Morphology Within the Multi-layered Annotation Scenario of the Prague Dependency Treebank

Magda Ševčíková(✉)

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics,
Charles University in Prague, Malostranské náměstí 25,
118 00 Prague, Czech Republic
sevcikova@ufal.mff.cuni.cz

Abstract. Morphological annotation constitutes a separate layer in the multi-layered annotation scenario of the Prague Dependency Treebank. At this layer, morphological categories expressed by a word form are captured in a positional part-of-speech tag. According to the Praguian approach based on the relation between form and function, functions (meanings) of morphological categories are represented as well, namely as grammateme attributes at the deep-syntactic (tectogrammatical) layer of the treebank.

In the present paper, we first describe the role of morphology in the Prague Dependency Treebank, and then outline several recent topics based on Praguian morphology: named entity recognition in Czech, formemes attributes encoding morpho-syntactic information in the dependency-based machine translation system, and development of a lexical database of derivational relations based partially on information provided by the morphological analyser.

Keywords: Annotation · Deep syntax · Lemma · Morphology · Multi-layered scenario · Part-of-speech tag · Surface syntax · Tagging

1 Introduction

The Prague Dependency Treebank (PDT) has a multi-layered scenario designed on the theoretical basis of Functional Generative Description (FGD). Though the theoretical framework itself focuses mainly on syntactic issues, the PDT annotation project started with annotation at the morphological layer. Information included at this layer was extensively used during annotation at both the layer of surface syntax and the deep-syntactic layer (tectogrammatics).

In the paper, the formal approach to Czech inflectional morphology is introduced first (see Sect. 2). An overview of tools for morphological analysis and disambiguation is followed by a description of the part-of-speech (POS) tags and morphological lemmas. The core of the paper presents annotation of morphological categories in PDT within the theoretical framework of FGD (Sects. 3.1

C. Mahlow and M. Piotrowski (Eds.): SFCM 2015, CCIS 537, pp. 1–26, 2015.
DOI: 10.1007/978-3-319-23980-4_1

and 3.2). A lemma and a positional POS tag capturing formally expressed inflectional categories were assigned manually to each token at the morphological layer (Sect. 3.3), and reinterpreted in a semi-automatic procedure during the annotation at the tectogrammatical layer; here, meanings of semantically relevant morphological categories were represented as values of special attributes (called grammatemes) assigned to nodes of the tectogrammatical tree (Sect. 3.4). PDT annotation scenario served as one of the resources for other treebanks mentioned in Sect. 3.5.

In Sect. 4, recent topics are outlined that are immediately connected with the presented approach to Czech morphology, namely named entity recognition in Czech, formemes encoding morpho-syntactic information in the dependency-based machine translation system, and development of a lexical database of derivational relations based partially on information provided by the morphological analyser.

2 Computational Morphology of Czech

2.1 Tools for Morphological Analysis and Disambiguation

Czech is a Slavic language with a complex system of both inflectional and derivational morphology. Though the traditional separation of inflections and derivations, which is documented in influential grammars of Czech, has been partially overcome in some NLP approaches to Czech, the main focus is still on inflectional morphology.

This section is limited to morphological analysis and morphological disambiguation (tagging) as two subtasks of morphological processing of Czech;[1] the former of them consists in assigning pairs of a tag and a lemma to an individual word form (usually regardless of the context) while the latter subtask is to select a single tag–lemma pair for the respective word form, mostly with respect to a (close) context.

Formulation of a computational approach to Czech morphology is dated back to the 1990s; cf. first experiments in automatic morphological analysis and disambiguation of Czech by Hladká and Hajič [13,18,23]. Morphological analysis was based on the Czech morphological dictionary (published now under the name MorfFlex CZ; [14]) which contains more than 350 thousand manually entered entries; the recogniser recognises about 12 million Czech word forms.

For first tagging experiments [23], it was possible to use manually annotated data, thanks to a pioneering corpus annotation project which was carried out at the Institute of the Czech Language of the Academy of Sciences of the Czech Republic from 1971 to 1985 (the corpus was called Korpus věcného stylu 'Practical Corpus' and, later on, converted into the Czech Academic Corpus with morphological and analytical annotation compatible with PDT; [24,66,67]).

[1] The issues of morphological synthesis, generation etc. go beyond the scope of the paper; see Hajič [11] for a complex description of computational approach to Czech morphology including formal definitions.

Table 1. Comparison of the taggers according to their accuracy on Czech (based on [51,56])

Tagger	Accuracy
Morče semi-supervised [51]	95.89 %
MorphoDiTa [56]	95.75 %
Combination of taggers [52]	95.70 %
Morče [68]	95.67 %
HMM [29]	94.82 %
Feature-based tagger [11]	94.04 %

The next, feature-based tagger was trained already on PDT data, which were manually annotated with positional POS tags and lemmas (Sects. 2.2 and 3.3). The tagger was based on a statistical algorithm with an exponential model [11], and distributed, along with a tool for morphological analysis, as a part of the PDT 2.0 release [16]. An implementation based on Hidden Markov Models is available as well [29].

In line with efforts to develop and to improve POS taggers for English and other languages inspired by Collins [6] and others, a tagger based on averaged perceptron, called Morče (an acronym of Morfologie češtiny 'Morphology of Czech'; [68]), was published in 2006. The Morče tagger was trained on manually annotated data of PDT, achieving a state-of-the-art performance on Czech,and later on, it was involved in experiments combining this tagger with the feature-based tagger, HMM tagger and a rule-based component [52], and in semi-supervised training experiments [51].[2] The semi-supervised version of the Morče tagger outperformed its original implementation as well as the combination with other taggers; see Table 1.

The most recent implementation, MorphoDiTa (Morphological Dictionary and Tagger; [53,56]), is an open-source tool for morphological analysis, tagging, and lemmatisation as well as for tokenisation and morphological generation; it is available along with trained linguistic models.

The feature-based tagger and the Morče tagger were used for morphological processing of large (100,000,000+ tokens) corpora of the SYN series, built at the Institute of Czech National Corpus.[3] Experiments with the rule-based disambiguation of large corpus data have been carried out [31,36,37,39]. Nevertheless, improvements in tagging have been reported recently by applying a combined disambiguation system including the Morče tagger and a rule-based component [40]; compare previous approaches to combining statistical and rule-based methods in [15,50], or [52].

[2] The semi-supervised version of Morče was published under the Compost project (http://ufal.mff.cuni.cz/legacy/compost/cz/). An implementation of the averaged perceptron algorithm was released in the Featurama project too (http://sourceforge.net/projects/featurama/).

[3] http://korpus.cz/.

Table 2. Positions of the positional POS tag

Position no.	Name	Description
1	POS	Part of speech
2	SUBPOS	Detailed part of speech
3	GENDER	Gender
4	NUMBER	Number
5	CASE	Morphological case
6	POSSGENDER	Possessor's gender
7	POSSNUMBER	Possessor's number
8	PERSON	Person
9	TENSE	Tense
10	GRADE	Degree of comparison
11	NEGATION	Negation
12	VOICE	Verbal voice
13	RESERVE1	Unused
14	RESERVE2	Unused
15	VAR	Variant, style, register, special usage

All the tools described above use compact tags or, predominantly, positional POS tags (both described in Sect. 2.2) as the output tag format.

An alternative system of encoding Czech morphology has been developed in the Natural Language Processing Centre at the Faculty of Informatics, Masaryk University in Brno, and implemented in the ajka analyser, which provides both inflectional and (to a limited extent) derivational analysis of Czech based on a large-coverage dictionary [44,45].

Last but not least a weakly-supervised (resource-light) approach to morphological analysis and tagging is to be mentioned, which substantially decreases requirements on cost-intensive manual input [8,20]. Though the weak supervision is often accompanied with a lower accuracy, the approaches are advantageous especially for underresourced languages.

