Manual for Morphological Annotation

Revision for the Prague Dependency Treebank 2.0

ÚFAL Technical Report No. 2005-27

Jiří Hana and Daniel Zeman

May 19, 2005

Preface to Version 2.0

Although the title of this report inherits the word ”Manual” from the previous version, it is no more intended to guide the annotators. Rather it attempts to describe the current state of the morphological annotation in PDT 2.0. Most of the added information resulted from several semi-automatic checks performed on the data before having released it. In some cases it was not manageable to bring the data to the desired state - if so, both the desired and the current state of the data are described.

PDT 2.0 contains 1,960,657 morphologically annotated tokens in 126,831 sentences. There are 168,454 distinct word forms, 71716 distinct lemmas, and 1740 morphological tags.

The final checking and analysis of the data as well as the work on this manual revision were supported by the Czech Academy of Sciences program called ”Information Society”, project No. 1ET101120503.

Preface to Version 1.0

We are pleased to publish the first version of the manual for morphological annotation of Czech sentences. We believe that such guidelines can be of use to the users of Prague Dependency Treebank 1.0 (PDT 1.0), as well as for preparation of new data.

Let us recall the most important steps we passed in order to get about two million morphologically annotated words (PDT 1.0). At the very beginning, we put together a team of eight annotators - we did introduce them to a system of morphological tags we designed to describe Czech morphological properties; we also used (as a preprocessing step) a morphological analyzer for processing isolated words, and, last but not least, we did rely on their knowledge of Czech morphology they have acquired while studying at secondary school, i.e. we did not offer them any annotation guidelines.

One can assume that this strategy is too hazardous - how to deal with discrepancies the annotators produce to ensure the consistency of annotation? First, two annotators annotated each text file. Then, by a ″blind″ automatic procedure (no matter what word is processed - just comparing two strings) we detected words annotated differently. Consequently, the only one annotator (as a member of just two-member team) handled these cases and, also, checked the morphological annotations against the syntactic-analytic annotations. This way we replaced the absence of annotation guidelines by sequential elimination of discrepancies across both the morphological and syntactic-analytic levels of annotation.

Along the way we were writing this annotation manual. It is not intended as a comprehensive guide to the morphological annotation of Czech sentences (in contrast to the manual for syntactic-analytic annotations). The authors concentrate ″only″ on those cases which caused the most ambiguities and problems while annotating PDT 1.0. The ongoing effort is directed to the treating of not- yet-solved problematic cases in accord with the conventions of the automatic morphological analyzer.

The morphological annotation of PDT 1.0 was carried out in the framework of experimental verification of the definition of formal representation of the analysis of Czech sentences (the project GAČR 405/96/0198, ″Formal representation of language structures″). The material obtained in this way (data) is used in many domains of research in computational linguistics, above all as basic (training) data in projects of the automatic language analysis, the MŠMT research project MSM113000006, the ″Laboratory for Language Data Processing″ (the MŠMT project VS961510) and the Center for Computational Linguistics (the MŠMT project LN00A063). These data have been also used as verification material for various partial projects within the complex program GAČR 405/96/K214 (″Czech Language in Computer Age″). The ″Center for Computational Linguistics″ project financially supported work on these morphological annotation guidelines.

Chapter 1

Introduction

We do not want to substitute a grammarbook of Czech. So we are not going to systematically define word classes and paradigms. All the annotators should understand the fundamentals of Czech morphology, as most native Czech speakers do (the stuff is being taught in elementary schools). What we are going to describe are the difficult or unusual phenomena. Most notably we will address the annotation of proper names, foreign words, and abbreviations. Such categories are rarely and sparsely covered by standard dictionaries. To get an idea what a foreign word, proper name etc. mean it is useful to try to find it using an internet portal, an encyclopedia etc. During annotation, we found the following internet links useful:

Portals.

- http://www.seznam.cz/[1] - for Czech products and companies
- <http://search.seznam.cz/search.cgi?mod=f&hlp=y> - for Czech companies
- http://www.google.com/[2]
- http://www.altavista.com/[3] (shop section for various searching products)

Encyclopedias.

- <http://cs.wikipedia.org/> and <http://en.wikipedia.org/>
- http://www.encyclopedia.com/[4]
- http://www.encarta.msn.com/[5]

Dictionaries.

- http://slovnik.seznam.cz/[6] - various dictionaries

Maps.

- http://mapy.atlas.cz/[7] - Czechia
- http://www.mapquest.com/maps/[8] - U.S.A and the world

[1] <http://www.seznam.cz>
[2] <http://www.google.com>
[3] <http://www.altavista.com>
[4] <http://www.encyclopedia.com>
[5] <http://www.encarta.msn.com>
[6] <http://slovnik.seznam.cz>
[7] <http://mapy.atlas.cz>
[8] <http://www.mapquest.com/maps>

Chapter 2

Lemma and tag structure

2.1 Lemma structure

Lemma in PDT 1.0 has two parts. First part, the lemma proper, has to be a unique identifier of the lexical item. Usually it is the base form (e.g. infinitive for a verb) of the word, possibly followed by a number distinguishing different lemmas with the same base forms. Second part (optional) is not part of the identifier and contains additional information about the lemma, e.g. semantic or derivational information.

The formal description of the lemma structure follows. Spaces were inserted between nonterminals to improve readability. Note however that no lemma contains any spaces. Capitalized multi-character symbols are nonterminals. All other symbols are terminals.

```
Lemma       ::= LemmaProper | LemmaProper AddInfo
LemmaProper ::= Word | Word - Number | Number | SpecialChar
Word        ::= Letter | Letter Word
Letter      ::= A | a | Á | á | Ä | ä | ... | Z | z | Ž | ž | ’
Number      ::= NonZero | NonZero Number0
Number0     ::= Digit | Digit Number0
NonZero     ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
Digit       ::= 0 | NonZero
SpecialChar ::= ! | " | # | $ | % | & | ’ | ( | ) | * | + | , |
                - | . | / | : | ; | < | = | > | ? | @ | [ | \ |
                ] | ˆ | _ | ‘ | { | | | } | ˜ | § | ◦
AddInfo     ::= Reference Category Term Style Comment
Reference   ::= <empty> | ‘ LemmaProper
Category    ::= <empty> | _: Category1 | _: Category1 Category
Term        ::= <empty> | _; Term1     | _; Term1 Term
Style       ::= <empty> | _, Style1    | _, Style1 Style
Comment     ::= <empty> | _ˆ Comment1
Category1   ::= N | J | A | Z | M | V | T | W | D | P | C | I | F | B | Q | X
Term1       ::= Y | S | E | G | K | R | m |
                H | U | L | j | g | c | y | b | u | w | p | z | o
Style1      ::= t | n | a | s | h | e | l | v | x
Comment1    ::= ( Explanation ) | ( Derivation ) |
                ( Explanation )_( Derivation )
Explanation ::= CommentChar | CommentChar Explanation
Derivation  ::= * Number Word | * Word
CommentChar ::= Letter | Digit |
                ! | " | # | $ | % | & | ’ | * | + | , | - | . |
                / | : | ; | < | = | > | ? | @ | [ | \ | ] | ˆ |
                _ | ‘ | { | | | } | ˜ | § | ◦
```

Notes on characters:

1. Any character that is letter in the Unicode standard[1] can appear in place of the Letter nonterminal. In the non-ASCII area this most frequently applies to the Czech accented characters: Á á Č č Ď ď É é Ě ě Í í Ň ň Ó ó Ř ř Š š Ť ť Ú ú Ů ů Ý ý Ž ž. However, other characters occur in names (e.g. German Ä ä Ö ö Ü ü, Serbo-Croatian Ć ć) and in foreign words (e.g. Slovak Ľ Ĭ Ĺ Í Ô ô Ŕ ŕ).

2. Standard HTML entities (such as `&amp;` for & or `&agrave;` 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 shall be encoded in UTF8, so few entities will be needed.

3. The single quote (′) is considered a Letter in some transcriptions of non-Latin alphabets (e.g. in Chinese *Mao C′-tung*, Hebrew *Be′er Sheva′*). If it marks deleted parts of words (e.g. English *don′t*, French *d′Artagnan*), it is considered a SpecialChar and it splits the string into three tokens (`d ′ Artagnan`). Even in these languages there are exceptions (e.g. the surname *Preud′homme* is one token).

Table 2.1: Lemma examples

Whole lemma	LemmaProper	AddInfo
`Chemik`	chemik	
`maso_^(jídlo_apod.)`	maso	_^(jídlo_apod.)
`Bonn_;G`	Bonn	_;G
`vazba-1_^(obviněného)`	vazba-1	_^(obviněného)
`vazba-2_^(spojení)`	vazba-2	_^(spojení)
`Martinův-1_;Y_^(*4-1)`	Martinův-1	_;Y_^(*4-1)

2.1.1 Base form and number

The Word in LemmaProper is the base form of the respective paradigm. This means nominative singular for nouns, the same plus masculine positive for adjectives, similarly for pronouns and numerals. Verbs are represented by their infinitive forms.

The Number in LemmaProper helps to distinguish several senses of a homonymous base form. It should neither be zero nor start with zero. The used numbers need not form a continuous sequence. Sometimes a particular number is repeatedly used for a special kind of word (e.g. the lemmas numbered ”-99” are almost invariantly authors′ signatures and their Category/Style part is ”_:B_;S”). Conventions of this kind exist solely for the convenience of a human reader but they are not meant to signal anything to a processing program. No conclusions should be ever drawn from the value of the lemma number! There is no warranty that an observed number ”semantics” holds anywhere else. Other sources of information, such as the AddInfo text, should be used instead.

