Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 70–78
Marseille, 24 June 2022

Advantages of a complex multilayer annotation scheme: The case of the Prague Dependency Treebank

Eva Hajičová, Marie Mikulová, Barbora Štěpánková, Jiří Mírovský
Institute of Formal and Applied Linguistics
Computer Science School, Faculty of Mathematics and Physics, Charles University, Prague
{hajicova,mikulova,stepankova,mirovsky}@ufal.mff.cuni.cz

Abstract

Recently, many corpora have been developed that contain multiple annotations of various linguistic phenomena, from morphological categories of words through the syntactic structure of sentences to discourse and coreference relations in texts. Discussions are ongoing on an appropriate annotation scheme for a large amount of diverse information. In our contribution we express our conviction that a multilayer annotation scheme offers to view the language system in its complexity and in the interaction of individual phenomena and that there are at least two aspects that support such a scheme: (i) A multilayer annotation scheme makes it possible to use the annotation of one layer to design the annotation of another layer(s) both conceptually and in a form of a pre-annotation procedure or annotation checking rules. (ii) A multilayer annotation scheme presents a reliable ground for corpus studies based on features across the layers. These aspects are demonstrated on the case of the Prague Dependency Treebank. Its multilayer annotation scheme withstood the test of time and serves well also for complex textual annotations, in which earlier morpho-syntactic annotations are advantageously used. In addition to a reference to the previous projects that utilise its annotation scheme, we present several current investigations.

Keywords: complex language description, multilayer annotation scheme, linguistically-based pre-annotation, linguistically-based checks, corpus-based study

1. Introduction

One of the aims of modern linguistic studies is to describe and explain the collection of language phenomena as a structured whole and at the same time to understand this structured whole as a functioning means of communication. In this context, several concepts of function should be distinguished; in one of these interpretations, function is opposed to form, which comes close to Saussure's binary understanding of sign (Saussure, 1916). This interpretation offers a basis for understanding language as a set of levels, which gave rise to several descriptive frameworks, from the original stratificational grammar of Lamb and Newell (1966) through Halliday's systemic grammar (Halliday, 1970) to Sgall's Functional Generative Description (Sgall, 1967; Sgall et al., 1986) or Mel'chukovian Meaning-Text Model (Mel'chuk, 1988), to name just a few that refer to strata or levels explicitly, and leaving aside those which acknowledge the existence of units with different status without giving them specific names (as is e.g. the case of the so-called construction grammar).

Following the multistratal descriptions of the language, various multilayer annotation schemes have been proposed, the purpose of which is to take into account multifarious linguistic phenomena from morphological categories of words through the syntactic structure of sentences to discourse and coreference relations in texts and other semantic features (such as temporal or spatial annotation), which allow for an assignment of labels to tokens, groups of tokens, sentences and entire sections of the raw texts.

In the present paper, we want to substantiate our conviction that such a complex annotation scheme offers to view the language system in its complexity, in the interaction of individual phenomena and thus contributes to the theoretical studies of this system. All aspects supporting a multilayer annotation scheme are demonstrated on the case of the Prague Dependency Treebank based on the language description framework known as Functional Generative Description. The annotation scheme of the Prague Dependency Treebank is described in Sect 3. In Sect. 4, annotation related aspects of a multilayer annotation scheme are demonstrated. The possibility to base linguistic research on a corpus search on more than a single layer of annotation has led to a series of studies which are introduced in Sect. 5.

2. Related Work

Recently, there has been an increased interest in the development of multilayer corpora, e.g. Groningen Meaning Bank (Bos et al., 2017) based on Discourse Representation Theory (Kamp, 1984; Kamp and Reyle, 1993) and Combinatory Categorial Grammar (Steedman, 2001), Manually Annotated Sub-Corpus (Ide, 2017) based on Linguistics Annotation Framework (Ide and Romary, 2004), Georgetown University Multilayer Corpus (Zeldes, 2017), OntoNotes (Hovy et al., 2006; Pradhan and Ramshaw, 2017), AnCora-UPF corpus (Mille et al., 2013), which is based on the Meaning-Text Model mentioned above. An overview of recently developed multilayer corpora with a proposal of a multilayer semantic annotation scheme is most recently presented by Silvano et al. (2021).