2.2 Tag Sets for Czech, Positional POS Tag and Morphological Lemma Used in the Prague Dependency Treebank

There have been several tag sets used for Czech. From the chronological perspective, the tag set used in the original annotation of the Czech Academic Corpus (CAC; see Sect. 2.1) should be mentioned first [66,67].

In the original CAC tag set,[4] tags of maximum eight positions were used. At the first and second position, the part-of-speech class of the token was specified; the remaining positions were associated with morphological categories that are relevant for the particular part-of-speech class. Thus, for instance, in the fourth

[4] http://ufal.mff.cuni.cz/rest/CAC/tOrig.html.

tag position, mood is encoded with verb forms while gender with noun, adjectives, pronouns, and numerals. The values to be filled in at a particular position were defined with respect to the part-of-speech class as well and encoded with digits. Therefore, for instance, the same digit in the same position is to be interpreted differently with adjectives and with verbs. Compare the original CAC tags to be assigned to the tokens *Pokládáte* '(you) find', *za* 'for', and *standardní* 'standard' (the first three tokens from the sentence analysed in Table 3) and their interpretation:

Pokládáte	5251_19	verb – imperfective – 2nd person plural – indicative present active – [imperative:default] – one-word form – gender not expressed
za	774	preposition – primary – with accusative
standardní	22_414	adjective – primary – [subclass:default] – neuter – singular – accusative

A system of compact tags was defined by Hajič [11], and used in compilation of the morphological dictionary (MorfFlex CZ; [14]) and in tagging experiments, e.g. [13]. This tag system works with positions, specifying a combination (a "pattern") of relevant morphological categories (each associated with a tag position) for each part-of-speech (sub)class.[5] Compact tags for the same three tokens should be interpreted as follows:[6]

Pokládáte	VPp2A	verb – indicative present – plural – 2nd person – affirmative
za	R4	preposition – with accusative
standardní	ANS41A	adjective – neuter – singular – accusative – no gradation – affirmative

As an alternative to compact tags, a system of positional POS tags was developed and gradually preferred to the former one; cf. Hajič [11].[7] Positional POS tags, along with two-component lemmas (described below), were assigned to the PDT data at the morphological layer; see Sect. 3.3.

A positional POS tag consists of 15 positions: The part of speech and a (functionally or formally delimited) subpart of it are encoded in the first and second positions of the tag, respectively. Positions 3 to 12 are each associated with a particular morphological category, positions 13 and 14 are reserved for a potential extension of the tag information, and the 15th position captures information of variants, register features etc.; see Table 2.[8] Part-of-speech classes

[5] http://ufal.mff.cuni.cz/pdt1/Morphology_and_Tagging/Doc/compact_tags.pdf.

[6] The tag of the verb form is composed according to the pattern for present indicative forms: VPnpa (i.e., verb – indicative present – number – person – negation).

[7] http://ufal.mff.cuni.cz/pdt1/Morphology_and_Tagging/Doc/hmptagqr.pdf.

[8] An extended version of 16 positions was used in corpora of the Czech National Corpus. The 16th position is associated with the category of aspect which is, when using the tag with 15 positions, encoded in the technical lemma suffix described below.

Table 3. Morphological lemma and positional POS tag assigned to tokens of the sentence *Pokládáte za standardní, když se s Mečiarovou vládou nelze téměř na ničem rozumně dohodnout?* (lit.: Find for standard, when REFL with Mečiar's government is-not-possible almost on nothing reasonably agree?) 'Do you find it standard when almost nothing can be reasonably agreed on with Mečiar's government?' at the morphological layer of PDT, and conversion of the positional POS tags into the Interset interlingua attribute–value pairs (last column)

Token	Morphological lemma	Positional POS tag	Interset
Pokládáte	pokládat_:T	VB-P---2P-AA---	pos="verb", negativeness="pos", number="plur", person="2", verbform="fin", mood="ind", tense="pres", voice="act"
za	za-1	RR--4----------	pos="adp", adpostype="prep", case="acc"
standardní	standardní	AAIP4----1A----	pos="adj", negativeness="pos", gender="masc", animateness="inan", number="plur", case="acc", degree="pos"
,	,	Z:----------	pos="punc"
když	když	J,----------	pos="conj", conjtype="sub"
se	se_^(zvr._zájmeno/částice)	P7-X4----------	pos="noun", prontype="prs", reflex="reflex", case="acc", variant="short"
s	s-1	RR--7----------	pos="adp", adpostype="prep", case="ins"
Mečiarovou	Mečiarův_;S_^(*2)	AUFS7M---------	pos="adj", poss="poss", gender="fem", number="sing", case="ins", possgender="masc"
vládou	vláda	NNFS7-----A----	pos="noun", negativeness="pos", gender="fem", number="sing", case="ins"
nelze	lze	VB-S---3P-NA---	pos="verb", negativeness="neg", number="sing", person="3", verbform="fin", mood="ind", tense="pres", voice="act"
téměř	téměř	Db----------	pos="adv"
na	na-1	RR--6----------	pos="adp", adpostype="prep", case="loc"
ničem	nic	PW--6----------	pos="adj—noun", prontype="neg", negativeness="neg", case="loc"
rozumně	rozumně_^(*1ý)	Dg-------1A----	pos="adv", negativeness="pos", degree="pos"
dohodnout	dohodnout_:W	Vf--------A----	pos="verb", negativeness="pos", verbform="inf"
?	?	Z:----------	pos="punc"

as well as values of morphological categories were delimited in accordance with their description in the academic grammar of Czech [25].

In spite of combinatorial restrictions implied by the language itself,[9] there is a considerable number of combinations of the category values attested in the language data; cf. 1,574 different positional POS tags (and 71,503 different morphological lemmas) assigned to 1,957,247 tokens of the PDT 3.0 data annotated at the morphological layer. The positional POS tag, which allows for a combination of values of single categories, enables thus to describe the rich inflection in an economical way (compare, for instance, the POS tag set used in the Penn Treebank project [32]).

Besides a positional POS tag, each token was assigned a morphological lemma composed of two parts at the morphological layer of PDT. The first part of the lemma (so-called lemma proper) is a string of characters mostly corresponding to the base form of the word (namely, nominative singular form of nouns, nominative singular masculine of pronouns and numerals, nominative singular masculine positive form of adjectives, infinitive form of verbs, and positive form of adverbs).[10] Since the lemma was proposed as a unique identifier, ambiguous base forms were disambiguated with a digit attached by a hyphen to the string of characters (cf. Lemmas assigned to prepositions *za, s,* and *na* in Table 3).

The second part of the lemma is a technical suffix. It is attached to the lemma proper by an underscore. Technical suffixes do not occur with most lemmas; however, if needed, more technical suffixes are possible with a single lemma. The suffix contains either a comment on verbal aspect (cf. the suffix of the verb lemma *pokládat* in Table 3), or a comment explaining the respective meaning (suffix of the pronoun *se*), a label identifying the named entity type (_;S with the lemma *Mečiarův* identifying surnames), or derivational information (namely, formally encoded changes to be carried out to arrive at the base word; cf. _^(*2) with the same lemma: two characters should be removed in order to get the base word *Mečiar*).

Motivated by the needs of parsing, machine translation and other NLP subtasks, a method for conversion of different sets of POS tags has been developed: Interset is a set of universal morpho-syntactic features to which tag sets used in different corpora can be converted; it has been proposed as a sort of interlingua for POS tags [71]. The most recent Interset version covers 64 different tag sets of 37 languages [70]. See the positional POS tags used in PDT converted into the Interset attribute–value structures in Table 3.

[9] Generally speaking, there are typical nominal categories, such as case and gender, which do not combine with verbal categories, such as person, tense, mood, and voice. However, for instance, some Czech verb forms (past participle, transgressive) are marked for gender.

[10] With pluralia tantum nouns and other words with an incomplete or deficient paradigm, other forms are used instead of the canonical one; for instance, the plurale tantum *kalhoty* 'trousers' is assigned the nominative plural form as a lemma.

