M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Manual for Morphological Annotation.
Revision for Prague Dependency Treebank – Consolidated
2020 release

Marie Mikulová, Jan Hajič, Jíří Hana, Hana Hanová, Jaroslava Hlaváčová,
Emil Jeřábek, Barbora Štěpánková, Barbora Vidová Hladká, Daniel Zeman

ÚFAL Technical Report
TR-2020-64

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Manual for Morphological Annotation Revision for Prague Dependency Treebank – Consolidated 2020 release

Marie Mikulová
Jan Hajič
Jíří Hana
Hana Hanová
Jaroslava Hlaváčová
Emil Jeřábek
Barbora Štěpánková
Barbora Vidová Hladká
Daniel Zeman

December 2020

Contents

1 Preface

Although the title of this report inherits the word "Manual" from the previous versions, it is no more intended to guide the annotators. Rather it attempts to describe the current state of the morphological annotation in the Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0).[1] We believe that the guidelines can be of use to the users of the PDT-C 1.0 data, as well as for possible preparation of new data.

PDT-C 1.0 consists of four different datasets coming from PDT-corpora of Czech published earlier:

- dataset of written texts (the core PDT corpus in version 3.5),[2]
- dataset of translated texts (Czech part of Prague Czech-English Dependency Treebank),[3]
- dataset of spoken texts (Prague Dependency Treebank of Spoken Czech),[4]
- datasets of user-generated texts (unpublished small treebank PDT-Faust).[5]

In the PDT-C project, we aim to provide all these treebanks with full manual annotation at the lower layers and unify and correct annotation at all layers. Specifically, the data in PDT-C 1.0 is (mainly) enhanced with a manual annotation at the morphological layer, consistently across all the four original treebanks. Altogether, the consolidated treebank contains almost 3,900,000 tokens with manual morphological annotation. The Czech morphological dictionary MorfFlex,[6] which is now an integral part of the PDT-C 1.0 release, consists of more than 1 million lemmas/paradigms.

Acknowledgment. The research and language resource work reported in the paper has been supported by the LINDAT/CLARIAH-CZ project funded by Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

[1] `https://ufal.mff.cuni.cz/pdt-c`
[2] `http://ufal.mff.cuni.cz/pdt3.5`
[3] `https://ufal.mff.cuni.cz/pcedt2.0`
[4] `https://ufal.mff.cuni.cz/pdtsc2.0`
[5] `https://ufal.mff.cuni.cz/grants/faust`
[6] `https://ufal.mff.cuni.cz/morfflex`

2 Introduction

We do not want to substitute a grammarbook of Czech. So we will not systematically define word classes, morphological categories, paradigms, etc. All the annotators and users of the data and dictionary should understand the fundamentals of the Czech morphology, as most native Czech speakers do (the stuff is being taught in elementary schools). What we will describe the main principles of morphological annotation and focus on difficult and unusual phenomena.

The morphological annotation is based on a manual disambiguation of an automatic, dictionary-based morphological analysis of the annotated texts. For such automatic preprocessing, we use the MorphoDiTa tool.[7] In the annotation, a lemma (see Sect. 4) and a tag (see Sect. 5) is assigned to each wordform. The lemma and the tag together uniquely identify the wordform (see Sect. 3.1). The annotation contains no syntactic structure, no attempt is even made to put together e.g. analytical verb forms or other types of multiword expressions (see Sect. 21).

Key element to annotation consistency is the Czech morphological dictionary **MorfFlex**, which is now an integral part of the PDT-C 1.0 release. MorfFlex (see Sect. 3) is a flat list of lemma-tag-wordform triples.[8] It is in fact only an (automatic) derivative of the original, so-called "source format", in which the dictionary is still being maintained. The source format is based on paradigm pattern system and a substantial part of the dictionary (65% lemmas/paradigms) is mapped onto so-called derivational patterns. If a word belongs to a derivational patterns, several other words can be automatically derived from it. All automatically derived lemmas have the derivational information stored as a technical suffix of the lemma (Sect. 4.2.5). The suffix is really technical, primarily, it carries information about the automatic creation of a lemma; the manifested word-formation relation may not be correct or complete.[9] This is important for annotation. Because the creating lemmas from derivational patterns is an automatic process, there is no possibility to manually annotate stylistic and other categories of the derived lemmas (see more in Sect. 7).

The dictionary itself has undergone a long development process. It has been developed gradually since 1988. During this long time period, some phenomena originally included in the dictionary (e.g. word-formation relations, detailed annotation of terms and names) has been delegated to separate projects. In the source format, this information is preserved, but it is inconsistent and/or incomplete and it did not make it into the currently released version of the dictionary. We do not describe the source format here,[10] occasional reference to the source format is made only if it is necessary to explain a phenomenon in the currently described version of the MorfFlex 2020. The the process of transformation the source format into the resulting dictionary, the description of which is, however, rather a technical matter, is not covered also in this document. We mention only some procedural aspects of this process (particularly the way of handling homonymy) that affect the way of representation of some phenomena in the dictionary, and therefore in the data.

The MorfFlex dictionary serves as a basis for annotation consistency. The goal of the annotation is full consistency between all the data and the dictionary. An inconsistency between the data and the dictionary indicates an annotation problem or error in the dictionary. All inconsistencies are corrected and there are only full matches now, except for a small amount of wordform occurrences in the data that are not in the dictionary (but have manual analysis in the data); this applies mostly to foreign wordfoms and non-standard, sparse forms of Czech. However, if a wordform is in the dictionary, the dictionary contains all its morphological analyses (all possible lemma-tag pairs) that have been found in annotated data (and in other sources). A paradigm included in the dictionary (identified by the lemma) contains not only Standard Czech wordforms, it also provides non-standard variants and contains all forms found in the data, even defective forms, misspelling, typos, which are properly marked (see Sect. 6 and 20). The morphological annotation of a wordform that is in the data but not in the dictionary follows the same principles as applied to the dictionary.

[7] `https://ufal.mff.cuni.cz/morphodita`

[8] The latest version of the dictionary contains 125,348,899 lemma-tag-wordform triples.

[9] The word-formation relations in Czech has been delegated to derivational data sources, such as Derinet: `https://ufal.mff.cuni.cz/derinet`

[10] The source format will be described in a separate document, which will be available in the first half of 2021.

3 Morphological Dictionary

The MorfFlex dictionary covers words (tokens) that occur in real Czech texts, i.e. Czech words, loan words and foreign words, proper nouns, abbreviations, isolated letters, part of words, numbers, and also punctuation and other non-alphanumeric characters (see also note on tokenization in Sect. 21). It captures both the singular and the plural set of wordforms of all inflected words, even of proper nouns. It is not focused only on standard Czech, the paradigms provide also non-standard variants and capture style characteristics of wordforms (see Sect. 6).

MorfFlex is a flat list of lemma-tag-wordform triples (see examples in Table 1). For each *wordform*, full inflectional information is coded in a *positional tag* (see Sect. 5). Wordforms are grouped into *paradigms* according to their formal morphological behavior. The paradigm is a set of all wordforms of the word. It is represented by a unique *lemma* (see Sect. 4). Apart from traditional morphological categories, the description also contains some semantic (see Sect. 8), stylistic (see Sect. 6) and derivational (see Sect. 7) information.

Wordform	Lemma	Tag
podle	`podle-1`	`Dg-------1A----`
nepodle	`podle-1`	`Dg-------1N----`
podleji	`podle-1`	`Dg-------2A----`
nepodleji	`podle-1`	`Dg-------2N----`
podlejc	`podle-1`	`Dg-------2A---6`
nepodlejc	`podle-1`	`Dg-------2N---6`
nejpodleji	`podle-1`	`Dg-------3A----`
nejnepodleji	`podle-1`	`Dg-------3N----`
nejpodlejc	`podle-1`	`Dg-------3A---6`
nejnepodlejc	`podle-1`	`Dg-------3N---6`
podle	`podle-2`	`RR--2----------`

Table 1: Example of the wordform-lemma-tag triples in the dictionary

3.1 Principle of unique analysis

The principle of unique analysis, so called "golden rule of morphology", is applied to the dictionary. The rule says that there must not exist more than one wordform with the same lemma and tag. Lemma and tag together uniquely identify the wordform. Two different wordforms always differ either in lemma or in tag. There are two means to ensure the principle is valid:

- **lemma numbering** (see Sect. 4.1.1),
- **tag numbering** at the 15^{th} position (see Sect. 5.15),

Wordform	Lemma	Tag
po lesích	`les`	`NNIP6-----A----`
po lesech	`les`	`NNIP6-----A---1`

Table 2: Example of tag numbering

We use these means mainly for capturing homonymy of lemmas and different types of wordform variants. Each of these problematic issues is addressed differently. The former one is solved by adding a numerical index to homonymous lemmas (see Tab. 3), the latter one by adding a numerical index to 15^{th} position of tag (see Tab. 2).[11]

[11] In the tables with examples, we present the analysis (lemma and tag) for the wordform given in the left column.

Wordform	Lemma	Tag
náš <u>stát</u>	`stát-1_^(státní_útvar)`	`NNIS1-----A----`
<u>stanu</u> se vojákem	`stát-2_^(stanu_staneš)`	`VB-S---1P-AAP--`
<u>stojím</u> tu už dlouho	`stát-3_^(stojím_stojíš)`	`VB-S---1P-AAI--`
sníh <u>staje</u>	`stát-5_^(sníh)`	`VB-S---3P-AAI--`

Table 3: Example of lemma numbering

3.2 Principle of morphological differentiation

MorfFlex captures primarily such phenomena that are of formal morphological nature. It does distinguish words with the same spelling (by adding numbers to lemmas; see Sect. 4.1.1), but different formal morphological characteristics.

Within one paradigm, all wordforms are allowed to appear with morphological tags of:

- **the same** `POS` (Sect. 5.1). The `POS` value is the same for the whole paradigm (identified by one lemma). E.g., there are two indexed lemmas for *drát*: `drát-1` for noun (meaning 'wire') and `drát-2` for verb with meaning 'to pluck' (see Tab. 4); or there are the lemmas `podle-1` and `podle-2` assigned in the dictionary to the string *podle* which can be either an adverb (meaning 'meanly') or a preposition ('along'; see Tab. 1).

  Similarly, homonymous forms of inflexible words (as well as inflexible loanwords) have as many lemmas in the dictionary as they express (inflexible) `POS` (see also Sect. 10.1). For example, *přece* which is in accordance to its function in a sentence interpreted as a conjunction (meaning 'despite') or as a particle ('after all'), has two lemmas with different indexes and with different `POS` values at tags (see Tab. 4).

- **the same** `SUBPOS` (Sect. 5.2). The `SUBPOS` value (detailed part of speech) is the same for the whole paradigm except for a few exceptions:

  - Paradigm of a verb is traditionally formed by several sets of forms (present/future forms, past participles, passive participles, transgressives, etc.) which are distinguished in the second `SUBPOS` position. However, all set of verbal wordforms are identified by one lemma.
  - Short (nominal) forms of adjectives (e.g. *mlád* 'young') have a special value at the second `SUBPOS` position but together with long forms (e.g. *mladý* 'young') they are part of one paradigm. See examples in Table 4.
  - In the paradigms of personal pronouns, a different value of `SUBPOS` indicates enclitic forms (cf. Sect. 11.1).

  The `SUBPOS` value serves as an indicator which tag positions are to be filled and which not (i.e. the categories of `GENDER`, `NUMBER`, `CASE`, `TENSE`, etc.; see combination in Sect. 22.2). We are using unique values for `SUBPOS` categry so that the value of the major speech category (the `POS` value) can be determined unambiguously from the value of the `SUBPOS` category. The only exceptions are abbreviations (`POS` = `B`) and segments (`POS` = `S`); potentially each `SUBPOS` value are possible for these POS (see more in Sect. 14 and Sect. 16).

- **the same** `GENDER` (Sect. 5.3) in case of nouns. The `GENDER` value of nouns is the same for the whole paradigm. For example, there are the lemma `rys-1` for masculine animate noun with meaning of an animal ('lynx'), and lemma `rys-2` for masculine inanimate noun with meaning of 'feature in face' or 'drawing'). Or there are two distinct paradigms of word

If there is context in the left column, then the wordform for which the analysis is given is underlined.

kredenc: `kredenc-1` as masculine and `kredenc-2` as feminine, even if they have the same meaning ('cupboard').[12]

The rule of different gender applies only to nouns (`NN`). For other part of speech with agreement gender (adjectives, etc.), a different value of `GENDER` does not necessarily mean a different paradigm.

- **the same** `ASPECT` (Sect. 5.13) in case of verbs. The value of `ASPECT` is the same for the whole paradigm. If a verb appears in two aspects, there should be two lemmas distinguished by the number. An example is the verb *pootevírat*. There must be two verbs with different aspects: `pootevírat-1` with imperfective aspect (meaning 'open slightly') and `pootevírat-2` with perfective aspect ('gradually open').

See examples in Table 4.

Wordform	Lemma	Tag
ostrý drát	drát-1	NNIS1-----A----
drát peří	drát-2	Vf--------A-I--
ne peníze, ale přece lásku	přece-1	J^----------
přece jen nelhal	přece-2	TT----------
je mlád	mladý	ACYS------A----
je mladý	mladý	AAMS1----1A----
dej mu	on-1	P5ZS3--3-------
jemu to nevadí	on-1	PEZS3--3-------
rys ostrovid	rys-1	NNMS1-----A----
rys v obličeji	rys-2	NNIS1-----A----
ošklivý pavouci	pavouk-1	NNMP1-----A----
pavouky dvouher v tenise	pavouk-2	NNIP1-----A----
pod kredencem	kredenc-1	NNIS7-----A----
pod kredencí	kredenc-2	NNFS7-----A----
moje dítě	dítě-1	NNNS1-----A----
moje děti	dítě-2	NNFP1-----A----
oko na polévce i lidské	oko-1	NNNS1-----A----
oka na polévce	oko-1	NNNP1-----A----
oči lidské	oko-2	NNFP1-----A----
pootevírat trochu dveře	pootevírat-1_^(otevírat_trochu)	Vf--------A-I--
postupně pootevírat dveře	pootevírat-2_^(postupně_otevírat)	Vf--------A-P--

Table 4: Examples: Principle of morphological differentiation

[12]The rule of same `GENDER` in noun paradigms is quite problematic is some cases. One group of problematic cases consists of words whose `GENDER` fluctuates even if they have the same meaning (cf. *ta kredenc* (fem.) - *ten kredenc* (masc.) 'cupboard'; *brambora* (fem.) - *brambor* (masc.) 'potato'; *ty bacily* (masc. inam.) - *ti bacilové* (masc. anim) 'bacillus'). Each `GENDER` variant is captured by a separate (indexed) paradigm/lemma. Particularly problematic are the cases of fluctuation between the masculine animate and non-animate gender. Some naturally inanimate words may have endings of the masculine animate in the genitive and accusative sigular (e.g. *mám nového forda* 'I have a new Ford', *drží toho hřiba* 'she holds the mushroom'). The genitive/accusative wordforms with an animate ending are captured as a wordform variant within the non-animate paradigm (with `I` value on `GENDER` position). However, this solution is not adequate, especially due to gender agreement of the dependent adjective or predicate.

Another problematic group are nouns that have different grammatical gender in the singular and plural set of wordforms (cf. *dítě* (neut.) - *děti* (fem.) 'child - children', including its derivatives *biodítě* 'bio-child', etc.). A similar problem is with words that have two sets of plural wordforms of two different genders (e.g. *ta oka* (neut.) - *ty oči* (fem.) 'loops - eyes'). To follow the rule of the same `GENDER`, we have separated the respective singular and/or the plural sets of wordforms into different paradigms. See examples in Tab. 4.

3.3 Principle of unique paradigm

MorfFlex does not capture any differences in meanings of homonymous or polysemous words. It means that there are no two identical paradigms (set of lemma-tag pairs) in the dictionary. It tries to avoid lexical distinctions that are not morphologically based: lexical homonymy and polysemy are not marked for words which do behave morphologically in the same way. Due to a large number of complicated cases, we do not take into account derivational, stylistic nor semantic differences.

Wordform	Lemma	Tag
přiletěl jeřáb	`jeřáb-1_^(pták)`	`NNMS1-----A----`
u silnice roste jeřáb	`jeřáb-2_^(strom;stroj)`	`NNIS1-----A----`
jeřáb na staveništi	`jeřáb-2_^(strom;stroj)`	`NNIS1-----A----`
vodovodní kohoutek	`kohoutek-2_^(květina;uzávěr)`	`NNIS1-----A----`
luční kohoutek	`kohoutek-2_^(květina;uzávěr)`	`NNIS1-----A----`
slepička a kohoutek	`kohoutek-1_^(pták)`	`NNMS1-----A----`
kovová matka	`matka`	`NNFS1-----A----`
matka spí	`matka`	`NNFS1-----A----`
palička a palič	`palička_^(*2)`	`NNFS1-----A----`
palička na maso	`palička_^(*2)`	`NNFS1-----A----`
pracuje jako ekonomka	`ekonomka_^(*2)`	`NNFS1-----A----`
studuje ekonomku	`ekonomka_^(*2)`	`NNFS4-----A----`
Barča a Helča	`Barča_;G_;Y`	`NNFS1-----A----`
sejdeme se na Barče	`Barča_;G_;Y`	`NNFS6-----A----`

Table 5: Examples: Principle of unique paradigm

Thus, we do not distinguish lemmas that have the same paradigm and that have:

- **different meaning**. If two (or more) words share all the morphological properties, there is only one lemma/paradigm in the dictionary. It means that we do not distinguish between the words *kohoutek* as 'a flower' and *kohoutek* as 'a tap' because both have the same inflectional model, namely for masculine inanimate noun. On the other hand, there is a different word *kohoutek* ('a small cock'), that has different inflectional model for masculine animate noun.

  Similarly, there is one paradigm/lemma for word *palička* which has two meaning. First meaning is the meat knocking tool 'tenderizer' and with this meaning, the word *palička* is derived from the word *palice* 'mallet'. Another meaning is 'arsonist-female' and with this meaning, the word *palička* is derived from the word *palič* 'arsonist-male'. These two meanings have the same inflectional paradigm, so there is only one paradigm in the dictionary. The lemma in the dictionary, `palička_^(*2)`, contains the derivational comment (see Sect. 4.2.5) indicating that the paradigm was automatically derived from the word *palič* 'arsonist-male' (see more in Sect. 7). This does not mean that the paradigm with lemma `palička_^(*2)` cannot be used to analyze the wordforms of the word *palička* derived from *palice* 'mallet'.

- **different derivational model**. If two (or more) words share all the morphological properties, there is only one lemma/paradigm, even if they differ in word-formation relations. E.g., the word *jeřábník* ('man who works with a crane-device') is derived from *jeřáb* as a 'device'. It is not possible to derive *jeřábník* from *jeřáb* as a 'tree', but we do not distinguish the meanings 'tree' and 'device', because they do not differ from the inflectional point of view. Another example is word *matka*, that has two different meanings, namely a 'nut' and 'mother'. These two meanings have the same inflectional paradigm, but their derivational behavior differs. It is possible to derive a possessive adjective only from word *matka* as a 'mother'. But these derivational differences are not captured in the dictionary.

- **different style value**. If two (or more) words share all the morphological properties, there is only one lemma/paradigm, even if there is a difference in style value. E.g., there is a

standard word *ekonomka* as 'female economist' (automatically derived from word *ekonom* 'male economist'; see derivational comment in the lemma) and a non-standard one with meaning 'school of economics'. These two meanings have the same inflectional paradigm, they only differ in terms of the stylistic characteristics, but we do not capture this difference in our description. There is only one lemma/paradigm without any style label. See more about style labeling in Sect. 6.

- **different semantic label**. If two (or more) words share all the morphological properties, there is only one lemma/paradigm, even if there is a difference in so-called name label (Sect. 4.2.2). E.g., there is a female name *Barča* (captured by name label `Y`) and also common name of Barricade House of Culture *Barča* (captured by label `G`). The names have the same inflectional paradigm, so there is only one lemma/paradigm with all semantic flags relevant.

See examples in Tab. 5.

4 Lemma Structure

Lemma represents the whole paradigm. It has two parts. The first part, the lemma proper (Sect. 4.1), has to be a unique identifier of the lexical item/paradigm. The second part, so-called AddInfo (Sect. 4.2), is optional, it is not part of the identifier and contains additional information about the paradigm, e.g. semantic or derivational information or style label. They are related to the whole paradigm (set of wordforms) belonging to the given lemma.

A lemma is the same for all wordforms in the paradigm. Two lemmas with different AddInfo must differ in lemma proper. To achieve that, numbers are used to distinguish the lemmas.

The formal description of the lemma structure is in Tab. 6. Spaces were inserted between nonterminals to improve readability. In fact, there are no spaces in lemmas. Capitalized multi-character symbols are nonterminals. All other symbols are terminals.[13]

```
Lemma ::= LemmaProper | LemmaProper AddInfo | Numeric
LemmaProper ::= Word | Word - Number | SpecialChar
Word ::= Letter | Letter Word
Letter ::= A | a | Á | á | Ä | ä | ...  | Z | z | Ž | ž | ’ | Other
Other ::= any character from the set covered by UNICODE, not listed before
Number ::= NonZero | NonZero Number0
Number0 ::= Digit | Digit Number0
NonZero ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
Digit ::= 0 | NonZero
SpecialChar ::= ! | " | # | $ | % | & | ’ | ( | ) | * | + | , |
                - | . | / | : | ; | < | = | > | ? | @ | [ | \ | ] |
                _ | ‘ | { | | | } | § | ○
Numeric ::= Number0 | Number0 NumSep Numeric NumSep ::= .  | , AddInfo ::=
Reference Name Style Comment
Reference ::= <empty> | ‘ LemmaProper
Name ::= <empty> | _; Name1 Name
Style ::= <empty> | _, Style1
Comment ::= <empty> | _^ Comment1
Name1 ::= Y | E | G | U | m | o
Style1 ::= n | a | s | h | e | l | v | i
Comment1 ::= CommentPart | CommentPart _^ Comment1
CommentPart ::= ( Explanation ) | ( Derivation ) | ( Variant ) Variant ::= ^
VariantName * * Word VariantName ::= D D | G C | D S Explanation ::= CommentChar
| CommentChar Explanation
Derivation ::= * Number Word | * Word
CommentChar ::= Letter | Digit |
                ! | " | # | $ | % | & | ’ | + | , | . | ~ |
                - | / | : | ; | < | = | > | ? | @ | [ | \ | ] |
                _ | ‘ | { | | | } | ^ | § | ○
```

Table 6: Formal structure of the lemma

[13] Any character that is Letter in the Unicode standard can appear in place of the Letter nonterminal. In the non-ASCII area this most frequently applies to the Czech accented characters: ÁáČčĎďÉéĚěÍíŇňÓóŘřŠšŤťÚúŮůÝýŽž. However, other characters occur in names (e.g. German ÄäÖöÜü, Serbo-Croatian Ćć) and in foreign words (e.g. Slovak ĽľĹĺÔôŔŕ). Standard HTML entities (such as & for & or à for à) are also allowed. PDT 1.0 was encoded in the ISO Latin 2 codepage, so representing any West European characters required using entities. PDT 2.0 and the current version of PDT-C 1.0 shall be encoded in UTF8, so few entities are needed.

The single quote (’) in some transcriptions of non-Latin alphabets (e.g. in Chinese *Mao C’-tung*, Hebrew *Be’er Sheva’*) and if it marks deleted parts of words (e.g. English *don’t*, French *d’Artagnan*), is considered a SpecialChar and it splits the string into three tokens (*d ’ Artagnan*). See also in Sect. 21.

4.1 Lemma proper

Lemma proper has to be a unique identifier of the paradigm. Usually it is the base form of the word (e.i. nominative singular for nouns, the same plus masculine positive for adjectives, similarly for pronouns and numerals; verbs are represented by their infinitive forms),[14] possibly followed by a number distinguishing different lemmas with the same spelling but different formal morphological behavior.

Lowercase and uppercase letters in lemma. Words that have to be always capitalized in writing, have their lemma capitalized as well (ex. *Král* as a 'surname', *Písek* as a 'surname' and as a 'town'; cf. Table 7 and Table 8). On the other hand, a lower-case lemma is automatically assumed to have also a wordform with uppercase letters; cf. Tab. 7. Capitalized lemmas are usually, but not always assigned a name label (see Sect. 4.2.2).

