in connection with another sentence element. As a consequence, function words in principle do not preserve their autonomous status on the tectogrammatical level (as represented in TGTS).

The contribution of function words basically concerns either the morphosemantic features of the related content word, or the syntactico-semantic relations between two content words in the sentence. In the former case, we speak about (morphological) grammatemes, in the latter about functors or, as the case may be, subfunctors. In PDT, both grammatemes and functors together with subfunctors are captured in the (complex) node labels of TGTS.

4.2 Reflection of function words on the tectogrammatical level

4.2.1 Morphosemantic features of predicates

On the tectogrammatical level, semantically relevant morphological features of the content words (such as number of nouns, tense of verbs) are stored in so-called grammatemes that are parts of the complex label of the respective node. The predicate is represented by one node representing the content verb and the semantic contribution of the function words that are parts of a multiword compound predicate is reflected in the grammatemes for modality and diathesis: *tense*, *factmod*, *deontmod* and *diatgram* (Panevová and Ševčíková 2010).[12]

[12] Predicates also have the grammateme of *aspect*. In this grammateme, the information about processual and complex meanings of the predicate is assigned with the possible values *proc* (corresponding to the imperfectivity of the verb), *cpl* (assigned to the perfective verbs), and *nr* (for the biaspectual verbs). Since *aspect*

The **tense** grammateme is a tectogrammatical correlate of the morphological category of tense of verbs, cf. (33)-(35) that are about the same activity but presented as simultaneous in (33), as preceding in (34), and as subsequent in (35). On the surface level, these morphological meanings are expressed by simple or complex verb forms of the past, future and present tense (Sect. 3.2.1, Fig. 2). The TGTS of the sentences (33)-(35) differ only in the value of the *tense* grammateme which is *sim* for (33), *ant* for (34), *post* for (35); cf. Fig. 14.

(33) *Píšu* článek. (Fig. 14)
write paper
'I *am writing* a paper.'

(34) *Psal* *jsem* článek. (Fig. 14)
wrote *am* paper
'I *wrote* a paper.'

(35) *Budu* *psát* článek. (Fig. 14)
will *to-write* paper
'I *will write* a paper.'

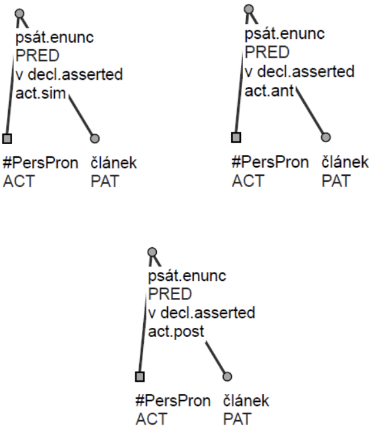

Fig. 14. TGTS of the sentences
(33) *Píšu článek.*
(34) *Psal jsem článek.*
(35) *Budu psát článek.*

is not expressed by means of function words in Czech, we consider it to be out of the scope of this article.

The **factmod** grammateme captures whether an activity is presented as given or as hypothetical. Such modal meanings are expressed by the morphological category of verbal mood in the surface structure of the sentence (Sect. 3.2.1). The indicative form of the verb (as in (33)-(35)) presents an activity as given (the value of the *factmod* grammateme is *asserted*). The conditional forms of verb express activities that could happen (*potential* value) or that are irreal (*irreal* value; (36)). The *appeal* value is for activities presented as requested (using an imperative form).

(36) *Byl bych býval psal* článek. (Fig. 15)
was would was wrote paper
'I *would have written* a paper (if I had had time).'

The **deontmod** grammateme is used to express the fact that the activity is understood as necessary, possible, permitted, etc. The value of the grammateme follows from the modal verb used (which has no separate node in TGTS), e.g., the *deb* value is used for an activity presented as necessary (expressed usually by the modal verb *muset* 'must,' (37), (39); cf. Sect. 3.2.1, Fig. 4), or the *vol* value is used for the event presented as wanted/intended (expressed typically by the modal verb *chtít* 'want'). If no marked modality is expressed (as in the (33)-(36) and (38)), the value of the *deontmod* grammateme is *decl*.

(37) *Musím psát* článek. (Fig. 15)
must to-write paper
'I must write a paper.'

The grammateme **diatgram** reflects the morphological meanings of active and passive voices, resultative, recipient, and dispositional diathesis, and of reflexive deagentive. For these meanings special syntactic requirements must be fulfilled, e.g., if the deagentive is expressed by the reflexive form *se* and the subject of the sentence is not the Actor of the event, the Actor is "general," cf. (38) and also (39), in which different morphological meanings are combined (cf. Sect. 3.2.1, Fig. 3).

(38) *Píše se* článek. (Fig. 15)
writes Refl paper
'The paper is being written.'

(39) *Bude se muset psát* článek. (Fig. 15)
will Refl must to-write paper
'A paper will have to be written.'

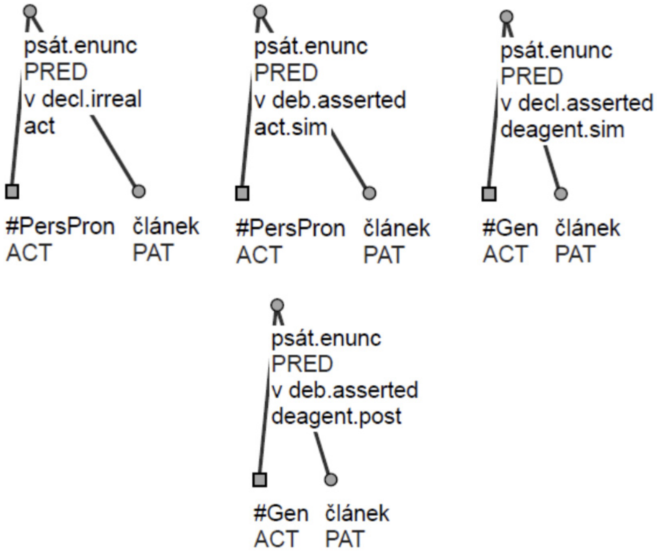

Fig. 15. TGTS of the sentences
(36) *Byl bych býval psal článek.*
(37) *Musím psát článek.*
(38) *Píše se článek.*
(39) *Bude se muset psát článek.*