The following rules shall hold for each group of lemmas sharing the same base form.

- **Rule 1:** If lemmas use numbers to distinguish lexical items with the same base form, they all have to use them - i.e. if there is the lemma X-2, the unnumbered lemma X should not exist. If more than one lemma share a base form, all of them must be numbered.
- **Rule 2:** If a lemma is numbered, its AddInfo should not be empty. The AddInfo must help to distinguish the lemma from other lemmas with the same base form but different numbers. Exception: if all but one lemmas with the same base form are foreign words, the domestic one need not have a non-empty AddInfo. All the foreign counterparts must have it, though.
- **Rule 3:** Two lemmas with different AddInfo must differ in numbers as well. Exceptions (see below): abbreviations (two lemmas differ in the presence of `_:B` but not in their numbers).
- **Rule 4:** Two lemmas with different number must differ in AddInfo as well.

Unfortunately many lemmas are not covered by our automatic morphological analyzer. Such lemmas were created by the annotators, and the administrator of the lexicon should later make their numbers and/or suffixes consistent and conformant to the above rules. In many cases it was not manageable to complete this task for PDT 2.0.

Base form in lemma is case-sensitive. Of course, words that have to be always capitalized in writing, have their lemma capitalized as well. As a consequence, *špaček* (starling) and *Špaček_;S* need not be

[1] <http://www.unicode.org/>

distinguished by numbers (or they can both use the same number). However, although not required, the unique numbering of such cases is recommended.

Sometimes the numbering of lemmas reflect that their base form is homonymous with another word, although the other meaning is not base form. For instance, *žena* is a noun (meaning woman) but it can also be transgressive form of the verb *hnát*. The morphological analyzer may assign different numbers to both meanings of *žena*, although the latter is not a base form. As a consequence, there may be lemma žena-2 even if there is no other lemma with the same base form. Such behavior is allowed but not required.

2.1.2 Reference

Some lemmas refer to other lemmas. A lemma can point at most to one other lemma. The reference is one of the means of explaining the meaning of the source lemma. Such mechanism is systematically used with spelled-out numbers (jeden'1, oba'2) and with abbreviations for various units (kWh'kilowatthodina). Occasionally a reference can occur elsewhere as well.

2.1.3 Category

Lemma category is indicated by ”_:” followed by a letter. Most categories correspond to parts of speech. They are rarely used because the part of speech is encoded in morphological tags as well (see below; note however that some parts of speech are encoded by different characters in the lemma than in the morphological tag). They should be used if the same lemma behaves as two or more parts of speech. No lemma is allowed to appear with morphological tags for two or more different parts of speech. For instance, *vedle* can be either adverb or preposition. There should be two lemmas, `vedle-1_:D`, and `vedle-2_:P`. Note however that in PDT 2.0 some lemmas, especially foreign words, occasionally appear with tags for different parts of speech, and if there are separate lemmas for each part of speech, it is often described verbally in the Comment part rather than formally using the Category field. In our example it would be `vedle-1_^(je_z_toho_vedle)`, and `vedle-2_^(vedle_něčeho)`. This will be corrected in future versions.

Three categories are used on a more systematical basis: _:T and _:W for verbal aspect, and _:B for abbreviations. Aspect has currently no representation in the morphological tags. It is treated as a lexical property - although there are some morphological implications, lots of irregularities could be expected if it was part of the verbal paradigm. The morphological analyzer covers aspect for some verbs while lacking the information for many others. If available, the aspect is indicated in the lemma. Note that there are biaspectual verbs, so `analyzovat_:T_:W` would be correct.

Abbreviations are exceptions to the Rule 3 (saying that different AddInfo implies different lemma numbers). There can be two lemmas with the same base form and number, if the only difference in their AddInfos is that one contains ”_:B” and the other does not. For more information on abbreviations see Chapter 4, “Abbreviations”.

Table 2.2: Lemma categories

Category	Explanation
N	noun
A, J	adjective
Z	pronoun
M	numeral
V	verb
T	imperfect verb
W	perfect verb
D	adverb
P	preposition
C	conjunction
I	particle
F	interjection
B	abbreviation
Q	???
X	do not use

2.1.4 Term

Lemmas of terms have categories of their own. The term type is indicated by ”_;” followed by a letter. More than one term type may apply to one lemma. Two groups of term types can be distinguished: the named entities and the scientific/professional terms. The former are mandatory, proper names must be categorized. The latter are optional, it is up to the lexicon administrator whether they decide that a term is so specialized that its branch shall be indicated.

Table 2.3: Term types

Type	Explanation, examples
Y	given name (formerly used as default): *Petr*, *John*
S	surname, family name: *Dvořák*, *Zelený*, *Agassi*, *Bush*
E	member of a particular nation, inhabitant of a particular territory: *Čech*, *Kolumbijec*, *Newyorčan*
G	geographical name: *Praha*, *Tatry* (the mountains)
K	company, organization, institution: *Tatra* (the company)
R	product: *Tatra* (the car)
m	other proper name: names of mines, stadiums, guerilla bases, etc.
H	chemistry
U	medicine
L	natural sciences
j	justice
g	technology in general
c	computers and electronics
y	hobby, leisure, travelling
b	economy, finances
u	culture, education, arts, other sciences
w	sports
p	politics, governement, military
z	ecology, environment
o	color indication

2.1.5 Style

Lemmas can be stylistically classified. The style flag is indicated by ”_,” followed by a letter. Standard lemmas have no stylistic flag but any lemma intended for special usage (bookish, colloquial language etc.) should be marked as such. It is necessary to distinguish between the style of the lemma and the style of the word form! For instance, *acht* is an archaic word meaning ”anathema”; its less archaic counterpart would be *klatba*. Its lemma should bear the archaic flag: `acht_,a`. On the other hand, *lvové* is just an archaic form of a non-archaic lemma *lev* (lion). In this case the archaicity should only be marked in the morphological tag describing the form (the tag would end in 3; see below for tag descriptions).

Table 2.4: Style flags

Style	Explanation
t	foreign word - see Chapter 6, “Foreign words and phrases”
n	dialect
a	archaic
s	bookish
h	colloquial
e	expressive
l	slang, argot
v	vulgar
x	outdated spelling or misspelling

2.1.6 Explanational comment

Any string in parentheses can be used as explanation of the lemma meaning. The string cannot contain spaces or parentheses. The underscore character is used to replace space, square brackets are used instead of parentheses. The meaning is described in Czech. Example of usage, synonym etc. can also be used or both a verbal description and an example can be mixed. Hint for English speakers: the word "example" can be abbreviated as *př.* or *např.* in the descriptions.

2.1.7 Comment on derivation

The morphological analyzer handles only inflection, not derivations - it means lemmas are rather shallow. However, sometimes the lemma contains information about lemmas it is derived from. For example lemmas of possessive adjectives contain information about the noun they are derived from (otcův ← otec). The information is encoded in the following way - how many characters you have to remove from the end, and what string you have to add to get the deeper lemma. Only the proper lemmas are both input and output of this process (but including the lemma number, if present).

Example 2.1.1: Following examples illustrate this:

- `kardinálův_^(*2)` - remove two letters: kardinál
- `Karlův_;Y_^(*3el)` - remove 3 characters, add "el": Karel
- `přijetí-2_^(např._návrh)_(*5mout-2)` - remove 5 characters, add "mout-2": přijmout-2
- `Martinův-1_;Y_^(*4-1)` - remove 4 characters, add "-1": Martin-1

Example 2.1.2: Other examples:

- `Sorosův_;S_^(*2)`
- `chlapcův_^(*3ec)`
- `Máchův_;S_^(*2a)`
- `Hlinkův-1_;S_^(*4a-1)`
- `podání_^(něco_[někomu]_[někam])_(*3at)`
- `prohlášení_^(*4sit)`
- `protiprávnost_^(*3ý)`

Note: Derivational comments of the form `barvicí_^(^IC**barvit)` occur occasionally in the current data. Cf. with `barvící_^(*3it)`.

2.2 Tag Structure

Lemma and tag together should uniquely identify the word form. Two different word forms should always differ either in lemmas or in morphological tags.

2.2.1 Positional tags

A positional tag is a string of 15 characters. Every positions encodes one morphological category using one character (mostly upper case letters or numbers).

Position	Name	Description
1	POS	Part of speech
2	SubPOS	Detailed part of speech
3	Gender	Gender
4	Number	Number
5	Case	Case

Position	Name	Description
6	PossGender	Possessor's gender
7	PossNumber	Possessor's number
8	Person	Person
9	Tense	Tense
10	Grade	Degree of comparison
11	Negation	Negation
12	Voice	Voice
13	Reserve1	Reserve
14	Reserve2	Reserve
15	Var	Variant, style

Some of the characters encode aggregation of more atomic values - for example: 'X' - means any value, Y means masculine animate (M) or inanimate (I). Dash ('-') means "not applicable" (e.g. tense for nouns).

Not all combinations of tag values are possible. There is about 4K tags.

- hraniční: `AAIS4----1A----` standard adjective, masc. inanimate, singular, accusative, positive
- potok: `NNIS4-----A----` noun, masc. inanimate, singular, accusative, positive
- karikaturistou: `NNMS7-----A----` noun, masc. animate, singular, instrumental, positive
- ODS: `NNFXX-----A---8` noun, feminine, any number, any case, positive, abbreviation
- podle: `RR--2----------` preposition (non vocalized), requiring genitive
- volen: `VsYS---XX-AP---` verb, passive participle, masculine, singular, any person, any tense, positive, passive

See also: <http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/docc0pos.pdf>

Or for quick reference:
<http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html>

1 - Part of speech

In fact, part of speech is rather lexical-syntactic than morphological property. It is practical to keep it in the tags but it would be more accurate to keep it in the lemmas. Anyway, no lemma is allowed to occur with two different parts of speech in the accompanying tags. If a word behaves syntactically as various parts of speech, several lemmas have to be reserved for it.