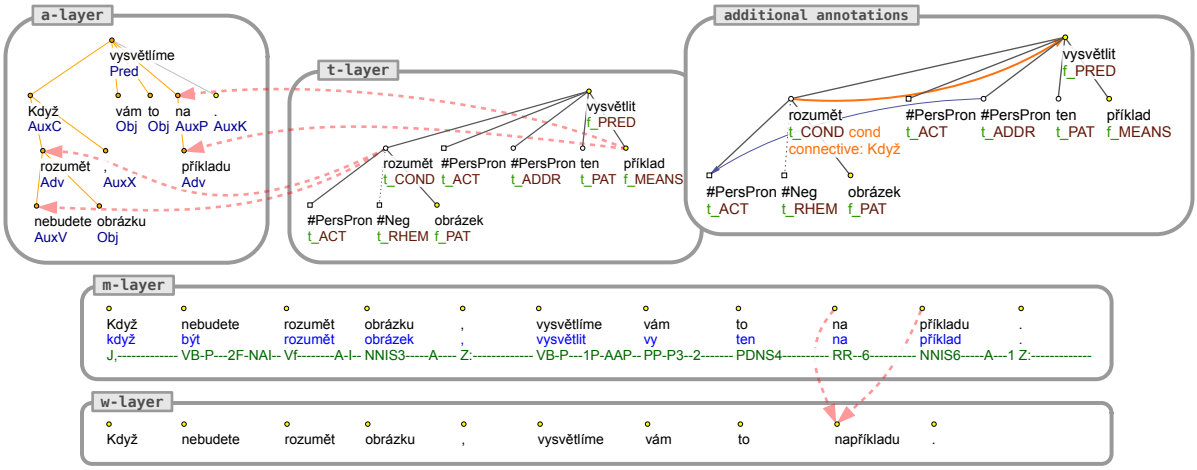

Figure 1: Multilayer annotation scheme of the PDT-treebank

In principle, a multilayer corpus is such a corpus that "contains mutually independent forms of information, which cannot be derived from one another reliably" (Zeldes, 2018). However, Ide et al. (2017) note that there are different types of multilayer annotation schemes. The layers may be defined in an independent way or a single scheme may be used that integrates all the layers; each layer may point directly into raw data, or each layer may define independently its units by referring to tokens in raw text, or to other units on some other annotation layer. The decision to represent annotation layers in this way does not automatically lead to a distribution of layers in separate data files or to a visualization of layers in separate graphs, etc.

3. The case of a complex multilayer annotation scheme: Prague Dependency Treebank

In the paper, we outline and illustrate the advantages of a multilayer corpus based on the hierarchical architecture of several annotation layers as applied in the family of the Prague Dependency Treebank (PDT). In order not to lose any piece of the original information, units (tokens, nodes) at a lower layer are explicitly referred to from the corresponding closest (immediately higher) layer. These links allow for tracing every unit of annotation all the way down to the original raw text. Thus, an annotation layer can provide information both about raw data and also about other annotations.

The annotation scheme of the PDT is inspired by and theoretically rooted in the stratificationally and dependency-based Functional Generative Description of language as proposed by Petr Sgall in the sixties (Sgall, 1967) and then developed and enriched by his students and followers up to the present time. The annotation scheme was first introduced at the end of the nineties of the last century (cf. (Hajič, 1998), more recently (Hajič et al., 2017; Hajič et al., 2020a) and the annotation manuals presented on the project web site[1]). The hierarchical multilayer architecture of PDT-annotation scheme is schematically illustrated in Fig. 1 on the example of the Czech sentence:

Když nebudete rozumět obrázku, vysvětlíme vám to například.
When you-will-not understand picture we-will-explain you it onexample.
'If you do not understand the picture, we will explain it to you on an example.'

In Fig. 1, each layer of the system is indicated by a separate box. The links between the layers are indicated by the red arrows. In fact, all nodes/tokens are linked. The original raw text is stored at the lowest layer of the system (**w-layer** box in Fig. 1). At this **raw text layer**, the text is segmented into documents and paragraphs and individual tokens are assigned unique identifiers.
Above the raw text layer, there are the following layers of annotations:

- **morphological layer** (**m-layer** box in Fig. 1): all tokens from the raw text (including punctuation marks) get a lemma and a (disambiguated) morphological tag and they are linearly structured. Also, typos and similar errors are corrected here. As we can see in Fig. 1, there is a typo in the original sentence: the preposition *na* 'on' is not separated – as it should be – from the following noun *příkladu* 'example' and at the m-layer, a correction is realized.
- **analytical layer** (**a-layer**) capturing surface syntactic dependency structure in the shape of a tree with the specification of the head for each node and the assignment of a syntactic function (so-called *afun*) that denotes the relation between the dependent node and its head (e.g. subject (`Sb`),

[1] `https//ufal.mff.cuni.cz/pdt-c/documentation`

object (Obj), adverbial (Adv)). In the PDT-style surface syntactic annotation, every token from the raw text of a sentence (including punctuation marks – cf. nodes for comma (AuxX) and terminal symbol of the sentence (AuxK) in Fig. 1) is represented by a node of the tree and at the same time, no additional nodes are allowed. The linear ordering of nodes corresponds to the word order of the tokens in the sentence.

- **tectogrammatical layer** (**t-layer**) representing the deep syntactic structure. The deep syntactic relations are captured by the so-called *functors*; cf. the value PRED for predicate, ACT for actor, PAT for patient, COND for adverbial with condition meaning, etc. in Fig. 1. The tectogrammatical dependency structure of a sentence consists of nodes only for the content (lexical) words; function words such as prepositions, subordinating conjunctions, auxiliary verbs, etc. are not present as separate nodes, their contribution to the meaning of the sentence is captured within the complex labels of the content words. Thus, there is for example only one node for the prepositional phrase *na příkladu* 'on example' in the tectogrammatical tree in Fig. 1. The red arrows indicate the links between the nodes at the tectogrammatical layer and corresponding nodes at the analytical layer. At the tectogrammatical layer, new nodes are also established for semantic units deleted on the surface; in Fig. 1 the restoration of deletions is illustrated by the *#PersPron* nodes for the Actors of the predicates in the main and dependent clause.