There are several groups of multiword compound predicates consisting of e.g., copula verbs (40) or light verbs (cf. type (iii) in Sect. 3.2.1), where these elements can hardly be considered as function words. These multiword predicates cannot be captured as a single node with the semantic contribution of the hidden parts reflected in the grammateme values because they consist not only of verbal forms, but also of words belonging to other POS (especially nouns, but also adjectives, etc.). These components carry most of the semantics of the compound and as such they are separate nodes in TGTS. There are special valency frames for copula and light verbs in the valency lexicon (Kettnerová et al. 2017, Urešová 2011).

(40) Silnice *je* hrbolatá. (Fig. 16)
road *is* bumpy
'The road is bumpy.'

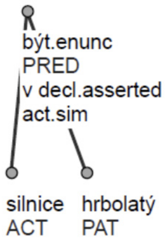

Fig. 16. TGTS of the sentence
(40) *Silnice je hrbolatá*.

4.2.2 Meanings of preposition groups

The semantic contribution of prepositions as well as of subordinate conjunctions (see below, Sect. 4.2.3) is reflected first of all in the values of the syntactico-semantic relations; these values are called functors and may be considered edge labels in the dependency trees. In the PDT annotation system, they are a part of the complex label of the dependent node.

As of the present status of the framework, about 60 functors are distinguished (the exact number may differ depending on whether we consider the present state-of-the art of the theoretical framework—the FGD, or its implementation in the current annotation scheme of the PDT). There are five functors for the so-called arguments (Actor, Patient, Addressee, Effect and Origin), determined as obligatory or optional and given in the valency frames of the particular lexical items, and several others for the so-called adjuncts ("free" modifications or adverbials),[13] which may be grouped into clusters having a general common semantic flavor such as temporal (distinguishing *when*, *since when*, *till when*, *how long*, *for how long*, *how often*, *during*, *from when*, *to when*), local (distinguishing *where*, *from where*, *which*

[13] The distinction between arguments and adjuncts roughly corresponds to Tesnière's (1959) distinction between *actants* and *circonstants*. For the criteria applied to this distinction in FGD, see Panevová (1974) and more recently Panevová et al. (2014). See also Lopatková et al. (2020) for the Czech valency dictionary and Urešová et al. (2021) for the PDT-based valency lexicon.

way, *where to*), causal (*cause*, *aim*, *condition*, *concession*, *intention*), manner and other related meanings (*accompaniment*, *beneficial*, *contradiction*, *comparison*, *criterion*, *difference*, *extension*, *heritage*, *means*, *regard*, *result*, *substitution*, *restriction*).

Some functors represent syntactico-semantic relations in a rather generalized way, which does not sufficiently capture the linguistically structured meaning as postulated for the tectogrammatical level. For example, the following modifications would all be classified as the same local (LOC) functor, as it corresponds to the general question *where*: *v domě* 'in the house,' *za domem* 'behind the house,' *před domem* 'in front of the house,' *blízko domu* 'near the house'; (41)-(43). In order to capture such more detailed meaningful distinctions, the functors have been subclassified into so-called subfunctors. The subfunctors thus offer a more specific choice of meaning within a single functor.

Thus, for example, in their detailed corpus-based study of local and temporal relations, Mikulová and Panevová (2021) postulate 25 subfunctors for the local functor LOC (*where*) some of which are illustrated here by (41)-(45), and 17 subfunctors for the temporal functor TWHEN (*when*) some of which are illustrated below by (46)-(49). With each example, we present the functor.subfunctor label and the typical forms, i.e., the prepositions and the morphological case(s) of the nouns within the prepositional phrases expressing the given syntactico-semantic relation.[14]

(41) Petr stojí *v domě / uvnitř domu*. (Fig. 17)[15]
'Peter is standing *in / inside the house*.'
`LOC.in`: *v*+6, *uvnitř*+2[16]

(42) Petr stojí *za domem*. (Fig. 17)
'Peter is standing *behind the house*.'
`LOC.behind`: *za*+7

[14] Note that functor.subfunctor meaning is expressed not only by prepositional cases, but can also be expressed by prepositionless cases of the noun.

[15] In the figures of TGTS, the subfunctor value is attached to the value of the functor.

[16] Here and in the following, we denote the morphological case of the given noun in the prepositional phrase by a number, in accordance with the Czech linguistic tradition, with 1 being Nominative, 2 Genitive, 4 Accusative, etc. We consider the prepositional case as a whole without distinguishing the semantic contribution of the preposition and the morphological case.

(43) Petr stojí *blízko / v blízkosti / nedaleko / opodál domu*. (Fig. 17)
'Peter is standing *near the house*.'
`LOC.near`: *blízko*+2, *v blízkosti*+2, *nedaleko*+2, *opodál*+2

(44) Pracuje *u divadla*.
'He works *at the theatre*.'
`LOC.at`: *u*+2

(45) *U divadla* nejsou žádná parkovací místa.
'There are no parking spots *by the theatre*.'
`LOC.by`: *u*+2

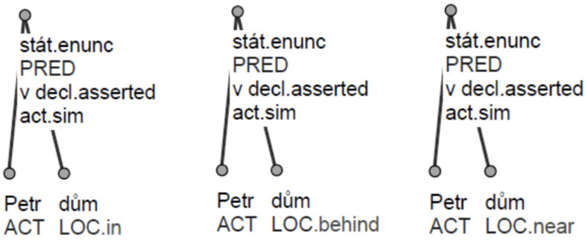

Fig. 17. TGTS of the sentences (41) *Petr stojí v domě/uvnitř domu*.
(42) *Petr stojí za domem*.
(43) *Petr stojí blízko / v blízkosti / nedaleko / opodál domu*.