3 Annotation of Morphological Categories in the Prague Dependency Treebank

3.1 Theoretical Background of the Prague Dependency Treebank: Functional Generative Description

Functional Generative Description is a theoretical linguistic framework formulated in Prague in the 1960s [48,49]. It is rooted in the structuralist approach of the Prague Linguistic Circle; however, it has responded to similar stimuli as foreign approaches with fundamentally different backgrounds.

FGD decomposes the language system into several levels;[11] the "lowest" of them corresponds to linear text (either spoken or written) whereas the "highest" level represents the linguistic meaning of the sentence and is modelled as a dependency tree structure.[12] Between these two levels (phonetic and tectogrammatical level, respectively), another three levels were discerned in the original proposal, namely the morphonological level, morphological level, and level of surface syntax.

The theoretical fundamentals of FGD, to which – besides multiple levels – the dependency approach to syntax and the theory of valency belong, served as a starting point for the design of the annotation scenario of PDT [5]. Out of the set of levels differentiated in FGD, three layers have been included in the PDT scenario: the morphological layer, surface-syntactic layer, and tectogrammatical layer. Differences between the layout of the PDT layers and levels in FGD were motivated by the needs of NLP tasks, e.g. parsing, and were analysed by Štěpánek [65].

The formalised approach to morphology as a separate level of the language system model and the description of the meanings of morphological categories at the tectogrammatical level is a stable part of the FGD framework[13] and has been adopted into the annotation scenario of PDT as well.

[11] The present paper draws a terminological distinction between a *level* as a concept of the theoretical framework of FGD and a *layer* as a part of the annotation scenario of PDT.

[12] An opposite perspective, i.e. the text as a surface string which is assigned a deeper analysis, is justifiable as well; however, we stick to the perspective from the text as a basis on the top of which analyses are built.

[13] There are considerable similarities in dealing with morphology between FGD (and PDT) and the Meaning-Text Theory (MTT). As in MTT even more levels are distinguished than in FGD, the morphological level in FGD corresponds mainly to the deep-morphological representation in MTT but shares several features with the surface-syntactic representation of this framework [34]. The function of morphological categories is then a part of the deep-syntactic representation in MTT (the attributes are called grammemes in MTT and grammatemes in FGD); see Žabokrtský [74] for a more detailed comparison of these frameworks.

3.2 History of the Prague Dependency Treebank

The Prague Dependency Treebank is a collection of Czech newspaper texts from 1990s, processed at four layers. At the first (non-annotation) layer, called word layer, the source text is segmented into documents and paragraphs, tokens are associated with unique identifiers. At the morphological layer, as the lowest annotation layer, each token is assigned a positional POS tag and a lemma, see Table 3. At the surface-syntactic (analytical) layer, the syntactic structure of each sentence is represented as a dependency-tree structure. Nodes of the analytical tree are in a one-to-one correspondence to tokens at the morphological layer and are labelled with surface-syntactic functions (such as subject Sb, object Obj etc.; Fig. 1). At the tectogrammatical layer (the highest layer of annotation), the

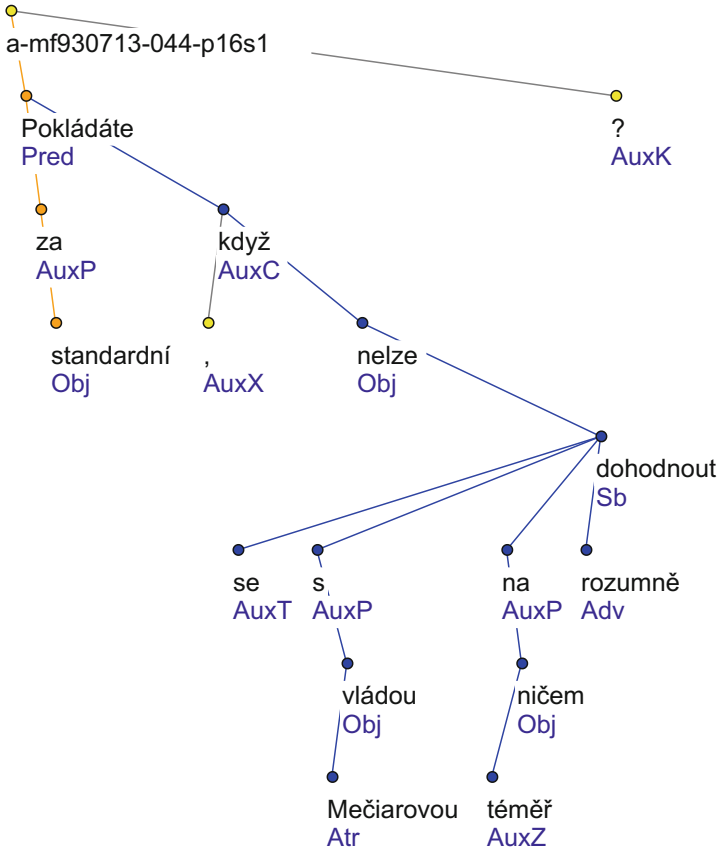

Fig. 1. Sentence *Pokládáte za standardní, když se s Mečiarovou vládou nelze téměř na ničem rozumně dohodnout?* 'Do you find it standard when almost nothing can be reasonably agreed on with Mečiar's government?' annotated at the analytical layer of PDT 3.0. Nodes are labelled with word forms and surface-syntactic functions (e.g., Sb for subject, Adv for adverbials, the Aux labels are assigned to different types of function words)

underlying syntactic structure of the sentence is also represented as a dependency tree, which, however, differs from the analytical one in several aspects.

While every token annotated at the morphological layer has exactly one corresponding node in the analytical tree, the correspondence between the nodes of the tectogrammatical tree and the analytical tree, which is nevertheless explicitly recorded in the data in the form of cross-layer references, is not always one-to-one, since only content words are represented as tectogrammatical nodes, and new nodes are constructed for deletions (cf. the node with the lemma #PersPron representing the pro-dropped subject pronoun of the verb *pokládáte* in Fig. 2) or for grammatical elements which do play a role in the syntactic structure but cannot be expressed in the surface shape of the sentence (see the #Cor node in Fig. 2, which is the subject of the infinitive *dohodnout se* and is relevant for coreference annotation). Nodes of the tectogrammatical tree were labelled with

- semantic roles (functors; e.g. ACT for Actor, PAT for Patient, MANN for Manner),
- labels defining the type of the respective node and its semantic part of speech (cf. the nodetype and sempos attributes),
- meanings of morphological categories (grammatemes), and
- labels identifying the node as an element of the topic or focus part of the sentence; see Fig. 2.

Non-dependency relations are annotated on the top of dependencies in the tectogrammatical tree; see the coreference arrow in Fig. 2. Annotation at the tectogrammatical layer is documented in [35].

There are four releases of the PDT data available: PDT 1.0, PDT 2.0, PDT 2.5, and PDT 3.0.[14] PDT 1.0 was published in 2001 and contains data annotated at the morphological layer and at the analytical layer [19]. Annotation of both types is available for 1,583 documents (containing 1,255,590 tokens in 81,614 sentences); there are also another 14 documents (469,652 tokens in 29,561 sentences) annotated at the morphological layer only and 314 documents (251,743 tokens in 16,649 sentences) with analytical annotation only. A sample of 3,490 tokens (in 203 sentences) with morphological and analytical annotation is annotated at the tectogrammatical layer as well.

The complete three-layer annotation is available for a large part of the data from PDT version 2.0 onwards. PDT 2.0, published in 2006 [16], contains 3,165 documents (with 833,195 tokens in 49,431 sentences) with morphological, analytical, and tectogrammatical annotations. Another 2,165 documents (with 670,544 tokens in 38,482 sentences) are annotated at the morphological and analytical layer, and for yet another 1,780 documents (with 453,508 tokens in 27,931 sentences) only morphological annotation is available in PDT 2.0. The data at each layer were divided into train data (app. 80 % of the data set with the respective annotation combination), development-test data (app. 10 %), and evaluation-test data (app. 10 %).