Base form in lemma is case-sensitive. The case of the initial letter (any different character in a base form) is a sufficient distinction to consider the two lemmas to be different. As a consequence, *král* ('king') and *Král* ('surname') need not be distinguished by numbers; cf. Tab. 7.[15]

Wordform	Lemma	Tag
král a královna	`král`	`NNMS1-----A----`
pan Král	`Král_;Y`	`NNMS1-----A----`
Král a královna odešli.	`král`	`NNMS1-----A----`
KRÁL a královna	`král`	`NNMS1-----A----`
pan KRÁL	`Král_;Y`	`NNMS1-----A----`

Table 7: Examples: Lowercase and uppercase letters in lemma

4.1.1 Lemma number

Lemma number distinguishes different lemmas with the same spelling but different formal morphological behavior (cf. Sect. 3.2). Thus, we have the lemma `pět-1'5` for number ('five') and lemma `pět-2_^(zpívat)` for verb ('to sing'); cf. Tab. 8.

However, there are unnumbered and numbered lemmas with the same base form in the dictionary and data. The used numbers also often do not form a continuous sequence. However, all different paradigms with lemmas of same base form are distinguished. Cf. lemmas for word *Písek* in Tab. 8.

Wordform	Lemma	Tag
pět dětí	`pět-1'5`	`Cn-S1----------`
nemohl pět	`pět-2_^(zpívat)`	`Vf--------A-I--`
ta sršeň	`sršeň-1`	`NNFS1-----A----`
ten sršeň	`sršeň-2`	`NNMS1-----A----`
pan Písek	`Písek_;Y`	`NNMS1-----A----`
do Písku	`Písek-2_;G`	`NNIS2-----A----`
přivezl písek	`písek`	`NNIS4-----A----`

Table 8: Examples: Lemma number

Special lemma index number. There are special lemma numbers used for a predefined word types (see ex. in Tab. 9 and respective sections):

[14] The cases where more than one "base form" is available are described in Sect. 9.3.2.

[15] The capitalization inside words (e.g. *AIDSu*, *McDonald*) are not sufficiently addressed in the dictionary and annotation.

-33 The lemma number `-33` is for an isolated letter (see more in Sect. 15),

-77 The lemma number `-77` is for a foreign word (see more in Sect. 17),

-88 The lemma number `-88` is for a special type of abbreviation (see more in Sect. 14),

-99 The lemma number `-99` is for an authors' signature (see more in Sect. 14.2.3).

Conventions of this kind exist solely for the convenience of a human reader.

Wordform	Lemma	Tag
písmeno A	`A-33`	`Q3----------`
cizí slovo black	`black-77`	`F%----------`
zkratka OP	`OP-88`	`BNXXX-----A----`
šifra mat	`mat-99_;Y`	`BNXXX-----A----`

Table 9: Examples: Special lemma index number

4.2 Additional information about the paradigm (AddInfo)

The second part of lemma is optional, it is not part of the identifier and contains additional information about the paradigm (AddInfo in sequel), namely semantic and derivational information, style label. The information attached to lemma is related to the whole paradigm belonging to the given lemma (e.g. a style label (see Sect. 4.2.3) attached to lemma is valid for all wordforms of the paradigm).

4.2.1 Reference

The reference is one of the means of explaining the meaning of the lemma. It is systematically used only with spelled-out numbers and with Roman numbers: in the additional lemma field, there is a corresponding number in Latin numerals. The numeric item is separated from the proper lemma by an apostrophe (cf. examples in Tab. 10).

The numerical reference is used also for the derivatives of the basic definite numerals (i.e. for words such as *šest* 'six', *šestý* 'sixth', but also *šestina* 'sixth', *šestkrát* 'six times'). The numerical reference only indicates which number the word refers to. Replacement of the word with the number in the reference is not assumed in all cases (e.g. *šestina* 'sixth' is not replaceable with *6*).

Wordform	Lemma	Tag
v šest hodin	`šest'3`	`Cl-S4----------`
šestina	`šestina'6_^(*3)`	`NNFS1-----A----`
římské číslo M	`M-3‘1000`	`C}----------`

Table 10: Examples: Reference

4.2.2 Name label

Words that have to be always capitalized in writing (names and other terms), have their lemma capitalized as well (e.g. *Novák* as a surname, *Praha* as a name of a town). Capitalized lemmas are usually (but not always) assigned a so-called name label.

The capitalized word type is indicated by `"_;"` followed by a letter (see Tab. 11). More than one name label may apply to one lemma. The name labels are sorted alphabetically except for *m* which is always the last. Examples can be found in the respective sections, mostly in Sect. 13.

Label	Description	Example
Y	personal name	*Petr, Novák, Aristoteles*
E	member of nation, inhabitant of territory	*Slovák, Newyorčan*
G	geographical name	*Praha, Tatry, Václavák*
U	scientific terminology	*Australopithecus, Rh, Hydrosulfit*
m	other proper name	*Madeta, Opel, Sázavafest*
o	color indication	*červený, modrobílý*

Table 11: Name labels

Color indication. Adjective color names are marked by a special label `_;o` (e.g. lemma `zelený_;o` for adjective *zelený*). This is mainly due to the specific morphological behavior of these words. Color names can be put together almost indefinitely into compound names (*červenomodrobílý* 'lit. red-blue-white' etc.). Compared to adjectives, they have different derivative behavior: adjective color names, besides having the usual adverb derivation (*zeleně* 'greenly'), have a very regular adverb derivation with the suffix *-o* and the prefix *na-*; (ex. *tam je zeleno* lit. 'there is greenly', *natřít nazeleno* lit. 'paint greenly').

4.2.3 Style label

Lemmas can be stylistically classified. The style flag is indicated by "_," followed by a letter (see the values in Tab. 12). Standard lemmas have no stylistic flag but any lemma intended for special usage (bookish, dialect, slang, etc.) should be marked as such. At most one style flag is attached to any lemma. For automatically derived lemmas (lemmas with Derivation Info; see Sect. 4.2.5), the style label is not inherited (see more in Sect. 7). More about style labeling is in Sect. 6.

Label	Description	Example
s	standard word, bookish	*asyl, kurs*
a	archaic word	*šlojíř, slout, these*
h	non-standard word	*Karlák, šutrák,*
i	distortion, typo	*součastník, intensívní*
n	dialect	*šufánek, čučkař*
e	expressive word	*kládička, sajtnička*
l	slang, argot	*genštáb, vertajmka*
v	vulgar word	*vlezdoprdelizmus*

Table 12: Style labels

4.2.4 Variant info

Orthographic and stylistic variants of a word (e.g. an archaic variant *these*, a standard variant *teze*, and a non-standard variant *téze* 'thesis') are captured in separate paradigms with different variant lemmas, but they are "interconnected" using the Variant Info field in the AddInfo part of lemma. One of the variants is selected as "basic" (the standard one, i.e. *teze*) and other variants (non-standard *téze* and archaic *these*) refer to it: in comment brackets (following the caret sign ^), the type of variant (see Tab. 13) is given after the other caret sign; after the two asterisks, there is the lemma of the basic variant. For example, lemma `these_,a_^(^DD**teze)` "interconnect" archaic variant *these* with the basic one and lemma `téze_,h_^(^GC**teze)` "interconnect" non-standard variant *téze*. More about capturing variants is in Sect. 9.1.

Type	Description
DD	standard variant
GC	non-standard variant
DS	distortion, typo

Table 13: Variant info

4.2.5 Derivation info

For regular derivations, the lemma contains information about base lemma it is derived from. For example, lemmas of possessive adjectives (e.g. `otcův_^(*3ec)`) contain information about the noun they are derived from (*otcův* 'father's' ← *otec* 'father'). The originating lemma is (for space saving reasons only) written in the form of a rule, which has two parts:

1. number of characters to remove from the end of the current lemma identification, expressed as a decimal number
2. characters to add to form the original lemma.

Only the proper lemmas are both input and output of this process (including the lemma number, if present). Each such rule must start with an asterisk to distinguish it from the explanation Comment (see next Sect. 4.2.6). Thus, for example, derivation info in lemma `otcův_^(*3ec)` means remove 3 characters, add *ec* to get `otec` or derivation info in lemma `koníčkův-2_^(*5ek-2)` means remove 5 characters, add *ek-2* to get `koníček-2`.

In the current version of the dictionary, only lemmas automatically derived from derivational patterns have filled-in derivation info. In the source format (from which the dictionary is made), there are also manually created comments on word-formation relation. However, due to their incompleteness and inconsistency, they were not transfered into the dictionary. More about derivations is in Sect. 7.

4.2.6 Explanational comment

Any string in parentheses can be used as explanation of the lemma/paradigm usage (meaning). The string cannot contain spaces or parentheses. The underscore character is used to replace space, square brackets are used instead of parentheses. The explanation is in Czech. Example of usage, synonym etc. can also be used or both a verbal description and an example can be mixed.

An explanational comment is optional in all cases. There is no rule when the comment can be used and when cannot. The lemmas automatically derived from derivational patterns (with Derivation info; see Sect. 4.2.5) do not have the explanational comment (see Sect. 7).

Note. The inconsistency in the comments is historically caused by how careful or active the dictionary administrator was. The comments were never intended to replace the interpretation of meaning, they are primarily perceived as "hints" for annotators and as such they can be very useful. In the context of the principle of a unique paradigm (Sect. 3.3), it must be emphasized that the paradigm/lemma is applicable to all cases with same morphological behavior, regardless of the explanation given in the comment.

5 Tag Structure

Every tag is represented as a string of 15 symbols. Each position in the string corresponds to one morphological category according to a more or less traditional system of formal morphology.

A value in each category is represented as a single symbol, mostly an uppercase letter of the English alphabet (for example, `P` for plural), sometimes also another symbol (`f` for an infinitive, `^` for conjunction). In some cases no distinction among "traditional" values is being made where the possibility of correctly distinguishing them is low based on local context. For example, possessive pronouns in third person plural are not distinguished in gender and number, nor in case; passive participles (both active and passive) in masculine are not distinguished in animateness, etc. Typically, a letter `X` is used where all possible values might be considered in a more detailed tagset, or a special letter is used with a more restricted choice (e.g. `Y` is used for masculine animate/masculine inanimate "non-distinction").

Non-applicable values are denoted by a single hyphen - (e.g. tense for nouns). Which categories are applicable/non-applicable is determined by the 2^{nd} position of the tag (`SUBPOS`; Sect. 5.2). The tables of applicability/non-applicability of the tag categories related to a `SUBPOS` value are in Sect. 22.2.

An overview of the 15 tag positions is in Table 14. The categories and their values are described in the following sections.

#	Category Name	Description in English	Description in Czech
1	`POS`	Part of Speech	Slovní druh
2	`SUBPOS`	Detailed Part of Speech	Slovní poddruh
3	`GENDER`	Gender	Rod
4	`NUMBER`	Number	Číslo
5	`CASE`	Case	Pád
6	`POSSGENDER`	Possessor's Gender	Rod vlastníka
7	`POSSNUMBER`	Possessor's Number	Číslo vlastníka
8	`PERSON`	Person	Osoba
9	`TENSE`	Tense	Čas
10	`GRADE`	Degree of Comparison	Stupeň
11	`NEGATION`	Negation (by prefix)	Negace
12	`VOICE`	Voice	Slovesný rod
13	`ASPECT`	Aspect	Vid
14	`AGGREGATE`	Aggregate	Agregát
15	`VAR`	Variant	Varianta

Table 14: Categories in positional tag

5.1 Part of speech (1st position)

The `POS` category denotes the main part of speech, according to the traditional Czech scheme known from both comprehensive as well as high-school grammars. However, the assignment of the `POS` values is driven mainly by the requirements of consistency in further processing, therefore it is not always in line with traditional grammars. In addition to the ten traditional parts of speech, we distinguish also the abbreviations (Sect. 14), foreign words (Sect. 17), segments (Sect. 16), isolated letters (Sect. 15) and punctuation (Sect. 21). An overview of the 1^{st} position values is in Table 15.[16]

[16] In morphological analysis using automatic tool, there can appear one more value of `POS`, namely `X`. In that case, the `SUBPOS` value is always `©`. It means that the wordform is not in the dictionary. In the PDT-C annotation, these values are not used; we analyze all wordforms according to the rules described here, even those that are not in the dictionary.

POS	Description	POS	Description
A	Adjective	P	Pronoun
B	Abbreviation	Q	Letter
C	Numeral	R	Preposition
D	Adverb	S	Segment
F	Foreign word	T	Particle
I	Interjection	V	Verb
J	Conjunction	Z	Punctuation
N	Noun		

Table 15: Values of POS category (1^{st} position)

5.2 Detailed part of speech (2nd position)

This category is the most detailed one; it contains values for fine-grained distinction of the major part of speech category. Its primary technical purpose, however, is to serve as an indicator of applicability/non-applicability of other categories (i.e. the categories GENDER, NUMBER, CASE, etc. up to the last category, VAR). We are using unique values so that the value of the major speech category can be determined unambiguously from the value of the SUBPOS category. There are two exceptions: abbreviations (B) and segments (S), where potentially each SUBPOS value are possible; see Sect. 14 and 16).

An alphabetical list of all 66 SUBPOS values is in the appendix (Sect. 22.1). There are also tables of applicability/non-applicability of other categories of tag related to SUBPOS value (Sect. 22.2). Here, for each of the POS values, a list of detailed part of speech (SUBPOS) values and examples is given (the tables are arranged alphabetically according to the POS value).

A Adjective

POS	Detailed part-of-speech used (SUBPOS)
A	A C G M O U

Examples:

POS & SUBPOS	wordform & translation	description
AA	*technický* 'technical'	adj. general
AC	*mlád* '(be) young'	adj. nominal
AG	*kouřící* 'smoking'	adj. derived from present transgressive
AM	*zvítězivší* 'having-won'	adj. derived from past transgressive
AO	*svůj* 'to be himself'	*svůj*, *nesvůj* in specific usage, *tentam*
AU	*Martinin* 'Martina's'	adj. possessive

B Abbreviation

For POS with a value of B, any SUBPOS value is allowed (see Sect. 14). In the tables, the SUBPOS values that occurred in the PDT-C data are listed.

POS	Detailed part-of-speech used (SUBPOS)
B	^ A b N

Examples:

POS & SUBPOS	wordform & translation	description
B^	*tzn* 'i.e.'	abbr. of conjunction
BA	*aj* 'and other'	abbr. of adjective phrase
Bb	*atd* 'and so on'	abbr. of adverb phrase
BN	*USA*	abbr. of noun phrase

C Numeral

POS	Detailed part-of-speech used (SUBPOS)
C	= a d h j k l n o r v w y z }

Examples:

POS & SUBPOS	wordform & translation	description
C=	*1.23* '1.23'	num. written using digits
C}	*XIV* 'XIV'	Roman numeral
Ca	*několik* 'several'	num. cardinal indef., non-adjectival declension
Cd	*čtverý* 'four-kinds'	num. generic, adjectival declension
Ch	*několikerý* 'several-kind'	num. generic indef., adjectival declension
Cj	*čtvero* 'four-kinds-of'	num. generic, noun usage
Ck	*několikero* 'several-kinds'	num. generic indef., noun usage
Cl	*čtyři* 'four', *pět* 'five'	num. cardinal, no gender
Cn	*jeden* 'one', *dva* 'two'	num. cardinal, agreement gender
Co	*několikrát* 'several-times'	num. multiplicative indef.
Cr	*druhý* 'second'	num. ordinal
Cv	*sedmkrát* 'seven-times'	num. multiplicative
Cw	*několikátý* 'several'	num. ordinal indef.
Cy	*nejeden* 'not-one'	num. cardinal indef., agreement gender
Cz	*sto* 'hundred'	num. cardinal, noun usage

More details about detailed part of speech of numerals are in Sect. 11.2.

D Adverb

POS	Detailed part-of-speech used (SUBPOS)
D	b g

Examples:

POS & SUBPOS	wordform & translation	description
Db	*nahoru* 'up'	no degrees of comparison, no negation
Dg	*rychle* 'quickly'	negation, degrees of c. possible

More details about detailed part of speech of adverbs are in Sect. 11.3.

`F` Foreign word

POS	Detailed part-of-speech used (SUBPOS)
F	%

Examples:

POS & SUBPOS	wordform & translation	description
F%	*The*	foreign word

`I` Interjection

POS	Detailed part-of-speech used (SUBPOS)
I	I

Example:

POS & SUBPOS	wordform & translation	description
II	*ach* 'oh!'	interjection

`J` Conjunction

POS	Detailed part-of-speech used (SUBPOS)
J	* , ^

Examples:

POS & SUBPOS	wordform & translation	description
J*	*krát* 'times'	binary math. operations
J,	*že* 'that'	conj. subordinate
J^	*a* 'and'	conj. coordinating

`N` Noun

POS	Detailed part-of-speech used (SUBPOS)
N	N

Example:

POS & SUBPOS	wordform & translation	description
NN	*robot* 'robot'	any noun incl. proper

P Pronoun

POS	Detailed part-of-speech used (SUBPOS)
P	1 4 5 6 7 8 9 D E H K L P Q S W Y Z

Examples:

POS & SUBPOS	wordform & translation	description
P1	*jehož, jejíž* 'whose'	p. relative possessive
P4	*jaký* 'what'	p. interrogative/relative with adj. declension
P5	*mu* 'him'	p. personal for 3^{rd} person, clitic
P6	*sebe* 'himself'	p. personal reflexive in long forms
P7	*se, si*	p. personal reflexive clitic
P8	*svůj* 'his'	p. personal reflexive possessive
P9	*jeho* 'his', *její* 'her'	p. personal possessive for 3^{rd} person
PD	*tento* 'this'	p. demonstrative
PE	*on* 'he', *ona* 'she'	p. personal for 3^{rd} person
PH	*mě* 'me', *ti* 'you'	p. personal, no gender, clitic
PK	*někdo* 'somebody'	p. indefinite, no gender
PL	*všechen* 'all', *sám* 'alone'	p. delimiting
PP	*já* 'I', *ty* 'you'	p. personal, no gender
PQ	*kdo* 'who'	p. interrogative/relative, no gender
PS	*můj* 'my'	p. personal possessive
PW	*ničí* 'nobody's'	p. negative, adjectival declension
PY	*nic* 'nothing'	p. negative, no gender
PZ	*nějaký* 'some'	p. indefinite with adj. declension

More details about detailed part of speech of pronouns are in Sect. 11.1.

Q Letter

POS	Detailed part-of-speech used (SUBPOS)
Q	3

Examples:

POS & SUBPOS	wordform & translation	description
Q3	*A*	isolated letter

R Preposition

POS	Detailed part-of-speech used (SUBPOS)
R	F R V

Examples:

POS & SUBPOS	wordform & translation	description
RF	*nehledě* 'regardless'	part of preposition
RR	*v* 'in'	prep. general (no vocalization)
RV	*ve* 'in'	prep. with vocalization

`S` Segment

For `POS` with a value of `S`, any `SUBPOS` value is allowed (see Sect. 16). In the table, the `SUBPOS` values that occurred in the PDT-C data are listed.

`POS`	Detailed part-of-speech used (`SUBPOS`)
`S`	`2 A b N n`

Examples:

`POS` & `SUBPOS`	wordform & translation	description
`S2`	*sci*	prefixal segm.
`SA`	*upový*	postfixal segm. of adjective
`Sb`	*line*	postfixal segm. of adverb
`SN`	*upista*	postfixal segm. of noun
`Sl`	*ti*	postfixal segm. of numeral

`T` Particle

`POS`	Detailed part-of-speech used (`SUBPOS`)
`T`	`T`

Examples:

`POS` & `SUBPOS`	wordform & translation	description
`TT`	*jen* 'only'	particle

`V` Verb

`POS`	Detailed part-of-speech used (`SUBPOS`)
`V`	`B c e f i m p q s t`

Examples:

`POS` & `SUBPOS`	wordform & translation	description
`VB`	*dělám* '(I) do'	present/future form
`Vc`	*bychom* '(we) would'	conditional of *být* 'to be'
`Ve`	*dělajíce* '(they-)doing'	transgressive present
`Vf`	*dělat* '(to) do'	infinitive
`Vi`	*dělejme* '(let's) do'	imperative
`Vm`	*udělav* 'having-done'	transgressive past
`Vp`	*dělali* '(they) did'	past participle
`Vq`	*dělalť* '(he) did'	archaic past participle with *-ť*
`Vs`	*děláno* '(it) was-being-done'	passive participle
`Vt`	*dělámť* '(I) do'	archaic present/future form with *-ť*

`Z` Punctuation

`POS`	Detailed part-of-speech used (`SUBPOS`)
`Z`	`:`

Examples:

`POS` & `SUBPOS`	wordform & translation	description
`Z:`	*, %*	punctuation, non-alphanumeric character

5.3 Gender (3rd position)

Grammatical gender is being described at the 3^{rd} position – both the lexical gender of nouns, as well as the agreement gender of verbs, adjectives, pronouns and numerals.

Czech grammatical gender is considered to have four different values: masculine animate, masculine inanimate, feminine, and neuter (`M`, `I`, `F` and `N`, respectively). The tags denoting ambiguous combinations of the four basic gender tags are: `H` (feminine or neuter), `Q` (feminine singular or neuter plural), `T` (masculine inanimate or feminine in plural), `Y` (for masculine, regardless of animateness), and `Z` (for not feminine forms). The tag `X` is used in its usual sense ("any gender"). Except for `X`, the other ambiguous tags are never used for nouns or long adjectival forms. An overview of the 3^{rd} position values is in Tab. 16.

GENDER	Description	Examples
F	Feminine	*píseň, malá,*
H	{F, N} Feminine or Neuter	*udělajíc, dvě, moje*
I	Masculine inanimate	*dům, malý*
M	Masculine animate	*učitel, mladí, oni*
N	Neuter	*město, malé, běhalo*
Q	Feminine (singular only) or Neuter (plural only) – only with participles and nominal forms of adjectives	*schopna, běhala, ráda*
T	Masculine inanimate or Feminine (plural only) – only with participles and nominal forms of adjectives	*schopny, běhaly, rády*
X	Any	*otcovic, jejich, mými*
Y	{M, I} Masculine (either animate or inanimate)	*schopen, běhal, jeden, on*
Z	{M, I, N} Not feminine – only for (some) pronoun forms and certain numerals	*mého, jednoho* (genitive)

Table 16: Values of `GENDER` category (3^{rd} position)

NUMBER	Description	Examples
D	Dual	*dvěma malýma nohama*
P	Plural	*dvě malé nohy*
S	Singular	*malá noha*
W	Singular for feminine gender, plural with neuter – only in participle, nominal adjective with `Q` gender	*schopna, běhala*
X	Any	*finále, jejich*

Table 17: Values of `NUMBER` category (4th position)

5.4 Number (4th position)

The number category mostly takes on only one of the two standard values: `S` for singular or `P` for plural. Nevertheless, Czech still uses so-called dual number for several nouns denoting symmetrical body parts: *oči* 'eyes', *ramena* 'arms', *nohy* 'legs', *uši* 'ears' (but not, for example, *kolena* 'knees'). Due to strong agreement rules, this distinction applies also to adjectives, pronouns and for numeral *dvě* 'two'. However, as the dual manifests itself only in the instrumental case, it is not necessary to make this distinction for verb number (the nominal agreement only applies to the nominative case). Also, as in all the other cases (nominative through locative) this distinction does not appear on the surface, we also do not make this distinction there. Moreover, all the nouns displaying the dual forms are in feminine (the plural of *prsa* 'breasts' (of neuter gender) is considered an ordinary

plural), thus the D value is actually used only in tags having also the feminine value of the gender category.

The special value W is used in connection with the GENDER value Q, in order to distinguish the combination of (feminine, singular) and (neuter, plural) from the simple cases (see ex. in Tab. 18).

The value X is used with undeclinable nouns and adjectives, and with pronouns with systematically ambiguous number. It is also used for abbreviations, where the grammatical number distinction is systematically hidden. An overview of the 4^{th} position values is in Tab. 17.

Wordform	Lemma	Tag
matka ja churava	churavý	ACQW------A----
děvčata jsou churava	churavý	ACQW------A----
matka přišla	přijít	VpQW----R-AAP--
děvčata přišla	přijít	VpQW----R-AAP--

Table 18: Examples: Special value W of NUMBER category

5.5 Case (5th position)

Czech traditionally distinguishes among seven cases: nominative, genitive, dative, accusative, vocative, locative and instrumental. Traditionally, the cases are numbered, see Tab. 19).