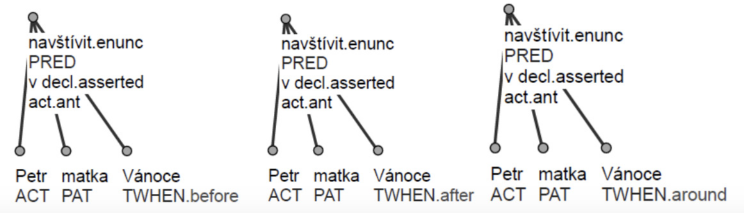

Fig. 18. TGTS of the sentences
(46) *Petr navštívil matku před Vánocemi*.
(47) *Petr navštívil matku po Vánocích*.
(48) *Petr navštívil matku okolo / kolem Vánoc*.

(46) Petr navštívil matku *před Vánocemi*. (Fig. 18)
'Peter visited his mother *before Christmas*.'
`TWHEN.before`: *před*+7

(47) Petr navštívil matku *po Vánocích*. (Fig. 18)
'Peter visited his mother *after Christmas*.'
`TWHEN.after`: *po*+6

(48) Petr navštívil matku *okolo / kolem Vánoc*. (Fig. 18)
'Peter visited his mother *around Christmas*.'
`TWHEN.around`: *okolo*+2, *kolem*+2

(49) Petr byl přes vánoce u matky.
'*During Christmas*, Peter was at his mother's.'
`TWHEN.during`: *přes*+4

Subfunctors are mostly assigned to functors expressing the meaning of adjuncts. The basic form of arguments (for example (usually) the Nominative case for Actor) is prescribed by the valency of the predicate and is given in the valency frame. However, there are also subfunctors assigned to the arguments indicating a quantitative specification of the relation expressed by the functor as demonstrated by (50)-(52) for the argument Actor.

(50) Demonstrace se zúčastnilo *tisíc* lidí.
'One *thousand* people took part in the demonstration.'
`ACT`: 1

(51) Demonstrace se zúčastnilo *na tisíc / okolo / kolem tisíce* lidí.
'*About a thousand* people took part in the demonstration.'
`ACT.approx`: *na*+4, *okolo*+2, *kolem*+2

(52) Demonstrace se zúčastnilo *přes tisíc* lidí.
'*More than a thousand* people took part in the demonstration.'
`ACT.more`: *přes*+4

As can be seen in these examples, prepositions play an important role for expressing distinct syntactico-semantic relations. It is a many-to-many relation: one preposition, or more precisely, one prepositional case, may express more than one subfunctor ((44) vs. (45) with the same preposition but different functor.subfunctor relation), and one subfunctor may be expressed by more than one prepositional case ((41), (43), or (48)).

4.2.3 Subordinate clauses

In TGTS, the semantic impact of subordinate conjunctions labelled on the analytical level as AuxC (Sect. 3.2.3) is reflected in a similar way as described above for the class of prepositions, namely by one of the functors accompanied, if needed, by a subfunctor. The functor of a clause which is an argument of the governing verb (traditionally called "content clause") follows from the valency frame of the governing verb (cf. (53) in which the clause attached by the conjunction *zda* 'whether' is in the position of the Patient of the governing verb *zeptat se* 'to ask').

(53) Zeptal se průvodce, *zda* je na zámku otevřeno.
'He asked the guide, *whether* the castle is open.'

Subfunctors are distinguished in the case of clauses that are in an adjunct position, cf. temporal clauses (55) and (56) which differ in the value of the subfunctor: in (55), there is the value *after* for the meaning "after the given time," and in (56), there is the value *as_soon_as* for the meaning "immediately after the given time"; the value of the functor is the same—TWHEN for the general temporal meaning "when."[17]

(54) Jan se oženil, *když* studoval univerzitu. (Fig. 19)
'John got married *when* he was studying (at the) university.'
`TWHEN.at`: *když*

[17] With each example, we present the functor.subfunctor label and a typical subordinate conjunction expressing the given syntactico-semantic relation.

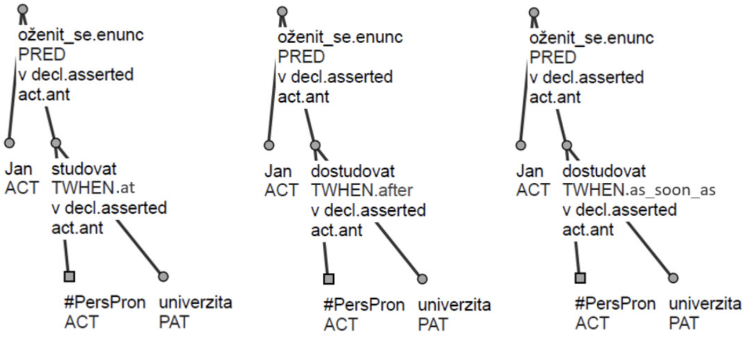

Fig. 19. TGTS of the sentences
(54) *Jan se oženil, když studoval univerzitu.*
(55) *Jan se oženil, když / poté co dostudoval univerzitu.*
(56) *Jan se oženil, jakmile dostudoval univerzitu.*

(55) Jan se oženil, *když / poté co* dostudoval univerzitu. (Fig. 19)
‘John got married *when* he finished his university studies.’
TWHEN.after: *když, poté co*

(56) Jan se oženil, *jakmile* dostudoval univerzitu. (Fig. 19)
‘John married *as soon as* he finished his university studies.
TWHEN.as_soon_as: *jakmile*

Similarly to prepositions, the relation between the particular conjunction and the meaning reflected in the functor.subfunctor value is a many-to-many relation: one conjunction may express more than one subfunctor (cf. (54) vs. (55) with the same conjunction but different functor.subfunctor relation), and one subfunctor may be expressed by more than one subordinate conjunction (55).