Value	Description
A	Adjective
C	Numeral
D	Adverb
I	Interjection
J	Conjunction
N	Noun
P	Pronoun
V	Verb
R	Preposition
T	Particle
X	Unknown, Not Determined, Unclassifiable
Z	Punctuation (also used for the Sentence Boundary token)

2 - Detailed part of speech

Further subcategorizes POS. The POS value is uniquely specified by SubPOS value.

Table 2.5: SUBPOS

Value	Description	POS
#	Sentence boundary	Z - punctuation
%	Author's signature, e.g. `haš-99_:B_;S`	N - noun
*	Word krát (lit.: times)	C - numeral
,	Conjunction subordinate (incl. aby, kdyby in all forms)	J - conjuction
}	Numeral, written using Roman numerals (XIV)	C - numeral
:	Punctuation (except for the virtual sentence boundary word ###, which uses the Section 2.2.1 #)	Z - punctuation
=	Number written using digits	C - numeral
?	Numeral kolik (lit. how many/how much)	C - numeral
@	Unrecognized word form	X - unknown
ˆ	Conjunction (connecting main clauses, not subordinate)	J - conjunction
4	Relative/interrogative pronoun with adjectival declension of both types (soft and hard) (jaký, který, čí, ..., lit. what, which, whose, ...)	P - pronoun
5	The pronoun he in forms requested after any preposition (with prefix n-: něj, něho, ..., lit. him in various cases)	P - pronoun

Table 2.5: *(continued)*

Value	Description	POS
6	Reflexive pronoun se in long forms (sebe, sobě, sebou, lit. myself / yourself / herself / himself in various cases; se is personless)	P - pronoun
7	Reflexive pronouns se (Section 2.2.1 = 4), si (Section 2.2.1 = 3), plus the same two forms with contracted -s: ses, sis (distinguished by Section 2.2.1 = 2; also number is singular only) This should be done somehow more consistently, virtually any word can have this contracted -s (cos, polívkus, ...)	P - pronoun
8	Possessive reflexive pronoun svůj (lit. my/your/her/his when the possessor is the subject of the sentence)	P - pronoun
9	Relative pronoun jenž, již, ... after a preposition (n-: něhož, niž, ..., lit. who)	P - pronoun
A	Adjective, general	A - adjective
B	Verb, present or future form	V - verb
C	Adjective, nominal (short, participial) form rád, schopen, ...	A - adjective
D	Pronoun, demonstrative (ten, onen, ..., lit. this, that, that ... over there, ...)	P - pronoun
E	Relative pronoun což (corresponding to English which in subordinate clauses referring to a part of the preceding text)	P - pronoun
F	Preposition, part of; never appears isolated, always in a phrase (nehledě (na), vzhledem (k), ..., lit. regardless, because of)	R - preposition
G	Adjective derived from present transgressive form of a verb	A - adjective
H	Personal pronoun, clitical (short) form (mě, mi, ti, mu, ...); these forms are used in the second position in a clause (lit. me, you, her, him), even though some of them (mě) might be regularly used anywhere as well	P - pronoun
I	Interjections	I - interjection
J	Relative pronoun jenž, již, ... not after a preposition (lit. who, whom)	P - pronoun
K	Relative/interrogative pronoun kdo (lit. who), incl. forms with affixes -ž and -s (affixes are distinguished by the category Table 2.16 (for -ž) and Section 2.2.1 (for -s))	P - pronoun
L	Pronoun, indefinite všechnen, sám (lit. all, alone)	P - pronoun
M	Adjective derived from verbal past transgressive form	A - adjective
N	Noun (general)	N - noun
O	Pronoun svůj, nesvůj, tentam alone (lit. own self, not-in-mood, gone)	P - pronoun
P	Personal pronoun já, ty, on (lit. I, you, he) (incl. forms with the enclitic -s, e.g. tys, lit. you're); gender position is used for third person to distinguish on/ona/ono (lit. he/she/it), and number for all three persons	P - pronoun
Q	Pronoun relative/interrogative co, copak, cožpak (lit. what, isn't-it-true-that)	P - pronoun
R	Preposition (general, without vocalization)	R - preposition
S	Pronoun possessive můj, tvůj, jeho (lit. my, your, his); gender position used for third person to distinguish jeho, její, jeho (lit. his, her, its), and number for all three pronouns	P - pronoun
T	Particle	T - particle
U	Adjective possessive (with the masculine ending -ův as well as feminine -in)	A - adjective
V	Preposition (with vocalization -e or -u): (ve, pode, ku, ..., lit. in, under, to)	R - preposition
W	Pronoun negative (nic, nikdo, nijaký, žádný, ..., lit. nothing, nobody, not-worth-mentioning, no/none)	P - pronoun
X	(temporary) Word form recognized, but tag is missing in dictionary due to delays in (asynchronous) dictionary creation	

Table 2.5: *(continued)*

Value	Description	POS
Y	Pronoun relative/interrogative co as an enclitic (after a preposition) (oč, nač, zač, lit. about what, on/onto what, after/for what)	P - pronoun
Z	Pronoun indefinite (nějaký, některý, číkoli, cosi, ..., lit. some, some, anybody's, something)	P - pronoun
a	Numeral, indefinite (mnoho, málo, tolik, několik, kdovíkolik, ..., lit. much/many, little/few, that much/many, some (number of), who-knows-how-much/many)	C - numeral
b	Adverb (without a possibility to form negation and degrees of comparison, e.g. pozadu, naplocho, ..., lit. behind, flatly); i.e. both the Section 2.2.1 as well as the Table 2.13 attributes in the same tag are marked by - (Not applicable)	D - adverb
c	Conditional (of the verb být (lit. to be) only) (by, bych, bys, bychom, byste, lit. would)	V - verb
d	Numeral, generic with adjectival declension (dvojí, desaterý, ..., lit. two-kinds/..., ten-...)	C - numeral
e	Verb, transgressive present (endings -e/-ě, -íc, -íce)	V - verb
f	Verb, infinitive	V - verb
g	Adverb (forming negation (XrefId[??] set to A/N) and degrees of comparison Table 2.13 set to 1/2/3 (comparative/superlative), e.g. velký, za\-jí\-ma\-vý, ..., lit. big, interesting	
h	Numeral, generic; only jedny and nejedny (lit. one-kind/sort-of, not-only-one-kind/sort-of)	C - numeral
i	Verb, imperative form	V - verb
j	Numeral, generic greater than or equal to 4 used as a syntactic noun (čtvero, desatero, ..., lit. four-kinds/sorts-of, ten-...)	C - numeral
k	Numeral, generic greater than or equal to 4 used as a syntactic adjective, short form (čtvery, ..., lit. four-kinds/sorts-of)	C - numeral
l	Numeral, cardinal jeden, dva, tři, čtyři, půl, ... (lit. one, two, three, four); also sto and tisíc (lit. hundred, thousand) if noun declension is not used	C - numeral
m	Verb, past transgressive; also archaic present transgressive of perfective verbs (ex.: udělav, lit. (he-)having-done; arch. also udělaje (Table 2.16 = 4), lit. (he-)having-done)	V - verb
n	Numeral, cardinal greater than or equal to 5	C - numeral
o	Numeral, multiplicative indefinite (-krát, lit. (times): mnohokrát, tolikrát, ..., lit. many times, that many times)	C - numeral
p	Verb, past participle, active (including forms with the enclitic - s, lit. 're (are))	V - verb
q	Verb, past participle, active, with the enclitic -ť, lit. (perhaps) -could-you-imagine-that? or but-because- (both archaic)	V - verb
r	Numeral, ordinal (adjective declension without degrees of comparison)	C - numeral
s	Verb, past participle, passive (including forms with the enclitic -s, lit. 're (are))	V - verb
t	Verb, present or future tense, with the enclitic -ť, lit. (perhaps) -could-you-imagine-that? or but-because- (both archaic)	V - verb
u	Numeral, interrogative kolikrát, lit. how many times?	C - numeral
v	Numeral, multiplicative, definite (-krát, lit. times: pětkrát, ..., lit. five times)	C - numeral
w	Numeral, indefinite, adjectival declension (nejeden, tolikátý, ..., lit. not-only-one, so-many-times-repeated)	C - numeral
y	Numeral, fraction ending at -ina; used as a noun (pětina, lit. one-fifth)	C - numeral
z	Numeral, interrogative kolikátý, lit. what (at-what-position- place-in-a-sequence)	C - numeral

Obsolete values:

Value	Description
!	Abbreviation used as an adverb
.	Abbreviation used as an adjective
˜	Abbreviation used as a verb
;	Abbreviation used as a noun
3	Abbreviation used as a numeral
x	Abbreviation, part of speech unknown/indeterminable

3 - Gender

In fact, gender is a truly morphological attribute only for adjectives, pronouns, numerals and verbs. For nouns, it is a lexical property. As a consequence, no noun lemma is allowed to occur with two different genders in the accompanying tags. If a word allows for more than genders, several lemmas have to be reserved for it.