[14] A preliminary, test version of the treebank (PDT 0.5), containing 450 thousand tokens in 26 thousand sentences, was compiled for the Summer Workshop on Language Engineering at the Johns Hopkins University in Baltimore in 1998.

In PDT 2.5 and PDT 3.0 (released in 2011 and 2013, respectively),[15] the texts of PDT 2.0 are enriched with new annotations at the tectogrammatical and analytical layer, but neither the size of the data nor the portions of the data annotated at individual layers have changed; particular mistakes were corrected in the recent releases as well [3,4]. The following annotations were new in the PDT 2.5 as compared to PDT 2.0:

- annotation of multiword expressions at the tectogrammatical layer,
- a new grammateme identifying a special usage of plural forms of nouns (pair/group meaning) at the tectogrammatical layer,
- clause segmentation at the analytical layer.

For the PDT 3.0 release, the tectogrammatical layer was further modified:

- changes in the modality grammatemes,
- an extended annotation of coreference and bridging anaphora,
- annotation of discourse relations,
- genre specification.

Table 4. Values of the nodetype attribute assigned to each tectogrammatical node

nodetype values	Description
complex	Complex nodes represent nouns, adjectives, verbs, adverbs, and pronouns and numerals; they are the only nodes assigned with grammatemes
root	The root of the tectogrammatical tree is a technical node labelled with a unique identifier of the sentence
atom	Atomic nodes represent rhematisers, modal modifications (with functors RHEM, MOD, respectively) etc.
coap	Roots of coordination and apposition constructions are, according to the FGD convention, assigned a lemma of the coordinating conjunction or an artificial lemma of a punctuation symbol (e.g. #Comma)
fphr	Nodes with the FPHR functor are parts of foreign phrases, i.e. they are components of phrases that do not follow rules of Czech grammar
dphr	Dependent parts of phrasemes represent words that constitute a single lexical unit with their parent node (with the DPHR functor); the meaning of this unit is not a sum of the meanings of its component parts
list	Roots of foreign and identification phrases (with lemmas #Forn and #Idph) were added into the tree as parent nodes of foreign phrases (i.e., nodes with nodetype=fphr) or as parents of a multi-word named entity
qcomplex	Quasi-complex nodes represent obligatory verbal complementations that are not present in the surface sentence (they are mostly labelled with the same functors as complex nodes but have a special lemma, e.g. #Gen)

[15] Syntactically annotated PDT data of the particular versions are publicly accesible via the PML Tree Query environment (https://lindat.mff.cuni.cz/services/pmltq/; [38]) for searching.

3.3 Morphology as a Layer of Annotation in the Prague Dependency Treebank

As one can see from the history of the PDT releases, data of PDT were annotated at the morphological layer first. Each token was assigned a positional POS tag and a morphological lemma within a manual procedure which was preceded by an automatic morphological analysis.

The manual annotation was carried out by eight annotators [21]. Each file was annotated by two annotators in parallel, their task was a manual disambiguation of results of the morphological analysis using the DA and LAW (Lexical Annotation Workbench) editors of morphological annotations.[16] When the lemma was not offered by the tagger, it was created manually by the annotator and, subsequently, included into the morphological dictionary. After the parallel annotation was finished, instances of disagreement were decided by a third annotator. See the morphological annotation of a sentence in Table 3.

Annotation at the morphological layer was used during annotation at the analytical and, more importantly, at the tectogrammatical layer, being the main source of information for automatic assignment of grammatemes.

Morphological annotation, after a separate checking at this layer, was involved in the cross-layer checking of analytical and tectogrammatical annotations before the public release of the data. Štěpánek [64] gives examples of rather simple comparisons of POS tag values with surface-syntactic functions at the analytical layer and with functors at the tectogrammatical layer (e.g. with conjunctions), and describes checking of named entity information involved in the technical suffix of the morphological lemma against the tectogrammatical annotation, or a complex verification whether all valency slots defined by the valency lexicon are filled in with tectogrammatical nodes representing the requested word forms.

Table 5. Frequency of the nodetype values in the PDT 3.0 data annotated at all three layers

nodetype value	Frequency
complex	550,909
root	49,431
qcomplex	45,995
coap	35,742
atom	34,032
fphr	4,553
list	2,515
dphr	1,283

[16] https://bitbucket.org/jhana/feat-morph/wiki/Home.

3.4 Morphological Meanings at the Tectogrammatical Layer

Following the Praguian tradition of distinguishing form and function, functions (meanings) of morphological categories are captured by grammateme attributes in the tectogrammatical tree. The inclusion of grammatemes into the tectogrammatical layer responds to the claim of self-containedness and unambiguity of the sentence representation at each layer. If, for instance, meanings conveyed by the grammatical number with nouns, degree of comparison with adjectives, or tense with verbs were not specified at the tectogrammatical layer, several semantically different sentences could be generated from a single tectogrammatical tree.

Since morphological meanings are conveyed only by some nodes of the tectogrammatical tree and, moreover, not all grammatemes are relevant for all nodes, tectogrammatical nodes were classified in two subsequent steps. First, eight general types of nodes were distinguished according to their functor and/or tectogrammatical lemma in a fully automatic procedure. Grammatemes are relevant for nodes of just one type (for complex nodes); cf. the nodetype values and their frequency in PDT 3.0 in Tables 4 and 5.

Second, complex nodes were subdivided into four groups, called semantic parts of speech (semantic nouns, semantic adjectives, semantic verbs, and semantic adverbs) within which 19 more specific subgroups were discerned automatically. Accordingly, the sempos attribute with 19 values was defined (Table 6). Each subgroup was associated with a set of relevant grammatemes.

Table 6. Frequency of the sempos values in the PDT 3.0 data annotated at all three layers

sempos value	Frequency	sempos value	Frequency
n.denot	236,890	n.pron.def.demon	4,760
adj.denot	101,057	adj.pron.indef	3,383
v	88,026	adv.pron.indef	3,107
n.pron.def.pers	32,938	adv.pron.def	2,928
adj.quant.def	19,428	adj.quant.grad	1,865
n.denot.neg	18,832	adv.denot.grad.nneg	1,139
n.pron.indef	11,342	adv.denot.grad.neg	1,073
adv.denot.ngrad.nneg	8,996	adv.denot.ngrad.neg	751
n.quant.def	7,993	adj.quant.indef	655
adj.pron.def.demon	5,745		

As annotation of grammatemes was the last task in the PDT 2.0 annotation procedure, it could profit from the annotation at lower layers as well as from annotations already done at the tectogrammatical layer (mainly from the tree structure, functors, and coreference).

Nearly 1,600,000 grammateme values in total (with more than 550 thousand complex nodes) were assigned at the tectogrammatical layer of PDT 2.0, most of them automatically. Manual annotation, carried out by two annotators in parallel, with a follow-up decision by a third annotator in cases of disagreement, is responsible for approximately 17,500 out of the grammateme values [42].

The set of grammatemes and values assigned at the tectogrammatical layer was based on the FGD framework [49]. However, the repertoire has been revisited and changed according to the recent linguistic research during the annotation of individual PDT releases. In this paper, we present the grammateme annotation which is available in PDT 3.0.

There are 15 grammatemes annotated at the tectogrammatical layer of PDT 3.0. Grammatemes number, gender, person, politeness, and typgroup were assigned to nodes classified as semantic nouns. The grammatemes degcmp, negation, numertype, and indeftype were annotated with semantic nouns and with semantic adjectives. Semantic adverbs were assigned grammatemes degcmp, negation, and indeftype. Semantic verbs were assigned a special subset of verbal grammatemes: tense, aspect, factmod, deontmod, diatgram, and iterativeness.