The variety of functors expressed by some subordinate conjunctions is very broad, therefore the formulation of contextual criteria is a very complicated task. A possible ambiguity of the conjunction can possibly be eliminated either based on the information from the valency frame of the governing verb (in the case of arguments; (53)) or using the contextual criteria given by the grammar (in the case of adjuncts). For example, in the case of the temporal conjunction *když* ‘when,’ the meaning “at the given time” (54) is conditioned by the presence of imperfective aspect with at least one of the predicates and the meaning "after the given time" (55) requires a perfective aspect of the dependent predicate.

To express the meaning in a clear way, dependent clauses are attached by various interchangeable multiword expressions the treatment of which in ATS is complicated (Sect. 3.2.6); however, in TGTS, sentences (57) and (58) (considered to have the same meaning) have the same representation as in Fig. 19 for sentence (54).

(57) Jan se oženil *tehdy*, *když* studoval univerzitu.
Jan Refl got-married *then* *when* studied university
'John got married *when* he was studying at university.'

(58) Jan se oženil *v době*, *když* studoval univerzitu.
John Refl got-married *in time* *when* studied university
'John got married *when* he was studying at university.'

4.2.4 Coordination and apposition

The representation of the coordinate structures in TGTS is similar to the representation in ATS (Sect. 3.2.4): a coordinate conjunction (or just a punctuation mark) is captured as the head of the structure, and the coordinated members formally depend on the head of the coordinate structure carrying the suffix _M (which corresponds roughly to the _Co and _Ap analytical function suffixes in ATS). However, unlike in ATS, in TGTS, the head of the coordinate structure does not have a unified Coord label, but the type of coordination relation (conjunction, adversative, exclusive, etc.) is distinguished by different labels (functors) attached to the formal head node of the coordinate structure.

Compare examples (59)-(61), in which two identical clauses are coordinated, but the type of coordination relation between them varies as reflected by the use of a different coordinate conjunction. In (59), the clauses are simply conjoined in a "logical conjunction" relation, in (60), there is an "adversative" relation and in (61), the clauses are have a form in a "consequence" relation. With each example, we present the functor label (describing the type of coordination relation) and some typical conjunctions expressing the given relation.

(59) Petr pracuje *a* Eva odpočívá. (Fig. 20)
'Peter works *and* Eva rests.'
CONJ: *a*

(60) Petr pracuje, *ale* Eva odpočívá. (Fig. 20)
'Peter works *but* Eva rests.'
ADVS: *ale, jenže, avšak*

(61) Petr pracuje, *a tak* Eva odpočívá. (Fig. 20)
'Peter works, *so* Eva rests.'
CSQ: *a tak, a tedy, a proto, tudíž*

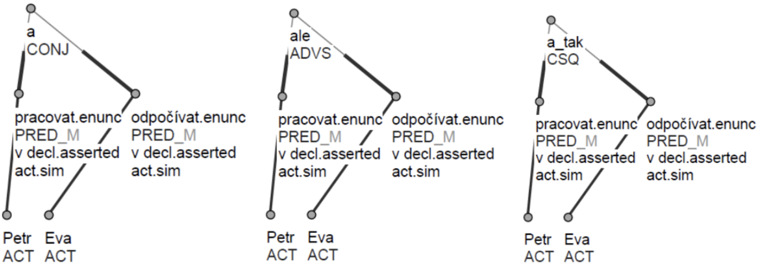

Fig. 20. TGTS of the sentences
(59) *Petr pracuje a Eva odpočívá.*
(60) *Petr pracuje, ale Eva odpočívá.*
(61) *Petr pracuje, a tak Eva odpočívá.*

The structure of the constructions with apposition is similar to that of coordinate constructions. The head of the apposition structure is assigned the functor APPS. Further distinctions, as to the type of the apposition, are not made (62).

(62) Chová chobotnatce, *neboli* slony. (Fig. 21)
'He breeds proboscideans, *i.e.* elephants.
APPS: *neboli, čili, tedy, to jest*

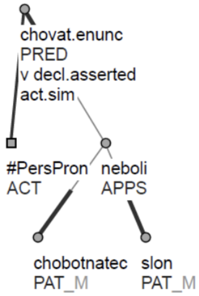

Fig. 21. TGTS of the sentence
(62) *Chová chobotnatce, neboli slony.*

4.2.5 Focalizers and other borderline cases

In Sect. 3.2.5 and elsewhere, we have referred to the heterogeneous character of function words. This character is reflected also in how we treat them on the tectogrammatical level. While some of them "disappear" as separate nodes from the TGTS dependency tree and their contribution to the meaning of the sentence is reflected in the complex labels of the respective content words, some keep their presence as nodes of their own. This is the case of the so-called focalizers (Sect. 3.2.5). Words belonging to this class (which contains function words as well as content words) have a specific function from the point of view of the information structure of the sentence.[18] The impact of focalizers on the meaning of the sentence is determined by their scope, i.e., the part of the sentence to which the focalizer applies. This is not an easy matter to decide since the scope of a focalizer need not directly correspond to its surface position and a single surface form can be semantically ambiguous between different scopes. The surface position, i.e., the word order, is only one possible indicator, the other one being prosody; first of all the placement of the intonation center indicated in the examples by capitals.