CASE	Description	Example
1	Nominative	*žena, ženy*
2	Genitive	*ženy, žen*
3	Dative	*ženě, ženám*
4	Accusative	*ženu, ženy*
5	Vocative	*ženo, ženy*
6	Locative	*ženě, ženách*
7	Instrumental	*ženou, ženami*
X	Any	*finále*

Table 19: Values of CASE category (5th position)

Once a case is relevant for some detailed part of speech category (i.e. for some SUBPOS), it can take any of the seven values (with negligible – and questionable – exceptions for vocative for personal pronouns). Virtually all nouns, adjectives, pronouns, and numerals express case, and for agreement purposes, we use the CASE category also for all prepositions. Having decided that verb valency is not part of morphological processing, there is no case agreement information linked to verbs. However, there is one case where CASE is present at a verb form: in passive participle (Vs). However, we distinguish only accusative of feminine forms which is unique. An accusative ending might be attached if the participle is used as a nominal adjective form (typically in a verbal attribute syntactic function). See ex. in Tab. 20. (A similar situation occurs with nominal adjectives with the tag starting with AC.)

Wordform	Lemma	Tag
už mám polévku uvařenu	uvařit	VsFS4---X-APP--
za Marie Terezie	za	RR--2----------
za týden	za	RR--4----------

Table 20: Examples: Special cases of the CASE category application

5.6 Possessor's gender (6th position)

At the 6^{th} position, the possessor's gender of possessive adjectives (with `AU` at `POS` and `SUBPOS` positions) and pronouns possessive for 3^{rd} person (with `P9` and `P1`) is captured. The values are in Table 21.

POSSGENDER	Description	Example
F	Feminine	*matčin, její*
M	Masculine animate	*otcův*
Z	{M, I, N} Not feminine	*jeho*
X	Any	jejich

Table 21: Values of `POSSGENDER` category (6th position)

Possessive adjectives and possessive pronouns (personal and relative) for the 3^{rd} person refer to a possessor, and a possessive agreement rule applies for number and gender of the possessor (cf. Tab. 22). Though it could be argued (similarly as we did for the categories `POS` and `SUBPOS`), that this category is more lexically than inflectionally based, it has been decided to treat it as all the other morphological categories.

For possessive adjectives, given the nature of possible possessors, it can separately take on only the values of masculine animate and feminine. For possessive pronouns, both masculine genders and the neuter gender are always homonymous. Thus (cf. also the method used for the `GENDER` category) we do not use any of the three basic values (`M`, `I`, `N`) – instead, the value of `Z` is used.

If a possessive pronoun usage relates to the subject of the sentence (i.e. wordforms of *svůj*), then there is no gender distinction at all, and therefore the category `POSSGENDER` and also `POSSNUMBER` is not used; cf. Tab. 22.

Wordform	Lemma	Tag
Petr opravil Janovi jeho auto.	jeho	P9XXXZS3-------
Petr opravil Janě její auto.	jeho	P9NS4FS3-------
Petr opravil jejich auto.	jeho	P9XXXXP3-------
Petr opravil svoje auto.	svůj-1	P8NS4----------

Table 22: Examples: Application of the `POSSGENDER` and `POSSNUMBER` categories

5.7 Possessor's number (7th position)

This category is closely related to the `POSSGENDER` category (cf. Sect. above). Similar reasons and rules apply here. This category is used only for possessive pronouns (personal and relative; with the values `P9` and `P1` at `POS` and `SUBPOS` positions). For possessive adjectives, it does not make sense to use it – there is no plural at least – the possessor, lexically present in the form, is always considered to be in singular. The values are in Table 23.

POSSNUMBER	Description	Example
P	Plural	*náš, váš, jejich*
S	Singular	*můj, tvůj, jeho*

Table 23: Values of `POSSNUMBER` category (7th position)

5.8 Person (8th position)

The PERSON category expresses the person of verb forms (if applicable), and person of personal pronouns. Personal pronouns, however, have a separate lemma for each person (and also for number; i.e. já for the 1^{st} person singular, my for the 1^{st} person plural, ty for the 2^{nd} person singular, vy for the 2^{st} person plural and on-1 for the 3^{rd} person; see more in Sect. 11.1). This redundancy has been introduced for better human readability, since these lemmas are traditionally considered to be separate words). The values are in Table 24.

There is no attempt to assign person to all participles, transgressives, etc. For capturing compound wordfoms with the auxiliary verb *být* ("to be"), e.g. *přišels, bych, abys*, see Sect. 18.

PERSON	Description	Example
1	1^{st} person	*píšu, píšeme, my*
2	2^{nd} person	*píšeš, píšete, ty*
3	3^{rd} person	*píše, píšou, ony*

Table 24: Values of PERSON category (8th position)

5.9 Tense (9th position)

The TENSE category belongs to verb forms only. The values are in Table 25.

Contrary to the traditional tense category assignment, tense is meant here in the purely morphological sense, without regard to the actual tense of an analytical verb form used within a particular sentence. Thus e.g. the verb form *pracoval* '(he) worked' is assigned only the past tense value R, even though it can appear as a part of present conditional (*pracoval by* '(he) would work' as well as the true past tense (or "perfective" tense, as Czech does not really distinguish these two): *pracoval jsem* '(I) (have) worked'.

In Czech, future tense is traditionally assigned also to present perfective verb forms. However, morphologically, these forms are simply present tense. On the contrary, there are real future tense forms (but only for a handful of verbs), such as *pojedu* '(I) will go', which is a future tense form of the verb *jet* 'to go'(vs. *jedu* '(I) go'), and of course *budu* '(I) will', *budeš* '(you) will' ... – future forms of the verb *být* 'to be' used both in the existential sense as well as an auxiliary verb form for analytically expressed future tense of imperfective verbs.

TENSE	Description	Example
F	Future	*pojedeme, budu*
P	Present	*napíšu, píšu, jsem*
R	Past	*psal, byl*

Table 25: Values of TENSE category (9th position)

5.10 Degree of comparison (10th position)

The category of degrees of comparison is used for adjectives and adverbs. Traditionally, numbers are being used for the degrees of comparison; see Table 26.

The GRADE category is considered to be an inflectional category (as opposed to a derivation). This means that the wordforms of the same stem in positive, comparative and superlative (e.g. *velká* 'big', *větší* 'bigger', *největší* 'biggest') are understood as the wordforms of one paradigm. If the graded wordforms are not of the same stem (e.g. *dobrý* 'good' and *lepší* 'better'), there are the separate paradigms represented by different lemmas. See examples in Tab. 27.

GRADE	Description	Example
1	Positive	*velká, pěkně*
2	Comparative	*větší, pekněji*
3	Superlative	*největší, nejpěkněji*

Table 26: Values of GRADE category (10th position)

For certain types of adjectives, such as possessive adjectives, this category is considered irrelevant (the value of Not applicable (-) is used). On the other hand, the value positive (1) is used for adjectives which do not form comparatives or superlatives only on semantic grounds (such as the often discussed adjective *optimální* 'optimal'). In most cases, however, the morphological analysis would not reject even the comparative or superlative forms of such adjectives, marking their degree of comparison correctly and consistently according to their prefix and suffix.

Wordform	Lemma	Tag
velký přítel	velký	AAMS1----1A----
největší přítel	velký	AAMS1----3A----
dobré podmínky	dobrý	AAFP1----1A----
lepší podmínky	lepší	AAFP1----2A----
nejlepší podmínky	lepší	AAFP1----3A----

Table 27: Examples: Degree of comparison

5.11 Negation (11th position)

Similarly to the degree of comparison category, the NEGATION category is treated as a fully inflectional category, as negation is expressed in Czech by a prefix. Negation can be attached to verbs, adverbs, adjectives, and in principle, also to nouns (although only verbal nouns are typically used with negation regularly). The values are in Table 28. (See more in Sect. 12.)

NEGATION	Description	Example
A	Affirmative (not negated)	*velká, pěkně, přišel, ochota*
N	Negated	*nevelká, nepěkně, nepřišel, neochota*

Table 28: Values of NEGATION category (11th position)

5.12 Voice (12th position)

The VOICE category is used for verb forms only (mainly for verb participles). It is not used for verbal adjectives, even if they are derived from passive participle forms. The values are in Table 29.

5.13 Verbal aspect (13th position)

At the 13^{th} position, the verbal ASPECT is coded. This category is used for verb forms only. It is not used for verbal adjectives and nouns. The values are in Table 30.

In fact, verbal aspect is rather lexical than morphological property but it is practical to keep it in the tags. Anyway, no paradigm/lemma is allowed to occur with two different verbal aspects in the accompanying tags.

VOICE	Description	Example
A	Active	*píšu, pojedu, psala*
P	Passive	*napsán*

Table 29: Values of VOICE category (12th position)

ASPECT	Description	Example
P	Perfect verb	*napsat*
I	Imperfect verb	*psát*
B	Biaspectual verb	*absolvovat*

Table 30: Values of ASPECT category (13th position)

5.14 Aggregate (14th position)

At the 14^{th} position, so-called aggregates are coded. An aggregate is a wordform created by combining two or more forms into one and cannot be simply assigned any POS category (e.g. wordform *proň* consists of a pronoun *on* 'he' and the preposition *pro* 'for'). The tag describes the main component of the aggregate (i.e. the pronoun *on*) and the joined components are coded at the 14^{th} position of the tag (for joined preposition *pro*, there is value p). The values for capturing aggregates are in Table 31. For more information about aggregates, see Sect. 18.

AGGREGATE	Joined component	Description	Example
d	*do-*	preposition *do*	*doň*
n	*na-*	preposition *na*	*nač, načpak, naň*
o	*o-*	preposition *o*	*oč, očpak, oň*
p	*pro-*	preposition *pro*	*proň*
v	*ve-*	preposition *ve*	*več*
z	*za-*	preposition *za*	*zač, začpak, zaň*
D	*do-* + *-s*	prep. *do-* + 2^{nd} pers. of *být*	*doňs*
N	*na-* + *-s*	prep. *na-* + 2^{nd} pers. of *být*	*načs, načpaks, naňs*
O	*o-* + *-s*	prep. *o-* + 2^{nd} pers. of *být*	*očs, očpaks, oňs*
P	*pro-* + *-s*	prep. *pro-* + 2^{nd} pers. of *být*	*proňs*
Z	*za-* + *-s*	prep. *za-* + 2^{nd} pers. of *být*	*začs, začpaks, zaňs*
c	*-ch/sem*	1^{st} pers. sg. of cond. verb *být*	*bych, abych, kdybysem*
s	*-s*	2^{nd} pers. sg. of verb *být*	*přišels, kdyžs, kdybys*
m	*-chom/sme*	1^{st} pers. pl. of cond. verb *být*	*abychom, kdybysme*
e	*-ste*	2^{nd} pers. pl. of cond. verb *být*	*byste, abyste, kdybyste*

Table 31: Values of AGGREGATE category (14th position)

5.15 Variant, style, abbreviation (15th position)

Variant information (cf. values in Tab. 32) is used whenever all morphological categories together with a given lemma can be expressed by more than one wordform.

For example, Czech masculine nouns can typically have one or two variants in the singular genitive, singular locative, plural nominative, and plural locative cases (the presence or absence and the number of variants differs in different paradigms, and often is just lexically based (e.g. both *orli* and *orlové* ('eagles') are the wordforms of the noun *orel* ('eagle') and express plural masculine nominative)).

VAR	Description	Example
-	Not applicable	*orlové, přijdeme, myslit*
	– basic variant, standard contemporary style	
1	Standard variant	*orli, myslet*
2	Standard variant	*mysliti*
3	Standard variant	*mysleti*
4	Standard variant	*pomažemť*
5	Non-standard variant	*přídeme*
6	Non-standard variant	*přijdem*
7	Non-standard variant	*přídem*
8	Non-standard variant	*přídjeme*
9	Non-standard variant, misspelling	*přídjem*
b	Abbreviated form	*s* (=sekunda)
a	Other abbreviated form	*sec* (=sekunda)
c	Other abbreviated form	*sek* (=sekunda)

Table 32: Values of VAR category (15th position)

Another important distinction is a style of the form. Many standard noun, adjectival, pronominal, numeral as well as some verbal endings have their non-standard counterpart, and the morphology must be able to handle them properly, i.e. to distinguish them from their contemporary counterparts. Numbers 1 to 4 mark standard variants, while numbers 5 to 9 relate to non-standard ones. The 9 value is also for misspellings, typos or another distortions (see Sect. 20). More information about variants can be found in Sect. 9.2.

We have also added new values – letter b, c and a – to the 15^{th} position of the tag for marking abbreviation of a (single) word which is captured as a special wordform of the paradigm of that word. For more information, see Sect. 14.

6 Stylistic Characteristics

Stylistic characteristics are captured to a limited extent. We used different means for capturing style of a paradigm and style of a particular wordform:

- Style label at AddInfo part of a lemma
- Variant Info at AddInfo part of a lemma
- Numbering of wordform variants at the 15^{th} `VAR` position of tag

For instance, *šlojíř* is an archaic word meaning 'veil'. Its lemma `šlojíř_,a` bears the archaic style label `a` at AddInfo part (Sect. 4.2.3). For the stylistic and orthographic variants of a whole word (for the full-paradigm variants; see Sect. 9.1), we also state their basic word variant at AddInfo part (Sect. 4.2.4). For example, for the non-standard word *mlejn* 'mill', there is the standard variant *mlýn* 'mill'. That fact is indicated at AddInfo part (`mlejn_,h_^(^GC**mlýn)`). On the other hand, *lvové* 'lions' is an archaic wordform of a standard lemma `lev` 'lion'. In this case, the variant of wordform is captured in the tag describing the wordform (there is the number `1` at the 15^{th} position). For wordforms, standard and non-standard variants are only distinguished (not archaic, dialect, etc. see Sect. 5.15).

Rules for style labeling are described below and there is a separate Sect. 9 dedicated to orthographic and stylistic variants of the paradigms and wordforms.

Wordform	Lemma	Tag
středověký šlojíř	`šlojíř_,a`	`NNIS1-----A----`
pěknej mlejn	`mlejn_,h_^(^GC**mlýn)`	`NNIS1-----A----`
lvové a psové	`lev`	`NNMP1-----A---1`

Table 33: Examples: Capturing style

6.1 Style labeling

A word can be stylistically classified with style label attached to the lemma. Only one style label (or none, i.e. no more) is assigned to a lemma, the most general one applicable to most of contexts. Possible style labels attached to a lemma are:

∅ – **no style label**. No style flag is applicable for:

- standard, non-marked words, i.e. usable in neutral texts with neutral stylistic feature, they do not appear inappropriate or expressive.
- literary words of higher, bookish style, the expressions that may appear slightly peculiar in some texts but not completely improperly or incorrectly (e.g. *trýzeň, blankyt*),
- words not yet incorporated into the language norm, occasionalisms, i.e. expressions well formed, meaning comprehensible, but authored for a particular occasion, unless they clearly belong to other group below (e.g. *blábolizmus*),
- special terms that cannot be considered non-standard, but they are marked because of their low frequency of use (e.g. *ikonostas*, *sakristie*),
- neologisms unless they clearly belong to other groups below (e.g. *europoslanec*).

`s` – **standard** style label. The style label with the value `s` is applied for:

- second (and possibly another) standard variant of the word (e.g. *kafeterie - kafetérie*, *kurs - kurz*). This style label is used only in connection with Variant Info (Sect. 4.2.4) when capturing orthographic and stylistic variants of word (see more in Sect. 9.1).

`a` – **archaic style label**. The style label with the value `a` is applied for:

- obsolete, archaic expressions (archaisms), which are not considered to be non-standard, but their use in neutral, standard texts is marked; they are almost with no evidence in a contemporary corpora, or they are only documented in historical texts; sometimes, they have been replaced by other (newer) words that are considered neutral (eg. *uštipáček*, *servít*, *šlojíř*, *slout*),
- words written according to the original or older spelling standard; there have usually variants written according to the newer spelling standard (e.g. *thema*, *these*),
- words (so-called historisms) that denote objects that no longer exist; their use in a neutral text is marked and their meaning is unclear to the speaker without detailed knowledge of the area (e.g. words from the area of military terminology of the Austro-Hungarian Empire, although they are passively known from fiction: *maršrúta, vachman*).

`h` – **non-standard style label**. The general non-standard label with the value `h` is used if a non-standard word may be annotated with more stylistic features or if it cannot be classified with any of the specific labels below; typically:

- expressions from common Czech which are not considered a dialect (e.g. *vocelárna, mlejn*),
- univerbations, i.e. words formed from collocations (e.g. *peďák, hotelovka, tirák, stadionovka*),

`n` – **dialect style label**. The style label with the value `n` is applied for:

- dialectal expressions; they might have marked phonic composition and/or there might be less awareness of the meaning among speakers, e.g. *šufánek*, *čučkař*.

`l` – **slang, argot style label**. the style label with the value `l` is applied for:

- non-standard expressions that are used by a very narrow group of people with the same professional or other special interests; e.g. *vertajmka, genšťáb*.

`e` – **expressive style label**. The style label with the value `e` is applied for:

- expressions with positive and negative emotional feature evident from their formation; expressive words may not be considered non-standard, but their use in neutral texts is marked; they can be diminutives from expressions that have a different style label (e.g. word *sajtnička* with the style label `e` is derived from the word *sajtna* with the style label `l`).

`v` – **vulgar style label**. From expressive expressions we only separate vulgar words, the use of which may have a more serious impact. The style label with the value `v` is applied for:

- non-standard expressions that have a negative emotional feature, very often with obscene and/or vulgar content, they usually have an expressive phonic composition, semantically often refer to taboo areas (e.g. *vlezdoprdelizmus*, *čurák*).

`i` – **special style label**. The style label with the value `i` is applied for:

- distorted words, misspelling which are frequently used (e.g. *součastník*, *vánoce*, *intensívní*).

Note. The lemmas automatically derived from derivational patterns (with Derivation info; see Sect. 4.2.5) do not have any style label (see Sect. 7).

7 Derivative Relations

The dictionary primary handles inflection, not word-formation (for modeling word-formation relations in the lexicon of Czech, there is another project - Derinet[17]). However, as mentioned in Introduction (Sect. 2), a substantial part of the dictionary is mapped onto so-called derivational patterns in the source format. If a lemma belongs to a derivational patterns, several other lemmas/paradigms can be derived from it. All lemmas created by the derivational pattern have the derivational information stored in AddInfo part of the lemma (Sect. 4.2.5). In the current version of the dictionary, information on derivation is attached only to the lemmas created by the automatic procedure.[18]

The Derivation Info is primarily of technical or procedural nature: it carries information about the automatic creation of a lemma; the manifested word-formation relation may not be correct or complete.

Originating Lemma	→	Derivative Lemma
vychovat	→	vychovaný_^(*2t)
vychovat	→	vychovatelný_^(*4)
vychovat	→	vychování_^(*3at)
vychovat	→	vychovací_^(*2t)
vychovat	→	vychovavší_^(*3t)
vychovat	→	vychovávat_^(*4at)
vychovaný_^(*2t)	→	vychovanost_^(*3ý)
vychovaný_^(*2t)	→	vychovaně_^(*1ý)
vychovatelný_^(*4)	→	vychovatelnost_^(*3ý)
vychovatelný_^(*4)	→	vychovatelně_^(*1ý)
vychovávat_^(*4at)	→	vychovávání_^(*3at)
vychovávat_^(*4at)	→	vychovávací_^(*2t)
vychovávat_^(*4at)	→	vychovávající_^(*4t)
vychovávat_^(*4at)	→	vychovávaný_^(*2t)
vychovávat_^(*4at)	→	vychovávatelný_^(*4)
vychovávaný_^(*2t)	→	vychovávanost_^(*3ý)
vychovávaný_^(*2t)	→	vychovávaně_^(*1ý)
vychovávatelný_^(*4)	→	vychovávatelnost_^(*3ý)
vychovávatelný_^(*4)	→	vychovávatelně_^(*1ý)

Table 34: Example of automatic derivation

The derivational subsystem is a very useful, effective and economical tool that saves space in the source format. Using derivation patterns, 65 percent of lemmas/paradigms are created in the dictionary. For example, 19 paradigms are automatically derived from the verb *vychovat* 'to bring up' with lemma vychovat. See examples in Table 34. In the source format, these paradigms are written using only one line: *vychov atd* =*vychovat*, where *atd* is a derivative pattern, the derivation rules of which are shown in Table 35.

On the other hand, the usage of derivation patterns has two important consequences for the structure of the dictionary:

- **Over-derivation**. Derivational patterns sometimes lead to only hypothetical derivatives, which are correctly formed but they are attested neither in the corpora nor in the dictionaries; e.g. lemma obchodovávatelnost_^(*3ý) with meaning possibility of repeatedly trade ('retradability'), or hypothetical iterative adjective převolovávatelný-2_^(*6-2) derived from verb převolovat-2-2_^(přemíra_volů) with meaning 'to cause an excess number of oxen'.

[17] https://ufal.mff.cuni.cz/derinet

[18] In the source format, there are also manually created comments on word-formation relation which did not make it into the currently described version for various reasons (inconsistency, incompleteness).

- **No manual labeling**. It is not possible to manually assign an additional information about the paradigm (i.e. AddInfo; e.g. value of term label or style labels) to automatically derived lemmas. During the automatic derivation process, derived lemmas can only automatically inherit values from originating lemma. See rules below.

atd 0,atd,r0,at,0,0,*
atd an,y,r0,aný,r0,at,-
atd ateln,y,r0,atelný,r0,at,-
atd ání,stn,r0,ání,r0,at,-
atd ac,i,r0,ací,r0,at,-
atd avš,iavv,r0,avší,r0,at,-
atd áv,atn,r0,ávat,r0,at,-
atd anost,kt1n,r0,anost,r0,aný,-
atd an,jev,r0,aně,r0,aný,-
atd atelnost,kt1n,r0,atelnost,r0,atelný,-
atd ateln,jev,r0,atelně,r0,atelný,-
atd ávání,stn,r0,ávání,r0,ávat,-
atd ávac,i,r0,ávací,r0,ávat,-
atd ávajíc,iavg,r0,ávající,r0,ávat,-
atd ávan,y,r0,ávaný,r0,ávat,-
atd ávateln,y,r0,ávatelný,r0,ávat,-
atd ávanost,kt1n,r0,ávanost,r0,ávaný,-
atd ávan,jev,r0,ávaně,r0,ávaný,-
atd ávatelnost,kt1n,r0,ávatelnost,r0,ávatelný,-
atd ávateln,jev,r0,ávatelně,r0,ávatelný,-

Table 35: Example of derivative pattern

7.1 Automatically derived lemmas

Automatically derived lemmas are identified by their derivation comment, which is created automatically. As mentioned above, automatically derived lemmas cannot be manually annotated. Values of AddInfo part of the lemma can be only automatically inherited from the original lemma. The rules are as follows. Automatically derived lemmas:

- **inherit reference** (Sect. 4.2.1). For example, from the base lemma *šest* 'six', the following derivatives are automatically derived: *šestkrát* 'six times', *šestina* 'sixth', *šestery* 'sise'. The reference to the number is valid for each derived word.

- **inherit name labels** (Sect. 4.2.2). Any name label added to the original lemma always applies to the derived word. For example, from male surname (e.g. *Hromádka* '(Mr.) Hromádka'), female surname (e.g. *Hromádková* '(Mrs.) Hromádka') and possesive adjective (e.g. *Hromádkův* 'Hromádka's') are derived. The name label `Y` belongs to the original lemma and also to all its derivatives.

- **do not inherit style label** (Sect. 4.2.3). Derived words often do not have the same stylistic value as the original lemma, therefore derived lemmas do not inherit stylistic labels. For example, the word *výtržnost* 'riot' is automatically derived from the word *výtržný* 'riotous'. While the word *výtržný* is an archaic word in the current Czech, the word *výtržnost* is not.

- **inherit variant info** (Sect. 4.2.4). An orthographic / phonetic change in the word stem also occurs in the derived words. Therefore, automatically derived words inherit variant info from the original lemma. The lemma of the basic variant in variant info is automatically

rewritten to the lemma of the corresponding derived variant. E.g. from the verb *votevřít* 'to open-informal' (which is a non-standard variant of verb *otevřít* 'to open'), the adjective *votevřený* 'open-informal' and the noun *votevřenost* 'openness-informal' (non-standard variants of adjective *otevřený* 'open' and noun *otevřenost* 'openness', respectively) are derived.

- **do not inherit explanational comment** (Sect. 4.2.6). Since there is no guarantee that the explanational comment added to the original lemma is also valid for the lemma derived, during the automatic derivation derivative lemma does not inherit the explanational comment.

See examples in Tab. 36.