  The green values t and f (in front of the functor values) stand for topic-focus articulation: t is for contextually bound nodes and f for contextually non-bound nodes. The ordering of nodes corresponds to the information structure of a sentence (cf. different position of pronouns *vám* 'you' and *to* 'it' at the analytical and tectogrammatical layer in Fig. 1.)

- The last box attached to the t-layer box in Fig. 1 indicates **additional annotations** such as textual coreference, bridging and discourse relations and other properties of the sentence such as genre specification, name entities which are technically also captured at the tectogrammatical layer of annotation. However, these phenomena are not a part of the tectogrammatical layer in the sense of the theoretical framework of Functional Generative Description. The additional annotation is exemplified here by the orange link between the predicates of the main and the dependent clause and is labelled as a discourse relation of condition, and by the blue coreferential link between the actor of the predicate *rozumět* 'understand' and the addressee of the predicate *vysvětlit* 'explain'.

In the paper, we use the data of the Prague Dependency Treebank (PDT) sub-corpus published within the consolidated release of the PDT-treebanks of Czech texts PDT-C 1.0 (Hajič et al., 2020),[2] and the Prague Czech-English Dependency Treebank PCEDT 2.0 (Hajič et al., 2012).[3] The whole Prague Dependency Treebank - Consolidated consists of over 2.2 million tokens, or 175 thousand sentences. The PDT sub-corpus consists of 675 thousand tokens, or 49 thousand sentences annotated at all annotation layers. The parallel corpus PCEDT sized over 1.2 million tokens in almost 50 thousand sentences for each part.
The corpora are encoded in the Prague Markup Language data format, PML (Pajas and Štěpánek, 2008), and the research was performed in the PML application framework: tree editor TrEd[4] for browsing and editing the PML data, btred for applying Perl scripts to the data and Prague Markup Language - Tree Query, PML-TQ (Pajas and Štěpánek, 2009), as a powerful graphically oriented query system.[5]

4. Annotation related aspects of a multilayer annotation scheme

A multilayer annotation scheme makes it possible to use relevant features of the existing annotation of a given layer to design a scheme for the annotation of other (not necessarily neighbouring) layers both conceptually and in a form of an automatic pre-annotation procedure (Sect. 4.1) or in a form of annotation checking rules (Sect. 4.2).

4.1. Automatic pre-annotation based on cross-layer relations

In the theoretical framework we subscribe to, information structure is considered to be a semantically relevant phenomenon and as such it should be represented at the layer of sentence meaning (tectogrammatical, in our terms). However, it has its reflection in the surface shape of the sentence, be it word order, prosody or similar means. This approach has led us to the idea to formulate and test a **pre-annotation module of information structure** in the PCEDT treebank assigning the features of contextual boundness (*t* for contextually bound nodes and *f* for contextually non-bound nodes) from which the global division of the sentence into its Topic and Focus can be derived based on several features present in the annotation of sentences on some of the lower layers (Mírovský et al., 2013). The pre-annotation procedure was able to mark over 40% of the text and the results of the application of such a pre-annotation procedure were evaluated face-to-face a sample of manually annotated sentences and the results were very encouraging: the average success rate was over 96%.

[2] https://ufal.mff.cuni.cz/pdt-c
[3] https://ufal.mff.cuni.cz/pcedt2.0/
[4] https://ufal.mff.cuni.cz/tred/
[5] http://ufal.mff.cuni.cz/pmltq

The benefits of tectogrammatical dependency structure used in the PDT for annotating language phenomena that cross the sentence boundary, namely coreference and bridging relations were also studied, described and applied by Nedoluzhko and Mírovský (2013). The authors use the detailed representation of deletions (important esp. in pro-drop languages), information on the deep syntactic relations (functors), and syntactic decisions for coordination and apposition structures at the tectogrammatical layer. These features make possible **to code basic coreference relations** in cases that are not so easy when annotating on the raw texts.

In distinction to methods based on raw texts, (Jínová et al., 2012a; Jínová et al., 2012b) formulated an automatic pre-annotation **procedure determining discourse relations, connectives and their arguments** directly on tectogrammatical trees, based on the assumption that certain syntactic features of a sentence analysis correspond to certain discourse-layer features. Hence, the authors looked for a possible analogy between intra-sentential syntactic relations already annotated in the corpus and intra-sentential discourse relations. An ideal case for the automatic detection based on tectogrammatical functors were those that directly correspond to some discourse type; this was the case e.g. of the functors of `REAS` (reason), `CSQ` (consequence), and `CAUS` (cause) all indicating the discourse relation reason-result; also the temporal functors were used as signals of a certain discourse relation. The scope of the discourse arguments was identified on the basis of the tectogrammatical tree structures. As a result, 9,991 tectogrammatical dependencies were converted into discourse relations, along with all properties of the relations (i.e. the position of arguments, the discourse type and the connective).