[18] The specific function of these words from the point of view of the bipartition of the sentence into theme and rheme (Topic and Focus) was noted first by Firbas (1957), who later called them "rhematizers." A detailed analysis of this function of focalizers is presented in Hajičová (1995, 2010). Some focalizers (especially *only, also, too, even*) have also been studied from the pragmatic and formal semantic point of view, see esp. Rooth (1985), who studied this class of words in relation to the prominence of the words that follow them. Since then, his approach has been followed by several specialists in formal semantics.

In order to capture the important consequences of the focalizers on the semantic interpretation of the sentence, they are represented by nodes labelled with the functor RHEM and their position in TGTS follows some fixed conventions, the most important of which is the principle that focalizers stand before the elements that are in their scope. More specifically, they are placed as the left sister of the first node of the subtree that is in the scope of the focalizer; (63)-(64).[19]

(63) Viděl to *jenom* JIRKA. (Fig. 22)
Saw it *only* Jirka
'It was seen only by JIRKA.'

(64) Jan dal *jenom* malý dárek MAMINCE. (Fig. 22)
John gave *only* small present mother
'Jan gave only a small gift to his MOTHER.'

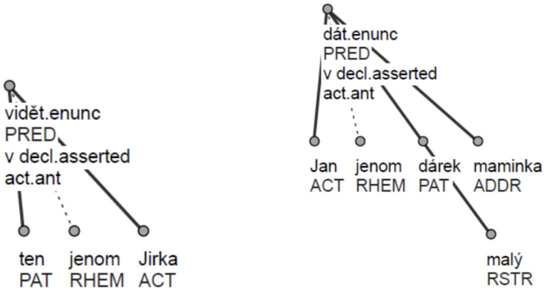

Fig. 22. TGTS of the sentences
(63) *Viděl to jenom Jirka.*
(64) *Jan dal jenom malý dárek mamince.*

This basic principle of the representation of focalizers together with the ordering of nodes in the TGTS reflecting the topic-focus articulation makes it possible to capture important consequences of the placement of the focalizers in TGTS for the semantic interpretation of the sentence (see the possible different interpretations of the sentences (65) with meaning "among other things she did," (66) with meaning

[19] In TGTS, the edge between a node with the functor RHEM and its mother node determines the position of the focalizer within TGTS and defines its scope. The edge is not actually a "true" dependency, which is indicated by the dotted line in the visualization.

“among other things she has cleaned” and (67) with the assumed intonation center on Eva and with meaning “Eva, in addition to other people”) and also makes it possible to adequately capture the scope of focalizers in cases where there are multiple nodes (or subtrees) in the scope (cf. (64) with meaning “gave nothing to nobody else“).

(65) Eva *také* vyčistila BOTY. (Fig. 23)
Eva *also* cleaned shoes
‘Eva also cleaned SHOES.’

(66) Eva vyčistila *také* BOTY. (Fig. 23)
Eva cleaned *also* shoes
‘Eva cleaned also SHOES.’

(67) *Také* EVA vyčistila boty. (Fig. 23)
also Eva cleaned shoes
‘Also EVA cleaned shoes.’

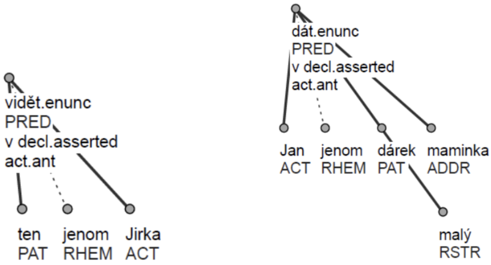

Fig. 23. TGTS of the sentences
(65) *Eva také vyčistila boty.*
(66) *Eva vyčistila také boty.*
(67) *Také Eva vyčistila boty.*

Similar properties as focalizers can be observed with function words that modify the whole utterance, such as *asi*, *snad* (meaning: *perhaps, maybe*, labelled by the functor MOD). The position of such an element in the sentence has a similar semantic consequence as a focalizer and has to be accounted for in an adequate way ((68) and (69)).

(68) Otec *asi* přijede domů zítra.
‘Father *perhaps* will arrive home tomorrow.’

(69) Otec přijede domů *asi* zítra.
'Father will arrive home *perhaps* tomorrow.'

(70) *Jenom* se opovaž!
'*Just* dare!'

(71) Můžeme *tedy* očekávat, že následky pandemie se ukážou i za několik let.
'*Thus* we can expect that the consequences of the pandemic will be evident even after several years'

There are also other more or less individual cases of function words that preserve their status as separate nodes in order to capture their semantic or contextual impact, such as those indicating some kind of evaluative or emotional attitude of the speaker (labelled ATT, (70)) or elements that signalize a continuity with the preceding context (labelled PREC, (71)).

4.2.6 Multiword expressions

As stated in Sect. 3.2.6 above, in ATS, each part of a multiword expression is represented by a node of its own. This not the case in TGTS, where the treatment of the multiword expression differs according to the type of the expression. Prepositions and subordinate conjunctions, be they single- or multiword expressions, do not have a counterpart in TGTS and their contribution to the meaning of the sentence is reflected in the complex labels of the respective content word (Sect. 4.2.2 and 4.2.3). The function words in verb groups are reflected in the morphological grammatemes of the content verb (Sect. 4.2.1).

Coordinate conjunctions, focalizers, and other particles have their counterparts in the corresponding TGTS; in case they are multiword, they are represented by a single node and placed in the structure as if they were a single-element units, cf. (72) with a multiword coordinate conjunction and (73) with a multiword focalizer (compare with Fig. 12 in Sect. 3.2.6).

(72) *Buď* jdi *nebo* mlč! (Fig. 24)
either go *or* shut-up
'Either leave or shut up!'

(73) *Přece* *jen* to uhnilo. (Fig. 24)
in-the-end *even* it rotted-away
'It has rotted away, in the end.'