Wordform	Lemma	Tag
šest	`šest'6`	`Cn-S1----------`
šestina	`šestina'6_^(*3)`	`NNFS1-----A----`
šestero	`šestero'6_^(*3)`	`CjNS1----------`
šestkrát	`šestkrát'6_^(*4)`	`Cv-------------`
pan Hromádka	`Hromádka-1_;Y`	`NNMS1-----A----`
paní Hromádková	`Hromádková-1_;Y_^(*5a-1)`	`NNFS1-----A----`
Hromádkův pes	`Hromádkův-1_;Y_^(*4a-1)`	`AUIS1M---------`
musíme votevřít	`votevřít_;h_^(^GC**otevřít)`	`Vf--------A-P--`
votevřený dveře	`votevřený_^(^GC**otevřený)_(*3ít)`	`AAFP1----1A---6`
votevřeně	`votevřeně_^(^GC**otevřeně)_(*1ý)`	`Dg-------1A----`
jeho votevřenost	`votevřenost_^(^GC**otevřenost)_(*3ý)`	`NNFS1-----A----`
výtržný fanoušek	`výtržný_,a`	`AAMS1----1A----`
výtržnosti na zápase	`výtržnost_^(*3ý)`	`NNFP1-----A----`

Table 36: Examples: Automatic derivation

7.2 Derivative relation types

Here, we present several typical derivation patterns that are applied regularly automatically to a great number of original lemmas.

- to verbs:
  - iterative verbs with suffix -[*áíé* .. etc.]*vat*
  - adjectives with suffix *-ný*, *-cí*, *-jící*, *-vaný*, *-vší*, *-telný*, *-vatelný*,
  - nouns with suffix *-ní*, *-vání*, *-nost*, *-telnost*, *-vatelnost*
  - adverbs with suffix *-ně*, *-telně*, *-vatelně*
- to nouns:
  - possessive adjectives with suffix *-ův*, *-in*
  - female counterparts with suffix *-kyně*, *-ka* are derived from the masculine animate nouns with suffixes *-∅*, *-ec*
  - female surname are derived from some male surnames
- to adjectives:
  - nouns with suffix *-skost*, *-ckost* are derived from adjectives with suffix *-ský*, *-cký*
  - adverbs with suffix *-ě* and nouns with suffix *-ost* are derived from adjectives with hard declension

– adverbs with suffix *-ě* are derived from adjectives with soft declension

- derivatives with suffix *-ina*, *-ery* and *–krát* are derived from basic numerals.

Examples of derivatives from verb are in Table 34 above. Examples of derivatives from nouns, adjectives and numerals are in Table 37.

Originating Lemma	→	Derivative Lemma
`Milan_;Y`	→	`Milanův_;Y_^(*2)`
`Eva_;Y`	→	`Evin_;Y_^(*2a)`
`chodec`	→	`chodkyně_^(*4ec)`
`doktor`	→	`doktorka_^(*2)`
`Nový_;Y`	→	`Nová_;Y_^(*1ý)`
`Konvička_;Y`	→	`Konvičková_;Y_^(*3a)`
`světský`	→	`světskost_^(*3ý)`
`vědecký`	→	`vědeckost_^(*3ý)`
`barevný`	→	`barevně_^(*1ý)`
`veliký`	→	`velikost_^(*3ý)`
`letní`	→	`letně_^(*1í)`
`sedm'7`	→	`sedmina'7_^(*3)`
`sedm'7`	→	`sedmkrát'7_^(*4)`
`sedm'7`	→	`sedmery'7_^(*3)`

Table 37: Examples: Derivative relation types

Derivative collisions. Sometimes, when deriving from several different originating lemmas, the morphologically same result is obtained. E.g. the same possessive adjective *Janův* 'John's' is derived from the names *Jan*, *Jano* and *Janus*. Derivative collisions can also occur between a derived lemma and non-derived one (e.g., the female surname *Černá* is automatically derived from the male surname *Černý* with `Y` name label. There is also an identical paradigm for the name of the village *Černá* with the name label `G`).

In this case, there is only one lemma/paradigm in the dictionary (according to the principle of a unique paradigm; see Sect. 3.3). When generating the dictionary from the source format, these identical lemmas are merged into one. All derivative comments are stored in the AddInfo part of the lemma. The "path" to all originating lemmas is thus preserved. The merged lemma (like all derivatives) inherits name labels and variant infos and does not inherit the style labels and explanational comments. Thus, all name labels and variant infos (from all originating lemmas) are stored in the AddInfo part of the merged lemma. See examples in Tab. 38.

Wordform	Lemma	Tag
Jan a jeho, Janův sen	`Janův_;Y_^(*2)_(*2o)_(*2us)`	`AUIS1M---------`
Jano a jeho, Janův sen	`Janův_;Y_^(*2)_(*2o)_(*2us)`	`AUIS1M---------`
Janus a jeho, Janův sen	`Janův_;Y_^(*2)_(*2o)_(*2us)`	`AUIS1M---------`
zboží dovážené z Německa	`dovážený_^(*2t)_(*3it)`	`AANS1----1A----`
zboží dovážené na váze	`dovážený_^(*2t)_(*3it)`	`AANS1----1A----`
paní Černá	`Černá_;G_;Y_^(*1ý)`	`NNFS1-----A----`
Černá v Pošumaví	`Černá_;G_;Y_^(*1ý)`	`NNFS1-----A----`

Table 38: Examples: Derivative collisions

8 Semantic Description

The description of word meanings and other semantic distinctions is the least pursued goal in building the morphological dictionary and in morphological annotation. Compare Sect. 3 which describes the main principles of the dictionary. We capture primarily the formal morphological behavior of words regardless of semantic differences. We build a dictionary of wordforms, not a dictionary of words or meanings.

However, partial semantic description is contained in the following attributes and values (their description is in the relevant sections):

- Explanational comment at AddInfo part of the lemma (Sect. 4.2.6),
- Numeric reference at AddInfo part of the lemma (Sect. 4.2.1),
- Name label at AddInfo part of the lemma (Sect. 4.2.2),
- `SUBPOS` position of the tag, namely of pronouns and numbers (Sect. 5.2).

9 Orthographic and Stylistic Variants

There are two types of orthographic and stylistic variants:

- **full-paradigm variant**: an orthographic, phonetic or stylistic change applies to the whole paradigm (e.g. non-standard *mlejn* 'mill' vs. the standard *mlýn* 'mill'),
- **wordform variant**: an orthographic, phonetic or stylistic change is manifested only in single wordforms (e.g. nominative singular of masculine inanimate *zelenej* 'green' vs. the standard *zelený*).

The full-paradigm variants are captured in the variant info field of the lemma (see more in Sect. 9.1), the wordform variants are indicated by the `VAR` position of the morphological tag (see more in Sect. 9.2).

9.1 Full-paradigm variants

If an orthographic, phonetic or stylistic change applies to the full paradigm, i.e. to all wordforms (cf. wordforms of *okno* 'window' in common Czech: *vokno, vokna, voknem* vs. in standard Czech: *okno, okna, oknem*), each set of wordforms is captured in separate paradigm with different lemma: we select one of the variants as "basic" (the standard one) and other variants (second standard, non-standard, archaic) refer to it in an additional descriptive element (variant info), attached to the lemma (Sect. 4.2.4). Non-basic paradigms (except derivatives, see below) also have a style label attached to the lemma (Sect. 4.2.3).

The basic variant is the one that is the least marked. Common alternations, that appear repeatedly, are solved the same way. For instance variants with the prosthetic consonant *v-* at the beginning of words have the basic variant without the prosthetic *v-* (e.g. *okno* - *vokno* 'window'). Cf. examples in Tab. 39. Types of the most common full-paradigm variants are listed below in Sect. 9.1.1.

Multiple variants. In the case of multiple full-paradigm variants, one basic variant is chosen and and other variants refer to it. Cf. example of variants of word *Afghánistán* in Tab. 39.

Variants and derivation. Paradigms derived automatically inherit variant info from the original lemma. This does not apply to style labels. The automatic derivatives have never any style label. Cf. examples of words *mlíkař* 'milkman' and *mlíkařka* 'milkwoman' in Tab. 39. See more information about derivatives in Sect. 7.

Wordform	Lemma	Tag
okno	`okno`	`NNNS1-----A----`
vokno	`vokno_,h_^(^GC**okno)`	`NNNS1-----A----`
mlíkař	`mlíkař_,h_^(^GC**mlékař)`	`NNMS1-----A----`
mlíkařka	`mlíkařka_^(^GC**mlékařka)_(*2)`	`NNFS1-----A----`
Afghánistán	`Afghánistán\_;G`	`NNIS1-----A----`
Afganistan	`Afganistan_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afgánistán	`Afgánistán_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afghánistán	`Afghánistán_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afganistán	`Afganistán_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afghanistán	`Afghanistán_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afghanistan	`Afghanistan_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afghánistan	`Afghánistan_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`
Afgánistan	`Afgánistan_;G_,s_^(^DD**Afghánistán)`	`NNIS1-----A----`

Table 39: Examples: Full-paradigm variants

9.1.1 Types of full-paradigm variants

There are a lot of types of variations in the Czech language. Alomorphs – both in the roots (*mlýn - mlejn*) and in prefixes and suffixes (*výlet - vejlet*) – are seen as cases of variation. Here, we list the most common ones. Examples of all listed types are in Table 40.

Wordform	Lemma	Tag
kurs	kurs_,s_^(^DD**kurz)	NNIS1-----A----
optimismus	optimismus_,s_^(^DD**optimizmus)	NNIS1-----A----
citron	citron_,s_^(^DD**citrón)	NNIS1-----A----
Abel	Abel_,s_^(^DD**Ábel)	NNMS1-----A----
parfumérie	parfumérie_,s_^(^DD**parfumerie)	NNFS1-----A----
archívní	archívní_,s_^(^DD**archivní)	AAIS1----1A----
príma	príma-2_,h_^(^GC**prima-2)	Db-------------
mlejn	mlejn_,h_^(^GC**plýtvat)	Vf--------A-I--
plejtvat	plejtvat_,h_^(^GC**mlýn)	NNIS1-----A----
vokolo	vokolo-1_,h_^(^GC**okolo-1)	RR--2----------
theorie	theorie_,a_^(^DD**teorie)	NNFS1-----A----
dyž	dyž_,h_^(^GC**když)	J,-------------
součastník	součastník_,i_^(^DS**současník)	NNMS1-----A----
vánoce	vánoce_,i_^(^DS**Vánoce)	NNMS1-----A----
management	management_,s_^(^DD**manažment)	NNIS1-----A----

Table 40: Examples: Full-paradigm variants

Standard variants of type `DD` in variant info. Style label of the variants is most usually `s`, alternatively `a`.

kurz – kurs. Variants with *-z/s-* spelling. The variant with *-z-* is captured as a basic one. A frequent subtype of this alternation are variants with *-zmus/smus* spelling (e.g. *optimizmus – optimismus*).

teorie – theorie. Consonant variation *-t/th-*. The variant with *-t-* is captured as a basic one.

citrón – citron. Variation in vowel length. The variant with long vowel is captured as a basic one. This solution applies to most vowel-length variation cases but not in general (see the types below).

parfumerie – parfumérie. Vowel-length variation in the suffixes *-erie/érie*, *-iv(ní)/ív(ní)* (e.g. *archiv(ní) – archív(ní)*), *-ivum/ívum* (e.g. *pasivum – pasívum*), *-emie/émie* (e.g. *leukemie – leukémie*), *ped/péd* (e.g. *logoped – logopéd*). The short-vowel spelling is captured as a basic variant.

manažment – management. Variation between Czechized spelling and original foreign language spelling. More domesticated variant is captured as a basic standard one.

Foreign-language names are an area where there are usually a lot of different spellings in Czech texts. Determining the basic variant can be difficult. In the complicated cases, we always just select one spelling as the basic; cf. variants of the word *Afghánistán* in Tab. 39.

Non-standard variants of type `GC` in variant info. Style label of non-standard variants is most usually `h`, alternatively `n`, `e`, `l` or `v`. Standard variant is captured as a basic one.

prima – príma. Non-standard variation in vowel length.

mlýn – mlejn. Change from *–ý/í–* to non-standard *–ej–*.

okolo – vokolo. Addition of non-standard prosthetic consonant *v-* to the beginning of a word.

když – dyž. Consonant group reduction in non-standard expressions.

Distortion variants of type `DS` in variant info. Style label of non-basic variants is `i`.

Vánoce – vánoce Distorted words, words with outdated spelling or misspelling which are frequently used, intentional typos, phonetic transcription of words, etc. are captured as a variant of standard spelling variant.

9.2 Wordform variants

For capturing orthographic and stylistic variants of a wordform manifested usually in the ending (e.g. both *orli* and *orlové* 'eagles' are the wordforms of the noun *orel* 'eagle' and express plural masculine nominative), we use the the 15^{th} position of the tag. No number at the the 15^{th} position indicates the basic wordform. Numbers `1` to `4` mark standard variants (e.g. *orli – orlové* 'eagles'), while numbers `5` to `9` relate to non-standard ones (e.g. *malý – malej* 'little'; cf. examples in Tab. 41). The number `9` is also used for distortions, typos, misspellings (e.g. *o mě* 'about me'). An overview of the values for 15^{th} tag position see in Sect. 4.2.3.

Wordform	Lemma	Tag
kde hnízdí orlové	`orel`	`NNMP1-----A----`
kde hnízdí orli	`orel`	`NNMP1-----A---1`
malý dům	`malý`	`AAIS1----1A----`
malej dům	`malý`	`AAIS1----1A---6`
přijdeme	`přijít`	`VB-P---1P-AAP--`
přídeme	`přijít`	`VB-P---1P-AAP-5`
přijdem	`přijít`	`VB-P---1P-AAP-6`
příjdeme	`přijít`	`VB-P---1P-AAP-7`
přídem	`přijít`	`VB-P---1P-AAP-8`
příjdem	`přijít`	`VB-P---1P-AAP-9`
o mě	`já`	`PP-S6--1------9`

Table 41: Examples: Wordform variants

The main function of wordform numbering at the 15^{th} tag position is to distinguish the forms according to the principle of unique analysis (see Sect. 3.1). Marking the stylistic value is secondary (if a form should have more standard variants and the values `1`, `2`, `3`, `4` would not be enough, the values primarily designated for the non-standard variants could be used and vice versa. However, so far there was no case for which the values `1-9` were not enough; cf. example of *přijdeme* 'we come' in Tab. 41.).

9.3 Boderline cases of full-paradigm and wordform variants

In this section, boundary cases – where it is difficult to decide whether a particular case is a full-paradigm variant of or wordform variant – are described.

9.3.1 Variation is not in the full paradigm

Some variations are not applied throughout the full paradigm, but only in some wordforms (e.g. within the verb *mýt* 'to wash', the change *-ý-* to *-ej-* occurs only in the present tense wordforms and in the infinitive). We consider such cases to be wordform variants. It means that all variant wordforms are captured within one paradigm and they are distinguished at the 15^{th} tag position.[19]

[19]The solution is based on the strong tradition, that lemma is represented by nominative for nouns and by infinitive for verbs, etc. In cases with no base form available, the variant wordforms are not perceived as wordforms

Wordform	Lemma	Tag
mýt	mýt	Vf--------A-I--
mejt	mýt	Vf--------A-I-6
myju	mýt	VB-S---1P-AAI--
myji	mýt	VB-S---1P-AAI-1
meju	mýt	VB-S---1P-AAI-3
být	být	Vf--------A-I--
nebejt	být	Vf--------N-I-6
jsem	být	VB-S---1P-AAI--
sem	být	VB-S---1P-AAI-6
myslit	myslit	Vf--------A-I--
myslet	myslit	Vf--------A-I-1
myslil	myslit	VpYS----R-AAI--
myslel	myslit	VpYS----R-AAI-1
začít	začít-1	Vf--------A-P--
začnout	začít-1	Vf--------A-P-1
přijal	přijmout	VpYS----R-AAP--
přijmul	přijmout	VpYS----R-AAP-1
svatější	svatý-1	AAIS1----2A----
světější	svatý-1	AAIS1----2A---1
světější	svatý-1	AAIS1----2A---1
rozepnuli	rozepnout	VpMP----R-AAP--
rozepjali	rozepnout	VpMP----R-AAP-1
rozpjali	rozepnout	VpMP----R-AAP-2
rozepli	rozepnout	VpMP----R-AAP-3

Table 42: Examples: Wordform variants

The main types are as follows (examples of all listed types are in Tab. 42

mýt – mejt. Within the verb *mýt* 'to wash', the change *-ý-* to *-ej-* (and similarly *-í-* to *-ej-* within the verb *sít*) occurs only in the present tense wordforms and in the infinitive. Within the verb *být* 'to be' the alternation *-ý-* to *-ej-* is only in the infinitive. There is only one paradigm with one lemma (the basic one: *mýt*, *sít*, *být* etc.) and all variant wordforms are marked at the 15^{th} tag position.

Similarly: *omýt – omejt*, *rýt – rejt*; *lít – lejt*, *nalít – nalejt*, etc.

myslit – myslet. Within the verb *myslit* 'to think', the suffix alternation *-e/i-* occurs only in the past participle wordforms and in the infinitive. There is only one paradigm with one lemma (with the basic one according to the Dictionary of Standard Czech by Czech Academy of Sciences, i.e *myslit*) and all variant wordforms are distinguished at the 15^{th} tag position.

Similarly: *bydlet – bydlit*, *bulet – bulit*; *bystřet – bystřit*; *bělet – bělit*, etc.

začít – začnout. Within the verb *začít* 'to begin', there is a suffixal alternation *-∅/nu-* and it does not occur with all wordforms, so there is only one paradigm with one lemma and variant wordforms are distinguished at the 15^{th} position of the tag.

Similarly: *přijal – přijmul*; *tnout – tít*; *najal – najmul*, etc.

of two different paradigms. Variants *přijal – přijmul* (for which there is only one infinitive wordform *přijmout*) are of the same kind as variants *začal – začnul* (for which there are two variant infinitive wordforms *začít – začnout*). Both *přijal – přijmul* and *začal – začnul* are captured in the same way. Dividing the variant forms to different paradigms (for example with numbered lemmas `přijmout-1` for the complete paradigm with forms *přijal*, etc. and `přijmout-2_,h` for the incomplete paradigm only including the non-standard wordforms; *přijmul*) would be very difficult, maybe impossible due to large amount of variant forms; cf. wordforms of *rozepnout* in Table 42.

zmrazen – zmražen. Within the verb *zmrazit* 'to freeze', there is an alternation *-z/ž-* and it does not occur with all wordforms (it occurs in the passive participle), so there is only one paradigm with one lemma and the wordforms are distinguished at the 15^{th} tag position.

svatější – světější. The root alternation *svat/svět* is only present in comparative and superlative. There is only one paradigm with one lemma (the basic one: *svatý*) and the variant wordforms are marked at the 15^{th} position of the tag.

Similarly: *bělejší – bílejší*.

9.3.2 Variation in the base form

If the wordform that is used as the lemma (i.e. mainly nominative, infinitive) has more variants in the respective paradigm, one of the variant wordforms, the one that is least marked, is chosen as the lemma and other forms are captured as variants at the 15^{th} tag position.

Wordform	Lemma	Tag
pracovat	pracovat	Vf--------A-I--
pracovati	pracovat	Vf--------A-I-2
péci	péci	Vf--------A-I--
péct	péci	Vf--------A-I-1
nit	nit	NNFS1-----A----
niť	nit	NNFS1-----A---6
konsenzus	konsenzus	NNIS1-----A----
konsenz	konsenzus	NNIS1-----A---1
virem	vir-1	NNIS7-----A----
virusem	vir-1	NNIS7-----A---1
Patricie	Patricie_;Y	NNFS1-----A----
Patricia	Patricie_;Y	NNFS1-----A---1
Avia	Avia_;m_^(vozidlo)	NNFS1-----A----
Avie	Avia_;m_^(vozidlo)	NNFS1-----A---1
Polanský	Polanský_;Y	NNMS1-----A----
Polanski	Polanský_;Y	NNMS1-----A---1
Polanskij	Polanský_;Y	NNMS1-----A---2
Rucký	Rucký_;Y	NNMS1-----A----
Ruckij	Rucký_;Y	NNMS1-----A---1
Ruckoj	Rucký_;Y	NNMS1-----A---2
Ruckej	Rucký_;Y	NNMS1-----A---6
nejvíce	více	Dg-------3A----
nejvíc	více	Dg-------3A---1
zase	zase-1	Db-------------
zas	zas-1_;s_^(^DD**zase-1)	Db-------------

Table 43: Examples: Wordform variants

The main types of this case are as follows (examples of all listed types are in Tab. 43):

pracovat – pracovati. There is *-t/ti* alternation in most infinitive forms. The *-t* variant is the basic one, it is captured as a lemma. Infinitive ending with *-ti* (which is less common, slowly disappearing, bookish, archaic) usually has value 2 at the 15^{th} tag position.

péci – péct. Another alternation in infinitive forms is *-ci/t* type. Here, infinitive ending with *-ci* is captured as a base form.

niť – nit. A noun can have multiple variants of the singular nominative (e.g. *nit* vs. *niť*). We marked variants at the 15^{th} position of the tag and one of them is used for the lemma.

konsenzus – konsenz. The masculine inanimate nouns with *-∅/us* alternation in ending of singular nominative and singular accusative (and sometimes also in all cases, e.g. *vir – virus*) are captured as one paradigm. The masculine animate nouns with *-∅/us* alternation (e.g. *Josephus – Joseph*) are captured as the two separate paradigms/lemmas.

Patricie – Patricia. There is an alternation of endings *-ia/ie* in the loan female personal names (e.g. *Patricie* vs. *Patricia*) and other names (e.g. *Istrie* vs. *Istria*). Nouns with variant nominative ending *-ia/ie* are represented in one paradigm with lemma ending *-ie*. The *-ia* form is only in nominative (with value `1` at the 15^{th} position of the tag). In some cases, however, the basic variant is the one with *-ia* ending (e.g. *Avia* vs. *Avie*).

Polanský – Polanski – Polanskij. Great variation in the endings also occurs in the transcription of Polish and Russian surnames into Czech, e.g. *Polanský* vs. *Polanski* vs. *Polanskij* or *Rucký* vs. *Ruckij* vs. *Ruckoj* vs. *Ruckej*. A variant with a Czech ending *-ý* is captured as the lemma.

více – víc. When it comes to inflexible words, the distinction between morphological, spelling and word-forming variants is not very clear. Close, similar words with a flexible `POS` (e.g. variants of multiple numbers *dvakrát* vs. *dvakráte*) and similar words that compare (e.g. variants of comparative adverb *nejvíce* vs. *nejvíc*) are captured as wordform variants (they are distinguished at 15^{th} tag position). In other cases, for non-comparative adverbs (e.g. *zase* vs. *zas*), variants are captured as full-paradigm (using a variant reference in AddInfo part of lemma). Formally or semantically distant words (e.g. *jen* vs. *jenom*, *aspoň* vs. *alespoň*) are not interlinked at all.

10 Part of Speech Determination (problematic cases)

10.1 Part of speech of inflexible words

The determination of traditional inflexible POS does not depend on morphological properties. Distinguishing some adverbs, particles, prepositions, conjunctions, and interjections is based on the syntactic function of a given word in a sentence.

Some inflexible words can perform various functions in a sentence; e.g. the word *tak* 'so' can be a conjunction (*pršelo, tak nešel* 'it was raining, so he didn't go'), a particle (*tak už jdeme* 'so here we go') or an adverb (*uděláme to tak* 'we'll do it that way'). In some cases, `POS` of inflexible words (like *tak* 'so', *totiž* 'that-is', *však* 'but', *vůbec* 'not-at-all', *jen* 'only', etc.) is difficult to determine. The decision on POS is left to the annotator in problematic cases. See examples in Tab. 44.

Wordform	Lemma	Tag
tak už jdeme	`tak-1`	`TT----------`
pršelo, tak nešel	`tak-2`	`J^----------`
uděláme to tak	`tak-3`	`Db----------`
legendy rocku	`rock-1`	`NNIS2-----A----`
rock festival	`rock-2`	`AAXXX----1A----`
rock'n'roll	`rock-3`	`S2--------A----`
skladba We will rock you	`rock-77`	`F%----------`
cyklo-oblečení	`cyklo-1`	`S2--------A----`
cyklo výlet	`cyklo-2`	`AAXXX----1A----`
pro příznivce cyklo	`cyklo-3`	`NNXXX-----A----`
v Sazka Aréně	`Sazka_;m`	`NNFS1-----A----`

Table 44: Examples: Part of speech of inflexible words

Similarly, according to a syntactic function of a word in a sentence, we also determine the part of speech of domesticated loanwords such as *online*, *cyklo* 'cyclo', *rock* (see also Sect. 17.3), and of inflexible words in general. E.g. the loanword *rock* has four lemmas in the dictionary that capture the usage of this word in various syntactic function (as we see in the Tab. 44, the loanword *rock* can be inflected if it is a noun).