Seven out of the 15 grammatemes correlate with morphological categories which are traditionally addressed in the grammatical description of Czech. Nevertheless, the grammateme values cannot be mostly interpreted from a single word form (its POS tag), but a more complex structure including auxiliaries had to be involved in the value assignment procedure (cf. grammatemes tense, factmod, deontmod, or diatgram described below), or manual annotation was needed, for instance, to assign number with pluralia tantum, absolute usage of comparative forms of adjectives and adverbs, or polite usage of 2nd person plural verbs.

- The number grammateme captures the number of entities to which the particular noun refers. In most cases, the value (sg or pl) correlates with the morphological category but is different, for instance, with pluralia tantum nouns (e.g., *otevřel dveře*.sg *na terasu* 'he opened the door to the terrace' vs. *několikery dveře*.pl 'several doors').
- Values of the gender grammateme (anim for animate masculines, inan for inanimates, fem and neut) correspond to the morphological gender of nouns, but if the grammatical gender does not coincide with the natural gender, the grammateme value was chosen according to the former one (cf. the neuter noun *děvče* 'girl').
- The person grammateme (values 1 for the speaker, 2 for the hearer, and 3 for a person/object it is talked about) was assigned with nodes representing pronouns. The grammateme values were non-trivially interpreted from agreement markers expressed by relevant verb forms.
- Values pos (positive), comp (comparative), and sup (superlative) of the degcmp grammateme correspond mostly to the category of degree of comparison, but comparative forms with an absolute (non-comparative) meaning were identified manually and assigned the third value acomp (e.g., *starší žena* 'an elder(ly) woman').

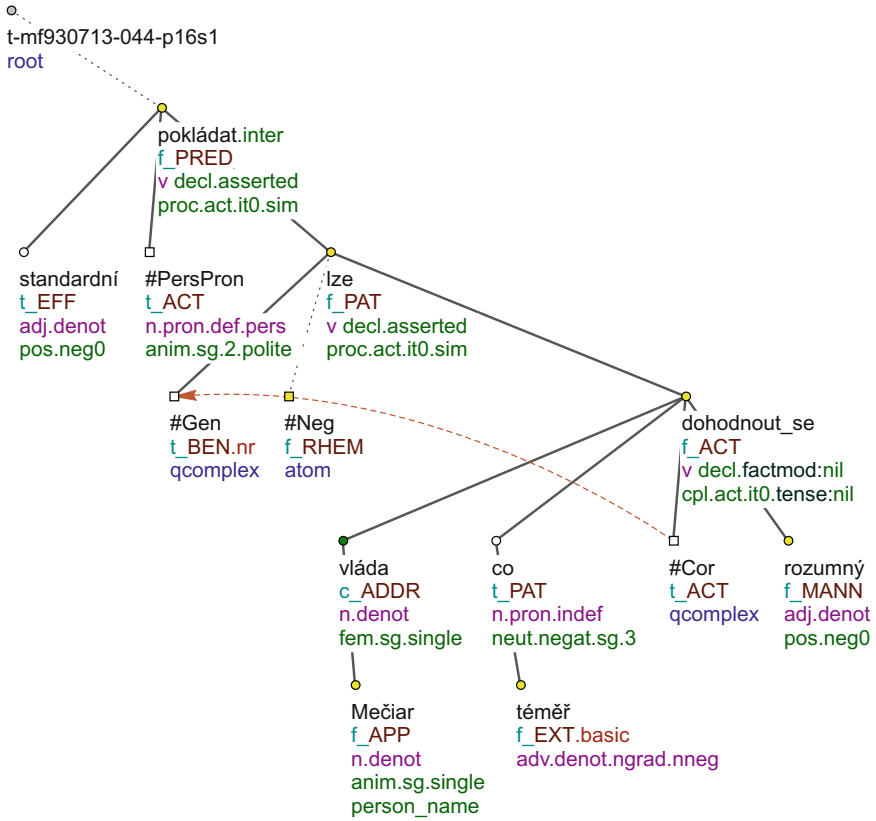

Fig. 2. Sentence *Pokládáte za standardní, když se s Mečiarovou vládou nelze téměř na ničem rozumně dohodnout?* 'Do you find it standard when almost nothing can be reasonably agreed on with Mečiar's government?' annotated at the tectogrammatical layer of PDT 3.0. Nodes are labelled with a tectogrammatical lemma, with a functor (e.g. ACT, MANN), topic-focus annotation (in front of the functor), a nodetype value (e.g., root or qcomplex), or a semantic part of speech and grammatemes (only with complex nodes, displayed under the functor). The predicate node of the tree (functor PRED) was assigned a sentence modality value (here, inter for interrogative sentences)

– Values of the tense grammateme distinguish the presented actions/states according to whether they preceded the moment of utterance or another action (ant), followed it (post), or happened simultaneously with it (sim). If the particular node represented a more complex verb form, the grammateme value had to be interpreted carefully. For example, future verbal tense in Czech is expressed by a simple inflected form (with perfectives; *dohodne se* '(he) will-agree'), or by an auxiliary verb (imperfectives; *bude pokládat* '(he) will find'), or by prefixing (lexically limited; cf. the future form *pojede* '(he) will-go' of the verb *jet* 'to go').

– For the factmod grammateme, four meanings (values) were distinguished according to the inner structure of the mood category in Czech, namely, asserted for actions/states presented as given (mostly by an indicative verb form), potential for potential events (expressed by a present conditional form), irreal for events expressed by a past conditional, and appeal for required events (conveyed by an imperative form).
– Values proc (processual/imperfective) and cpl (complex/perfective) of the aspect grammateme correlate with the aspect information captured by the technical lemma suffix at the morphological layer.

Another four grammatemes are considered grammaticalised meanings in the FGD framework as well:

– Values polite and basic of the politeness grammateme were assigned to personal pronouns to distinguish the polite form (*Vy*.polite *jste se už přihlásil?* 'Have you.polite logged in already?') from a common usage (*Vy*.basic *jste se už přihlásili?* 'Have you.basic logged in already?').
– The typgroup grammateme was included into the grammateme system to capture the pair/group meaning (like in *koupil si boty* '(he) bought a-pair-of shoes') expressed by plural forms; the pair/group meaning was delimited as another meaning of plural in Czech (besides the common usage reffering to several single entities, cf. *vystaveny byly jen pravé boty* 'only right shoes were displayed'; [58]).
– The diatgram grammateme captures meanings subsumed under grammaticalised diatheses, which are expressed by different verbal forms with a scale of auxiliaries: act for active voice, pas for passive voice, res1, res2.1 and res2.2 for different types of resultative forms, recip for recipient diathesis, disp for verb forms expressing dispositional modality, and deagent for deagentive verb forms.
– The deontmod grammateme was used to represent modal verbs as auxiliaries at the tectogrammatical layer; seven values were delimited according to modal meanings of necessity, possibility etc.

Even subsumed under the term of grammatemes, the following attributes capture derivational morphology,[17] rather than inflections:

– The iterativeness grammateme enables to represent an iterative verb by the tectogrammatical lemma of its non-iterative counterpart.
– The negation grammateme represents the negative meaning (expressed mostly by the *ne-* prefix) of nouns, adjectives and adverbs.[18]
– The indeftype grammateme made it possible to reduce pronouns and pronominal adverbs to a small set of lemmas at the tectogrammatical layer, exploiting the semantically relevant regularities within this closed class [62]. Cf. the node

[17] These derivations are subtypes of lexical derivation according to Kuryłowicz [30].
[18] Negated verb forms are analysed differently at the tectogrammatical layer, namely, they are decomposed into two nodes; cf. the verbal node with the lemma *lze* and node with the artificial lemma #Neg representing the negation in Fig. 2.

with the lemma *co* 'what' in Fig. 2, which represents the pronoun *(na) ničem* '(on) nothing' (the negative semantic feature was captured by the negat value).
– Similarly, the numertype is used to capture the specific meanings of different types of numerals (e.g. ordinal numerals, multipliers) that are represented by the tectogrammatical lemma of the corresponding cardinal numeral.