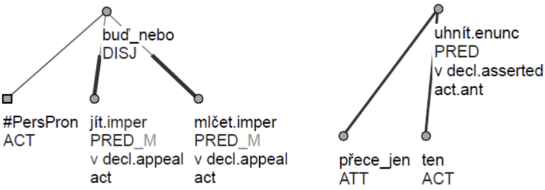

Fig. 24. TGTS of the sentences
(72) *Buď jdi, nebo mlč!*
(73) *Přece jen to uhnilo.*

As mentioned in Sect. 3.2.1, in one of the meanings of the reflexive forms *se*/*si* (labelled AuxT) these reflexives co-form a verbal lexical unit with the verbs they are attached to. In TGTS, these reflexive forms have no counterparts of their own and are represented as a part of the verb lemma, cf. Fig. 25 for the sentence (74) and the ATS for the same sentence (quoted as (6) in Sect. 3.2.1) in Fig. 3.

(74) Virus *se* rychle *šíří*. (Fig. 25)
Virus *Refl* quickly *spreads*.
'Virus *spreads* quickly.'

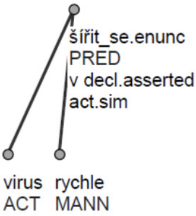

Fig. 25. TGTS of the sentence
(74) *Virus se rychle šíří.*

4.2.7 Contribution of punctuation to the meaning of the sentence

Punctuation marks present in ATS are not usually represented in any way in TGTS: there is no node that corresponds to them and they do not affect attribute values. However, there are several cases in which

a punctuation mark (i) replaces a (content) word and (ii) contributes to the meaning of a sentence. In the former case, the punctuation mark is captured as a separate node, in the latter, the contribution to the meaning of the sentence is reflected in some attribute.
(i) A punctuation mark is captured as a separate node in TGTS especially if it fulfills the function of the head of the coordination or apposition structure (75) (cf. Sect. 3.2.7., Fig. 13). Dashes and colons sometimes serve as a copula verb (76).

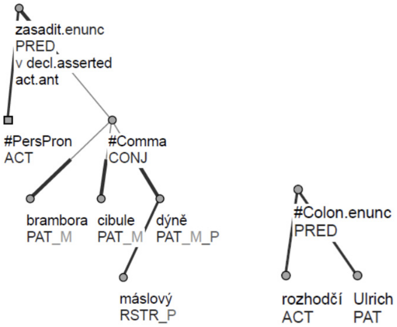

Fig. 26. TGTS of the sentence
(75) *Zasadili: brambory, cibuli (máslovou dýni).*
(76) *Rozhodčí: Ulrich.*

(75) Zasadili: brambory, cibuli (máslovou dýni). (Fig. 26)
‘They planted: potatoes, onion (butter pumpkin).’

(76) Rozhodčí: Ulrich. (Fig. 26)
‘Judge: Ulrich.’

(ii) There are several attributes in TGTS whose values are fully or partially determined by the presence and/or type of punctuation mark—parentheses indicate which words are a part of a parenthetical expression; parenthetical construction is marked in TGTS with the suffix _P attached to all nodes representing the parenthesis (75). Similarly, the terminal symbol of the sentence affects the value of the *sentmod* attribute. This attribute contains the information on sentential modality. The exclamation modality (value *excl*) or interrogative mood (value *inter*) can be often determined only by the terminal punctuation mark; (77) vs. (78).

(77) Vyhráli jsme! (Fig. 27)
‘We won!’

(78) Vyhráli jsme? (Fig. 27)
‘Did we win?’

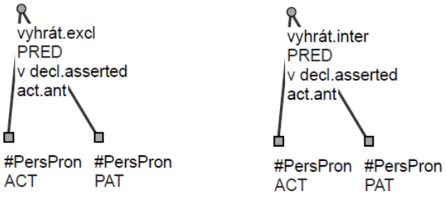

Fig. 27. TGTS of the sentences
(77) *Vyhráli jsme!*
(78) *Vyhráli jsme?*

5. Conclusion

In this article we present the treatment of so-called function words within the framework of the dependency-based Functional Generative Description proposed in Prague by Petr Sgall and his team and its reflection in the Prague Dependency Treebank, an original annotated corpus of Czech. Both the theoretical framework and the treebank work with a stratificational model of language, a part of which are two levels of dependency-based syntactic structure—one oriented towards the syntactic structure of the sentence on the surface level called *analytical,* and the other oriented towards the underlying, deep sentence structure called *tectogrammatical*.

Such a two-level approach makes it possible to make the distinction between content words and function words quite explicit. The dependency tree structure of the sentence on the analytical level contains all the words present in the sentence as separate nodes. A distinction is made between different classes of function words, the main boundary being between the function words occurring within verb groups (i.e., auxiliaries), and those being parts of nominal groups (prepositions) or connecting clauses (or, in some cases, parts of clauses) into one whole, i.e., a complex sentence (conjunctions). This distinction is substantiated by the status of the function words within the language system: those within verb groups contain information about the morphosemantic properties of verbs, while those within nominal groups or connecting clauses express relations between heads and their dependents. This is the reason why in the annotation of the Prague Dependency Treebank, auxiliaries are (mostly) considered to be dependents on the verb that is their governor and to which they "belong," and prepositions and conjunctions, on the contrary, are considered to be the heads of the nouns or clauses the form of which they it may be said,"control" or "govern." Auxiliaries, prepositions and conjunctions are the most pronounced classes of function words, though there are some groups of words such as particles that stand on the borderline between function words and content words and to which we also pay attention in this study.

A different situation exists on the tectogrammatical level. By definition, the dependency representation of a sentence on this level is conceived of as a linguistically structured meaning of the sentence and as such the dependency tree contains only content words at its nodes. The semantic contribution of the function words to the meaning of the sentence is not lost; it is reflected by information attached to the nodes of the tree in the form of complex labels, namely as grammatemes covering the morphosemantic properties of verbs and functors and the subfunctors covering the information on the semantics of the syntactic relations expressed by prepositions and (subordinate) conjunctions.