Note. Commonly inflected names in an attributive position (e.g. *Gambrinus* in *Gambrinus liga* 'Gambrinus league', *Sazka* in *Sazka Aréna* 'Sazka Arena') have a noun (not an adjective) tag. Cf. Tab. 44.

10.1.1 Frozen wordforms (*krážem*, *bycha*, *domácku*)

In the case of frozen wordforms of words that occur in only one type of collocation (such as *krážem* 'cross' in idiom *křížem krážem* 'cross by cross'; *bycha* in idiom *pozdě bycha honit*; *domácku* 'home' in *po domácku* 'home-made'), the lemma of paradigms equals the frozen form, and there is only a single form in the paradigm. See examples in Tab. 45.

Wordform	Lemma	Tag
křížem krážem	`krážem`	`Db----------`
pozdě bycha honit	`bycha`	`NNIS4-----A----`
po domácku	`domácku`	`NNNS6-----A----`

Table 45: Examples: Frozen wordforms

10.2 Nouns from adjectives

In accordance with the tradition, we capture the so-called substantiated adjectives as nouns (the value of the SUB/POS tag position is NN), although their inflection is adjectival and they fulfill the role of the noun mainly syntactically. Due to the large number of unclear cases, only a limited number of cases are captured in this way (i.e. clear historical examples *vrátný* 'porter', *průvodčí* 'conductor' which in contemporary language already appear only as nouns; further *cestující* 'traveller', *zemřelý* 'dead', *milá* 'girlfriend', *nevidomí* 'the-blind', but for example not *zaměstnaný* 'employee', *milovaný* 'beloved', *dojíždějící* 'commuting', *místní* 'local', *nastávající* 'groom-to-be'.)

Cf. examples in Tab. 46.

Wordform	Lemma	Tag
náš vrátný	vrátný-1_^(osoba)	NNMS1-----A----
to je moje milá	milá-2_^(*3ý-2)	NNFS1-----A----
moje milá žena	milý-1_^(příjemný)	AAFS1----1A----
nevidomí maséři	nevidomý-1	AAMP1----1A----
telefon pro nevidomé	nevidomý-2	NNMP4-----A----
lidé dojíždějící do zaměstnání	dojíždějící_^(*4t)	AGMP1-----A----
zpráva pro dojíždějící	dojíždějící_^(*4t)	AGMP4-----A----

Table 46: Examples: Nouns from adjectives

10.3 Part of speech of predicatives (words with suffix *-o*)

Deadjectival and deverbal words with suffix/ending *-o* (e.g. *teplo* 'warm', *zataženo* 'cloudy') which serve as a predicative in a sentence are captured as adverbs. However, these words can be homonymous with nouns and also with nominal forms of adjectives / passive participles and prefixal segments. Cf. Tab. 47.

Wordform	Lemma	Tag
období tepla a sucha	teplo-1	NNNS2-----A----
je velmi teplo	teplo-2__^(být_někomu_teplo)	Dg-------1A----
teplo-vodní topení	teplo-3	S2--------A----
je tu docela hořko	hořko-1	Db----------
nálada přešla v hořko	hořko-2	NNNS4-----A----
hořko-sladké vzpomínky	hořko-3	S2--------A----
je zataženo	zataženo-2_^(být_zataženo)	Dg-------1A----
fotím i v zataženu	zataženo-1	NNNS6-----A----
obloha je zatažena	zatáhnout	VsQW----X-APP--
bylo tam prázdno	prázdno-1	Dg-------1A----
má v hlavě prázdno	prázdno-2	NNNS4-----A----
síně byly pusty a prázdny	prázdný	ACTP------A----

Table 47: Examples: Words with suffix *-o*

In some contexts, it can be difficult to decide on morphological analysis. Here are guidelines for determining the part of speech of these derivatives.

Adverb. As adverb (with the values Dg or Db at the 1^{st} and the 2^{nd} tag positions), we capture the derivatives with *-o* in predicative and adverbial positions, unless they follow a preposition. Cf. examples:

Dg *Je velmi teplo.* 'It's very warm.'; *Bylo mi teplo.* 'I was warm.'; *Máte tu lacino.* 'You're cheap here.'; *Je zataženo a bude zataženěji.* ' It is cloudy and it will be more cloudy.'

Db *Země medu, v níž je docela hořko.* 'A land of honey in which it is quite bitter.'; *známe se krátko 'We know each other for a short time.'*; *draho prodat* 'sell dearly'.

Deciding whether the adverb is of the Dg type (i.e. it forms a comparative and superlative) or whether it does not form degrees of comparision (Db type; cf. Sect. 11.3) is relatively difficult. In addition, in many cases, there is competition between adverbs ending in *-o* and adverbs ending in *-e*. We evaluate wordforms of comparative and superlative preferably as adverbs ending in *-e*. Cf. examples in Tab. 48.

Wordform	Lemma	Tag
bylo mu lehko po těle	lehko-1	Dg-------1A----
bylo mu lehce u srdce	lehce	Dg-------1A----
to se lehko řekne	lehko-1	Dg-------1A----
to se lehce řekne	lehce	Dg-------1A----
to se lehčeji řekne	lehce	Dg-------2A----
cestuje na lehko	lehko-2	NNNS4-----A----
lacino získat	lacino-1	Dg-------1A----
lacině získat	lacině_^(*1ý)	Dg-------1A----
lacině̱ji získat	lacině_^(*1ý)	Dg-------2A----
prodává dost nelacino	lacino-1	Dg-------1N----
draho prodat	draho	Db--------------
draze prodat	draze	Dg-------1A----
prodat nejdráž	draze	Dg-------3A----

Table 48: Examples: Adverbs ending in *-o* vs. adverbs ending in *-e*

Noun. As nouns with neutrum gender (with the NNN value at the first three tag positions), we capture the words ending *-o* if they are in subject or object position, if there is an adjectival or pronominal modifier or if they follow a preposition (and are inflected). Cf. examples:

NNN *Je velké teplo.* 'It's very hot.'; *To jsou velká tepla.* 'These are the heat.'; *Pojďme do tepla.* 'Let's get warm.'; *Dobrá nálada přešla v hořko.* 'The good mood turned bitter.'; *Cítil to úzko, jež ho obcházelo.* 'He felt the tightness around him.'; *Jsem právě naladěna na něžno a milo.* 'I'm just in the mood for being gentle and kind.'; *Prodat za lacino.* 'Sell for cheap.'; *Fotím i v zataženu.* 'I take photos even in the cloudy.'

Nominal adjective or passive participle. If a derivative has an adjectival agreement with a governing noun (usually in a subject position), it is a short form of the adjective (with the value AC at the 1^{st} and the 2^{nd} tag positions) or a passive participle (with the value Vs at the 1^{st} and the 2^{nd} tag positions). Cf. examples:

AC *Pusty, prázdny byly síně ostatní.* 'The other halls were desolate and empty'; *Je daleka toho, aby mu šla opět zachraňovat život.* 'She is far from saving his life again.'; *Žena ležela na nosítkách blízka kómatu.* 'The woman lay on a stretcher near a coma.'

Vs *Obloha je zatažena a zamračena.* 'The sky is cloudy and overcast.'

In this case, the word with *-o* is a wordform of respective adjective or verb. Therefore, in the case of adjectival nominal wordfom, the lemma is a respective long adjective. In the case of passive participle, the lemma is an infinitive of the respective verb.

Prefixal segment. Derivatives with suffix *-o* which are part of hyphenated composites (Sect. 19) are captured as a prefixal segment (with the value S2 at the 1^{st} and the 2^{nd} tag positions; Sect. 16).

S2 *černo-bílý svět* 'black-and-white world'; *hořko-sladké vzpomínky* 'bitter-sweet memories'

11 Detailed Part of Speech

In this section we describe criteria according to which the words within the given POS are classified into subtypes (described by a value on the 2^{nd} tag position - SUBPOS). We describe subtypes for pronouns (Sect. 11.1), numerals (Sect. 11.2) and adverbs (Sect. 11.3).

11.1 Subtypes of pronouns

We identify several features that can serve as criteria for dividing pronouns into various subtypes:

- morphological behavior
- semantic function
- possession
- reflexivity
- clitichood

So called agreement gender and semantic function are chosen as the main criteria. The criterion of agreement gender affects which morphological categories (especially GENDER and NUMBER) are relevant to determine for a given pronoun. Semantic classification is based on the traditional division of pronouns into personal, indefinite, demonstrative, negative, etc. The subtypes of pronouns are described by a value on the 2^{nd} tag position - SUBPOS, see overview in Sect. 5.2 and Tab. 49 here.

Type & Subtype	**Gender**	**No Gender**
Personal	PE *on*, *něj*; clitic P5 *mu*	PP *já*, *ty*; clitic PH *mi*
- Reflexive	-	P6 *sebe*; clitic P7 *se*, *si*
- Possessive	PS *můj*, *náš*	-
- Possessive, 3^{rd} pers.	P9 *jeho*, *její*, *jejich*	-
- Reflexive possessive	P8 *svůj*	-
Relative	P4 *který*, *jaký*, *čí*, *jenž*	PQ *kdo*, *co*, *kdož*, *copak*
- Possessive	P1 *jehož*, *jejíž*, *jejichž*	-
Indefinite	PZ *nějaký*, *čísi*, *sotvakterý*	PK *někdo*, *cosi*, *nevímco*
Negative	PW *nijaký*, *ničí*, *žádný*	PY *nikdo*, *nic*
Demonstrative	PD *ten*, *tentýž*, *takový*	-
Delimiting	PL *všechen*, *sám*, *veškerý*	-

Table 49: Subtypes of Pronouns

Morphology. The morphological criterion divides pronouns into two groups: gender pronouns and no gender pronouns.

Gender pronouns express variable values of the gender (and also number) depending on the gender (and number) of the governing noun (cf. *nějaký dům* 'some house' (masc. sg), *nějaká žena* 'some woman' (fem. sg.), *nějaké dítě* 'some child' (neut. sg.), *nějaké domy* 'some houses' (masc. pl.)) or according to the sense (*on* 'he', *ona* 'she', *ono* 'it', *oni* 'they'). All forms are represented by one lemma (nom. sg. masc. anim.); similarly to adjectives. The GENDER and NUMBER tag positions are filled. Cf. examples in Tab. 50.

No gender pronouns are pronouns with no gender and number variation (e.g. *někdo*, *kdeco*; cf. Tab. 50). The GENDER and NUMBER tag positions are not filled (although we are aware that most of them could be classified as masculine (e.g. *kdo* and other various personal pronouns) or neutrum (*co* and other various non-personal pronouns)). No gender pronouns behave as syntactic nouns in sentences.

Wordform	Lemma	Tag
nějaká žena	nějaký	PZFS1----------
někdo tu je	někdo	PK--1----------
mně to dal	já	PP-S3--1-------
dal mi to	já	PH-S3--1-------
ona přišla	on-1	PEFS1--3-------
jeho si vážím	on-1	PEYS4--3-------
vážím si ho	on-1	P5ZS4--3-------
bojím se o něho	on-1	PEZS4--3------1
stará se o sebe	se_^(zvr._zájmeno/částice)	P6--4----------
nestarej se	se_^(zvr._zájmeno/částice)	P7--4----------
můj dům	můj	PSYS1-S1-------
váš dům	váš	PSYS1-P2-------
svůj dům	svůj-1	P8YS1----------
do jeho domu	jeho	P9XXXZS3-------
do jejího domu	jeho	P9ZS2FS3-------
do její chalupy	jeho	P9FXXFS3-------
jejich dům	jeho	P9XXXXP3-------
žena, jejíž dům vidím	jehož	P1IS4FS3-------

Table 50: Examples: Subtypes of pronouns

Semantic function. We classify six main semantic groups of pronouns, largely following the Czech grammar tradition. In each group, we use the unique SUBPOS value to distinguish between gender and no gender pronouns.

- **personal**:

  PE gender (*on* 'he', *ona* 'she', *ono* 'it', *jim* 'them)

  PP no gender (*já* 'I', *ty* 'you', *my* 'we', *vám* 'to you')

- **relative/interogative**:

  P4 gender (*jaký* 'what', *který* 'which', *čí* 'whose', *jenž* 'who', *kterýžto* 'which')

  PQ no gender (*kdo* 'who', *co* 'what', *cožpak* 'isn't-it-true-that', *kdož* 'who')

- **indefinite**:

  PZ gender (*nějaký* 'some', *čísi* 'somebody's', *číkoli* 'anybody's', *sotvakterý* 'hardly-some')

  PK no gender (*někdo* 'somebody', *bůhvíkdo* 'whoever', *cosi* 'something')

- **negative**:

  PW gender (*nijaký* 'no/none', *ničí* 'nobody's', *žádný* 'no/none')

  PY no gender (*nic* 'nothing', *nikdo* 'nobody')

- **demonstrative**:

  PD gender (*ten* 'this', *tamten* 'that', *onen* 'that-over-there', *tentýž* 'same', *takový* 'such')

- **delimiting**:

  PL gender (*všechnen* 'all', *sám* 'alone', *veškerý* 'whole')

Unlike Czech grammars, we do not distinguish interrogative as a separate subtype because of its unclear distinction from the relative type (i.e. its enormous homonymy).

Other values of the SUBPOS category distinguish possessive, reflexive and clitic pronouns (see below).

Clitichood. Several personal pronouns have a clitic (short) form. They have a special value of the SUBPOS:

P5 personal gender clitic (e.g. *mu*, *ho* 'him');

PH personal no gender clitic (e.g. *mě* 'me', *mi* 'me', *ti* 'you');

P7 personal (no gender) reflexive clitic (e.g. *se*, *si*).

Distinguishing clitic forms at SUBPOS tag position violates the principle that the 2^{nd} position is the same for the whole paradigm (cf. in Sect. 3.2).

The pronouns in the forms requested after any preposition (with prefix *n-*: *něj*, *něho* 'him') are not distinguished in the SUBPOS position; these wordfoms are distinguished at the 15^{th} position of the tag (cf. in Tab. 50).

Reflexivity and Possession. Futher subtypes of personal (and also relative) pronouns are introduced based on the feature of reflexivity and possession:

PS personal possessive pronouns (e.g. *můj* 'my', *tvůj* 'your', *náš* 'our' and *váš*);

P6 personal reflexive pronoun (long forms; i.e. *sebe*, *sobě*);

P8 personal possessive reflexive pronoun (i.e *svůj* 'my' / 'your' / 'her' / 'his').

Furthermore, two groups are specially divided:

P9 personal possessive pronouns for the 3^{rd} person (e.g. *jeho* 'his', *její* 'her', *jejich* 'their');

P1 relative possessive pronouns (*jehož*, *jejíž* 'whose').

Except the gender and number of an object, these pronouns (P9 and P1) express also the gender and number of a subject, i.e. possessor (e.g. *jeho dům* 'his house' (masc. sg., possessor: masc. sg.), do jejího domu 'to her house' (masc. sg., possessor: fem. sg.), do její chalupy 'to her cottage' (fem. sg. possesor: fem. sg.), jejich dům 'their house' (masc. sg., possessor: pl.). That's why the POSSGENDER and POSSNUMBER tag positions are also filled here. Cf. examples in Tab. 50.

11.2 Subtypes of numerals

Within the category of numerals, words are associated mainly on the basis of numerical meaning. From a formal point of view, there are words that behave as nouns (*nula* 'zero', *milión* 'million'), pronouns (*dvě* 'two', *jedny* 'one-kind'), adjectives (*první* 'first', *paterý* 'five-kinds', *několikátý* 'umpteenth') and adverbs (*třikrát* 'three-times'). We respect the traditional semantic concept of the POS of numerals, however, the formal behavior is the basic for the subdivision of numerals into individual subtypes.

Numerals (except the numbers written by arabic or roman numeral symbols) are distinguisthed into several subtypes (captured at the 2^{nd} tag position – SUBPOS, see overview in Sect. 5.2 and Tab. 51 here) according various combination of the two basic features:

- morphological behavior
- semantics including definiteness

Type & Subtype		**Adjectival**	**Non-adjectival**
Cardinal	def.	`Cn` *jeden*, *dva*, *oba*	`Cl` *tři*, *čtyři*, *pět*, *půl*
			`Cz` *sto*, *tisíc*, *miliarda*, *nula*
	indef.	`Cy` *nejeden*	`Ca` *mnoho*, *málo*, *t/kolik*, *pár*
Ordinal	def.	`Cr` *první*, *druhý*, *šestsetdruhý*	-
	indef.	`Cw` *k/tolikátý*, *bůhvíkolikátý*	-
Multiplicative	def.	-	`Cv` *třikrát/e*
	indef.	-	`Co` *k/tolikrát/e*, *několikrát*
Generic	def.	`Cd` *jedny*, *dvoje/í*, *paterý/y*, *obé*	`Cj` *patero*, *dvé*, *tré*
	indef.	`Ch` *nejedny*, *několikerý/y*	`Ck` *několikero*
Arabic numerals		`C=` *1.24*	
Roman numerals		`C}` *MXV*	

Table 51: Subtypes of numerals

Morphology. The morphological criterion is primary. We distinguish between numerals with adjectival declension and numerals with another declension.

The **adjectival numerals**, i.e. with agreement gender, express variable values of the gender (and also number) depending on the gender (and number) of the governing noun (e.g. *jeden člověk* 'one man' (masc. sg.), *jedna žena* 'one woman' (fem. sg.), *jedno dítě* 'one child' (neut. sg.) or *několikátý problém* 'multiple problem' (masc. sg.), *několikáté problémy* 'multiple problems' (masc. pl.)). All wordforms are represented by one lemma (nom. sg. masc. anim.); similarly to adjectives. However, there are not degrees of comparison in adjectival declension of numerals, and some paradigms have only singular (e.g. *jeden* 'one') or only plural wordforms (*dva* 'two'). In the tag, the `GENDER`, `CASE` and `NUMBER` positions are filled. Cf. examples in Table 52.

There are several types of **non-adjectival numerals**. First, their morphological behavior is similar to that of nouns (e.g. *sto* 'hundred', *nula* 'zero', *patero* 'five-kinds-of'). The position of `GENDER` is filled (one value for the whole paradigm).

In the other types, the morphological behavior is quite specific and expresses no gender; they can be inflected (e.g. *tři domy* 'three houses' (Nom.), *bez tří domů* 'without three houses' (Gen.), *mnoho domů* 'many houses' (Nom.), *bez mnoha domů* 'without many houses' (Gen.)), or they are inflexible (e.g. *několikrát* 'several-times' with only variant wordform *několikráte* 'several-times'). The tag position of `GENDER` is not filled. Other tag positions (`CASE`, `NUMBER`) are filled in relation to other features. Only the subtype of cardinal non-adjectival definite numerals covers both the types, i.e. distinguishes "noun" and "no-gender" type by a specific value of `SUBPOS` in the tag (cf. value `Cl` and `Cz` in the Tab. 51).

In certain contexts, numerals do not display grammatical relations with other words in a sentence by means of declension, they are used in a rigid (indeclinable) form. Cf. *ke čtyřem stům dětem* 'to four hundred children' (inflectional form) vs. *ke čtyři sta dětem* 'to four hundred children' (indeclinable form); cf. also *sto* 'hundred' in *do sto lidí* 'to a hundred people' or *v pouhých sto výtiscích* 'in just a hundred copies',[20] *tisíc* 'thousand' in *až po stovky tisíc let* 'for hundreds of thousands of years', *jedna* 'one' in *sedmdesát jedna občanů* 'seventy-one citizens', *čtvrtě* 'quarter' in *před tři čtvrtě rokem* 'three quarters of a year ago', *pár* 'few' in *o pár stech tisících* 'about a few hundred thousand', etc. So for most (cardinal) numerals we introduce a subspecified analysis (with the value `X` for `GENDER`, `NUMBER`, `CASE`). Cf. examples in Tab. 52.

[20]Note that in *v pouhých sto výtiscích* 'in just a hundred copies' the adjective *pouhých* 'just' is in agreement not with the following numeral *v pouhých sto výtiscích* 'hundred', but with the noun in the genitive *výtiscích* 'copies' depending on the numeral.

Wordform	Lemma	Tag
jeden muž	jeden‘1	CnYS1----------
jedna žena	jeden‘1	CnFS1----------
jedno dítě	jeden‘1	CnNS1----------
několikátý problém	několikátý	CwYS1----------
několikáté problémy	několikátý	CwIP1----------
sto lidí	sto-1‘100	CzNS1----------
nulou dělit nelze	nula	CzFS7----------
patero přikázání	patero‘5	CjNS1----------
tři domy	tři‘3	Cl-P1----------
do mnoha zemí	mnoho-1	Ca--2----------
několikrát zazvonil	několikrát	Co----------
několikráte zazvonil	několikrát	Co----------1
ke čtyřem stům dětem	čtyři‘4	Cl-P3----------
ke čtyři sta dětem	čtyři‘4	Cl-XX----------
ke čtyři sta dětem	sto-1‘100	CzNXX---------1
do sto lidí	sto-1‘100	CzNXX----------
až po stovky tisíc let	tisíc‘1000	CzIXX----------
sedmdesát jedna občanů	jeden‘1	CnXXX----------
před tři čtvrtě rokem	čtvrt	CzFXX---------1

Table 52: Examples: Subtypes of numerals

Semantics. We classify the following subtypes of numerals:

- **Cardinal** – express quantity.
  - `Cn` adjectival definite: *jeden* ‘one’, *dva* ‘two’, *oba* ’both’.
  - `Cy` adjectival indefinite: *nejeden* ‘not-only-one’.
  - `Cl` non-adjectival no gender definite: *tři* ‘three’, *čtyři* ‘four’, *pět* ’five’, *půl* ‘half’.
  - `Ca` non-adjectival no gender indefinite: *kolik* ‘how much’, *mnoho* ‘much/many’, *málo* ‘little/few’, *tolik* ‘that much/many’, *několik* ‘some (number of)’, *kdovíkolik* ‘who-knows-how-much/many’, *pár* ‘some (number of)’.
  - `Cz` non-adjectival "noun" definite: *sto* ‘hundred’, *milion* ‘million’, *nula* ’zero’, *čtvrt* ‘quarter’.
- **Ordinal** – express position in a sequential order.
  - `Cr` adjectival definite: *třetí* ‘third’, *pátý* ‘fifth’.
  - `Cw` adjectival indefinite: *kolikátý* ‘at-what-position-in-a-sequence’, *tolikátý* ‘at-that-position-in-a-sequence’, *několikátý* ‘umpteenth’.
- **Multiplicative** – express how many times/folds.
  - `Cv` definite: *pětkrát* ‘five-times’, *sedmkrát* ‘seven-times’.
  - `Co` indefinite: *kolikrát* ‘how-many-times’, *mnohokrát* ‘many-times’, *tolikrát* ‘that-many-times’, *několikrát* ‘several-times’, *nejednou* ‘not-only-one-time’.
- **Generic** – express number of different kinds, types.
  - `Cd` adjectival definite: *jedny* ‘one-kind’, *dvojí* ‘two-kinds’, *desaterý* ‘ten-kinds’, *patery* ‘five-kinds’, *obé*.

`Ch` adjectival indefinite: *kolikerý* 'how-many-kinds', *nejedny* 'not-only-one-kind', *tolikerý* 'that-many-kinds', *několikerý* 'several-kinds', *několikery* 'several-kinds'.

`Cj` non-adjectival "noun" definite: *čtvero* 'four-kinds-of', *desatero* 'ten-kinds-of', *dvé*, *tré*.

`Ck` non-adjectival "noun" indefinite: *kolikero* 'how-many-kinds', *několikero* 'several-kinds-of', *tolikero* 'that-many-kinds'.

Unlike Czech grammars, we do not distinguish interrogative numerals as a separate type: the interrogative numerals are included in the corresponding types of indefinite numerals (e.g the interrogative numeral *kolikátý* 'at-what-position-in-a-sequence' is included in the ordinal indefinite type, the interrogative numeral *kolikrát* 'how-many-times' is included in the multiplicative indefinite type).

Words composed from a numeral and a preposition, e.g. *podruhé* 'the-second-time', *poněkolikáté* 'several-times', *zadruhé* 'the-second-time', *naněkolikrát* 'several-times' are considered compound adverbs.

11.3 Subtypes of adverbs

We divide adverbs into two groups differentiated in the second position in the tag (i.e. `SUBPOS`):

`Db` Adverb without a possibility to form negation and degrees of comparison, i.e. positions of the `NEGATION` and `GRADE` are not filled. E.g. *pozadu* 'behind', *naplocho* 'flatly', *včera* 'yesterday', *kde* 'where'.

`Dg` Adverbs forming negation or/and degrees of comparison; positions of the `GRADE` and `NEGATION` are filled. E.g. *rychle* 'fast', *nerychle* 'not-fast' *zajímavější* 'more interesting'. Some adverbs only form negative worfforms, but they do not form degrees of comparison; e.g. *mnoho* 'many'.