In addition to the approach described above for selected derivational relations captured by grammatemes, two types of highly regular derivatives, namely possessive adjectives and deadjectival adverbs, were converted into their base words, i.e., into nouns and adjectives, respectively. Since both these types of derivatives differ from their base words just in the function they play within the tectogrammatical structure,[19] it is sufficient to use the functor to encode the difference between the derived word and the base; see the nodes with the lemma *Mečiar* and *rozumný* 'reasonable' in Fig. 2.

Possible extension of the annotation of derivational morphology at the tectogrammatical layer is discussed in Sect. 4.3.

3.5 PDT-Style Annotations in Other Treebanks

Czech Academic Corpus, mentioned above in Sect. 2, has been converted from the original annotation (carried out in the 1970s and 1980s) into the PDT annotation scheme after the PDT 2.0 release; cf. CAC 1.0 [67] and CAC 2.0 [66]. CAC 2.0 contains morphological and analytical annotation for nearly 500 thousand tokens (and another data portion with morphological annotation only) which is now fully compatible with PDT.

Besides CAC, PDT annotation scenario has been used also for Arabic [17] and English [12], and has served as one of the resources for annotation schemes for Slovak (Slovak Treebank, which is a part of the Slovak National Corpus), Slovenian (Slovene Dependency Treebank),[20] Ancient Greek and Latin (Ancient Greek and Latin Dependency Treebanks),[21] and as an inspiration for other treebanking projects.

In 2011, an important project of bringing treebanks of different languages (some of them just mentioned) under a common annotation scheme has been proposed under the acronym HamleDT (HArmonized Multi-LanguagE Dependency Treebank). Treebanks were harmonised into the Prague Dependencies annotation style (based on analytical PDT annotation; [73]) and, recently, converted into Stanford Universal Dependencies [33]. Thirty treebanks are available in HamleDT 2.0 [43,72].[22]

[19] They belong to syntactic derivation as defined by Kuryłowicz [30].
[20] http://nl.ijs.si/sdt/.
[21] http://nlp.perseus.tufts.edu/syntax/treebank/.
[22] Stanford Universal Dependencies, the Interset interligua (mentioned in Sect. 2.2), and Google universal POS tags [41] served as a basis for the annotation scheme of the Universal Dependencies treebank project, the current version of which (Universal Dependencies 1.1; [1]) contains dependency annotated data for 18 languages including Czech.

4 Morphology in Named Entity Recognition, Dependency-Based Machine Translation, and in a Database of Derivational Relations in Czech

4.1 Named Entity Recognition in Czech

In a pilot approach to named entity (NE) classification and recognition, started only in 2007 [60], technical suffixes of morphological lemmas were used as an important resource for this task. Based on a survey of previous NE research using a low number of coarse-grained categories (such as [9]) on the one hand, or detailed categories (preferred in semantically oriented tasks, cf. [47]) on the other, a two-level classification has been proposed for Czech, which is convenient for both a robust processing and research interested in more subtle categorisation.

At the first level of the classification, ten rough categories were distinguished and, at the second level, further subclassified into 62 detailed categories. For instance, within the category of geographical names, subcategories of names of continents, states, towns, hydronyms etc. were discerned. This classification was used in the Czech Named Entity Corpus (CNEC), which consists of 6 thousand sentences with more than 150 thousand tokens manually assigned with NE categories [57,61]. The data were used for development of several recognisers of NE in Czech texts; cf. [26–28,55,60], and the most recent of them, NameTag [54,56], which is an open-source tool for NE recognition, distributed along with trained linguistic models.

4.2 Formemes in Dependency-Based Machine Translation

The complex dependency deep-syntactic analysis has been used as a transfer layer in a machine translation system developed at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague. The MT system, originally called TectoMT [75], has been extended with a number of modules into a modular NLP framework Treex, which is either available for installation from CPAN,[23] or can be run on-line under the LINDAT/CLARIN repository [46]. Recently, the Treex framework has been used, for instance, in the QTLeap European machine translation project.[24]

The deep-syntactic analysis provided by the Treex framework has introduced a special type of attributes, called formemes, into the deep-syntactic tree. Formemes are node attributes in which the form of the word represented by respective node is encoded by a combination of morphological and syntactic features. Taking the example of the prepositional phrase *s (Mečiarovou) vládou* in Fig. 2 and its English equivalent *with (Mečiar's) government*, the formeme n:with+X is to be assigned to the tectogrammatical node representing the (source) phrase *with government* within the English-to-Czech machine translation, while the node representing the (target) phrase *s vládou* is assigned the formeme n:s+7

[23] See http://ufal.mff.cuni.cz/treex.
[24] http://qtleap.eu/.

in which the morphological case (7 for instrumental) is specified in addition to the particular preposition. A complete list of formemes implemented in Treex can be found in [7].

From the perspective of the PDT annotation scheme, information encoded in formemes is a combination of information involved in POS tags at the morphological layer and in surface-syntactic functions at the analytical layer of PDT with selected auxiliary words (e.g., prepositions).

4.3 Derivational Morphology in Czech

Besides a basic NE annotation, the technical suffix of the morphological lemma provides information on regular derivational relations as well.[25] In PDT, derivational information involved in the lemma suffix at the morphological layer was extended by derivational information captured in selected grammatemes or in functors at the tectogrammatical layer (see Sect. 3.4).

This rather preliminary approach to interconnection of Czech derivational morphology with inflections on the one hand, and with syntax on the other has indicated the way how to overbridge the separation of derivations from inflectional morphology which is documented in all representative grammars of Czech.[26]

In order to put the annotation of derivations in PDT on a solid basis but, primarily, to build a reliable resource of derivational data for Czech, a lexical network of derivationally related words (DeriNet; [59]) is being developed. The current version DeriNet 0.9 contains more than 305 thousand lexemes which were connected with more than 117 thousand links that correspond to derivational relations between pairs of lexemes (i.e., between a base lexeme and a lexeme derived from it).[27] The pairs of derivationally related lexemes can be arranged into a tree graph; see the derivational tree with the root *standard* 'standard' (displayed by DeriNet Viewer)[28] in Fig. 3.

The network was initialised with a set of lexemes whose existence was supported by corpus evidence. As the data were morphologically processed by the Morče tagger, technical suffixes including derivational information were available, and were extensively used in creating derivational links in the network. This starting annotation phase has been followed by several rounds of semi-automatic annotation within which special attention had to be devoted to vowel and consonant alternations that occur very frequently during derivation in Czech. Since some of the alternations are involved in the inflectional paradigm as well, recent efforts in exploiting the inflectional morphological dictionary seem to make it

[25] A limited derivational analysis is carried out also by the ajka analyser (see Sect. 2.1).

[26] In Czech linguistics, derivation is separated from inflectional morphology, being described as the core part of word-formation, which is kept apart from the grammatical module; only inflectional morphology and syntax are supposed to constitute the grammatical structure of Czech.

[27] http://ufal.mff.cuni.cz/derinet.

[28] http://ufal.mff.cuni.cz/derinet/viewer.

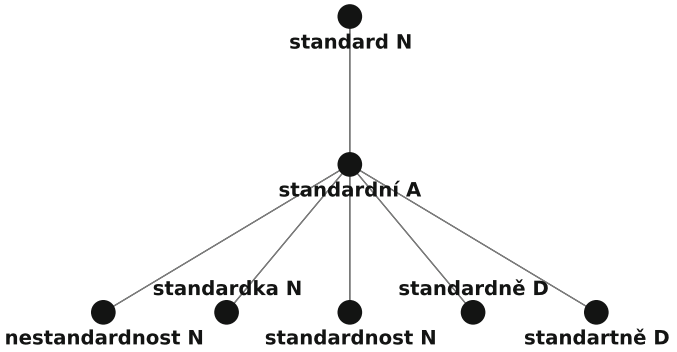

Fig. 3. The derivational tree of the noun *standard* 'standard' in the lexical network DeriNet

possible to build a model of alternations which will enable to couple derivationally related lexemes automatically with a high precision even if they differ substantially due to the alternations.[29]

Though DeriNet is still being developed (besides exploitation of the inflectional data, the main focus is on addition of new edges and correction of mistakes),[30] it is, to the best of our knowledge, the most complex and the only freely available resource of derivational data for Czech, and it belongs to a relative small number of derivational resources in general (cf. CELEX [2] for English, German and Dutch, DerivBase for German [69], DerivBase.Hr for Croatian [63], or most recently, the Démonette network for French [22]).