We believe that such a two-level representation offers a consistent way to represent function words that captures both their morphosyntactic as well as syntactico-semantic contribution to the structure of the sentence and its meaning.

Acknowledgements

The work described herein has been supported by the LINDAT/CLARIAH-CZ Research Infrastructure (project No. LM2023062 funded by the Ministry of Education, Youth and Sports of the Czech Republic), and by the Czech Science Foundation grants No. GX20-16819X and GA23-05238S.

Works cited

1. Bresnan, J., Ash, A., Toivonen, I. and S. Wechsler. 2016. *Lexical-Functional Syntax*. 2nd Edition. Chichester: Wiley-Blackwell.
2. Firbas, J. 1957. K otázce nezákladových podmětů v současné angličtině. Příspěvek k teorii aktuálního členění větného. Časopis pro moderní filologii 39: 22-42, 165-173. An abbreviated and modified version of this contribution was published as Non-thematic Subjects in Contemporary English, *Travaux Linguistiques de Prague* 2. 1966, 239-256.
3. Hajič, J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová,* ed. by Hajičová, E., 106-132. Prague: Karolinum.
4. Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Štěpánek, J., Pajas, P. and J. Kárník. 1999. *Annotations at Analytical Level. Instructions for annotators*. Prague: Institute of Formal and Applied Linguistics, Charles University. (Also available at https://ufal.mff.cuni.cz/pdt-c/documentation and as part of https://hdl.handle.net/11234/1-3185.)
5. Hajič, J., Hajičová, E., Mikulová, M., Mírovský, J., Panevová, J. and D. Zeman. 2015. Deletions and node reconstructions in a dependency-based multilevel annotation scheme. In *Lecture Notes in Computer Science*, *16th International Conference on Computational Linguistics and Intelligent Text Processing*, 17-31, Berlin / Heidelberg: Springer.
6. Hajič, J., Hajičová, E., Mikulová, M. and J. Mírovský. 2017. Prague Dependency Treebank. In *Handbook of Linguistic Annotation*, ed. by Ide, N. and J. Pustejovsky, 555-594. Dordrecht: Springer.
7. Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J. and B. Štěpánková. 2020a. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 5208-5218, Marseille: European Language Resources Association.
8. Hajič, J., Bejček, E., Bémová, A., Buráňová, E, Fučíková, E., Hajičová, E., Havelka, J., Hlaváčová, J., Homola, P., Ircing, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mareček, D., Mikulová, M., Mírovský, M., Nedoluzhko, A, Novák, M., Pajas, P., Panevová, J., Peterek, N., Poláková, L, Popel, M., Popelka, J., Romportl, J., Rysová, M., Semecký, J., Sgall, P., Spoustová, J., Straka, M., Straňák, P., Synková, P., Ševčíková, M., Šindlerová, J., Štěpánek, J.,

Štěpánková, B., Toman, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š. and Z. Žabokrtský. 2020b. *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*. Data/software, Prague: LINDAT/CLARIAH-CZ digital library, Institute of Formal and Applied Linguistics, Charles University, URL: https://hdl.handle.net/11234/1-3185.

9. Hajičová, E. 1995. Postavení rematizátorů v aktuálním členění věty [The Position of Rhematizers in the Information Structure]. *Slovo a slovesnost* 56: 241-251.
10. Hajičová, E. 2010. Rhematizers Revisited. *Philologica Pragensia* 20(2): 57-70.
11. Hajičová, E, Havelka, J., Sgall, P., Veselá, K. and D. Zeman. 2004. Issues of Projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics* 81:5-22.
12. Hajičová, E., Mikulová, M. and J. Panevová. 2015. Reconstructions of Deletions in a Dependency-Based Description of Czech: Selected Issues. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015),* 131-140, Uppsala.
13. Havelka, J. 2005. Projectivity in totally ordered rooted trees: An alternative definition of projectivity and optimal algorithms for detecting non-projective edges and projectivizing totally ordered rooted trees. *The Prague Bulletin of Mathematical Linguistics* 84: 13–30.
14. Havelka, J. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 608–615, Prague.
15. Kettnerová, V., Kolářová, V. and A. Vernerová. 2017. Deverbal Nouns in Czech Light Verb Constructions. In *Lecture Notes in Computer Science, Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017. London, UK*, 205-219 Cham: Springer.
16. Lopatková, M., Kettnerová, V., Vernerová, A., Bejček, E. and Z. Žabokrtský. 2020. *Vallex 4.0*. Data/software, LINDAT/CLARIAH-CZ digital library, Prague: Institute of Formal and Applied Linguistics, Charles University, URL: http://hdl.handle.net/11234/1-3524.
17. Marcus, S. 1965. Sur la notion de projectivite. *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik* 11:181-192.