See examples in Tab. 53.

Wordform	Lemma	Tag
pozadu	`pozadu`	`Db----------`
zítra	`zítra`	`Db----------`
kde	`kde`	`Db----------`
rychle	`rychle_^(*1ý)`	`Dg-------1A----`
nerychle	`rychle_^(*1ý)`	`Dg-------1N----`
nejrychleji	`rychle_^(*1ý)`	`Dg-------3A----`
pracuje nemnoho	`mnoho-2`	`Dg-------1N----`

Table 53: Examples: Subtypes of adverbs

12 Negation

Negated words, which are only the opposite of positive words (e.g. *neherec* non-actor', *nevelký* 'non-big', *nepracovat* 'do not work') are captured in the same paradigm (with a positive lemma) together with non-negated forms. Negated forms have the value `N` at the 11^{th} `NEGATION` position in the tag (Sect. 5.11), affirmative forms have the value `A`. Cf. examples in table 54.

Negated forms of adjectives, adverbs and verbs are generated into the dictionary regularly within inflective patterns. Negated forms of nouns are added to the dictionary selectively.

A negated lemma (and thus a separate paradigm for negated forms) occurs when the negated forms of the word are not merely a negation of the positive form (there is a noticeable shift in meaning; e.g. *nemoc* 'disease', *nebývalý* 'unprecedented'). Cf. examples in Tab. 54.

For numerals (with the value `C` of `POS`), pronouns (with the value `P` of `POS`) and possessive adjectives (with the value `AU` at the `POS` and `SUBPOS` positions), the `NEGATION` position is not filled. Negated forms of these words are rare (e.g. *nemnoho* 'non-much', *neúčastníkův* 'non-participant's', *nemůj* 'non-mine'). If they occur, these negated forms are captured as a separate paradigm with the lemma in the negated form. Cf. examples in Tab. 54.

Wordform	Lemma	Tag
neherec	`herec`	`NNMS1-----N----`
nevelký dům	`velký`	`AAIS1----1N----`
zase nepracuje	`pracovat`	`VB-S---3P-NAI--`
má moc	`moc-1`	`NNFS4-----A----`
má nemoc	`nemoc`	`NNFS4-----A----`
nebývalý zájem	`nebývalý`	`AAIS1----1A----`
neúčastníkův zájem	`neúčastníkův_^(*2)`	`AUIS1M---------`
nemnoho lidí	`nemnoho`	`Ca--1----------`

Table 54: Examples: Negation

13 Names and Terms

Proper names in Czech always start with a capital letter. From morphological point of view, they are "normal" words, but it appears to be useful to give them a simple semantic label. Thus, name labels are attached to their lemmas. Named entities are very often multiple-word expressions and as such cannot be optimally described at wordform level. Therefore, lemma of a proper name bears name label only if it is a proper name itself, i.e. not in combination with another word. We respect the original idea that the name label explains the meaning of the lemma, not the context it appears in. Thus for instance *Nový* 'New' is lemmatized as `nový` in *Nový Bydžov*, not `Nový_;G` and *Bydžov* is lemmatized `Bydžov_;G`. For details see below.

For overview of the labels for names and terms see Sect. 4.2.2.

13.1 Personal names

Personal names are assigned the label `Y`. Given names and surnames are not distinguished by the name label in their lemmas. The value `Y` is used for all personal names, also for nicknames, names of horses, pets, etc. Names of members of a particular nation, inhabitants of a particular territory are labeled with `E`. See examples in Tab. 55.

Wordform	Lemma	Tag
Pythagoras	`Pythagoras_;Y`	`NNMS1-----A----`
Jiří Včelař Kotas	`Včelař_;Y`	`NNMS1-----A----`
Alík	`Alík_;Y`	`NNMS1-----A----`
Dán	`Dán_;E`	`NNMS1-----A----`
fotbalista Petr Čech je Čech	`Čech_;E_;Y`	`NNMS1-----A----`
Pcheng je čínské jméno i příjmení	`Pcheng_;Y`	`NNMS1-----A----`
Italčino vítězství	`Italčin_;E_^(*3ka)`	`AUNS1F---------`
František Palacký	`Palacký_;Y`	`NNMS1-----A----`
Milena Jesenská	`Jesenská_;Y`	`NNFS1-----A----`
paní Nováková	`Nováková_;Y`	`NNFS1-----A----`
Angelina Jolieová	`Jolieová_;Y`	`NNFS1-----A----`
Angelina Jolie	`Jolie_;Y`	`NNFXX-----A----`
s Georgem Washingtonem	`George_;Y`	`NNMS7-----A----`
s George Washingtonem	`George_;Y`	`NNMXX-----A----`

Table 55: Examples: Personal names

If a name can serve both as a personal name and as a members of a particular nation (e.g. name *Čech* as a surname and nation) and if there is a same declension for the both usages, then there is one lemma/paradigm covering all usages and all relevant name labels are attached to it (cf. ex. in Tab. 55). Possessive adjectives derived from proper names (*Novákův*, *Italčin*) inherit the name labels from the original lemma.

Personal names homonymous with a normal Czech word always have a lemma of their own. Thus *Zeman* (surname) is lemmatized as `Zeman-1_;Y`, not `zeman` 'squire' (cf. also ex. of *Písek* in Tab. 8 in Sect. 4.1.1).

Personal names are always tagged as nouns, even if they have an adjectival form (cf. ex. *Palacký*, *Jesenská*, *Nováková*, *Jolieová* in Tab. 55).

Foreign personal names are not marked as foreign words because in Czech texts, they are usually declined according to the Czech grammar; e.g. *Bill Clinton, bez Billa Clintona, Billu Clintonovi, s Billem Clintonem....* Even if a name allows for a frozen (undeclined) form, there usually is a context in which it can be declined: *s George Washingtonem* vs. *s Georgem Washingtonem*. Some foreign names, such as *Steffi*, *Jolie* are never declined. Cf. Tab. 55.

Names of horses, pets, etc. have all kinds of names (e.g. *Vinná réva, He Shall Reign*, *La Paloma*, *Lučina*, *Areál*). Quite often one does not know whether it is male or female (sometimes even female-like names belong to a male horse). If any reasonable analysis is possible it should be used regardless the lemma is marked as a name or not. It will be marked as a name within a separate project on named entity recognition. However, if the name is a word that has no other meaning or if it has different gender (declension), a new lemma with the label `Y` is introduced.

Prepositions and other function words in personal names. Prepositions, conjunctions and other determiners form parts of personal names that indicate geographical roots of the family (*Jiří z Poděbrad*, *Kryštof Harant z Polžic a Bezdružic*). They are analyzed as normal words. It may not be always clear whether the part after the preposition shall be annotated as a surname or a geographical name. If the Czech preposition *z* is present, the following word is a geographical name (even if it is a foreign location as in *Blanka z Valois*). In the foreign personal names (e.g. *Ludwig van Beethoven*, *Miguel de Cervantes y Saavedra*, *Hans van den Broek*), the foreign prepositions (*von*, *van* and *de*) are analyzed as foreign words. The original geographical meaning is usually less obvious for a Czech reader and the following word is annotated as surname. See ex. in Table 56.

Wordform	Lemma	Tag
Jiří z Poděbrad	`z-1`	`RR--2----------`
Jiří z Poděbrad	`Poděbrady_;G`	`NNIP2-----A----`
František Lobkovic	`Lobkovic_;Y`	`NNMS1-----A----`
František z Lobkovic	`Lobkovice_;G`	`NNFP2-----A----`
František z Lobkovic	`z-1`	`RR--2----------`
Blanka z Valois	`Valois_;G`	`NNNXX-----A----`
Ludwig van Beethoven	`van-77`	`F%-------------`
Ludwig van Beethoven	`Beethoven_;Y`	`NNMS1-----A----`

Table 56: Examples: Function words in personal names

Chinese and other Asian names can consist of one syllable or of two (or more) syllables, often connected with a dash (however sometimes separated by a space). For multi-syllable (hyphenated) names (e.g. *Siao-Pching*, *Ir-Sen*), the tag usually specifies `NUMBER` and `CASE` position only for the last syllable. If the name is in the nominative according to the context (and therefore it is not possible to recognize whether it is used as a declinable word), the `NUMBER` and `CASE` positions are filled as a nominative (i.e. value `X` is not use). However, if a foreign name is never declined in Czech, the unspecified tag is used (with `X`-value at `NUMBER` and `CASE` positions). The last syllable in a hyphenated composite can also be captured as a segment, especially if it is written with lower case and expresses a case and a number (e.g. *Chuang-timu*). Cf. ex. in Tab. 57.

Wordform	Lemma	Tag
Teng Siao-Pching odchází.	`Teng_;Y`	`NNMS1-----A----`
Teng Siao-Pching odchází.	`Siao-3_;Y`	`NNMXX-----A----`
Teng Siao-Pching odchází.	`Pching_;Y`	`NNMS1-----A----`
podobá se Čhin Š' Chuang-timu	`Čchin-1_;Y`	`NNMXX-----A----`
podobá se Čchin Š' Chuang-timu	`Š-1_;Y`	`NNMXX-----A----`
podobá se Čchin Š' Chuang-timu	`Chunag-1_;Y`	`NNMXX-----A----`
podobá se Čchin Š' Chuang-timu	`ti-2_;Y`	`SNMS3-----A----`

Table 57: Examples: Chinese and other Asian names

13.2 Geographical names

Geographical names are assigned label `G`.

13.2.1 Countries, cities, rivers, mountains

The main word (head) in a multi-word geographical name is a noun; the same holds for a one-word city name. If it is homonymous with an adjective, a new noun lemma/paradigm is created for the name. Thus *Hluboká* in the name *Hluboká nad Vltavou* is captured as a noun with lemma `Hluboká_;G` (not as an general adjective `hluboký` 'deep').

Nouns that are frequently used in names (such as *Újezd*, *Ústí* may have their own geographical lemmas even if they are homonymous with a normal word. For homonymous pairs where the non-geographical usage is much more common (such as *voda* 'water', *ves* 'village', *město* 'city') it is recommended to capture them with the non-geographical lemma even in geographical usages. Other words in multi-word names (adjectives, prepositions, conjunctions, etc.) are represented as ordinary words. Other nouns in names can be represented as geographical names only if they are themselves geographical names (names of rivers and mountains in city names, etc.). See Tab. 58.

Wordform	Lemma	Tag
Hluboká nad Vltavou	`Hluboká_;G`	`NNFS1-----A----`
Ústí nad Labem	`Ústí_;G`	`NNNS1-----A----`
Ústí nad Labem	`nad-1`	`RR--7----------`
Ústí nad Labem	`Labe_;G`	`NNNS7-----A----`
Ohrada u Hluboké	`ohrada`	`NNFS1-----A----`
Ohrada u Hluboké	`Hluboká_;G`	`NNFS2-----A----`
Kostelec nad Černými lesy	`Kostelec_;G`	`NNIS1-----A----`
Kostelec nad Černými lesy	`černý_;o`	`AAIP7----1A----`
Kostelec nad Černými lesy	`les`	`NNIP7----1A----`
Karlovy Vary	`Karlův_;Y_^(*3el)`	`AUIP1M---------`
Karlovy Vary	`Vary_;G`	`NNIP1-----A----`
Orlické hory	`orlický`	`AAFP1----1A----`
Orlické hory	`hora`	`NNFP1-----A----`
Divoká Orlice	`divoký`	`AAFS1----1A----`
Divoká Orlice	`Orlice_;G`	`NNFS1-----A----`
na Mont Blanku	`Blanc-1_;G`	`NNIS6-----A----`
v Cincinnati	`Cincinnati_;G`	`NNNXX-----A----`
v Los Angeles	`Los-77`	`F%-------------`
v Los Angeles	`Angeles-77`	`F%-------------`

Table 58: Examples: Names of countries, cities, rivers, mountains

Foreign geographical names are mostly annotated as an undeclined noun (e.g. *Cincinnati*) or foreign words (mainly in the case of multi-word names, e.g. *Los Angeles*). If they are declined in Czech, they have analysis according to the Czech morphology (no matter how morphologically they behave in the original language). For instance, *blanc* is adjective in French *Mont Blanc* but it behaves as a noun in *na Mont Blanku*. *Mont* is annotated as a foreign word. See Tab. 58. See Sect. 17 for more information on foreign words.

Adjectives derived from geographical names. Adjectives derived from geographical names (e.g. *africký* 'African' from *Afrika* 'Africa') are not marked as geographical (no `G` label in lemma). These adjectives are not capitalized in Czech, while the original nouns are.

13.2.2 Streets, squares, stations

In street, station and square names, we suppose that a word such as *ulice* 'street', *náměstí* 'square', etc. is always present, even if elided on the surface. Therefore the tagging of the name of the street and square is not altered. Cf. Table 59.

Wordform	Lemma	Tag
Bydlím v Horské ulici	horský	AAFS6----1A----
Bydlím v ulici Štěpánská	štěpánský	AAFS1----1A----
Bydlím v ulici Mezi Zahrádkami	mezi-1	RR--7----------
Bydlím v ulici Mezi Zahrádkami	zahrádka	NNFP7-----A----
Palackého náměstí	Palacký_;Y	NNMS2-----A----
Sejdme se na Palackého	Palacký_;Y	NNMS2-----A----
Bydlím na náměstí Míru	mír_^(opak_války)	NNIS2-----A----
stanice Staroměstská	staroměstský	AAFS1----1A----
ve směru od Dejvické	dejvický	AAFS2----1A----
zastávka Na Knížecí	Knížecí_;G	NNNS6-----A----
stanice Anděl	Anděl-2_;G	NNIS1-----A----

Table 59: Examples: Names of streets, squares, stations

13.2.3 Buildings

Building names can be annotated as a geographic name. On the other hand, many building names are composed of common words (apelatives) and we do not capture such names as geographical names. Cf. Tab. 60.

Wordform	Lemma	Tag
Rudolfinum	Rudolfinum_;G	NNNS1-----A----
Pražský hrad	pražský	AAIS1----1A----
Pražský hrad	hrad	NNIS1-----A----
Chrám sv. Barbory	chrám	NNIS1-----A----
Chrám sv. Barbory	svatý-1	AAXXX----1A---b
Chrám sv. Barbory	Barbora_;Y	NNIS2-----A----

Table 60: Examples: Names of buildings

13.3 Scientific terminology

Scientific terms from chemistry, medicine, natural science (written with uppercase letter) are assigned U-label. Cf. Tab. 61.

Wordform	Lemma	Tag
Acylpirin	Acylpirin_;U	NNIS1-----A----
Australopithecus	Australopithecus_;U	NNMS1-----A----
Hydrosulfit	Hydrosulfit_;U	NNIS1-----A----

Table 61: Examples: Scientific terminology

13.4 Other proper names

Other proper names, names of companies, foundations, shops, clubs, sport clubs, restaurants, unique product names, names of events, works of art etc. have lemmas flagged `m`.

However, "words", the usage of which is not limited to the name, get their "normal" lemmas (and they are not annotated as a name). Only if a word cannot be explained another way or if its meaning has nothing to do with the company or in the case of well-known name (e.g. *Škoda*), the label and capitalized lemma is used.

The border between personal and other names is fuzzy: if it is clear that a surname is part of a company name (e.g. *Uzenářství Novák a syn*) it is lemmatized with personal name label. On the other hand, *Škoda* is annotated as a company no matter that it was also named after a person. This name is mostly known as a company name. The same applies to geographic and other names if they are part of a company name (cf. *Stavební společnost Šumava, s.r.o* and other examples in Tab. 62).

Wordform	Lemma	Tag
Škoda-auto, a. s.	`Škoda-1_;m`	`NNFS1-----A----`
Škoda-auto, a. s.	`auto`	`NNNS1-----A----`
Uzenářství Novák a syn	`Novák_;Y`	`NNMS1-----A----`
restaurace U Medvídků	`u-1`	`RR--2----------`
restaurace U Medvídků	`medvídek`	`NNMP2-----A----`
TJ Sokol	`Sokol-2_;m_^(organizace)`	`NNIS1-----A----`
Stavební společnost Šumava	`Šumava_;G`	`NNFS1-----A----`
televize Nova	`Nova_;m_^(televize)`	`NNFS1-----A----`
Tatra, a. s.	`Tatra_;m_^(auto;;mléko)`	`NNFS1-----A----`
nový Renault	`Renault_;m`	`NNIS1-----A----`
časopis Sluníčko	`sluníčko`	`NNNS1-----A----`
výstava Habitat 2019	`Habitat_;m`	`NNIS1-----A----`
Mistrovství světa v hokeji	`mistrovství`	`NNNS1-----A----`

Table 62: Examples: Other proper names

14 Abbreviations

There are two ways to capture abbreviations:

- as a special (abbreviated) wordform of a non-abbreviated (single) word paradigm: there is a special value (`b`, `a` or `c`) at the 15^{th} position of the tag,
- as a special `POS` Abbreviation with the code `B` at the 1^{st} position of the tag.

In the first way, fixed abbreviations of a single word are captured (e.g. abbreviation *s.* 'p.' of word *strana* 'page'; see more in Sect. 14.1). Other abbreviations (e.g. *USA*) are captured in the second way (see more in Sect. 14.2).

14.1 Fixed abbreviations of a single word

Fixed, well-known abbreviations of a single word usually followed by a period in Czech text (e.g. abbreviation *s.* 'p.' of word *strana* 'page') are captured as a special wordform of the paradigm of the non-abbreviated word. The fact that it is an abbreviated form is expressed at the 15^{th} position by the letters `b` (`a` or `c` in the case of the other fixed abbreviation of the same word (e.g *s.* and *str.* 'p.' are both used as the abbreviation for *strana* 'page'). Only those categories that are valid for each use of the abbreviation are coded in the tag. Usually, the `NUMBER` and `CASE` positions are not specified.

Wordform	Lemma	Tag
délka 3 m	`metr`	`NNIXX-----A---b`
3m tyč	`metrový`	`AAXXX----1A---b`
ve 2 h	`hodina`	`NNFXX-----A---b`
ve 2 hod.	`hodina`	`NNFXX-----A---a`
na s. 8	`strana`	`NNFXX-----A---b`
na str. 8	`strana`	`NNFXX-----A---a`
za 2 s	`sekunda`	`NNFXX-----A---b`
za 2 sec	`sekunda`	`NNFXX-----A---a`
za 2 sek.	`sekunda`	`NNFXX-----A---c`
kWh	`kilowatthodina`	`NNFXX-----A---b`
800 m n. m.	`nad-1`	`RR--7---------b`
800 m n. m.	`moře`	`NNNS7-----A---b`
č. 5	`číslo`	`NNNXX-----A---b`
300 n. l.	`náš`	`PSZS2-P1------b`
300 n. l.	`letopočet`	`NNIS2-----A---b`
např.	`například`	`TT------------b`
It.	`Itálie_;G`	`NNFXX-----A---b`

Table 63: Examples: Fixed abbreviations of a single word

Units of measurements. In this way, the abbreviations of units of measurements are captured, too (e.g. *km* is abbreviated form of the word *kilometr*, *C* is abbreviated form of *Celsius*, etc.). This type of abbreviations usually appears without the period. Cf. examples in Tab. 63.

Note. It is impossible to capture the abbreviation of more than one word (e.g. abbreviation *atd.* 'etc.' of the words *a tak dále* 'et cetera/and so on') in the way described here. These abbreviations belong to the part of speech Abbreviations (with the `B` value at `POS` position). They are described in the following section 14.2. cf. Tab. 64.

Wordform	Lemma	Tag
atd	atd_^(a_tak_dále)	Bb----------
apod	apod_^(a_podobně)	Bb----------
ap	ap_^(a_podobně)	Bb----------
aj	aj-1_^(a_jiný/á/é)	BAXXX----1A----

Table 64: Examples: Fixed abbreviations of multiple words

14.2 Other abbreviations

Abbreviations which abbreviate at least two words (e.g. *atd* 'and so on'), or are composed of uppercase letters e.g. *USA*) are captured as a special POS with the value B at the 1^{st} tag position. The lemma of such an abbreviation is the abbreviation itself. The belonging of an abbreviation to a traditional POS – although this may be difficult to determine in a number of cases (e.g. multi-word abbreviations; cf. abbreviation *atd.* in Tab. 64) – is reflected at the 2^{nd} position of the tag. The value at the 2^{nd} position also determines which other positions of the tag are filled.

Most abbreviations are nouns and can be used with more than one gender. Of course, abbreviations have no endings but the surrounding context can reveal their underlying gender whenever gender agreement is required by the Czech grammar (e.g. *staronový NKÚ* as masculine inanimate vs. *slavné NKÚ* as neuter). However, the GENDER, NUMBER and CASE positions are usually not specified in the tag of abbreviations (cf. examples in Tab. 65 and 66).

14.2.1 Well-known abbreviations composed of uppercase letters

The lemma of well-known abbreviations composed of uppercase letters (e.g. *USA*; everyone knows what this abbreviation means even without context) is accompanied with a name label (e.g. G in case of geographical name abbreviation). There is a semantic explanation attached to the lemma. Cf. examples in Tab. 65.

Similarly, the symbols of chemical elements and compounds (e.g. abbreviation *N* for nitrogen, *CO* for carbon monoxide) as well as academic titles (e.g. *JUDr*) are captured in this way.

Wordform	Lemma	Tag
USA	USA_;G_^(United_States_of_America)	BNXXX-----A----
NKÚ	NKÚ_;m_^(Nár._kontrolní_úřad)	BNXXX-----A----
CD	CD-1_^(Audio/Data,_Compact_Disc)	BNXXX-----A----
KB	KB_;m_^(Komerční_banka)	BNXXX-----A----
N	N-1_;U_^(zn._dusíku)	BNXXX-----A----
JUDr	JUDr_^(doktor_práv)	BNXXX-----A----

Table 65: Examples: Well-known abbreviations composed of uppercase letters

14.2.2 Less familiar abbreviations and abbreviations with many meanings

In the case of less familiar abbreviations and abbreviations with many meanings (cf. HK in Tab. 66) we do not distinguish partial meanings; there is only one analysis for all usages of the abbreviated form. These abbreviations have the lemma with the number -88. The tag is always BNXXX-----A----. Cf. Tab. 66.

Note: One-letter abbreviations with many meanings (e.g. *V. Havel* vs. *V. Mrštík*) are captured as a special POS for isolated letters (the tag starts with Q3), see Sect. 15.

Wordform	Lemma	Tag
HK je horní končetina	HK-88	BNXXX-----A----
HK je Hradec Králové	HK-88	BNXXX-----A----
HK je Hospodářská komora	HK-88	BNXXX-----A----
OU je Oxford University	OU-88	BNXXX-----A----
OU je odborné učiliště	OU-88	BNXXX-----A----
pr. volno	pr-88	BNXXX-----A----

Table 66: Examples: Less familiar abbreviations and abbreviations with many meanings

14.2.3 Author's signature

The author's name abbreviations used in newspapers (e.g. *ber, mas*, etc. in "sentences" like *PRAHA (ČTK, ber)*) have the lemma equal to the wordform. They are numbered `-99` as a human-readable indication and flagged `Y`. Cf. Tab. 67.

Wordform	Lemma	Tag
Praha (haš)	haš-99_;Y	BNXXX-----A----
Praha (ČTK, ber)	ber-99_;Y	BNXXX-----A----

Table 67: Examples: Author's signature

15 Isolated Letters

Isolated letters stand for many meanings. We do not distinguish between the usage as an abbreviation (e.g. *A. Franklin*), a label (e.g. *skupina A* 'group A'), *A-konto* 'A-account'), an item in a list (e.g. *a)*, a separator in a text (e.g. *o o o o o o o o*), or even other meanings.

Any isolated letter is represented as a special `POS` with the value `Q3` at the first and second position of the tag. The lemma has the number `-33` as a human-readable indication. This analysis exists for all letters of the Czech alphabet (upper- and lower-case). Cf. examples in Tab. 68.

Note. Not every one-letter token in a text is annotated as an isolated letter. We distinguish letters as well-known abbreviations in the case of chemical symbols (e.g. *O* for *oxygen*; see Sect. 14.2.1), letters as one-syllable personal names (especially Chinese; e.g. *S* in *Wang S'tching*), and letters as foreign words (e.g. *I* in *I love you*; see Sect. 17) .