Table 2.6: Gender

Value	Description
F	Feminine
H	{F, N} - Feminine or Neuter
I	Masculine inanimate
M	Masculine animate
N	Neuter
Q	Feminine (with singular only) or Neuter (with plural only); used only with participles and nominal forms of adjectives
T	Masculine inanimate or Feminine (plural only); used only with participles and nominal forms of adjectives
X	Any
Y	{M, I} - Masculine (either animate or inanimate)
Z	{M, I, N} - Not fenimine (i.e., Masculine animate/inanimate or Neuter); only for (some) pronoun forms and certain numerals

4 - Number

Table 2.7: Number

Value	Description
D	Dual , e.g. nohama
P	Plural, e.g. nohami
S	Singular, e.g. noha
W	Singular for feminine gender, plural with neuter; can only appear in participle or nominal adjective form with gender value Q
X	Any

5 - Case

Table 2.8: CASE

Value	Description
1	Nominative, e.g. žena
2	Genitive, e.g. ženy
3	Dative, e.g. ženě
4	Accusative, e.g. ženu
5	Vocative, e.g. ženo

Table 2.8: *(continued)*

Value	Description
6	Locative, e.g. ženě
7	Instrumental, e.g. ženou
X	Any

6 - Possessor's Gender

Table 2.9: Possessor's Gender

Value	Description
F	Feminine, e.g. matčin, její
M	Masculine animate (adjectives only), e.g. otců
X	Any
Z	{M, I, N} - Not feminine, e.g. jeho

7 - Possessor's Number

Table 2.10: Possessor's Number

Value	Description
P	Plural, e.g. náš
S	Singular, e.g. můj
X	Any, e.g. your

8 - Person

Table 2.11: PERSON

Value	Description
1	1st person, e.g. píšu, píšeme
2	2nd person, e.g. píšeš, píšete
3	3rd person, e.g. píše, píšou
X	Any person

9 - Tense

Table 2.12: Tense

Value	Description
F	Future
H	{R, P} - Past or Present
P	Present
R	Past
X	Any

10 - Degree of Comparison

Table 2.13: GRADE

Value	Description
1	Positive, e.g. velký
2	Comparative, e.g. větší
3	Superlative, e.g. největší

11 - Negation

Table 2.14: NEGATION

Value	Description
A	Affirmative (not negated), e.g. možný
N	Negated, e.g. nemožný

12 - Voice

Table 2.15: Voice

Value	Description
A	Active, e.g. píšící
P	Passive, e.g. psaný

15 - Variant

Table 2.16: VAR

Value	Description
-	Basic variant, standard contemporary style; also used for standard forms allowed for use in writing by the Czech Standard Orthography Rules despite being marked there as colloquial
1	Variant, second most used (less frequent), still standard
2	Variant, rarely used, bookish, or archaic
3	Very archaic, also archaic + colloquial
4	Very archaic or bookish, but standard at the time
5	Colloquial, but (almost) tolerated even in public
6	Colloquial (standard in spoken Czech)
7	Colloquial (standard in spoken Czech), less frequent variant
8	Abbreviations
9	Special uses, e.g. personal pronouns after prepositions etc.

2.2.2 Compact tags

For most (but not all cases) just omit the dashes from positional tags.

For more information, see
<http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/compact_tags.pdf>

2.2.3 Informal abbreviations

In certain cases (including some places in this manual), the following tag abbreviations are used. Most of them are self-evident (dashes and rarely used fields dropped), as you can see in the following list:

- Ngnc - noun; NFS1 = `NNFS1-----A----`
- Aagnc - adjective; AAXXX = `AAXXX----1A----`
- Db - adverb; Db = `Db-------------`
- Dg - adverb; Dg = `Dg-------1A----`
- Dgd - adverb; Dga2 = `Dg-------2A----`
- J^ - conjunction; J^ = `J^-------------`
- J, - conjunction; J, = `J,-------------`
- Rc, RRc - preposition, RR7 = `RR--7----------`
- RVc - vocalized preposition, RV7 = `RV--7----------`
- TT - particle; TT = `TT-------------`
- Ng-8, NNgXX-8 - noun abreviation; NFXX-8 = `NNFXX-----A---8`
- AX-8, AAXXX-8 - adjective abreviation; AAXXX-8 = `AAXXX----1A---8`

- Db-8 - adverb abreviation; Db-8 = `Db------------8`
- Rc-8, RRc-8 - preposition abreviation; RR7-8 = `RR--7---------8`

Chapter 3

Names

Unlike in version 1.0, it is now preferred to separate named entity tagging from morphology. Named entities (often multiple-word) should be marked and categorized as special *phrases* on a layer other than morphological; this is a separate project that has not been included in PDT 2.0. Lemmas of proper names will still bear information on the name category. Nevertheless, we respect the original idea that the term suffixes shall explain the meaning of the lemma, not the context it appears in. Thus for instance *New* should be lemmatized as `new_,t` in *New York*, not `New_;G`. *York* should be lemmatized `York_;G` even in *New York Times* where it was previously `York_;K`. For details see below.

Unfortunately, it was not manageable to enforce the desired lemmatization in PDT 2.0. The annotation is still inconsistent in this respect. We plan to correct it in a future version.

Table 3.1: Name types

Type	Explanation, examples
Y	given name (formerly used as default): *Petr*, *John*
S	surname, family name: *Dvořák*, *Zelený*, *Agassi*, *Bush*
E	member of a particular nation, inhabitant of a particular territory: *Čech*, *Kolumbijec*, *Newyorčan*
G	geographical name: *Praha*, *Tatry* (the mountains)
K	company, organization, institution: *Tatra* (the company)
R	product: *Tatra* (the car)
m	other proper name: names of mines, stadiums, guerilla bases, etc.

The lemma should start with upper case if the word is always in upper-case in names (`Špaček_;S` is always capitalized, `špaček` is not).

3.1 Personal names

Given names and surnames are distinguished by the term field in their lemmas (`_;Y` vs. `_;S`). Note that we do not use the terms *first name* and *last name* because in some cultures the surname (family name) comes first and, more importantly, sometimes the original order is respected in Czech texts. If a name can serve both as given and family name, the preferable solution is to reserve two lemmas (for instance, *Pavel Pavel* would be lemmatized as `Pavel-1_;Y` `Pavel-2_;S`. However, in some cases there is currently one lemma covering both usages (such as `Pavel_;Y_;S`).

If a person has only one name, it usually is a given name: `Aristoteles_;Y` (Aristotle).

Personal names homonymous with a normal Czech word should always have a lemma of their own. Thus *Zeman* (surname) is lemmatized as `Zeman-1_;S`, not `zeman` (squire).

Personal names are always tagged as nouns, even if they have an adjectival form (true for many Slavic surnames): `Palacký_;S / NNMS1-----A----`.

Czech female surnames are usually derived from (but not equal to!) a male surname. Their form strongly resembles a possessive adjective: *paní Nováková* (Mrs. Novák) differs from *Novákova žena* (Novák's wife) just in the length of the final *a/á*. However, *Nováková* will neither be analyzed as `Novákův_;S_^(*2) / AUFS1M---------` (a surname cannot be adjective), nor as `Novák_;S / NNMS1-----A----` (this lemma implies the masculine gender). The correct analysis would be `Nováková_;S_^(*3) / NNFS1-----A----` (but it lacks the derivational information in the current data).

Foreign surnames of women are usually ”femalized” in Czech texts (*Condoleeza Riceová*). In such cases they are treated as normal Czech female surnames. If they are left intact (*Condoleeza Rice*), their lemma must indicate their foreign origin and their tag must tell that their gender and case are unknown: `Rice_;S_,t` / `NNXSX-----A----`.

Otherwise, foreign personal names are rarely marked as foreign words because in Czech texts, they are usually declined according to the Czech grammar: *Bill Clinton, bez Billa Clintona, Billu Clintonovi, s Billem Clintonem...* Thus *Bill* is lemmatized as `Bill_;Y`, not `Bill_;Y_,t`. (See also Chapter 6, “Foreign words and phrases”.) Even if a name allows for a frozen (undeclined) form, there usually is a context in which it can be declined: *kniha o Willie Nelsonovi* vs. *kniha o Williem Nelsonovi*; *zvolili Teng Siao-pchinga* vs. *zvolili pana Tenga*. Some foreign names, such as *Steffi*, are never declined.

3.1.1 von, van, etc.

Prepositions, conjunctions and (foreign) determiners form parts of personal names that indicate geographical roots of the family (*Ludwig van Beethoven, Jiří z Poděbrad, Kryštof Harant z Polžic a Bezdružic, Miguel de Cervantes y Saavedra, Hans van den Broek...*) Both Czech and foreign words of that kind are lemmatized as *normal words*, not as given or family names: `z-1`, `von-2_,t`, `de_,t`.

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 case of *von*, *van* and *de*, the original geographical meaning is usually less obvious for a Czech reader and the following word is annotated as surname.

Example 3.1.1: Personal names with *von, van* etc.

- *Ludwig van Beethoven* - `Ludwig_;Y van-2_,t_^(v_hol._jménech) Beethoven_;S`
- *František Lobkovic* - `František_;Y Lobkovic_;S`
- *František z Lobkovic* - `František_;Y z-1 Lobkovice_;G`
- *Kryštof Harant z Polžic a Bezdružic* - `Kryštof_;Y Harant_;S z-1 Polžice_;G a-1 Bezdružice_;G`

3.1.2 Chinese and Korean names

Usage. The surname precedes the given name. In most cases, the whole name is used (not just the family name). The thing is complicated by the fact, that many Chinese living abroad often change the order of their name or use their given name as a surname, etc. The discussion below can help you to determine, which part of a name is the given name and which part is the surname. If you are in doubt annotate them all as given names (Y).