After arriving at a final version of the DeriNet data, semantic labelling of the derivational relations is proposed as the next step. Here, dealing with ambiguity and homonymy is expected to be the biggest challenge.[31]

The DeriNet network enriched with semantic labels is then envisaged to be used as the main resource for an extension of the derivational annotation of tectogrammatical data in PDT. Nevertheless, it is expected that only the most frequent semantic classes of derivatives with a transparent derivational meaning will be processed in order not to "overload" the data and to keep them usable for both NLP tasks and linguistic research.

[29] For instance, one of the changes occurring during derivation of the adjective *sněžný* 'snowy' from the noun *sníh* 'snow' is present in the inflectional paradigm of the noun (*sníh*.nom.sg – *sněhu*.gen.sg).

[30] One of the current mistakes is documented in the tree in Fig. 3: the noun *nestandardnost* 'non-standardness' is to be captured as derived either from the noun *standardnost* 'standardness', or from the adjective *nestandardní* 'non-standard' (which is not included in the network, though).

[31] For instance, the suffix *-ka* is used both in diminutives and female nouns (e.g. *skříň* 'cupboard' > *skříňka* 'small cupboard', *učitel* 'teacher' > *učitelka* 'female teacher'), and, on the other hand, several meanings are expressed by formally different affixes in Czech (e.g. female nouns are derived by the suffixes *-ka*, *-yně*, *-ice*, *-ovna* and several others).

5 Conclusions

The aim of the present paper was to put together a complex picture of the role of morphology in the richly annotated data of the Prague Dependency Treebank. Morphological annotation constitutes a separate layer in the treebank, nevertheless, it has been used as a source of information encoded at the higher, structural layers of annotation. Correlations between morphological categories captured at the morphological layer and grammateme attributes included in the tectogrammatical tree were analysed in detail.

Though tagging has been discussed to be a sort of solved task for at least "sufficiently resourced" languages [10], probably including Czech, it is still an interesting and appealing task since, particularly in a morphologically rich language like Czech, a high-quality lemmatisation and POS tagging are considered a common prerequisite of most NLP tasks.

In the paper we briefly outlined several topics that are based on morphological tools, and on morphologically annotated data as well. An outlook, concerning the proposed extension of the tectogrammatical annotation with derivations, documents the importance of morphology in efforts to deepen the syntactic analysis of language data.

Acknowledgements. The research reported on in the paper has been supported by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (LM2010013).

References

1. Agić, Ž., Aranzabe, M.J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A.T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., Lynn, T., Manning, C., Martínez, H.A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., Zeman, D.: Universal Dependencies 1.1. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2015). http://hdl.handle.net/11234/LRT-1478
2. Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX lexical database (release 2), Data/software. Linguistic Data Consortium, Philadelphia (1995)
3. Bejček, E., Hajič, J., Panevová, J., Mírovský, J., Spoustová, J., Štěpánek, J., Straňák, P., Šidák, P., Vimmrová, P., Šťastná, E., Ševčíková, M., Smejkalová, L., Homola, P., Popelka, J., Lopatková, M., Hrabalová, L., Kluyeva, N., Žabokrtský, Z.: Prague Dependency Treebank 2.5. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2011). http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8

4. Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: Prague Dependency Treebank 3.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2013). http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3
5. Böhmová, A., Hajič, J., Hajičová, E., Hladká, B.: The Prague dependency treebank: a three-level annotation scenario. In: Abeillé, A. (ed.) Treebanks: Building and Using Syntactically Annotated Corpora, pp. 103–128. Kluwer Academic Publishers, Dordrecht (2003)
6. Collins, M.: Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), vol. 10, pp. 1–8. Association for Computational Linguistics, Philadelphia (2002)
7. Dušek, O., Žabokrtský, Z., Popel, M., Majliš, M., Novák, M., Mareček, D.: Formemes in english-czech deep syntactic MT. In: Proceedings of the Seventh ACL Workshop on Statistical Machine Translation, pp. 267–274. Association for Computational Linguistics, Montréal (2012)
8. Feldman, A., Hana, J.: A Resource-Light Approach to Morpho-Syntactic Tagging. Rodopi, Amsterdam (2010)
9. Fleischman, M., Hovy, E.: Fine-grained classification of named entities. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING), vol. I, pp. 267–273. Association for Computational Linguistics, Taipei (2002)
10. Giesbrecht, E., Evert, S.: Part-of-speech tagging - a solved task? an evaluation of POS taggers for the Web as corpus. In: Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, pp. 27–35 (2009)
11. Hajič, J.: Disambiguation of Rich Inflection: Computational Morphology of Czech. Karolinum, Prague (2004)
12. Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z.: Prague Czech-English Dependency Treebank 2.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2012). http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4
13. Hajič, J., Hladká, B.: Probabilistic and rule-based tagger of an inflective language - a comparison. In: Proceedings of the 5th Conference on Applied Natural Language Processing, pp. 111–118. Association for Computational Linguistics, Washington, DC (1997)
14. Hajič, J., Hlaváčvá, J.: MorfFlex CZ. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (1990). http://hdl.handle.net/11858/00-097C-0000-0015-A780-9
15. Hajič, J., Krbec, P., Oliva, K., Květoň, P., Petkevič, V.: Serial combination of rules and statistics: a case study in Czech tagging. In: Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL 2001), pp. 260–267. Association for Computational Linguistics, Tolouse (2001)
16. Hajič, J., Panevová, J., Hajičcová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., Urešová, Z.: Prague Dependency Treebank 2.0. Data/software. Linguistic Data Consortium, Philadelphia (2006)

17. Hajič, J., Smrž, O., Zemánek, P., Pajas, P., Šnaidauf, J., Beška, E., Kracmar, J., Hassanová, K.: Prague Arabic Dependency Treebank 1.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2009). http://hdl.handle.net/11858/00-097C-0000-0001-4872-3
18. Hajič, J., Vidová Hladká, B.: Czech language processing - PoS tagging. In: Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998), pp. 931–936. ELRA, Granada (1998)
19. Hajič, J., Vidová Hladká, B., Panevová, J., Hajičcová, E., Sgall, P., Pajas, P.: Prague Dependency Treebank 1.0. Data/software. Linguistic Data Consortium, Philadelphia (2001)
20. Hana, J., Feldman, A.: Resource-light approaches to computational morphology. Part 1: monolingual approaches. Lang. Linguist. Compass **6**, 622–634 (2012)
21. Hana, J., Zeman, D., Hajič, J., Hanová, H., Hladká, B., Jeřábek, E.: Manual for Morphological Annotation, Revision for the Prague Dependency Treebank 2.0. Technical report no. 2005/TR-2005-27, FAL MFF UK, Prague (2005)
22. Hathout, N., Namer, F.: Démonette, a French derivational morpho-semantic network. Linguist. Issues Lang. Technol. **11**, 125–168 (2014)
23. Hladká, B.: Software Tools for Large Czech Corpora Annotation. Master thesis. MFF UK, Prague (1994)
24. Hladká, B., Králík, J.: Proměny Českého akademického korpusu. Slovo a Slovesnost **67**, 179–194 (2006)
25. Komárek, M., Kořenský, J., Petr, J., Veselková, J., et al.: Mluvnice češtiny 2. Tvarosloví. Academia, Prague (1986)
26. Konkol, M., Konopík, M.: Maximum entropy named entity recognition for czech language. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 203–210. Springer, Heidelberg (2011)
27. Konkol, M., Konopík, M.: CRF-based Czech named entity recognizer and consolidation of Czech NER research. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS, vol. 8082, pp. 153–160. Springer, Heidelberg (2013)
28. Kravalová, J., Žabokrtský Z.: Czech named entity corpus and SVM-based recognizer. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp. 194–201. Association for Computational Linguistics, Suntec (2009)
29. Krbec, P.: Language Modelling for Speech Recognition of Czech. Ph.D. thesis. MFF UK, Prague (2005)
30. Kuryłowicz, J.: Dérivation lexicale et dérivation syntaxique. Bull. de la Société de Linguistique de Paris **37**, 79–92 (1936)
31. Květoň, P.: Rule-based Morphological Disambiguation. Ph.D. thesis. MFF UK, Prague (2006)
32. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building A Large Annotated Corpus of English: The Penn Treebank. Technical reports (CIS), Paper 237 (1993). http://repository.upenn.edu/cis_reports/237/
33. de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.: Universal stanford dependencies: a cross-linguistic typology. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 4585–4592. ELRA, Reykjavík (2014)
34. Mel'čuk, I.A.: Dependency Syntax: Theory and Practice. State University of New York Press, New York (1988)