4.2. Automatic annotation checking rules based on cross-layer relations

Currently, the Prague team is in the process of extending the fully manual mid-layer syntax annotation to all parts of the PDT-C. To get a higher quality and a greater consistency of annotated data, a set of automatic checking procedures has been proposed and created in accordance with the annotation guidelines and incorporated into the annotation process to prevent the annotators from making accidental mistakes. When building the annotation rules, we also utilize the multilayer structure of PDT and use relevant information from the finished manual annotation of the lower morphological layer.

There are two main groups of rules which exploit the already established morphological tagging. The first group includes rules that take advantage of the fact that some surface syntactic functions (afuns) strictly correspond to the word-type of the word or token (part of speech (abbreviated POS), or type of punctuation), which is contained in the morphological tag of the corresponding node at the morphological layer, cf. the following examples:

- Prepositions are assigned afun `AuxP`: a node with the afun `AuxP` corresponds to a node with the tag for preposition at the morphological layer (the tag has the letter `R` in the first position).
- Subordinate conjunctions are assigned afun `AuxC`: a node with the afun `AuxC` corresponds to a node with the tag for subordinate conjunction at the morphological layer (the tag has the letters `J,` in the first two positions).
- Auxiliary verbs are assigned afun `AuxV`: a node with the afun `AuxV` corresponds to a node with the tag for verb at the morphological layer (the tag has the letter `V` in the first position).
- Punctuation marks are assigned afun `AuxX` (in case of comma), `AuxG` (in case of colon, slash, bracket, etc.), or `AuxK` (in case of final punctuation mark): a node with the afun `AuxX`, `AuxG` or `AuxK` corresponds to a node with the tag for a non-alphanumeric character at the morphological layer (the tag has the letter `Z` in the first position).

The second group of rules is based on the fact that dependency relations are defined with respect to the POS characteristics of the head node (e.g. attribute depends on a noun), see the following examples:

- Attribute (afun `Atr`) never depends on a verb, i.e. the corresponding node at the morphological layer does not have a tag for a verb (the tag with the letter `V` in the first position).
- Adverb (afun `Adv`) never depends on a noun, i.e. the corresponding node at the morphological layer does not have a tag for a noun (the tag with the letter `N` in the first position),
- Nominal part of a predicate (`Pnom`) depends on the verb *být* 'to be', i.e. the corresponding node at the morphological layer has the lemma *být*.

Annotators run the checking procedure after annotation of every single tree and consequently check and fix possible errors.[6] The experiment we performed at the beginning of the annotation project (Mikulová et al., 2022) showed that the control rules considerably contribute to the quality of the resulting annotation. Moreover, the rules formulated on the basis of current knowledge about language not only contribute to the improvement of annotation, but also point to insufficiently described phenomena and refine knowledge of language.

[6] A similar automatic linguistically-based (rule-formulated) checks have been used with advantage also in previous annotation projects (for example, the annotation at the tectogrammatical layer of PCEDT corpus; (Mikulová and Štěpánek, 2010)).

5. Corpus studies based on features across the layers

The possibility to base linguistic research on a corpus search on more than a single layer of annotation has led to a number of findings that take advantage of the multilayer annotation scheme in PDT-corpora to more accurately describe language phenomena. E.g., a valency dictionary of Czech (Urešová et al., 2021) is based on the relation between tectogrammatical, morphological and analytical annotation. It not only describes the participants of predicates, but also provides a detailed list of their formal realizations including surface syntactic structure. A lexicon of Czech discourse connectives (Mírovský et al., 2021) also draws on information from multiple layers. Similarly, a detailed description of the forms and functions of adverbials of place (Mikulová et al., 2017) and time (Panevová and Mikulová, 2020) exploit the cross-layer information of the adverbials. Different principles of annotation at the syntactic layers (see Sect. 3) are also an invaluable basis for examining the surface deletion (Hajičová et al., 2015). A monograph devoted to the syntax of Czech was also created on the basis of the PDT (Panevová et al., 2014).

A substantial part of the cross-layer studies examines the information structure and discourse structure in relation to the form of its expression at the analytical layer and at the tectogrammatical layer, also in comparison of Czech and English (in the PCEDT). There belongs e.g. our study on the variability of the position of adverbials of time and place in Czech and English word order and their function in the information structure of the sentence (Hajičová et al., 2019).

The annotation of Czech and English at the analytical and tectogrammatical layers has also allowed us to examine the morpho-syntactic properties of three representatives of the class of the so-called focalizers, namely *only*, *also* and *even*, and their word order position in English compared with Czech and their semantic scope vis-à-vis the information structure (Hajičová and Mírovský, prep).

The corpus served as a basis for another study of the contribution of expressions having the functor of focalizers at the tectogrammatical layer to the discourse structure as discourse connectors and for determination which kinds of discourse relations they indicate (Hajičová et al., 2020).

In the following two sections, we present two recent case studies as an illustration of the use of a multilayer annotation for a comprehensive description of a linguistic phenomenon performed on the data presented in Sect. 3. In the first study (Sect. 5.1), the syntactic annotations are used for a more precise part-of-speech determination of uninflected word types. Sect. 5.2 contains a comparative corpus study of Czech and English with regard to the relation between the analytical syntactic functions, the surface word order and the information structure as captured at the tectogrammatical layer.