Surnames. There are relatively few surnames in China (200 most common surnames account for >96% of all surnames). Most of them consist of one syllable (Wang, Li, Chen, etc.) Only few surnames consist of two syllables (Ou-yang, Mo-qi, Si-ma, Pu-yang). Married women do not get their husband’s surname.

Given names. Mostly two syllables, often connected with a dash (however sometimes separated by a space).[1] Some given names can be widely used, some are unique. Often it is impossible (for a non-Chinese speaker) to say whether it is a name of a male or a female. The second syllable is usually used in informal addressing. The first syllable can be shared by all siblings. In traditional China a person had several given names during his/her life.

Most common Chinese surnames (in Pinyin / Czech transcription):. *Cai / Cchaj, Chen / Čchen, Deng / Teng, Gao / Kao, Guo / Kuo, He / Che, Hu / Chu, Huang / Chuang, Li, Liang, Lin, Lü, Ma, She / Še, Sun, Tang / Tchang, Wang, Wu, Xie / Sie, Xu / Sü, Yang / Jang, Ye / Jie, Zhang / Čang, Zhao / Čao, Zheng / Čeng, Zhu / Ču*

Links.

- <http://www.geocities.com/Tokyo/3919/atoz.html> - Alphabetical Index of Chinese Surnames (incl. Pinyin, Anglicized and other versions)

[1] Chinese names are usually transcribed using a Chinese-Czech transcription system (a mutation of Wade-Giles). Pinyin is rarely used. In pinyin, the given name would be concatenated to one token instead of three (two words and the dash).

Korean names. Most Korean names look and behave similarly to Chinese names. The most common Korean surnames (45% of the population) are *Kim, Lee* (often spelled as *Rhee, Yi, Li*), and *Park*.

> Note
>
> Analogical annotation may be suitable for other Far-Eastern names as well (e.g. Vietnamese). It does not apply to Japanese. Japanese are similar in their preference to indicate surname in the first position and given name in the second but the order is usually swapped in Czech texts and if not, non-Japanese speakers have little clues to decide. Both names usually use one to two Chinese characters each but they may be pronounced (and transcribed) using much more syllables (packed in two words, one for the given name and the other for the surname). One clue is that given names of Japanese women often take the suffix *-ko*.

Example 3.1.2: Chinese and Korean names

- *Teng Siao-pching* - `Teng_;S Siao_;Y - pching_;Y`
- *Kim Ir-sen* - `Kim_;S Ir_;Y - sen-2_;Y`

3.1.3 Foreignized Czech names

Sometimes you can encounter names that are Czech in their origin, but are somehow altered to fit other languages (accents omitted, female and male surnames are the same - e.g. *Judy Sedivy*, from Czech *Šedivý*).

Use the following guidelines to decide the lemma and tag for such a name:

- A name that does not distinguish female and male variant should have just one lemma and a tag with the `X` (unknown) gender: `Sedivy_;S_,t` / `NNXXX-----A----`
- A name that has the same spelling as in Czech, should use the Czech lemma: `Jane_;Y Janda_;S`
- A name with altered spelling has its own lemma (with the `_,t` suffix): `Judy_;Y Sedivy_;S_,t`

3.2 Geographical names

3.2.1 Countries, cities, rivers, mountains

Main noun. The main word (head) in a multi-word name of a city is always noun; the same holds for a one-word city name. If it is homonymous with an adjective, a new noun lemma is created for the name. Thus *Hluboká* is lemmatized as `Hluboká_;G / NNFS1-----A----` rather than `hluboký / AAFS1----1A----` (lit. 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 stick with the non-geographical lemma even in geographical usages.

Modifiers in multi-word names. Attributive adjectives, prepositions, conjunctions etc. should be lemmatized as normal words. Other nouns may be lemmatized as geographical if they are nested geographical names (e.g. names of rivers or mountains in names of cities).

Part of speech of foreign words. Original part of speech of the word in the source language is used unless there is a good reason not to do so. Besides not knowing the original part of speech, a very good reason is that the word behaves as a different part of speech in Czech texts. For instance, *blanc* is adjective in French *Mont Blanc* but it behaves as a noun in *na Mont Blanku*. *Mont* can be annotated as an undeclined noun. See Chapter 6, “Foreign words and phrases” for more information on foreign words.

Table 3.2: Examples of geographical names

Name	Type	Morphological annotation
Česká republika	country	český / AAFS1----1A---- // republika / NNFS1-----A----
Ústí nad Labem	city	Ústí_;G / NNNS1-----A---- // nad-1 / RR--7---------- // Labe_;G / NNNS7-----A----
Karlovy Vary	city	Karlův_;Y_^(*3el) / AUIP1M--------- // Vary_;G_^(Karlovy_Vary) / NNIP1-----A----
Dobrá Voda	city	dobrý / AAFS1----1A---- // voda / NNFS1-----A----
Odolena Voda	city	Odolena_;G_^(Odolena_Voda) / AAXXX----1A---- // voda / NNFS1-----A----
Černá v Pošumaví	city	Černá_;G / NNFS1-----A---- // v-1 / RR--6-----A---- // Pošumaví_;G / NNNS6-----A----
Ohrada u Hluboké	city	ohrada / NNFS1-----A---- // u-1 / RR--2---------- // Hluboká_;G / NNFS2-----A----
Hradec Králové	city	Hradec_;G / NNIS1-----A---- // králová_^(královna) / NNFS2-----A----
Kostelec nad Černými Lesy	city	Kostelec_;G / NNIS1-----A---- // nad-1 / RR--7---------- // černý_;o / AAIP7----1A---- // les / NNIP7-----A----
New York	city	new_,t_^(angl._nový) / AAXXX----1A---- // York_;G / NNIS1-----A----
A Coruña	city	o-10_,t_^(port._člen) / AAFSX----1A---- // Coruña_;G / NNFS1-----A----
São Paulo	city	são_,t_^(port._svatý) / AAMSX----1A---- // Paulo_;Y / NNMS1-----A----
Rio de Janeiro	city	Rio_;G / NNNS1-----A---- // de_,t / RR--X---------- // Janeiro_;G / NNNS1-----A----
Le Havre	city	le_,t_^(fr._člen) / AAISX----1A---- // Havre_;G / NNIS1-----A----
Krems an der Donau	city	Krems_;G / NNIS1-----A---- // an_,t / RR--3---------- // der_,t_^(něm._člen) / AAFS3----1A---- // Donau_;G / NNFSX-----A----
San Juan de la Rambla	city	san_,t_^(šp._a_it._svatý) / AAMSX----1A---- // Juan_;Y / NNMS1-----A---- // de_,t / RR--X---------- // el_,t_^(šp._člen) / AAFSX----1A---- // Rambla_;G / NNFSX----1A----
Kao-hsiung	city	Kao_;G / AAXXX----1A---- // - / Z:------------- // hsiung_;G_^(př._Kao-hsiung) / NNXXX-----A----
Wu-lu-mu-čchi	city	Wu_;G / NNXXX-----A---- // - / Z:------------- // lu_;G / NNXXX-----A---- // - / Z:------------- // mu_;G / NNXXX-----A---- // - / Z:------------- // čchi_;G / NNXXX-----A----
Gerlachovský štít	mountain	gerlachovský / AAIS1----1A---- // štít / NNIS1-----A----
Divoká Orlice	river	divoký / AAFS1----1A---- // Orlice_;G / NNFS1-----A----

3.2.2 Streets

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 is not altered.

3.3 Companies and institutions

Companies, foundations, shops, clubs, sport clubs, restaurants, etc. all can have lemmas flagged _;K. However, "normal words" (those the usage of which is not limited to the company name) should get

Example 3.2.1: Street names

- *Dlouhá* - dlouhý / AAFS1----1A----
- *Dlouhá ulice* - dlouhý / AAFS1----1A---- // ulice / NNFS1-----A----
- *Palackého* - Palacký_;S / NNMS2-----A----

their normal lemmas. Only if a word cannot be explained another way or if its meaning has nothing to do with the company (e.g. Škoda_;K), the flag should be used. The border between personal and company names is fuzzy: if it is clear that a surname is part of a company name (e.g. *Uzenářství Novák_;S a syn*) it should be lemmatized as a surname. On the other hand, *Škoda* should be lemmatized as a company no matter that it was also named after a person. This name is mostly known as a company name. Abbreviations and acronyms are frequent company names - see also Chapter 4, "Abbreviations".

Table 3.3: Examples of company names

Name	Annotation
Škoda auto, a.s.	Škoda_;K / NNFS1-----A---- // auto / NNNS1-----A---- // , / Z:------------- // akciový_:B / AAFXX----1A---8 // . / Z:------------- // společnost_:B / NNFXX-----A---8 // . / Z:-------------

3.3.1 Restaurants

Table 3.4: Examples of restaurant names

Name	Annotation
Bar Viola	bar / NNIS1-----A---- // Viola_;K / NNFS1-----A----
U Medvídků	u-1 / RR--2---------- // medvídek / NNMS2-----A----
La cambusa	le_,t_^(fr._člen) / AAFSX----1A---- // cambusa_;K_,t / NNFS1-----A----
Restaurant HaPi	restaurant / NNIS1-----A---- // HaPi_;K / NNXXX-----A----
Čínská restaurace S'-ČCHUAN	čínský / AAFS1----1A---- // restaurace / NNFS1-----A---- // S'_;G / AAXXX----1A---- // - / Z:------------- // čchuan_;G / NNIS1-----A---- (Note: the restaurant has been named after the Sichuan province in China.)
Francouzská restaurace v Obecním domě	francouzský / AAFS1----1A---- // restaurace / NNFS1-----A---- // v-1 / RR--6---------- // obecní / AAIS6----1A---- // dům / NNIS6-----A----
Hospůdka U vylitýho mrože	hospůdka / NNFS1-----A---- // u-1 / RR--2---------- // vylitý / AAMS2----1A---6 // mrož / NNMS2-----A----

3.3.2 Sport clubs

Names of sporting clubs are often combined of the proper club name and a geographical name of the location the club comes from. The former should have `_;K` in lemma, the latter should have `_;G`.