18. Marneffe de, M.-C, Manning, C. D., Nivre, J. and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2): 255-308.
19. Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. New York: State University of New York Press.
20. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K. and Z. Žabokrtský. 2006. *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual*. Technical report no. TR-2006-30, Prague: Institute of Formal and Applied Linguistics, Charles University. (Also available at https://ufal.mff.cuni.cz/pdt-c/documentation and as part of https://hdl.handle.net/11234/1-3185.)
21. Mikulová, M. and J. Panevová. 2021. *Formy a funkce okolnostních určení v češtině. Určení časová a prostorová* [Forms and Functions of Adverbials in Czech. Temporal and Local Adverbials]. Prague: Institute of Formal and Applied Linguistics, Charles University.
22. Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D. and D. Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659-1666, Portorož: European Language Resources Association.
23. Osborne, T. and K. Gerdes. 2019. The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics* 4(1): 17-28.
24. Panevová, J. 1974. Verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics* 22, Part I: 3-40.
25. Panevová, J., Hajičová, E., Kettnerová, V., Kolářová, V., Lopatková, M., Mikulová, M. and M. Ševčíková. 2014. *Mluvnice současné češtiny. 2. Syntax češtiny na základě anotovaného korpusu* [Grammar of Present Day Czech. 2. Syntax of Czech Based on an Annotated Corpus]. Prague: Karolinum.
26. Panevová, J. and P. Karlík, 2016. Reflexívní sloveso [Reflexive verb]. In *Nový encyklopedický slovník češtiny*, ed. by Karlík, P., Nekula, M. and J. Pleskalová, 1536-1541, Prague: Lidové noviny.
27. Panevová, J. and M. Ševčíková. 2010. Annotation of Morphological Meanings of Verbs Revisited. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 1491–1498, Valletta: European Language Resources Association.

28. Rooth, M. 1985. *Association with Focus*. PhD Thesis. Amherst: Univ, of Massachusetts.
29. Sgall, P. 1967. *Generativní popis jazyka a česká deklinace* [Generative Description of Language and Czech Declension]. Prague: Academia.
30. Sgall, P. and E. Hajičová. 1987. The Ordering Principle. *Journal of Pragmatics* 11: 435-454.
31. Sgall, P., Nebeský, L., Goralčíková, A. and E. Hajičová. 1969. *A Functional Approach to Syntax in Generative Description of Language*. New York: American Elsevier Publ. House.
32. Sgall, P., Hajičová, E. and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Prague-Dordrecht: Academia-Reidel.
33. Štěpánková, B. 2014. *Aktualizátory ve výstavbě textu, zejména z pohledu aktuálního členění* [Focalizers in the Structure of Text, Especially from the Point of View of Topic-Focus Articulation]. Prague: Institute of Formal and Applied Linguistics, Charles University.
34. Tesnière, L. 1959. Élements de syntaxe structurale. Paris: Klinksieck. [English translation: Tesnière, L. 2015. Elements of Structural Syntax (translated by Osborne, T. and S. Kahane). Amsterdam: John Benjamins].
35. Urešová, Z. 2011. *Valence sloves v Pražském závislostním korpusu* [The valency of verbs in the Prague Dependency Treebank]. Prague: Institute of Formal and Applied Linguistics, Charles University.
36. Urešová, Z., Bémová, A., Fučíková, E., Hajič, J., Kolářová, V., Mikulová, M., Pajas, P., Panevová, J. and J. Štěpánek. 2021. *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.0*. Data/software, Prague: LINDAT/CLARIAH-CZ digital library, Institute of Formal and Applied Linguistics, Charles University, URL: http://hdl.handle.net/11234/1-3499.
37. Veselá, K., Havelka, J. and E. Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of Coling 2004*, 289-295, Geneva.
38. Zeman, D. 1998. A Statistical Approach to Parsing of Czech. *The Prague Bulletin of Mathematical Linguistics* 69: 29-37.
39. Zeman, D. 2004. *Parsing with a Statistical Dependency Model*. PhD Thesis, Prague: Charles University.

List of abbreviations

ATS	Analytical Tree Structure
FGD	Functional Generative Description
PDT	Prague Dependency Treebank
TGTS	Tectogrammatical Tree Structure
1^{st}	1st person
3^{rd}	3rd person
F	Feminine
M	Masculine
POS	Part of speech
Refl	Reflexives *se*, *si*
Sg	Singular

List of labels used in the figures (also in text)

Labels in ATS:

Adv	Adverbial
_Ap	Member of apposition structure
Apos	Head of apposition structure
Atr	Attribute
AuxC	Subordinate conjunction
AuxG	Other punctuation mark
AuxK	Terminal symbol of sentence
AuxP	Preposition
AuxR	Reflexive as indication of diathesis
AuxT	Reflexive as a part of verb
AuxV	Auxiliary verb *to be*
AuxX	Comma
AuxY	Part of multiword expression
AuxZ	Particle
_Co	Member of coordination structure
Coord	Head of coordination structure
Obj	Object
_P	Head of parenthesis
Pnom	Nominal part of copula predicate
Pred	Predicate
Sb	Subject

Labels in TGTS:

act	Active voice
ACT	Actor
ADDR	Addressee
ADVS	Adversative
after	“After the given time”
ant	Preceding (anterior) activity
appeal	Activity presented as requested
approx	“Approximately”
APPS	Apposition
around	“Around the given place/time”
asserted	Activity presented as given
as_soon_as	“Immediately after the given time”
at	“At the given place/time”
ATT	Attitude
before	“Before the given time”
behind	“Behind the given place”
by	“In the immediate vicinity of the given place”
CONJ	Conjunction
CSQ	Consequence
deagent	Diathesis of deagentive
deb	Activity presented as necessary
decl	Unmarked modality
DIR3	Spatial modification where-to
during	“During the given time”
EFF	Effect
#Gen	General participant
enunc	Indicative mood
excl	Exclamation
imper	Imperative modality
in	“Inside the given place”
inter	Interrogative mood
irreal	Irreal activity
near	“Near the given place”
LOC	Spatial modification where
_M	Member of coordination/apposition
MOD	Expression of modality
more	“More than the given amount”
ORIG	Origin
P	Member of parenthesis

PAT	Patient
#PersPron	Personal pronoun
post	Subsequent (posterior) activity
potential	Activity that could happen
PREC	Expression referring to the preceding text
PRED	Predicate
RHEM	Rhematizer (focalizer)
RSTR	Restrictive adnominal modification
sim	Simultaneous activity
TWHEN	Temporal modification when
v	Verb
vol	Activity presented as wanted