5.1. Case study I: Distinguishing homonymous uninflected words

One of the traditional language phenomenon that combines morphological, syntactic and semantic features is the part of speech category (POS). In our study, we focussed primarily on uninflected word types and tried to show how sentence representation at the different layers can be useful for their better description, classification and annotation.

At the morphological layer, all tokens of a sentence are traditionally assigned the part of speech value within the morphological tag (e.g. `Dg` for adverbs forming negation and degrees of comparison, `Db` for the other adverbs, `J` for conjunctions, `R` for prepositions, `T` for particles). We are aware that from the strictly morphological point of view a subcategorization of uninflected words might be considered questionable because in case of uninflected words the morphological criterion is rather irrelevant; moreover, together with the frequent homonymy of these words such a subcategorization leads to issues concerning disambiguation. However, there are several practical reasons for POS tagging of uninflected words: The most important one is that the structure and also the content of the PDT-C morphological layer is unified with the MorfFlex - the Morphological Dictionary of Czech (Hajič et al., 2020b). MorfFlex, among other things, serves for tagging and lemmatization of other synchronic corpora of Czech, which have a one-layer structure and therefore the morphological tag is their main/only source of linguistic information.

Thanks to the fact that each syntactic layer of the PDT scheme captures different aspects of syntactic behaviour of a word, we postulated a hypothesis that the annotation of values of an afun (at the analytical layer) and of a functor (at the tectogrammatical layer) might be helpful in the disambiguation of homonymous uninflected words at the morphological layer.

We identified 12 types of homonymous uninflected POS combinations at the morphological layer. The most frequent are homonyms used as a preposition and an adverb (30 different words; e.g. *kolem*: *šel kolem domu* 'he walked *around* the house' (preposition) vs. *šel kolem* 'he walked *by*' (adverb)), then a non-graded adverb and a particle (25 words; e.g. *hned*: *přijď hned* 'come *immediately*' (adverb) vs. *hned ze dvou důvodů* '*even* for two reasons' (particle)), graded adverb and particle (14 words; e.g. *prostě*: *oblékl se prostě* 'he dressed *plainly*' (adverb) vs. *prostě to udělej* '*just* do it' (particle)), and a conjunction and a particle (11 words; e.g. *přece*: *prší, a přece šli* 'it's raining and *yet* they did go' (conjunction) vs. *tomu přece nevěříš* '*surely* you don't believe that' (particle)). Due to the lack of distinctive features, the most problematic are the words used as adverbs and particles, in contrast to prepositions and adverbs, which differ in the presence of a valency potential, or in contrast to particles and conjunctions, which usually differ in the position in a sentence.

Functor	Afun	Tag	Freq
RHEM	AuxZ	TT----------	247
CM	AuxZ	TT----------	63
CM	Apos	TT----------	3
CM	Adv	TT----------	1
ATT	AuxZ	TT----------	1
RHEM	Adv	TT----------	1

Table 1: Annotation of the homonymous particle *zejména* 'particularly' at the three layers of PDT

Functor	Afun	Tag	Freq
ATT	AuxY	Dg-------1A----	33
ATT	Adv	Dg-------1A----	18
ATT	AuxY	TT----------	15
ATT	Adv	TT----------	3
MANN	Adv	Dg-------1A----	3
ATT	AuxZ	Dg-------1A----	1

Table 2: Annotation of the homonymous particle *prostě* 'plainly, just, simply' at the three layers of PDT

At the analytical layer, particles are assigned the afun AuxY (for modifying words) or the afun AuxZ (for emphasizing words) and at the tectogrammatical layer, these words are assigned the functor RHEM or CM (for words with a rhematizing function), PREC (for words linking the sentence to its preceding context), MOD (for words expressing modality), and ATT (for words expressing attitude). Thus we expect that a co-occurrence of a "particle" afun and a "particle" functor with a single word in a text indicates also the particle value of the POS at the morphological layer (T). Cf. the results of annotation of a non-homonymous particle *zejména* 'particularly' in Tab. 1. We can observe that with the exception of two errors (the combination of "particle" functors RHEM or CM with "adverb" afun Adv), the tectogrammatical functor, analytical afun, and morphological tag are consistent.

As an example of homonymous particle annotation see the analysis of the words *prostě* and *hned*. The word *prostě* is understood as a graded adverb Dg (plainly) or a particle T (just). According to the annotated data, the frequency of the particle *prostě* is quite higher, there are only 3 examples of the adverb *prostě* in the annotated dataset (cf. Tab. 2). The combination of the tectogrammatical functor ATT (attitude) and the analytical afun AuxY (modifying word) seems to be annotated correctly (ex. (1)), as well as the combination of functor MANN (adverbial of manner) and afun Adv (adverbial) function (2).

(1) ATT+AuxY: Padělek chtěl mladík *prostě* vyměnit.
'The boy *just* wanted to exchange the fake.'