Of course, it may be difficult tell whether a word in a foreign club name is a location. If you do not know, annotate it as a company. To determine, whether something is a name of a town or a club, you can try to find that name on a map (eg. <http://www.expedia.com/pub/agent.dll?qscr=mmfn>) or to find the club (e.g. http://www.soccerage.com/[2]).

Table 3.5: Examples of sport club names

Name	Annotation
SKP Union Cheb	`SKP_:B_;K / NNNXX-----A---- // Union_;K / NNIS1-----A---- // Cheb_;G / NNIS1-----A----`
Chelsea FC	`Chelsea_;G / NNFS1-----A----` (part of London, UK) `FC-1_:B_;K_;w_,t_^(football_club)`
Sparta Praha	`Sparta-2_;K Praha_;G` (Although there is a town of Sparta in Greece, it has nothing to do with the football club located in Praha, Czechia.)
Viktoria Žižkov	`Viktoria-2_;K_^(jméno_sportovního_klubu) Žižkov_;G`
Udinese	`Udinese_;K / NNNSX-----A----` It is an adjective derived from *Udine* (a city in Italy), the official name of the club is *Udinese Calcio* (Football of Udine). However, the name is perceived in Czech as a noun.

Names of sport clubs often contain abbreviations. Some are common and present in the analyzer's lexicon (e.g. FC, AC) some are quite unusual (e.g. EV, ERC, EC, ERC, EG, VS, AS). If they are not present in the lexicon, enter them suffixing the lemma by `_:B_;K_;w` and tag them by `NNNXX-----A---8`.

3.4 Horses, DJ's etc.

Horses have all kind of names (e.g. *Vinná réva*, *Deprivace*, *He Shall Reign*, *La Paloma*, *Monitor*, *Frýdlant*, *Gold End*, *Lučina*, *Green Peace*, *Areál*, *First*, *Bounty*). Quite often one does not know whether it is male or female (sometimes even female-like names belong to a male horse). One clue is, that in an Oak (a horse contest type), all horses are young mares - females.

If any reasonable analysis is possible it should be used regardless the lemma is marked as 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, a new lemma with the `_;Y` flag should be introduced.

Example 3.4.1: Names of horses

- *Vinná réva* - `vinný / AAFS1----1A---- // réva / NNFS1-----A----`
- *Deprivace* - `Deprivace_;Y / NNFS1-----A----`
- *He Shall Reign* - `he_,t / PPYS1--3------- // shall_,t / VB-S---3P-AA--- // reign_,t / Vf--------A----`

Most of the horse names were not annotated correctly in PDT 1.0 - simply any available name was selected. (Otherwise, a new lemma with category Y inserted in each case: e.g. Deprivace would be Deprivace_;Y, annotated as deprivace, He Shall Reign annotated as a normal English phrase: he_,t, shall_,t reign_,t).

Similar problem is with the names of musical groups and DJ's. For famous groups and DJ's enter separate lemmas, for others use normal available lemmas.

[2] <http://www.soccerage.com>

3.5 Products

Similarly to companies, only words that are uniquely product names (or they have a homonym but its meaning has nothing to do with the product) have their lemmas flagged `_;R`.

If there is a company and a product of the same name, there should be two lemmas, e.g. `Tatra-1_;K` in *Tatra, a.s.*, and `Tatra-2_;R` in *Tatra 613*.

3.6 Sporting and other events

There is no special lemma term flag for events but the `_;m` for generic proper names can be used (`_;m_;w` for sporting events). Similarly to companies, only words that are uniquely event names (or they have a homonym but its meaning has nothing to do with the event) have their lemmas flagged `_;m`.

If there is a company and an event of the same name, there should be two different lemmas.

Table 3.6: Examples of event names

Name	Annotation
Paris Indoor	`Paris_;G_,t / NNIXX-----A---- // Indoor_;m_,t / NNIXX-----A----`
US Open	`US-2_:B_,t_^(americký) / AAXXX----1A---8 // Open-1_;m_;w_,t_^(otevřený_[turnaj],_v_názvu) / NNIXX-----A----`
akce Stop milión	`akce / NNFS1-----A---- // stopit_:W_^(úplně_spotřebovat_topením) / Vi-S---2--A---- // milión‘1000000 / NNIS4-----A----`
Pohár mistrů	`pohár / NNIS1-----A---- // mistr / NNMP2-----A----`
Mistrovství světa	`mistrovství / NNNS1-----A---- // svět / NNIS2-----A----`

3.7 Other

3.7.1 Buildings

If a name of a building cannot be annalyzed other way, it should be a geographical name (`Parthenón_;G`). However, most building names are made of normal words (`tančící_^(*3it) dům`, `pražský hrad`, `kostel svatý_:B . kříž`) or other names (`chrám svatý_:B . Barbora_;Y`).

3.7.2 Televisions

Generally televisions are annotated as institutions (`_;K`). Only when a company runs several channels, then the channels are annotated as products (`_;R`). It is currently used only with the Czech(oslovak) public television (*ČT1*, *ČT2* and *F1*).

Example 3.7.1: TV company names

- *ČT - ČT_:B_;K*
- *ČT1 - ČT1_:B_;R*
- *Nova - Nova_;K*
- *NBC - NBC-4_:B_;K*
- *CNN - CNN-1_:B_;K_;y_;b_,t*

3.7.3 News and magazines

All names of periodicals shall be annotated as products (`_;R`) even if their publishing company has the same name.

Example 3.7.2: Names of periodicals

- *Sme* - `Sme_;R_^(noviny) / NNNSX-----A----`
- *Zeitung* - `Zeitung-1_;R_,t_^(souč._názvu_něm._novin) / NNISX-----A----` (originally feminine gender in German but perceived as masculine inanimate in Czech)

3.7.4 Song names

Songs, TV programs etc. are in fact products. Their names usually consist of more than one word and the component words mostly have meaning of their own (not unique to the song name). Thus the `_;R` flag will rarely be used.

3.8 Adjectives derived from names

Possessive adjectives derived from personal names (or names of nation members, territory inhabitants) retain the name flags in their lemmas: `Karlův_;Y_^(*3el)`, `Mariin_;Y_^(*2e)`, `Novákův_;S_^(*2)`, `Číňanův_;E_^(*2)`.

Adjectives derived from geographical names are *not* marked as geographical (no `_;G` flag in lemma). They do not even show the derivational information. These adjectives are not capitalized in Czech, while the original nouns are. So if we used the usual mechanism to describe derivation we would have to replace the whole lemma: `africký_^(*7Afrika)`, not `africký_^(*3ka)`.

Chapter 4

Abbreviations

Abbreviations of a single word should use the lemma of the word, augmented with the `_:B` flag. This is the only acceptable situation in which two lemmas share LemmaProper, are not distinguished by numbers, but differ in their AddInfo. For instance, the three letters (separate tokens) in *s.r.o.* are lemmatized as `společnost_:B` (company), `ručení_:B` (liability), `omezený_:B_^(*3it)` (limited).

Abbreviations consisting of a single capital letter represent names. Lots of names can be represented by a letter, and we often do not know the name. In such cases, the abbreviation uses itself as a lemma (augmented with the appropriate flags). For instance, in *G. Bush* it would be `G_:B_;Y` (despite the fact that in this particular case we know that most probably the *G* stands for *George*).

Acronyms and abbreviations of multi-word expressions use themselves as lemmas (again, flagged `_:B`). If possible, the comment should explain the abbreviation. For instance, *FIDE* would be `FIDE_:B_;K_;w_,t_^(Fédération_Internationale_des_Échecs)`.

Morphological tags of abbreviations should always end in `8`.

Table 4.1: Examples of abbreviations

Abbreviation	Full expression	Annotation
např.	*například*	`například_:B / Db------------8`
P.S.	*post scriptum*	`post-2_:B_,t_^(lat._po,_např._P.S.) / RR--X---------8 / / scriptum_:B_,t_^(lat.,_např._P.S.) / NNNXX-----A---8`
n.L.	*nad Labem*	`nad-1_:B / RR--7---------8 // Labe_:B_;G / NNNS7-----A---8`
r. 1998	*rok/roku/roce 1998*	`rok_:B / NNIXX-----A---8`
r.:	*režie:*	`režie_:B / NNFXX-----A---8`
rež.:	*režie:*	`režie_:B / NNFXX-----A---8` Note: This and the previous example violate the rule that each lemma/tag pair leads to no more than one word form. Numbering the lemmas is not appropriate in this case but no suitable solution has been devised so far.

4.1 Gender

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. Neuter is always possible. Besides that, the author may use the gender of the main word of the abbreviated expression. The matter can become further complicated with foreign expressions if their Czech gender does not correspond to the gender in the original language.

In order to keep the rule of a noun lemma not having more than one gender, tags of abbreviations should use the `X` gender code. This is often broken in PDT 2.0 and abbreviations are the most frequent nouns to have two different genders.