Wordform	Lemma	Tag
kružnice A	`A-33`	`Q3----------`
jedeme do p.	`p-33`	`Q3----------`
V. Havel	`V-33`	`Q3----------`
V. Mrštík	`V-33`	`Q3----------`
Karlovy V.	`V-33`	`Q3----------`
Ch. Dickens	`Ch-33`	`Q3----------`

Table 68: Examples: Isolated Letters

16 Segments

Segments are incomplete words. They are parts of words; in order to understand them, they must be joined with another string or word to create a complete word. They are usually joined with a separator, most often with a hyphen. Note that combinations of two complete words connected with a separator e.g *Praha-Hradčany* 'Prague-Hradcany', *Anna-Marie*, *propan-butan* 'propane-butane') are not segments (see more in Sect. 19). The segments are treated as a special `POS` with the value `S`. According to their position in the complete word, we distinguish:

- prefixal segments,
- postfixal segments.

Wordform	Lemma	Tag
česko-ruská kniha	`česko`	`S2--------A----`
nepoliticko-politická diskuse	`politicko`	`S2--------N----`
hudebně-zábavný pořad	`hudebně-2`	`S2--------A----`
pěti- až sedmiletý chlapec	`pěti'5`	`S2--------A----`
ultra-moderní	`ultra-1`	`S2--------A----`
ultra jemný prášek	`ultra-2`	`AAXXX----1A----`
mini-sukně	`mini-1`	`S2--------A----`
mini sukně	`mini-3`	`AAXXX----1A----`
ne/souhlasím	`ne-2`	`S2--------A----`
s manželem/kou	`ka`	`SNFS7-----A----`
n-tice	`tice_^(n-tice)`	`SNFS1-----A----`
řekl(a)	`a-2`	`SpQW----R-AA---`
uslyší to už po x-té	`tý-2_^(x-tý)`	`SAFS6----1A----`
přikládáme soubor/y	`y-2_^(soubor/y)`	`SNIP4-----A----`
12-ti hodinová směna	`ti_^(10-ti)`	`Sl-XX----------`
na Tchaj-wanu	`wan`	`SNIS6-----A----`

Table 69: Examples: Segments

Prefixal segments are strings that appear at the beginning of words. They are usually followed with a separator, most often with a hyphen (e.g. segment *česko* in example *česko-ruská kniha* 'Czech-Russian book'), but also with a space or another separator (e.g. segment *pěti* in example *pěti až sedmiletý chlapec*).

A typical prefixal segment has a special form ending with *-o* (e.g. *česko-ruský* 'Czech-Russian') but there are other types of segments (cf. segments *kvazi* 'quasi', *hydroxy* in *kvazi-valenční* 'quasi-valency', *hydroxy-sloučeniny* 'hydroxy-compounds'). The form of prefixal segment can be homonymous with an adverb or wordforms belonging to other `POS` (cf. *hudebně* 'musically' as a segment in *hudebně-zábavný* 'musically-entertaining' and *hudebně* as an adverb in *hudebně nadaný* 'musically gifted' or *mini* as a segment in *mini-sukně* 'mini-skirt' and as a noun in *nosit mini* 'wear a mini'). Homonymous wordforms are common particularly in the case of inflexible loanwords (cf. *super* as a segment in *super-moderní* 'super-modern', as an adjective in *super zábava* 'super fun' and as an adverb in *mám se super* 'I am super'; see other examples in Tab. 69).

If a possible segment is not expressed in a special "segment" form like *česko*, *anglo*, *tří*, then we tag it as a segment only in those cases where the wordform is attached to the next word by a non-space separator (cf. *pop*, *mini* as segments in *pop-kultura* 'pop-culture', *mini-sukně* 'mini-skirt' and as adjectives in *pop kutura* 'pop culture', *mini sukně* 'mini skirt').

Lemma of a prefixal segment is the string itself, unless it is in negative form. In that case, the positive form (without the negative prefix *ne-* 'non') is considered to be the lemma (cf. *nepoliticko* 'nonpolitical' in Tab. 69). The tag of all prefixal segments has the code `2` at the 2^{nd} position. Moreover, we specify for them also the 11^{th} position concerning negation (see examples in Tab. 69).

Postfixal segments are strings that appear at the end of a wordform. They are usually attached directly to the word they combine with (cf. segment *ka* in example *manžel/ka* 'husband/wife') or segment *kou* in example *s manželem/kou* 'with husband/wife'). The separator is most often a hyphen, a parenthesis or a slash.

The postfixal segments express an affiliation to a specific `POS`. Thus, all the inflectional categories that describe the whole wordform, except for the first one (the code for `POS`, which is `S`), are filled in the tag (with the exception of the `ASPECT` for verbs). The lemma is the closest "basic wordform". See examples in Tab. 69.

17 Foreign Words

Foreign words enter Czech texts in three different ways:

- in citations (i.e. whole phrase in a foreign language, multi-word foreign name),
- as individual words (i.e. single foreign word),
- as domesticated words of foreign origin.

17.1 Citation use

Whole phrases in a foreign language are sometimes inserted into Czech texts as citations. Multi-word foreign names, multi-word Latin nomenclature, names of songs and other works belong to this category, too (e.g. *European market research center*, *University of Colorado*, *Monomorium floricola*).

Wordform	Lemma	Tag
práce v European market research center	European-77	F%----------
práce v European market research center	market-77	F%----------
práce v European market research center	research-77	F%----------
práce v European market research center	center-77	F%----------
na University of Colorado	University-77	F%----------
na University of Colorado	of-77	F%----------
na University of Colorado	Colorado-77	F%----------
v Colorado Springs	Colorado-77	F%----------
v Coloradu Springs	Colorado-1_;G	NNNS6-----A----
v Coloradu Springs	Springs-77	F%----------
chování mravence Monomorium minimum	Monomorium-77	F%----------
chování mravence Monomorium minimum	minimum-77	F%----------
Peugeot 306 Grand Prix	Peugeot-77	F%----------
Peugeot 306 Grand Prix	306	C=----------
Peugeot 306 Grand Prix	Grand-77	F%----------
Peugeot 306 Grand Prix	Prix-77	F%----------
The FIA Cup	The-77	F%----------
The FIA Cup	FIA-88	BNXXX-----A----
The FIA Cup	Cup-77	F%----------
Třetí ročník Škoda Hockey Cupu	Škoda-1_;m	NNFS1-----A----
Třetí ročník Škoda Hockey Cupu	Hockey	F%----------
Třetí ročník Škoda Hockey Cupu	cup-1_^(pohár)	NNIS2-----A----
najlepšie slovenské piesne	najlepšie-77	F%----------
najlepšie slovenské piesne	slovenský	AAFP1----1A----
najlepšie slovenské piesne	piesne-77	F%----------

Table 70: Examples: Foreign words in citations

All wordfoms in such a foreign piece of text are analyzed as the foreign word POS coded F% at the 1^{st} and 2^{nd} tag positions. Lemma is the same as the wordform, including uppercase and lowercase initial letter, it has an index of 77 and no other labels. The whole tag is always F%----------. There are two exceptions to this rule: Abbreviations of uppercase letters (cf. *FIA* in *The FIA Cup*) are analyzed as abbreviations (see Section 14) and numbers writtten in digits have always tag C=----------. See examples in Table 70.

If an apparently foreign word is used with a Czech suffix, we analyse it as a Czech word. In other words, its `POS` is not "foreign word". For instance the wordform *Cupu* in *Třetí ročník Škoda Hockey Cupu* has an analysis based on the Czech morphology. This rule applies mostly to multi-word foreign geographical names (e.g. *v Tel Avivu*, *v Coloradu Springs*, etc.). See examples in Table 70.

Slavic languages and Czech dialects. Slavic languages (most prominently Slovak, but also Czech dialects and old Czech) are related to contemporary standard Czech. Citations may contain words that are identical to their Czech counterparts. When a word has a foreign suffix it must be annotated as a foreign word even if its baseform belongs to Czech. If all words in a phrase are identical in their forms and meanings to Czech, the phrase should be annotated as Czech, even if we know that it is in fact Slovak or other language. For instance, if a Slovak song was named *Drahý otec*, there is no need to annotate it as foreign. However, if a single word does not fit the Czech grammar or vocabulary, the best would be to annotate whole citation as foreign words. See example *Vracaja sa dom* in Table 70.

17.2 Single word use

In word usages, Czech morphology takes precedence.

Wordform	Lemma	Tag
s Georgem Washingtonem	`George_;Y`	`NNMS7-----A----`
s George Washingtonem	`George_;Y`	`NNMXX-----A----`
to je George Washington	`George_;Y`	`NNMS1-----A----`
pracuje v Coloradu	`Colorado-1_;G`	`NNNS6-----A----`
pracuje v Colorado	`Colorado-1_;G`	`NNNXX-----A----`
prapaguje nový business	`business`	`NNIS4-----A----`
v novém businessu	`business`	`NNIS6-----A----`
mluvil s Hillary	`Hillary-1_;Y`	`NNFXX-----A----`
následuje písnička Girls	`Girls-77`	`F%-------------`

Table 71: Examples: Single foreign words

The basic rule applies to separately used foreign words in a Czech sentence (often names, terms; e.g. *George*, *Colorado*): If a wordform takes a Czech suffix (e.g *s Georgem*, *v Coloradu*), it is not captured as a foreign word. It is analyzed according to the Czech morphology. See examples in Table 71. If a single foreign name or term does not take Czech suffixes (e.g *s George Washingtonem*, *pracuje v Colorado*), it is usually captured as an inflexible noun (with value `X` in `NUMBER` and `CASE` position). However, in the case of little known words (e.g *následuje písnička Girls*), the wordform is captured as the foreign word `POS`.

Wordform	Lemma	Tag
online služby	`online-1`	`AAXXX----1A----`
pracuje online	`online-2`	`Db-------------`
před úvodním buly	`buly`	`NNNXX-----A----`
s pop artem	`art-1`	`NNIS7-----A----`
art zóna	`art-2`	`AAXXX----1A----`
Museum of Art	`Art-77`	`F%-------------`
láska, jídlo a faux pas	`faux-77`	`F%-------------`
láska, jídlo a faux pas	`pas-77`	`F%-------------`

Table 72: Examples: Domesticated words of foreign origin

17.3 Domesticated words of foreign origin

Foreign words constantly enter Czech language, take Czech endings, settle with Czech declension paradigms and become normal Czech words. Words that entered Czech long ago are not felt as foreign any more (e.g *kakao*). Nevertheless, even newer words (e.g *hardware*) should not be treated as foreign if they fit into this category.

Domesticated loanwords are analyzed according to the Czech morphology. However, some loanwords are inflexible, they do not settle with any Czech declension paradigms, but they are part of the Czech vocabulary (e.g *online, buly, wi-fi*). These domesticated loanwords are not captured with the foreign word `POS`, they are treated as inflexible nouns, adjectives, adverbs, etc. (cf. also Sect. 10.1). Multi-word inflexible loanword (e.g *faux pas, de iure, de facto, play off, ad hoc, cash flow*), though domesticated, are captured as foreign word `POS`. The boundary between inflexible loanwords and foreign words is very blurred. See examples in Table 72.

18 Aggregates

An aggregate is a wordform that is created by joining two or more wordforms (components of the aggregate) into one and cannot be simply assigned any `POS`. Aggregates are common especially in agglutinative languages, but there are some aggregate types in Czech, too. The following types of aggregates are captured:

- **pronominal aggregate**: aggregate consisting of a preposition and the pronoun *on* 'he' or *co, copak* 'what' (e.g. *pro* + *on* → *proň*; *za* + *co* → *zač*).
- **verbal aggregate**:
  - aggregate containing the contracted *-s* which stands for the wordform *jsi* 'you are'. It can be appended to the end of a wordform of almost any `POS` (e.g. *promluvil* + *jsi* → *promluvils*; *dobře* + *jsi* → *dobřes*).
  - aggregate consisting of the conditional verbal wordform *by* 'would' or conditional conjunction such as *aby, kdyby* 'so that, if-would' and contracted form of the auxiliary verb *být* 'to be': *-ch* for the 1^{st} person singular, *-s* for the 2^{nd} person singular, *-chom* for the 1^{st} person plural, *-ste* for the 2^{nd} person plural (e.g.*kdyby* + *byste* → *kdybste*; *aby* + *bychom* → *abychom*).

The lemma of a pronominal aggregate is the lemma of the pronoun. The lemma of a verbal aggregate (containing the contracted form of the present tense of the auxiliary verb *být* 'to be') is the lemma of its first component.

The fact that a wordform is an aggregate is coded at the 14^{th} position of the tag. The code of pronominal aggregates corresponds to the initial letter of the preposition that forms their first component. Verbal aggregates are coded with the letter `c` for *-ch*, `s` for *-s*, `m` for *-chom* and `e` for *-ste*; see also Section 5.14. Verbal and pronominal aggregates can combine; such aggregates are marked with the initial letter of the preposition, but in an uppercase letter (see the example *začs* in Table 73). The lemma of such (combined) aggregates is the pronoun involved.

Wordform	Lemma	Tag
zač	co	PQ--4--------z-
začs	co	PQ--4--------Z-
doň	on-1	P5ZS2--3-----d-
načpaks	copak	PQ--4--------N-
dobřes	dobře	Dg-------1A--s-
promluvils	promluvit	VpYS----R-AAPs-
bych	být	Vc----------Ic-
bysme	být	Vc----------Im6
kdybychom	kdyby	J,----------m-
dybychom	dyby_,h_^(^GC**kdyby)	J,----------m-
dybysme	dyby_,h_^(^GC**kdyby)	J,----------m6
abyste	aby	J,----------e-

Table 73: Examples: Aggregates

19 Hyphenated Composites

Words written with a hyphen (e.g. *Praha-Hradčany* 'Prague-Hradcany') are tokenized into three tokens in the annotation: the part before the hyphen, the hyphen and the part after the hyphen (see Sect. 21). Each part is analyzed separately. We distinguish:

- hyphenated composite of single words,
- hyphenated compound word,
- hyphenated foreign words.

Wordform	Lemma	Tag
Praha-Hradčany	`Praha_;G`	`NNFS1-----A----`
Praha-Hradčany	`Hradčany_;G`	`NNIP1-----A----`
na propan-butanovém hořáku	`propan`	`NNIS1-----A----`
Karel-Ferdinandova univerzita	`Karel_;Y`	`NNMS1-----A----`
Karlo-Ferdinandova univerzita	`Karlo`	`S2--------A----`
do e-mailu	`e-2_^(e-mail)`	`S2--------A----`
do e-mailu	`mail`	`NNIS2-----A----`
hraju ping-pong	`ping_^(ping-pong)`	`S2--------A----`
hraju ping-pong	`pong_^(ping-pong)`	`SNIS4-----A----`
do Tchaj-peje	`Tchaj`	`S2--------A----`
do Tchaj-peje	`pej-1`	`SNFS2-----A----`
tae-kwon-do	`tae_^(tae-kwon-do)`	`S2--------A----`
tae-kwon-do	`kwon_^(tae-kwon-do)`	`S2--------A----`
tae-kwon-do	`do-2_^(tae-kwon-do)`	`SNNS1-----A----`
cinéma-vérité	`cinéma-77`	`F%-------------`
cinéma-vérité	`vérité-77`	`F%-------------`
v Hanty-Mansijsku	`Hanty-77`	`F%-------------`
v Hanty-Mansijsku	`Mansijsk_;G`	`NNIS6-----A----`

Table 74: Examples: Hyphenated composites

Hyphenated composite of single words is a composite of two or more hyphenated words with a parataxis relation between them. A typical example is the composite of two or more personal or geographical names (e.g. *Anna-Marie*, *Praha-Hradčany* 'Prague-Hradcany', *Clam-Gallasův palác* 'Clam-Gallas palace'), but also composites like *metyl-alkohol* 'methyl-alcohol' or *Hewlett-Packard*. The hyphenated parts are analyzed as if they were separate single words. If the first part (before a hyphen) in an adjectival composite is not an adjective (e.g. *propan* 'propane' in *propan-butanový hořák* 'propane-butane burner', *Karel* 'Charles' in *Karel-Ferdinandova univerzita* 'Charles-Ferdinand University'), it is captures as a noun. See examples in Tab. 74.

Hyphenated compound word consists of two or more parts which are hyphenated to create a new word. The part before the hyphen is (usually) an incomplete word (e.g. *anglicko-česká kniha* 'Czech-English book') and it is captured as a prefixal segment. If the part after the hyphen is a meaningful word, it is analysed as that wordform. If a hyphenated compound word cannot be decomposed into meaningful words (wordforms), i.e. a hyphenated word only has meaning in its entirety, all parts are captured as segments - this is typically a case of loanwords (e.g. *sci-fi*, *ping-pong*, *wi-fi*, *Tchaj-pej* 'Taipei'). There can be more than one segment in a hyphenated compound word (e.g. *tae-kwon-do*). See Tab. 74. See more about capturing segments in Sect. 16.

Foreign-language composites (e.g. *cinéma-vérité*, *flos-cuculi*) are represented as foreign words (see Section 17). See examples in Tab. 74.

20 Typo, Distortion, Misspelling

Intentional misspellings, typos, distortions, as well as frequent errors caused by ignorance of the rules or new codification rules are analyzed as follows.

In the case of a wordform error (e.g. *vidím jí* instead of correct *vidím ji* 'I see her' with a short vowel), the wrong wordform is captured as a special wordform of the corresponding paradigm and it is captured at the 15^{th} `VAR` tag position (by the numbers `5-9`, but especially `9`; Sect. 5.15). If an error, misspelling or distortion is in the whole paradigm (e.g. *vánoce* 'christmas' instead of correct *Vánoce* 'Christmas' with the capital letter), there is style label `i` in the paradigm lemma. There can also be a reference of the `DS` type (see Sect. 4.2.4) to the basic variant. See examples in Tab. 75.

Wordform	Lemma	Tag
vidím jí	`on-1`	`PPFS4--3------6`
odvezl mi sem	`já`	`PH-S4--1------5`
pro Božíkovi hosty	`Božíkův_;Y_^(*2)`	`AUMP4M--------9`
u Lukášové maminky	`Lukášův_;Y_^(*2)`	`AUFS2M--------9`
místo lokomotiva řekl lomokotiva	`lomokotiva_,i_^(^DS**lokomotiva)`	`NNFS1-----A----`
o vánocích	`vánoce_,i_^(^DS**Vánoce)`	`NNFP6-----A----`

Table 75: Examples: Typo, distortion, misspelling

21 Note on Tokenization

Data in the PDT-C corpora are tokenized from delimiter to delimiter. Delimiters are spaces and all non-alphanumeric characters except for the decimal point and decimal comma. Also wordforms containing numbers are tokenized at seams between a string of characters and a string of numbers.

All numbers written with digits are assigned the tag: `C=----------`.

Non-alphanumeric characters are assigned the tag: `Z:----------`.

From this simple rule it follows that even units that we normally understand as one word break down into more tokens. E.g. the single quote (') or hypen (-) splits the string such as *C'tung* or *wi-fi* into three tokens. Because of the number inside the word, the word *12bodový* '12-point' is divided into tokens *12* and *bodový* 'point' or the non-standard name of the car *V3ska* is divided into tokens *V*, *3* and *ska*. No attempt is made to put together these separate tokens within morphological annotation. Each token is subject to morphological annotation, although determining the lemma and morphological tag of these "pieces of words" can be difficult (see also Sect. 16 about segments and Sect. 19 about hyphenated composites).

Cf. examples in Tab. 76.

Wordform	Lemma	Tag
5	5	C=----------
35	35	C=----------
3.5	3.5	C=----------
?	?	Z:----------
)	)	Z:----------
%	%	Z:----------
Mao C'tung	C-1_;Y	NNMXX-----A----
Mao C'tung	tung-1	SNMS1-----A----
12bodový	12	C=----------
12bodový systém	bodový	AAIS1----1A----
V3ska	V-33	Q3----------
V3ska	3	C=----------
V3ska	ska-2	SNFS1-----A----

Table 76: Examples: Tokenization

22 Appendix

22.1 Detailed part of speech (SUBPOS): Quick reference

`#` (POS: Z) Sentence boundary (for the "virtual" word/token ###) This SUBPOS does not appear in the dictionary, but it is used in annotated text for manually or automatically recognized sentence boundaries, if they need to be inserted in the form of a separate token.

`*` (POS: J) Binary mathematical operations as a conjunction (e.g. *plus*; *krát* 'times')

`,` (POS: J) Conjunction subordinate (e.g. *protože* 'because'; *že* 'that'; incl. *aby* 'in order to', *kdyby* 'if' in all forms)

`:` (POS: Z) Punctuation, non-alphanumeric character (e.g. *, %*)

`=` (POS: C) Number written using digits (e.g. *38*, *3.5*)

`^` (POS: J) Conjunction connecting main clauses (*a* 'and'; *ale* 'but')

`%` (POS: F) Foreign word (e.g. *home made*)

`@` (POS: X) Unrecognized word form, unknown (used only by automatic tagger, not in the dictionary and manual annotation)

`}` (POS: C) Numeral, written using Roman numerals (e.g. *XIV*)

`1` (POS: P) Relative possessive pronoun *jehož* and *jehožto* 'whose' including wordforms *jehož*, *jejíž*, *jejížto*, *jejichž* 'whose', etc.

`2` (POS: S) Prefixal segment (e.g. *černo-* 'black-')

`3` (POS: Q) Isolated letter

`4` (POS: P) Relative/interrogative pronoun with agreement GENDER (*jaký* 'what', *který* 'which', *čí* 'whose', *jenž* 'who')

`5` (POS: P) Clitical form of personal pronoun *on* 'he' (only *mu*, *ho* 'him')

`6` (POS: P) Personal reflexive pronoun *se* in its long forms (only *sebe*, *sobě*, *sebou* 'myself' /'yourself' /'herself' /'himself' in various cases)

`7` (POS: P) Personal reflexive pronoun *se* in its short (clitic) forms (only wordforms *se*, *si*, *ses*, *sis* 'myself' /'yourself' /'herself' /'himself')

`8` (POS: P) Personal pronoun reflexive possessive *svůj* 'my' / 'your' / 'her' / 'his', the POSSGENDER and POSSNUMBER positions are not filled

`9` (POS: P) Personal pronoun possessive for the 3rd person *jeho* 'his', including wordforms *její* 'her', *jejich* 'their' etc. with the POSSGENDER and POSSNUMBER positions filled

`A` (POS: A) Adjective, general (e.g. *velký* 'big', *dlouhý* 'long')

`B` (POS: V) Verb, present (e.g. *pracuje* 'he-works') or future form (e.g. *bude* 'will', *pojedu* 'I-willl-go')

`C` (POS: A) Adjective, nominal (short, participial) form (e.g. *rád* 'pleased', *schopen* 'able')

`D` (POS: P) Demonstrative pronoun (*ten* 'this', 'that', *onen* 'that over there')

`E` (POS: P) Personal pronoun *on* 'he' for the 3^{rd} person (including wordforms *ona* 'she', *jim* 'them' etc.), for which the GENDER position is filled

`F` (`POS: R`) Preposition, part of; never appears isolated, always in a phrase (e.g. *nehledě (na)* 'regardless', *vzhledem (k)* 'because of')

`G` (`POS: A`) Adjective derived from present transgressive form of a verb (e.g. *dělající* 'working')

`H` (`POS: P`) Clitical (short) form of personal pronouns *já* 'I' and *ty* 'you', for which `GENDER` position is not filled (e.g. *mě* 'me', *mi* 'me', *ti* 'you')

`I` (`POS: I`) Interjection (e.g. *ach*)

`K` (`POS: P`) Indefinite pronoun for which `GENDER` position is not filled (e.g. *někdo* 'somebody', *něco* 'something', *bůhvíkdo* 'whoever', *cosi* 'something')

`L` (`POS: P`) Delimiting pronoun (e.g. *všechnen* 'all', *sám* 'alone')

`M` (`POS: A`) Adjective derived from verbal past transgressive form (e.g. *udělavší* 'done')

`N` (`POS: N`) Noun (e.g. *dům* 'house', *Jan* 'John')

`O` (`POS: A`) Adjective *svůj* 'own self', *nesvůj* 'not-in-mood' and *tentam* 'gone' with agreement `GENDER`

`P` (`POS: P`) Personal pronoun *já* 'I', *ty* 'you', *my* 'we' and *vy* 'you', for which `GENDER` position is not filled

`Q` (`POS: P`) Relative/interrogative pronoun for which `GENDER` position is not filled (*kdo* 'who', *co* 'what', *cožpak* 'isn't-it-true-that')

`R` (`POS: R`) Preposition (general, without vocalization; e.g. *v* 'in', *pod* 'under')

`S` (`POS: P`) Personal pronoun possessive *můj* 'my', *tvůj* 'your', *náš* 'our' and *váš* 'your' for which the `POSSGENDER` position is not filled

`T` (`POS: T`) Particle (e.g *ano* 'yes')

`U` (`POS: A`) Adjective possessive, with the masculine ending *-ův* (e.g. *otcův* 'father's') as well as *-in* (e.g. *matčin* 'mother's')

`V` (`POS: R`) Preposition with vocalization *-e* (e.g. *ve* 'in', *pode* 'under') or *-u* (e.g. *ku* 'to')

`W` (`POS: P`) Negative pronoun with agreement `GENDER` (e.g. *nijaký* 'no/none', *ničí* 'nobody's', *žádný* 'no/none')

`Y` (`POS: P`) Negative pronoun for which `GENDER` position is not filled (e.g. *nic* 'nothing', *nikdo* 'nobody')

`Z` (`POS: P`) Indefinite pronoun with agreement `GENDER` (e.g. *nějaký* 'some', *některý* 'some', *něčí* 'somebody's', *číkoli* 'anybody's')

`a` (`POS: C`) Cardinal numeral indefinite for which `GENDER` and `NUMBER` position is not filled, incl. interrogative numeral *kolik* 'how much' (e.g. *mnoho* 'much/many', *málo* 'little/few', *tolik* 'that much/many', *několik* 'some (number of)', *kdovíkolik* 'who-knows-how-much/many', *pár* 'some (number of)')

`b` (`POS: D`) Adverb without a possibility to form negation and degrees of comparison (e.g. *pozadu* "behind", *naplocho* "flatly', *včera* 'yesterday'); i.e. positions of the `NEGATION` and `GRADE` are not filled

`c` (`POS: V`) Conditional of the verb *být* 'to be' (e.g. *by*, including *bych*, *bys*, *byste* 'would') that are treated as aggregates

`d` (`POS: C`) Generic numeral definite with agreement `GENDER` (e.g. *jedny* 'one-kind', *dvojí* 'two-kinds', *desaterý* 'ten-kinds', *patery* 'five-kinds')

`e` (`POS: V`) Verb, transgressive present (endings *-e/-ě*, *-íc*, *-íce*; e.g. *dělaje* 'doing'), also archaic present transgressive of perfective verbs (e.g. *udělaje* '(he-)having-done'; `VAR: 4`)

`f` (`POS: V`) Verb, infinitive (e.g. *dělat* 'to do')

`g` (`POS: D`) Adverbs forming negation and degrees of comparison (e.g. *dobře* 'well', *zajímavě* 'interestingly'); positions of the `GRADE` and `NEGATION` are filled

`h` (`POS: C`) Generic numeral indefinite with agreement `GENDER` incl. interrogative numeral *kolikerý* 'how-many-kinds' (e.g. *nejedny* 'not-only-one-kind', *tolikerý* 'that-many-kinds', *několikerý* 'several-kinds', *několikery* 'several-kinds')

`i` (`POS: V`) Verb, imperative form (e.g. *dělej* 'do!')