(2) MANN+Adv: Veronique, *prostě* oblečena, típla cigaretu.
'Veronique, *plainly* dressed, lit the cigarette.'

These two cases represent two main meanings of the word *prostě* (its particle and adverb function). However, in Tab. 2, we can observe that the POS annotation at the morphological layer is not consistent with the annotation at the syntactic layers. The combination of the functor ATT (attitude) and the afun Adv (adverbial) represents transitional cases (3); combination of functor ATT and afun AuxZ is a mistake made by an annotator at the analytical layer. (Mistakes of this kind should be detected using the automatic checking rules described above in Sect. 4.2.)

(3) ATT+Adv: Roku 1981 předvedla světu svůj první osobní počítač, nazvaný *prostě* IBM PC.
'In 1981, it introduced to the world first personal computer, called *simply/just* IBM PC.'

From this, we can deduce: if a word *prostě* is annotated at the syntactic layers by a combination of the functions ATT and AuxY, then the word *prostě* belongs to the POS of the particle (T) at the morphological layer; if a word *prostě* is annotated at the syntactic layers by a combination of functions MANN and Adv, then the word *prostě* belongs to the POS of the adverb (Dg) at the morphological layer.

Functor	Afun	Tag	Freq
RHEM	AuxZ	Db----------	58
TWHEN	Adv	Db----------	36
RHEM	Adv	Db----------	13
TWHEN	AuxZ	Db----------	9
RHEM	AuxZ	TT----------	2

Table 3: Annotation of the homonymous word *hned* 'immediately, even' at the three layers of the PDT

Similarly, the word *hned* is considered an adverb Db with a temporal or spatial meaning (immediately, soon) and an emphasizing particle T (even, right). The PDT data show that in most cases there is a match between the assignment of the given functor and the assignment of the afun (cf. Tab. 3) with 60 matches of the particle combination of functor RHEM (rhematizer) and afun AuxZ (emphasizing word; (4)), and 36 matches of the combination of functor TWHEN (temporal adverbial answering the question "when?") and the afun Adv (adverbial; (5)). The other combinations point again to questionable cases.

(4) RHEM+AuxZ: Spolupráce pediatra s obchodníkem je výhodná *hned* z několika důvodů.
'The cooperation of a pediatrician with a businessman is beneficial *even* for several reasons.'

(5) TWHEN+Adv: Přináší blaho *hned*, hoře z něj později.
'It brings happiness *immediately*, grief later.'

The analysis of words *prostě* and *hned* shows that manual annotation at the two syntactic layers can be relevant for the tagging of the POS information at the morphological layer. While the annotation at the morpho-

logical layer appears to be quite inconsistent, in most cases there is an agreement in the annotation of functions at the syntactic layers. A minority of cases where there is no consistence between functions at the two syntactic layers points to a transitional area between the two possible POS values. In these cases, the determination of the POS is always questionable.

5.2. Case study II: Information structure and its expression in Czech and English

At present, our attention is focussed on a comparative corpus study of Czech and English with regard to the relation between the analytical syntactic functions, the surface word order and the information structure as captured at the tectogrammatical layer. It is commonly assumed by traditional comparative grammars of Czech and English (see e.g. Dušková et al. (1971) following up Mathesius (1947) pioneering observations) that if the information structure of sentences in Czech and English is to be preserved (which is a precondition consistent with the assumption of the semantic relevance of information structure we subscribe to), the grammatically fixed English word order (the preverbal position of subject, in our case) and the relatively free word order in Czech (the subject may principally occur in any word order position) makes it necessary to use some means other than word order to express information structure in English.

For our study, we used the part of the parallel English-Czech corpus PCEDT containing 267 documents with 4,826 sentences annotated also for information structure. We searched for Czech sentences in which the noun with the afun `Sb` (subject) at the analytical layer was placed after the governing predicate and at the tectogrammatical layer it was annotated as belonging to the Focus. There were 328 cases fulfilling these conditions. We then searched in the English part of the corpus for corresponding (i.e. aligned) English sentences and we have found 133 cases in which the noun corresponding to the subject in the search in the Czech part was also placed after the predicate though in any syntactic function and at the tectogrammatical layer annotated as belonging to the Focus.

Based on this search, we have found that there were three most frequent English constructions corresponding to the Czech subject in Focus: (i) the use of passive voice (6), (ii) the use of a *there*-construction (7), and (iii) the use of a different verb allowing for a change of the Czech subject into another analytical afun that can be placed in English after the predicate (8). In addition, our material has allowed us to identify some special contexts where the subject was placed after the verb in English. This was the case of structures in which the predicate is the verb *to be* ((9) and (10)).

(6) Most of the picture *is taken up* with endless *scenes* of many people.
Většinu filmu *zabírají* nekonečné *scény* velkého množství lidí.

(7) Moreover, *there have been* no *orders* for the Cray-3 so far.
Kromě toho *nepřišly* dosud na Cray-3 žádné *objednávky*.

(8) Besides Messrs. Cray and Barnum, other management at the company *includes* Neil Davenport.
Kromě pánů Craye a Barnuma *je* hlavní řídící pracovník ve společnosti Neil Davenport.