35. 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., Žabokrtský, Z.: Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical report no. 2006/30, ÚFAL MFF UK, Prague (2006)
36. Oliva, K., Květoň, P., Ondruška, R.: The computational complexity of rule-based part-of-speech tagging. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 82–89. Springer, Heidelberg (2003)
37. Oliva, K., Hnátková, M., Petkevič, V., Květoň, P.: The linguistic basis of a rule-based tagger of Czech. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2000. LNCS (LNAI), vol. 1902, pp. 3–8. Springer, Heidelberg (2000)
38. Pajas, P., Štěpánek, J., Sedlák, M.: PML Tree Query. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2009). http://hdl.handle.net/11858/00-097C-0000-0022-C7F6-3
39. Petkevič, V.: Reliable morphological disambiguation of Czech: rule-based approach is necessary. In: Šimková, M. (ed.) Insight into the Slovak and Czech Corpus Linguistics, pp. 26–44. Veda, Bratislava (2006)
40. Petkevič, V.: Problémy automatické morfologické disambiguace češtiny. Naše řeč **97**, 194–207 (2014)
41. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 2089–2096. ELRA, Istanbul (2012)
42. Razímová, M., Žabokrtský, Z.: Annotation of grammatemes in the prague dependency treebank 2.0. In: Proceedings of the LREC Workshop on Annotation Science, pp. 12–19. ELRA, Genova (2006)
43. Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., Žabokrtský, Z.: HamleDT 2.0: thirty dependency treebanks stanfordized. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 2334–2341. ELRA, Reykjavík (2014)
44. Sedláček, R.: Morfologický analyzátor češtiny. Master thesis. FI MU, Brno (1999)
45. Sedláček, R., Smrž, P.: A new Czech morphological analyser ajka. In: Matoušek, V., Mautner, P., Mouček, R., Tauser, K. (eds.) TSD 2001. LNCS (LNAI), vol. 2166, pp. 100–107. Springer, Heidelberg (2001)
46. Sedlák, M.: Treex::Web. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11858/00-097C-0000-0023-44AF-C
47. Sekine, S.: Sekine's Extended Named Entity Hierarchy (2003). http://nlp.cs.nyu.edu/ene/
48. Sgall, P.: Generativní Popis Jayzka a Česká Deklinace. Academia, Prague (1967)
49. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Reidel Publishing Company, Dordrecht (1986)
50. Spoustová, D.: Kombinované statisticko-pravidlové metody značkování češtiny. Ph.D. thesis. MFF UK, Prague (2007)
51. Spoustová, D., Hajič, J., Raab, J., Spousta, M.: Semi-supervised training for the averaged perceptron POS tagger. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 763–771. Association for Computational Linguistics, Athens (2009)
52. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The best of two worlds: cooperation of statistical and rule-based taggers for Czech. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, pp. 67–74. Association for Computational Linguistics, Prague (2007)

53. Straka, M., Straková, J.: MorphoDiTa: Morphological Dictionary and Tagger. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0
54. Straka, M., Straková, J.: NameTag. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E
55. Straková, J., Straka, M., Hajič, J.: A new state-of-the-art Czech named entity recognizer. In: Habernal, I. (ed.) TSD 2013. LNCS, vol. 8082, pp. 68–75. Springer, Heidelberg (2013)
56. Straková, J., Straka, M., Hajič, J.: Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014): System Demonstrations, pp. 13–18. Association for Computational Linguistics, Baltimore (2014)
57. Straková, J., Straka, M., Ševčíková, M., Žabokrtský, Z.: Czech Named Entity Corpus. In: Ide, N., Pustejovsky, J. (eds.) Handbook of Linguistic Annotation. Springer, Heidelberg (in press)
58. Ševčíková, M., Panevová, J., Smejkalová, L.: Specificity of the number of nouns in Czech and its annotation in prague dependency treebank. Prague Bull. Math. Linguist. **96**, 27–47 (2011)
59. Ševčíková, M., Žabokrtský, Z.: Word-formation network for czech. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 1087–1093. ELRA, Reykjavík (2014)
60. Ševčíková, M., Žabokrtský, Z., Krůza, O.: Named entities in Czech: annotating data and developing NE tagger. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 188–195. Springer, Heidelberg (2007)
61. Ševčíková, M., Žabokrtský, Z., Straková, J., Straka, M.: Czech Named Entity Corpus 1.1. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11858/00-097C-0000-0023-1B04-C
62. Ševčíková Razímová, M., Žabokrtský, Z.: Systematic parameterized description of pro-forms in the prague dependency treebank 2.0. In: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006), pp. 175–186. Institute of Formal and Applied Linguistics, Prague (2006)
63. Šnajder, J.: DerivBase.Hr: a high-coverage derivational morphology resource for croatian. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 3371–3377. ELRA, Reykjavík (2014)
64. Štěpánek, J.: Post-annotation checking of prague dependency treebank 2.0 data. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 277–284. Springer, Heidelberg (2006)
65. Štěpánek, J.: Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat). Ph.D. thesis. MFF UK, Prague (2006)
66. Vidová Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., Raab, J.: Czech Academic Corpus 2.0. Data/software. Linguistic Data Consortium, Philadelphia (2008)
67. Viová Hladká, B., Hana, J., Hajič, J., Hlaváčová, J., Mírovský, J., Votrubec, J.: Czech Academic Corpus 1.0. Data/software. Karolinum, Prague (2007)
68. Votrubec, J.: Volba vhodné sady rysů pro morfologické značkování češtiny. Master thesis. MFF UK, Prague (2005)

69. Zeller, B., Šnajder, J., Padó, S.: DerivBase: inducing and evaluating a derivational morphology resource for German. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pp. 1201–1211. Association for Computational Linguistics, Sofia (2013)
70. Zeman, D.: Lingua: Interset 2.026. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11234/1-1465
71. Zeman, D.: Reusable tagset conversion using tagset drivers. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 213–218. ELRA, Marrakech (2008)
72. Zeman, D., Mareček, D., Mašek, J., Popel, M., Ramasamy, L., Rosa, R., Štěpánek, J., Žabokrtský, Z.: HamleDT 2.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2014). http://hdl.handle.net/11858/00-097C-0000-0023-9551-4
73. Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J.: HamleDT: to parse or not to parse? In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 2735–2741. ELRA, Istanbul (2012)
74. Žabokrtský, Z.: Resemblances between meaning-text theory and functional generative description. In: Proceedings of the 2nd International Conference of Meaning-Text Theory, pp. 549–557. Slavic Culture Languages Publishers House, Moskva (2005)
75. Žabokrtský, Z., Ptáček, J., Pajas. P.: TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In: Proceedings of the Third ACL Workshop on Statistical Machine Translation, pp. 167–170. Association for Computational Linguistics, Columbus (2008)