There is a similar problem with abbreviations of personal names (`J_:B_;Y` can mean both *Jan* and *Jana*). The difference is that here the neuter interpretation is not plausible. Nevertheless, the tagset does not provide any code for `{M+F}` genders, so the best bet is to stick with `X`.

Table 4.2: Gender of abbreviations

Abbreviation	Full expression	Possible genders
UK	*Univerzita Karlova*	`FN`
FBI	*Federal Bureau of Investigation*	`N` (default), `F` (probably à la *CIA*)
CIA	*Central Intelligence Agency*	`FN`

4.2 Isolated letters

Most isolated letters (e.g. *A-konto*) are handled as abbreviations. Only if they do not form part of a name they are lemmatized as `_^(označení_pomocí_písmene)`: *zápas skupiny B*.

The following is a prototype of lemmas, their numbers and AddInfos for an isolated letter. There should be such lemmas for all letters of the Czech alphabet. Note that numbering a lemma by zero is not used anywhere else and might be deprecated in future. Anyway, no program should ever rely that the numbers will be as indicated. Lemma numbers serve to distinguish between homonymous lemmas but they are not meant to bear any semantic information.

- `K-0_:B_;Y` - given names
- `K-4_:B_;K` - names of institutions
- `K-5_:B_;G` - geographical names
- `K-6_:B_;R` - names of products
- `K-7_:B_;m` - other names (sporting events etc.)
- `K-9_:B_;S` - surnames
- `k-8_:B_^(ost._zkratka)` - other abbreviations (not names) - should not be used if the annotator knows the abbreviated word - then the `word_:B` lemma should be used instead
- `k-3_^(označení_pomocí_písmene)` - other isolated letters (not abbreviations, not in names)

Table 4.3: Examples of isolated letters

Expression	Annotation of the letter
A-mužstvo	`a-3_^(označení_pomocí_písmene) / NNXXX-----A----` (Note: Adjective would be more appropriate in this particular case but noun is plausible as well and no lemma is allowed occur with more than one part of speech.)
§ 27 odst. 1 písm. d	`d-3_^(označení_pomocí_písmene) / NNXXX-----A----`
16 A	`A-1‘ampér_:B / NNIXX-----A---8`
A-konto	`A-6_:B_;R / NNXXX-----A---8`
ABC, a.s.	`akciový_:B / AAXXX----1A---8`
na s. 128	`strana-4_:B_^(v_knize,_rukopise...)  / NNFXX-----A---8`

4.3 Units of measurements

Unlike most abbreviations, standard unit abbreviations are not followed by a period in Czech texts. In PDT 2.0, they often use a lemma equal to the abbreviated form, referring to the unabbreviated lemma via `‘`: `V-1‘volt_:B`. Unfortunately, this approach is not taken consistently, so for instance *Celsius* uses directly the target lemma instead of a reference to it: `Celsius_:B`.

Units called after male persons (*V - volt, A - ampér*, etc.), have the masculine *inanimate* gender. However, units using degrees (*C, F*) have masculine *animate* gender, because the word *stupeň* (degree) is always present (even if omitted in the written text). Absolute temperature uses the unit called *Kelvin (K)*, not *degree of Kelvin*. Therefore the unit has the masculine inanimate gender. The author may use it errorneously as degrees but we cannot correct them because the gender of a noun is implied by its lemma, not its context.

Table 4.4: Examples of units

Expression	Annotation of the unit abbreviation
Ráno byly 3°C.	`Celsius_:B / NNMXX-----A---8`
Ráno byly 3 C. (read as *Ráno byly tři stupně Celsia.*)	`Celsius_:B / NNMXX-----A---8`
teplota 5000 K (read as *teplota pět tisíc kelvinů*)	`K-1‘kelvin_:B / NNIXX-----A---8`

If the C character is preceded by some character trying to look like the degree symbol ° (eg. -C, o C, O C), it should be marked as an error. The form attribute should be "°", while the origf attribute retains the original character.[1] The lemma shall be `stupeň_:B`, the tag `NNIXX-----A---8`.

4.4 Authors' signatures

The authors' name abbreviations used in newspapers (e.g. *ber, mas, jst...* in "sentences" like *PRAHA (ČTK, ber)* -) have the base form in the lemma equal to the word form, they are numbered -99 and AddInfo-ed `_:B_;S`. Their tag has a special SUBPOS character, `%`. For instance, *ber* is annotated as `ber-99_:B_;S / N%XXX-----A---8`. Again, no program should rely on the number being always 99.

4.5 Academic titles

The morpohological analyzer currently distinguishes genders in titles, generating one lemma for men and another for women (`JUDr-1_:B_^(doktor_práv) / NNMXX-----A---8` vs. `JUDr-2_:B_^(doktorka_práv) / NNFXX-----A---8`). It is possible that the lemmas will be merged in future, using an indefinite gender: `JUDr_:B_^(doktor_práv) / NNXXX-----A---8`.

[1] On Czech keyboards usually Shift+<key-on-the-left-from-1>, followed by Space. On any keyboard under MS Windows: Alt+0176.

Chapter 5

Colloquial Czech

The annotation should distinguish between colloquial lemmas (e.g. *Rusák* (Russian) instead of the standard *Rus*) and colloquial forms of standard lemmas (e.g. *zelenej* (green) instead of the standard *zelený*). The former should be marked in the AddInfo of the lemma (`Rusák_;E_,h`), the latter should be indicated by the VAR field of the morphological tag. The values of `6`, `5`, `7`, and sometimes also `3` may be applicable; in most common cases, `6` is used (`zelený / AAIS1----1A---6`). See also Section 2.2.1.

5.1 *Cos, kdys, jaks...*

A set of Czech words can take the suffix *-s* representing deleted auxiliary verb *jsi* (2nd person). For instance, *"To je dobře, že jsi přišel."* ("It is good that you came.") can be shortened to *"To je dobře, žes přišel."*

These words are only slightly colloquial if at all. Moreover, the reflexive pronouns *ses, sis* were constructed the same way but are perfectly standard while the alternative *jsi se, jsi si* is poor style. *ses* is distinguished from *se* by the 2nd person and by the singular number in tag (`P7-S4--2-------` vs. `P7-X4----------`). Similarly, *kdos* is tagged `PKM-1--2-------` while *kdo* (who) is tagged `PKM-1----------`. *žes* is tagged `J,-S---2-------` while *že* (that) is tagged `J,-------------`. It is questionable whether it is a good solution to let tags of various classes sometimes indicate the person and sometimes not. Nevertheless, the current morphological analyzer behaves so, and the approach should be extended to words not covered by the analyzer (e.g. *cos*, *kdys*).

5.2 Suffix *-é* in plural of neuter

It is officially ungrammatical to say **malé koťata* instead of *malá koťata*. However, the number of people doing the error is constantly growing.

The phenomenon should not be treated as misspelling. It should be annotated as a colloquial variant of the official *-á* form (`VAR = 5`).

Table 5.1: Colloquial examples

Expression	Annotation
koťata, které	`který / P4NP4---------5`
Novákovic pes	`Novákův_;S_^(*2) / AUXXXM--------6` It is sometimes obsoletely tagged `AUMS1M--------6` in PDT 2.0. If the tag system allowed such tags, `AUXXXXP-------6` might be even more appropriate.
takovejhlema	`takovýhle / PDFD7---------6` (Correct - but rarely used - is *takovýmahle*.)
hovadinama	`hovadina_,h / NNFP7-----A---6` (Both lemma and suffix are colloquial. The current morphological analyzer does not mark the lemma but it should do so.)
pro naší atletiku	`můj_^(přivlast.)  / PSFS4-P1------6` (Short *-i*, *naši* is the correct ending in accusative.)

Chapter 6

Foreign words and phrases

Foreign words enter Czech texts in three different ways:

Citation use. Whole phrases in foreign languages can be inserted into Czech texts as citations. Besides real citations of something someone said or wrote, also names of songs and other works belong to this category. If a foreign verb is present, it is most probably a citation use. Single words can be cited as well but the rule is that a word in a cited phrase never takes Czech suffixes.

Word use. Single words or short phrases (usually noun phrases), supplying a term. This ought to be a rather tiny category. If a foreign word does not take Czech suffixes, it might be a citation. And if it does, the possible domestication of the word should be considered carefully.

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* (cocoa)). Nevertheless, even newer words should not be treated as foreign if they fit into this category. For instance, the current morphological analyzer marks *management* (Czech *vedení*, sometimes also Czechized spelling *manažment*) as a foreign word (`management_,t_^(vedení,_manažment;_angl.)`). According to the word's usage, the `_,t` flag should be omitted.

Despite the uncertainty whether some words shall be marked `_,t`, the following rule affects also domesticated expressions of foreign origin, some names that do not have a Czech equivalent etc. (e.g. *Mont Blanc*).

General rule

1. In citations, the original morphology of the source language shall be described to the extent possible with respect to our tags, and to the annotator's knowledge about the foreign word.

2. In word usages and domesticated expressions, Czech morphology takes precedence. For instance, abovementioned *Mont Blanc* is noun + adjective according to French morphology but *Blanc* has to be tagged as noun because the Czech locative of the phrase reads *na Mont Blanku* (i.e., *Blanc* is declined according to a noun paradigm). Unless there is such a conflict between the original and the Czech morphology, the original part of speech shall be preserved.