(9) Behind all the hoople *is* some heavy-duty competition.
Za vším tímto nadšením *je* velmi tvrdá soutěž.

(10) The one *character* at least somewhat interesting *was* Irving Louis *Lobsenz*.
Jednou alespoň trochu zajímavou *postavou je* Irving Louis *Lobsenz*.

It is important to note that besides these special contexts we have found no case in which the two languages would differ in the information structure with regard to the postverbal placement of the nominal subject in Czech at the analytical layer and its functioning as (a member of) Focus at the tectogrammatical layer.

The observation presented in this case study offers an additional support for our thesis that information structure is a semantically relevant phenomenon, which may be expressed on the surface structure of sentences by different means such as word order (esp. in the so-called free word order languages), prosody (the position of the intonation centre or the intonational contour), or special constructions (such as the particles *wa* and *ga* in Japanese, or the above mentioned *there*-construction in English). The nature and the extent of the use of these means depend largely on the type of language concerned.

6. Conclusion

In our contribution, we argue that a complex multilayer annotation of a corpus provides an invaluable resource for both an in-depth study of different language phenomena in their relationships as well as for the automatic pre-annotation and checking procedures. We have supported these claims by several case studies based on the multilayer annotation scheme of the Prague Dependency Treebank.

This scheme is a hierarchical architecture of several annotation layers where units (tokens, nodes) at a lower layer are explicitly referred to from the corresponding higher layer. These links allow for tracing every unit of annotation all the way down to the original raw text. The annotation scheme was introduced at the end of the nineties of the last century, and to this day, the annotation scheme has proven to withstand the test of time: it can serve very well for complex annotations of discourse and coreference relations over whole texts, in which earlier morpho-syntactic annotations are advantageously used.

7. Acknowledgements

The research and language resources reported in the paper have been supported by the LINDAT/CLARIAH-CZ project funded by Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101) and by the project No. GX20-16819X funded by the Grant Agency of the Czech Republic.

8. Bibliographical References

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The Groningen Meaning Bank. In N. Ide et al., editors, *Handbook of linguistic annotation*, pages 463–496. Springer, Dordrecht.

Dušková, L., Caha, J., and Bubeníková, L. (1971). *Stručná mluvnice angličtiny [A concise grammar of English]*. Academia, Praha.

Hajič, J., Hajičová, E., Mikulová, M., and Mírovský, J. (2017). Prague Dependency Treebank. In N. Ide et al., editors, *Handbook of linguistic annotation*, pages 555–594. Springer, Dordrecht.

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., and Štěpánková, B. (2020a). Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France, May. European Language Resources Association.

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., and Štěpánková, B. (2020b). *MorfFlex CZ*. LINDAT/CLARIAH-CZ, Prague, URL: http://hdl.handle.net/11234/1-3186.

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

Hajičová, E. and Mírovský, J. (prep). Focalizers through the lens of a parallel english–czech corpus. a case study.

Hajičová, E., Mikulová, M., and Panevová, J. (2015). Reconstruction of deletions in a dependency-based description of czech: Selected issues. In Eva Hajičová et al., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 131–140, Uppsala, Sweden. Uppsala University, Uppsala University.

Hajičová, E., Mírovský, J., and Rysová, K. (2019). Ordering of adverbials of time and place in grammars and in an annotated english–czech parallel corpus. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 51–60, Paris, France. Université Paris Sorbonne Nouvelle, Association for Computational Linguistics.

Hajičová, E., Mírovský, J., and Štěpánková, B. (2020). Focalizers and discourse relations. *The Prague Bulletin of Mathematical Linguistics*, (115):187–197.

Halliday, M. A. (1970). Language structure and language function. *New horizons in linguistics*, 1:140–165.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, Stroudsburg.

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225.

Ide, N., Chiarcos, C., Stede, M., and Cassidy, S. (2017). Designing annotation schemes: From model to representation. In N. Ide et al., editors, *Handbook of linguistic annotation*, pages 73–111. Springer, Dordrecht.

Ide, N. (2017). Case study: The manually annotated sub-corpus. In N. Ide et al., editors, *Handbook of Linguistic Annotation*, pages 497–519. Springer, Dordrecht.

Jínová, P., Mírovský, J., and Poláková, L. (2012a). Analyzing the most common errors in the discourse annotation of the Prague Dependency Treebank. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 127–132. Edicoes Colibri, Lisboa.

Jínová, P., Mírovský, J., and Poláková, L. (2012b). Semi-automatic annotation of intra-sentential discourse relations in PDT. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, pages 43–58. Coling 2012 Organizing Committee, Mumbai.

Kamp, H. and Reyle, U. (1993). Tense and aspect. In *From Discourse to Logic*, pages 483–689. Springer, Dordrecht.

Kamp, H. (1984). A theory of truth and semantic representation. In J. Groenendijk, et al., editors, *Truth, interpretation and information*, pages 1–41. Foris, Dordrecht.

Lamb, S. M. and Newell, L. E. (1966). *Outline of Stratificational Grammar: With an Appendix by LE Newell*. Georgetown University Press, Georgetown.