`j` (`POS: C`) Generic numeral definite used as a syntactic noun, with lexical `GENDER` (e.g. *čtvero* 'four-kinds-of', *desatero* 'ten-kinds-of')

`k` (`POS: C`) Generic numeral indefinite used as a syntactic noun, with lexical `GENDER` incl. interrogative numeral *kolikero* 'how-many-kinds' (e.g. *několikero* 'several-kinds-of', *tolikero* 'that-many-kinds')

`l` (`POS: C`) Cardinal numeral definite for which `GENDER` position is not filled (e.g. *tři* 'three', *čtyři* 'four', *pět* 'five', *půl* 'half')

`m` (`POS: V`) Verb, past transgressive (e.g. *udělav* '(he-)having-done')

`n` (`POS: C`) Cardinal numeral definite with agreement `GENDER` (only *jeden* 'one', *dva* 'two' and *oba* 'both')

`o` (`POS: C`) Multiplicative numeral indefinite incl. interrogative numeral *kolikrát* 'how-many-times' (e.g. *mnohokrát* 'many-times', *tolikrát* 'that-many-times', *několikrát* 'several-times', *nejednou* 'not-only-one-time')

`p` (`POS: V`) Verb, past participle, active (e.g. *pracoval* '(he)-worked')

`q` (`POS: V`) Verb, past participle, active with the archaic enclitic *-ť* (e.g. *pracovalť* '(he)-worked-could-you-imagine-that?')

`r` (`POS: C`) Ordinal numeral definite, adjective declension without degrees of comparison (e.g. *třetí* 'third', *pátý* 'fifth')

`s` (`POS: V`) Verb, past participle, passive (e.g. *udělán* 'done')

`t` (`POS: V`) Verb, present or future tense, with the archaic enclitic *-ť* with meanings (perhaps) "could-you-imagine-that?" or "but-because" (e.g. *dělámeť* '(we)-do-could-you-imagine-that?')

`v` (`POS: C`) Multiplicative numeral definite (e.g. *pětkrát* 'five-times', *sedmkrát* 'seven-times')

`w` (`POS: C`) Ordinal numeral indefinite, adjective declension without degrees of comparison, incl. interrogative numeral *kolikátý* 'at-what-position-in-a-sequence' (e.g. *tolikátý* 'at-that-position-in-a-sequence', *několikátý* 'umpteenth')

`y` (`POS: C`) Cardinal numeral indefinite with agreement `GENDER` (only *nejeden* 'not-only-one')

`z` (`POS: C`) Cardinal numeral definite with lexical `GENDER` (e.g. *sto* 'hundred', *milion* 'million', *nula* 'zero', *čtvrt* 'quarter')

22.2 Categories relevant for POS and SUBPOS combinations

The tables of applicability/non-applicability of the tag categories related to SUBPOS value are presented here. For each value of the SUBPOS category (see the list in Sect. 22.1 above and also Sect. 5.2), the major part-of-speech value (POS) is given to which the current subcategory value uniquely belongs. Only abbreviation (B) and segment (S) POS can potentially be associated with any SUBPOS value. The combinations that have occurred in the current version of the dictionary are also listed here.

The SUBPOS category serves as an indicator of applicability/non-applicability of other tag categories (i.e. the categories GENDER, NUMBER, CASE, etc. up to the last category, VAR). Thus each subsequent row of the table shows the applicable tag categories for the given SUBPOS. The values of such categories that occur in the dictionary are listed in the right-hand column.

If a tag category is used for a given SUBPOS, then the tag position never bears the non-applicable value (-). There are only two exceptions: the AGGREGATE and VAR categories (see Sect. 5.14 and 5.15).These two categories are optional for all SUBPOS values. The AGGREGATE and VAR tag positions can bear the non-applicable value (-).

Sentence boundary (for the "virtual" word/token ###) This SUBPOS does not appear in the dictionary, but it is used in annotated text for manually or automatically recognized sentence boundaries, if they need to be inserted in the form of a separate token.

Category	Values used
POS	Z

: Punctuation, non-alphanumeric character (e.g. *,* *%*)

Category	Values used
POS	Z

@ Unrecognized word form, unknown (used only by automatic tagger, not in the dictionary and manual annotation)

Category	Values used
POS	X

* Binary mathematical operations as a conjunction (e.g. *plus*; *krát* 'times')

Category	Values used
POS	J
VAR	- 1

% Foreign word (e.g. *home made*)

Category	Values used
POS	F

= Number written using digits (e.g. *38*, *3.5*)

Category	Values used
POS	C

} Numeral, written using Roman numerals (e.g. *XIV*)

Category	Values used
POS	C
VAR	- 1 2

, Conjunction subordinate (e.g. *protože* 'because'; *že* 'that'; incl. *aby* 'in order to', *kdyby* 'if' in all forms)

Category	Values used
POS	J
AGGREGATE	- c e m s
VAR	- 6 7 8

^ Conjunction connecting main clauses (*a* 'and'; *ale* 'but')

Category	Values used
POS	J
AGGREGATE	- s
VAR	- 2

1 Relative possessive pronoun *jehož* and *jehožto* 'whose' including wordforms *jehož*, *jejíž*, *jejížto*, *jejichž* 'whose', etc.

Category	Values used
POS	P
GENDER	F I M N X Z
NUMBER	D P S X
CASE	1 2 3 4 6 7 X
POSSGENDER	F X Z
POSSNUMBER	P S
PERSON	3
VAR	- 2

2 Prefixal segment (e.g. *černo-* 'black-')

Category	Values used
POS	S
NEGATION	A N

Category co-occurrence for SUBPOS = 2

3 Isolated letter

Category	Values used
POS	Q

4 Relative/interrogative pronoun with agreement GENDER (*jaký* 'what', *který* 'which', *čí* 'whose', *jenž* 'who')

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
AGGREGATE	- s
VAR	- 1 2 3 4 6 7 8 9

5 Clitical form of personal pronoun *on* 'he' (only *mu*, *ho* 'him')

Category	Values used
POS	P
GENDER	Z
NUMBER	S
CASE	2 3 4
PERSON	3

6 Personal reflexive pronoun *se* in its long forms (only *sebe*, *sobě*, *sebou* 'myself' /'yourself' /'herself' /'himself' in various cases)

Category	Values used
POS	P
CASE	2 3 4 6 7

7 Personal reflexive pronoun *se* in its short (clitic) forms (only wordforms *se*, *si*, *ses*, *sis* 'myself' /'yourself' /'herself' /'himself')

Category	Values used
POS	P
CASE	3 4
AGGREGATE	- s

8 Personal pronoun reflexive possessive *svůj* 'my' / 'your' / 'her' / 'his', the POSSGENDER and POSSNUMBER positions are not filled

Category	Values used
POS	P
GENDER	F H I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 1 6 7

9 Personal pronoun possessive for the 3rd person *jeho* 'his', including wordforms *její* 'her', *jejich* 'their' etc. with the POSSGENDER and POSSNUMBER positions filled

Category	Values used
POS	P
GENDER	F I M N X Z
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
POSSGENDER	F X Z
POSSNUMBER	P S
PERSON	3
VAR	- 6

A Adjective, general (e.g. *velký* 'big', *dlouhý* 'long')

Category	Values used
POS	A
GENDER	F I M N X
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
GRADE	1 2 3
NEGATION	A N
VAR	- 1 3 5 6 7 8 9 a b

B Verb, present (e.g. *pracuje* 'he-works') or future form (e.g. *bude* 'will', *pojedu* 'I-willl-go')

Category	Values used
POS	V
NUMBER	P S
PERSON	1 2 3
TENSE	F P
NEGATION	A N
VOICE	A
ASPECT	B I P
VAR	- 1 2 3 4 5 6 7 8 9

C Adjective, nominal (short, participial) form (e.g. *rád* 'pleased', *schopen* 'able')

Category	Values used
POS	A
GENDER	F M N Q T Y
NUMBER	P S W
CASE	- 4
NEGATION	A N

D Demonstrative pronoun (*ten* 'this', 'that', *onen* 'that over there')

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	D P S X
CASE	1 2 3 4 6 7 X
AGGREGATE	- s
VAR	- 1 2 4 5 6 7 b

E Personal pronoun *on* 'he' for the 3^{rd} person (including wordforms *ona* 'she', *jim* 'them' etc.), for which the GENDER position is filled

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	P S
CASE	1 2 3 4 6 7
PERSON	3
AGGREGATE	- d n o p z
VAR	- 1 2 6 7

F Preposition, part of; never appears isolated, always in a phrase (e.g. *nehledě (na)* 'regardless', *vzhledem (k)* 'because of')

Category	Values used
POS	R

G Adjective derived from present transgressive form of a verb (e.g. *dělající* 'working')

Category	Values used
POS	A
GENDER	F I M N
NUMBER	D P S
CASE	1 2 3 4 5 6 7
NEGATION	A N
VAR	- 6

H Clitical (short) form of personal pronouns *já* 'I' and *ty* 'you', for which GENDER position is not filled (e.g. *mě* 'me', *mi* 'me', *ti* 'you')

Category	Values used
POS	P
NUMBER	S
CASE	2 3 4
PERSON	1 2
VAR	- 5 6

I Interjection (e.g. *ach*)

Category	Values used
POS	I
VAR	- 1 6

K Indefinite pronoun for which GENDER position is not filled (e.g. *někdo* 'somebody', *něco* 'something', *bůhvíkdo* 'whoever', *cosi* 'something')

Category	Values used
POS	P
CASE	1 2 3 4 5 6 7
VAR	- 6

L Delimiting pronoun (e.g. *všechnen* 'all', *sám* 'alone')

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 1 3 4 5 6 7 8

M Adjective derived from verbal past transgressive form (e.g. *udělavší* 'done')

Category	Values used
POS	A
GENDER	F I M N
NUMBER	D P S
CASE	1 2 3 4 5 6 7
NEGATION	A N
VAR	- 1 6 7

N Noun (e.g. *dům* 'house', *Jan* 'John')

Category	Values used
POS	N
GENDER	F I M N X
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
NEGATION	- A N
VAR	- 1 2 3 4 5 6 7 8 9 a b c

O Adjective *svůj* 'own self', *nesvůj* 'not-in-mood' and *tentam* 'gone' with agreement GENDER

Category	Values used
POS	A
GENDER	F I M N Y
NUMBER	P S
VAR	- 1 6

P Personal pronoun *já* 'I', *ty* 'you', *my* 'we' and *vy* 'you', for which GENDER position is not filled

Category	Values used
POS	P
NUMBER	P S
CASE	1 2 3 4 5 6 7
PERSON	1 2
AGGREGATE	- s
VAR	- 6 9

Q Relative/interrogative pronoun for which GENDER position is not filled (*kdo* 'who', *co* 'what', *cožpak* 'isn't-it-true-that')

Category	Values used
POS	P
CASE	1 2 3 4 6 7
AGGREGATE	- N O V Z n o s v z
VAR	- 1 6 9

R Preposition (general, without vocalization; e.g. *v* 'in', *pod* 'under')

Category	Values used
POS	R
CASE	1 2 3 4 6 7 X
VAR	- 6 7 a b c

S Personal pronoun possessive *můj* 'my', *tvůj* 'your', *náš* 'our' and *váš* 'your' for which the POSSGENDER position is not filled

Category	Values used
POS	P
GENDER	F H I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
POSSNUMBER	P S
PERSON	1 2
VAR	- 1 6 7 9 b

T Particle (e.g *ano* 'yes')

Category	Values used
POS	T
AGGREGATE	- s
VAR	- 1 2 3 6 7 a b

`U` Adjective possessive, with the masculine ending *-ův* (e.g. *otcův* 'father's') as well as *-in* (e.g. *matčin* 'mother's')

Category	Values used
POS	A
GENDER	F I M N X
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
POSSGENDER	F M
VAR	- 1 2 5 6 7 8 9 a b

`V` Preposition with vocalization *-e* (e.g. *ve* 'in', *pode* 'under') or *-u* (e.g. *ku* 'to')

Category	Values used
POS	R
CASE	2 3 4 6 7
VAR	- 1

`W` Negative pronoun with agreement `GENDER` (e.g. *nijaký* 'no/none', *ničí* 'nobody's', *žádný* 'no/none')

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 6 7

`Y` Negative pronoun for which `GENDER` position is not filled (e.g. *nic* 'nothing', *nikdo* 'nobody')

Category	Values used
POS	P
CASE	1 2 3 4 6 7
VAR	- 2 6

`Z` Indefinite pronoun with agreement `GENDER` (e.g. *nějaký* 'some', *některý* 'some', *něčí* 'somebody's', *číkoli* 'anybody's')

Category	Values used
POS	P
GENDER	F I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 1 5 6 7

`a` Cardinal numeral indefinite for which `GENDER` and `NUMBER` position is not filled, incl. interrogative numeral *kolik* 'how much' (e.g. *mnoho* 'much/many', *málo* 'little/few', *tolik* 'that much/many', *několik* 'some (number of)', *kdovíkolik* 'who-knows-how-much/many', *pár* 'some (number of)')

Category	Values used
POS	C
CASE	1 2 3 4 5 6 7 X
AGGREGATE	- s
VAR	- 1

`b` Adverb without a possibility to form negation and degrees of comparison (e.g. *pozadu* "behind", *naplocho* "flatly', *včera* 'yesterday'); i.e. positions of the `NEGATION` and `GRADE` are not filled

Category	Values used
POS	D
AGGREGATE	- s
VAR	- 1 2 3 4 6 7 8 a b

`c` Conditional of the verb *být* 'to be' (e.g. *by*, including *bych*, *bys*, *byste* 'would') that are treated as aggregates

Category	Values used
POS	V
ASPECT	I
AGGREGATE	- c e m s
VAR	- 6

`d` Generic numeral definite with agreement `GENDER` (e.g. *jedny* 'one-kind', *dvojí* 'two-kinds', *desaterý* 'ten-kinds', *patery* 'five-kinds')

Category	Values used
POS	C
GENDER	F I M N X Y
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 1 2 6 7

`e` Verb, transgressive present (endings *-e/-ě*, *-íc*, *-íce*; e.g. *dělaje* 'doing'), also archaic present transgressive of perfective verbs (e.g. *udělaje* '(he-)having-done'; `VAR`: `4`)

Category	Values used
POS	V
GENDER	H X Y
NUMBER	P S
NEGATION	A N
ASPECT	B I P
VAR	- 1 2 4 6

`f` Verb, infinitive (e.g. *dělat* 'to do')

Category	Values used
POS	V
NEGATION	A N
ASPECT	B I P
VAR	- 1 2 3 4 6 7 b

`g` Adverbs forming negation and degrees of comparison (e.g. *dobře* 'well', *zajímavě* 'interestingly'); positions of the `GRADE` and `NEGATION` are filled

Category	Values used
POS	D
GRADE	1 2 3
NEGATION	A N
VAR	- 1 2 3 4 6 7 b

`h` Generic numeral indefinite with agreement `GENDER` incl. interrogative numeral *kolikerý* 'how-many-kinds' (e.g. *nejedny* 'not-only-one-kind', *tolikerý* 'that-many-kinds', *několikerý* 'several-kinds', *několikery* 'several-kinds')

Category	Values used
POS	C
GENDER	F I M N X Y
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 1 6 7

`i` Verb, imperative form (e.g. *dělej* 'do!')

Category	Values used
POS	V
NUMBER	P S
PERSON	1 2 3
NEGATION	A N
ASPECT	B I P
VAR	- 1 2 3 4 5 6 7 8 9 b

`j` Generic numeral definite used as a syntactic noun, with lexical `GENDER` (e.g. *čtvero* 'four-kinds-of', *desatero* 'ten-kinds-of')

Category	Values used
POS	C
GENDER	N
NUMBER	P S X
CASE	1 2 3 4 5 6 7 X
VAR	- 1

`k` Generic numeral indefinite used as a syntactic noun, with lexical `GENDER` incl. interrogative numeral *kolikero* 'how-many-kinds' (e.g. *několikero* 'several-kinds-of', *tolikero* 'that-many-kinds')

Category	Values used
POS	C
GENDER	N
NUMBER	P S X
CASE	1 2 3 4 5 6 7 X

`l` Cardinal numeral definite for which `GENDER` position is not filled (e.g. *tři* 'three', *čtyři* 'four', *pět* 'five', *půl* 'half')

Category	Values used
POS	C
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
VAR	- 1 2 6

`m` Verb, past transgressive (e.g. *udělav* '(he-)having-done')

Category	Values used
POS	V
GENDER	H X Y
NUMBER	P S
NEGATION	A N
ASPECT	B I P
VAR	- 1 2 6

`n` Cardinal numeral definite with agreement `GENDER` (only *jeden* 'one', *dva* 'two' and *oba* 'both')

Category	Values used
POS	C
GENDER	F H I M N X Y Z
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
VAR	- 1 6 8

`o` Multiplicative numeral indefinite incl. interrogative numeral *kolikrát* 'how-many-times' (e.g. *mnohokrát* 'many-times', *tolikrát* 'that-many-times', *několikrát* 'several-times', *nejednou* 'not-only-one-time')

Category	Values used
POS	C
VAR	- 1

`p` Verb, past participle, active (e.g. *pracoval* '(he)-worked')

Category	Values used
POS	V
GENDER	F M N Q T Y
NUMBER	P S W
TENSE	R
NEGATION	A N
VOICE	A
ASPECT	B I P
AGGREGATE	- s
VAR	- 1 2 3 6 7 8 9

`q` Verb, past participle, active with the archaic enclitic *-ť* (e.g. *pracovalť* '(he)-worked-could-you-imagine-that?')

Category	Values used
POS	V
GENDER	M N Q T Y
NUMBER	P S W
TENSE	R
NEGATION	A N
VOICE	A
ASPECT	B I P
VAR	2 3 4 5 6

`r` Ordinal numeral definite, adjective declension without degrees of comparison (e.g. *třetí* 'third', *pátý* 'fifth')

Category	Values used
POS	C
GENDER	F I M N
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 6 7

`s` Verb, past participle, passive (e.g. *udělán* 'done')

Category	Values used
POS	V
GENDER	F M N Q T Y
NUMBER	P S W
CASE	- 4
TENSE	H X
NEGATION	A N
VOICE	P
ASPECT	B I P
AGGREGATE	- s
VAR	- 1 2 5 6 7 8

`t` Verb, present or future tense, with the archaic enclitic *-ť* with meanings (perhaps) "could-you-imagine-that?" or "but-because" (e.g. *dělámeť* '(we)-do-could-you-imagine-that?')

Category	Values used
POS	V
NUMBER	P S
PERSON	1 2 3
TENSE	F P
NEGATION	A N
VOICE	A
ASPECT	B I P
VAR	1 2 3 4 5 6 7 8 9

`v` Multiplicative numeral definite (e.g. *pětkrát* 'five-times', *sedmkrát* 'seven-times')

Category	Values used
POS	C
VAR	- 1 7

`w` Ordinal numeral indefinite, adjective declension without degrees of comparison, incl. interrogative numeral *kolikátý* 'at-what-position-in-a-sequence' (e.g. *tolikátý* 'at-that-position-in-a-sequence', *několikátý* 'umpteenth')

Category	Values used
POS	C
GENDER	F I M N X Y Z
NUMBER	D P S
CASE	1 2 3 4 5 6 7
VAR	- 6 7

`y` Cardinal numeral indefinite with agreement `GENDER` (only *nejeden* 'not-only-one')

Category	Values used
POS	C
GENDER	F I M N Y Z
NUMBER	S
CASE	1 2 3 4 5 6 7

`z` Cardinal numeral definite with lexical `GENDER` (e.g. *sto* 'hundred', *milion* 'million', *nula* 'zero', *čtvrt* 'quarter')

Category	Values used
POS	C
GENDER	F I N
NUMBER	P S X
CASE	1 2 3 4 5 6 7 X
VAR	- 1 2 6 b

SUBPOS co-occurrence tables for POS B (Abbreviations)

`A` Abbreviation of adjective (`AA`)

Category	Values used
POS	B
GENDER	X
NUMBER	X
CASE	X
GRADE	1
NEGATION	A

`N` Abbreviation of noun (`NN`)

Category	Values used
POS	B
GENDER	F I M N X
NUMBER	X
CASE	X
NEGATION	A

`b` Abbreviation of adverb (`Db`)

Category	Values used
POS	B

`^` Abbreviation of conjunction (`J^`)

Category	Values used
POS	B

SUBPOS co-occurrence tables for POS S (Segments)

A Postfixal segment of adjective (AA)

Category	Values used
POS	S
GENDER	F I M N X
NUMBER	D P S X
CASE	1 2 3 4 5 6 7 X
GRADE	1 2
NEGATION	A
VAR	- 6 7

N Postfixal segment of noun (NN)

Category	Values used
POS	S
GENDER	F I M N X
NUMBER	P S X
CASE	1 2 3 4 5 6 7 X
NEGATION	A
VAR	- 1 6

b Postfixal segment of adverb (Db)

Category	Values used
POS	S

l Postfixal segment of numeral (Cl)

Category	Values used
POS	S
NUMBER	X
CASE	X

2 Prefixal segment

Category	Values used
POS	S
NEGATION	A N

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová, *Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description*

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman, *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies*

ÚFAL TR-2018-61 Lukáš Kyjánek, *Morphological Resources of Derivational Word-Formation Relations*

ÚFAL TR-2019-62 Zdeňka Urešová, Eva Fučíková, Eva Hajičová, *CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting (CzEngClass: Kontextová synonymie a valence sloves v bilingvním prostředí)*

ÚFAL TR-2019-63 Ján Faryad,
Identifikace derivačních vztahů ve španělštině

ÚFAL TR-2020-64 Marie Mikulová, Jan Hajič, Jiří Hana, Hana Hanová, Jaroslava Hlaváčová, Emil Jeřábek, Barbora Štěpánková, Barbora Vidová Hladká, Daniel Zeman,
Manual for Morphological Annotation. Revision for Prague Dependency Treebank – Consolidated 2020 release