Mathesius, V., (1947). *Čeština a obecnỳ jazykozpyt*, chapter O funkci podmětu [On the Function of Subject], pages 277–285. Melantrich, Praha.

Mel'chuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press, New York.

Mikulová, M. and Štěpánek, J. (2010). Ways of evaluation of the annotators in building the Prague Czech-English Dependency Treebank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1836–1839. European Language Resources Association, Valletta.

Mikulová, M., Bejček, E., Kolářová, V., and Panevová, J. (2017). Subcategorization of adverbial meanings

based on corpus data. *Jazykovedný časopis / Journal of Linguistics*, 68(2):268–277.

Mikulová, M., Straka, M., Štěpánek, J., Štěpánková, B., and Hajič, J. (2022). Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*. European Language Resources Association, Marseille.

Mille, S., Burga, A., and Wanner, L. (2013). AnCora-UPF: A multi-level annotation of Spanish. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 217–226. Charles University, Prague.

Mírovský, J., Rysová, K., Rysová, M., and Hajičová, E. (2013). (Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 55–63. Asian Federation of Natural Language Processing, Nagoya.

Mírovský, J., Synková, P., Poláková, L., Kloudová, V., and Rysová, M. (2021). *CzeDLex 1.0*. LINDAT/CLARIAH-CZ, Prague, URL: http://hdl.handle.net/11234/1-4595.

Nedoluzhko, A. and Mírovský, J. (2013). How dependency trees and tectogrammatics help annotating coreference and bridging relations in Prague Dependency Treebank. In *Proceedings of the Second International Conference on Dependency Linguistics, Depling 2013*, pages 244–251. Charles University, Prague.

Pajas, P. and Štěpánek, J. (2008). Recent Advances in a Feature-rich Framework for Treebank Annotation. In Donia Scott et al., editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester. The Coling 2008 Organizing Committee.

Pajas, P. and Štěpánek, J. (2009). System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec. Association for Computational Linguistics.

Panevová, J. and Mikulová, M. (2020). Subcategorization of adverbials (the case of temporal meanings). *Korpus – gramatika – axiologie*, (22):16–30.

Panevová, J., Hajičová, E., Kettnerová, V., Lopatková, M., Mikulová, M., and Ševčíková, M. (2014). *Mluvnice současné češtiny 2, Syntax na základě anotovaného korpusu*, volume 2. Karolinum, Praha.

Pradhan, S. and Ramshaw, L. (2017). Ontonotes: Large scale multi-layer, multi-lingual, distributed annotation. In N. Ide et al., editors, *Handbook of linguistic annotation*, pages 521–554. Springer, Dordrecht.

Saussure, F. d. (1916). *Cours de linguistique générale*. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger (eds.), Payot, Lausanne and Paris.

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

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

Silvano, P., Leal, A., Silva, F., Cantante, I., Oliveira, F., and Mario Jorge, A. (2021). Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13. Association for Computational Linguistics, Groningen.

Steedman, M. (2001). *The Syntactic Process*. The MIT Press, Cambridge.

Urešová, Z., Bémová, A., Fučíková, E., Hajič, J., Kolářová, V., Mikulová, M., Pajas, P., Panevová, J., and Štěpánek, J. (2021). *PDT-Vallex: Valenční slovník češtiny propojený s korpusy 4.0*. LINDAT/CLARIAH-CZ, Charles University, Prague, URL: http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.

Zeldes, A. (2017). The Gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeldes, A. (2018). *Multilayer corpus studies*. Routledge, New York and London.

9. Language Resource References

Hajič, J. and Bejček, E. and Bémová, A. and Buráňová, E. and Fučíková, E. and Hajičová, E. and Havelka, J. and Hlaváčová, J. and Homola, P. and Irc-ing, P. and Kárník, J. and Kettnerová, V. and Klyueva, N. and Kolářová, V. and Kučová, L. and Lopatková, M. and Mareček, D. and Mikulová, M. and Mírovský, J. and Nedoluzhko, A. and Novák, M. and Pajas, P. and Panevová, J. and Peterek, N. and Poláková, L. and Popel, M. and Popelka, J. and Romportl, J. and Rysová, M. and Semecký, J. and Sgall, P. and Spoustová, J. and Straka, M. and Straňák, P. and Synková, P. and Ševčíková, M. and Šindlerová, J. and Štěpánek, J. and Štěpánková, B. and Toman, J. and Urešová, Z. and Vidová Hladká, B. and Zeman, D. and Zikánová, Š. and Žabokrtský, Z. (2020). *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*. LINDAT/CLARIAH-CZ digital library, Charles University, Prague, http://hdl.handle.net/11234/1-3185.

Hajič, J. and Hajičová, E. and Panevová, J. and Sgall, P. and Cinková, S. and Fučíková, E. and Mikulová, M. and Pajas, P. and Popelka, J. and Semecký, J. and Šindlerová, J. and Štěpánek, J. and Toman, J. and Urešová, Z. and Žabokrtský, Z. (2012). *Prague Czech-English Dependency Treebank 2.0*. LINDAT/CLARIAH-CZ digital library, Charles University, Prague, http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.