Table 6.1: Examples of foreign phrases

Expression	Annotation	Comments
V kostele zpívala Musica Bohemica.	`musica_,t_^(lat._hudba) / NNFS1-----A---- / / bohemica_,t_^(lat._česká) / NNFS1-----A----`	*Bohemica* is adjective in Latin but noun in Czech. It is declined according to the Czech noun pattern *žena*. For the same reason, the base form is not converted to masculine gender.
To je trochu ad hoc.	`ad_,t / RR--X---------- // hoc_,t / NNXXX-----A----`	*hoc* is adverb in Latin but it is annotated as a noun in Czech.

6.1 Articles

Unlike in many other languages, there are no articles in Czech. Articles in foreign phrases are annotated as adjectives.

In some languages, articles distinguish gender, number and case. Analogically to Czech, their lemma should reflect the masculine singular nominative form, the morphological tag should encode the real word form in the text. However, sometimes this approach is not possible due to a different gender or number in Czech: *La Manche* is feminine in French, masculine inanimate in Czech; *Los Angeles* is plural in Spanish, singular in Czech (and in English). There has to be a special lemma for each such frozen article. Thus, *los* would be annotated `el-3_,t_^(šp._člen) / AAMSX----1A----` in *"do Prahy přijeli Los Paraguayos"* but `los-3_,t_^(šp._člen) / AAXXX----1A----` in *"pracuje v Los Angeles"*.

NOTE

The separate lemma reflects the fact that the word form is frozen since it was ported to other languages. However, it might not be needed. Articles are annotated as adjectives and adjectives (unlike nouns) are not required to stick with one gender.

Articles merged with a preposition (e.g. French *du*, Italian *della*, German *aufs, beim, vom, zur, im, am...*) are treated as prepositions.

Table 6.2: Articles in common foreign languages

Language	Form	Lemma	Tag
English	*the*	`the-1_,t_^(angl._urč._člen)`	`AAXXX----1A----`
English	*a*	`a-2_,t_^(angl._neurč._člen)`	`AAXXX----1A----`
English	*an*	`a-2_,t_^(angl._neurč._člen)`	`AAXXX----1A---1`
German	*der*	`der-1_,t_^(něm._člen)`	`AAMS1----1A---- AAFS2----1A---- AAFS3----1A---- AAXP2----1A----`
German	*die*	`der-1_,t_^(něm._člen)`	`AAFS1----1A---- AAFS4----1A---- AAXP1----1A---- AAXP4----1A----`
German	*das*	`der-1_,t_^(něm._člen)`	`AANS1----1A---- AANS4----1A----`
German	*des*	`der-1_,t_^(něm._člen)`	`AAMS2----1A---- AANS2----1A----`
German	*dem*	`der-1_,t_^(něm._člen)`	`AAMS3----1A---- AANS3----1A----`
German	*den*	`der-1_,t_^(něm._člen)`	`AAMS4----1A---- AAXP3----1A----`
Dutch	*de*	`de-2_,t_^(niz._člen)`	`AAMSX----1A---- AAFSX----1A---- AAXPX----1A----`
Dutch	*het*	`de-2_,t_^(niz._člen)`	`AANSX----1A----`
Dutch	*den*	`de-2_,t_^(niz._člen)`	`AAMS3----1A---5 AANS3----1A---5`
French	*le*	`le-1_,t_^(fr._člen)`	`AAMSX----1A----`
French	*la*	`le-1_,t_^(fr._člen)`	`AAFSX----1A----`
French	*l*	`le-1_,t_^(fr._člen)`	`AAXSX----1A----`

Table 6.2: *(continued)*

Language	Form	Lemma	Tag
French	*les*	`le-1_,t_^(fr._člen)`	`AAXPX----1A----`
Italian	*il*	`il-1_,t_^(it._člen)`	`AAMSX----1A----`
Italian	*la*	`il-1_,t_^(it._člen)`	`AAFSX----1A----`
Italian	*gli*	`il-1_,t_^(it._člen)`	`AAMPX----1A----`
Italian	*le*	`il-1_,t_^(it._člen)`	`AAFPX----1A----`
Spanish	*el*	`el-1_,t_^(šp._člen)`	`AAMSX----1A----`
Spanish	*la*	`el-1_,t_^(šp._člen)`	`AAFSX----1A----`
Spanish	*los*	`el-1_,t_^(šp._člen)`	`AAMPX----1A----`
Spanish	*las*	`el-1_,t_^(šp._člen)`	`AAFPX----1A----`
Portuguese	*o*	`o-10_,t_^(port._člen)`	`AAMSX----1A----`
Portuguese	*a*	`o-10_,t_^(port._člen)`	`AAFSX----1A----`
Portuguese	*os*	`o-10_,t_^(port._člen)`	`AAMPX----1A----`
Portuguese	*as*	`o-10_,t_^(port._člen)`	`AAFPX----1A----`
Arabic	*al, ad, an, ar, as, az*	`al-5_,t_^(arab._člen)`	`AAXXX----1A----`
Arabic	*el, ed, en, er, es, ez*	`el-5_,t_^(arab._člen)`	`AAXXX----1A----`
Hebrew	*ha*	`ha-2_,t_^(hebr._člen)`	`AAXXX----1A----`

6.2 English noun clusters

The original approach taken in PDT was that all attributively used nouns were annotated as adjectives. That was quite problematic because virtually all English nouns can be used as attributes of other nouns while they never take Czech adjectival suffixes in Czech texts. Now it is preferred to tag such words as foreign nouns in unknown case. In PDT 2.0, it is still annotated inconsistently.

NOTE

English-like attributive use of nouns has been imported to Czech (*Staropramen Extraliga*, *Český Telecom Cup* etc.)

6.3 Nouns

English nouns in plural form usually preserve the plural perception in Czech. However, terms that were imported in singular are rarely pluralized according to English grammar when the surrounding text requires plural. If a Czech plural ending cannot or is not added, the singular form is used as plural.

Therefore, and for the sake of simplicity, all English nouns should be annotated with unknown number (`X`), unless they have a Czech ending.

English (and most other non-Slavic) nouns have unknown (`X`) case in citations but they can be sometimes declined in word use.

Table 6.3: Number and case of English nouns

Expression	Annotation
oba dva cash flow (oficiální i skutečný)	`flow_,t / NNIXX-----A----`
v cash flow statementu	`statement_,t / NNIS6-----A----`
Beatles: Girl	`girl_,t / NNFXX-----A----`
A teď zahrajeme písničku Girls.	`girl_,t / NNFXX-----A----`

6.4 Verbs

6.4.1 English verbs

The following tags are applied:

- Infinitive *(go)*: `Vf--------A----`
- Present other than 3rd person singular *(go)*: `VB-X---XP-AA---`
- Present 3rd person singular *(goes)*: `VB-S---3P-AA---`
- Imperative *(go)*: `Vi-X---X--A----`
- Past tense *(went)*: `Vp-X---XR-AA---`
- Perfect / passive participle *(gone)*: `Vs-X---XX-AP---`

If it is difficult to determine the base form usage, annotate it as infinitive. If it is difficult to decide between past tense and passive participle, use past tense.

Table 6.4: Examples of English verbs

Expression	Annotation
to be or not to be	`be_,t_^(angl._být,_v_názvech_apod.) / Vf--------A----`
Do it right now!	`do-2_,t / Vi-X---X--A----`

6.5 Slavic languages and Czech dialects

Slavic languages (most prominently Slovak) are related to 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 is identical 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. It would be strange if a ”Czech” word intervened in the middle of a foreign phrase. Nevertheless, this is not always kept in PDT 2.0.

Examples: *ulica kapitána Nálepku* - `Nálepka_;S_,t / NNMS2-----A----`; *ste v Bratislave* - `byť_,t / VB-P---2P-AA--- // v-2_,t / RR--6---------- // Bratislava-2_;G_,t / NNFS6-----A----`

Sometimes a Slovak-like phrase is in fact just a Moravian dialect of Czech, as in *Slovácko sa nesúdí*. The lemmas should be flagged `_,n` instead of `_,t` in such cases.

Chapter 7

Errors

Sometimes the author of a PDT 2.0 text uses a word incorrectly - e.g. a name of a woman as a man's name etc. In such cases, the real usage should be annotated, not the should-be usage.

The texts can contain errors. It is reasonable to correct some of them (but the original - errorneous - word form should always be preserved in the `origf` attribute). However, only low-level errors (spelling and morphology) should be corrected. We do not want to correct Engels' text into Heidegger's. Never replace a colloquial form with an official one (e.g. **zelené města* → *zelená města*, **bez noh* → *bez nohou*), even if the analyzer does not know the form[1].

7.1 Characters

If the author of the text misspelled a foreign name (e.g. converted a non-Czech character to a Czech one, say *Miloševic̈́* to *Miloševič*), it is a low-level error that should be corrected.

Sometimes, foreign characters had been be screwed (e.g. Fran?oise), which may not only lead to an unknown word, it may mislead the tokenizer, resulting in three tokens. Since most work until the release of PDT 2.0 has been done in the ISO Latin 2 encoding, there is a problem with letters not contained in Latin 2. HTML entities should be used but the corresponding accent-free character is also acceptable.

7.2 Separators

Sometimes, the text contains *o* or *I* in place of bullets or separators. *o* should be annotated `o-4_^(graf._oddělovač) / Z:-------------`.

[1] You have to insert a new lemma and/or tag - see Chapter 13, "Insertion" for more details.

Chapter 8

Hard to decide

8.1 až

- ***až-1 + Jˆ***
- *2 až 3 (but not od 2 až do 3 - see až-3)*
- *nabízí přiblížení až přijetí*
- ***až-2 + J,***
- *tak .. až: Nabízí se tak okatě, až je to hanba.*
- *.. začnou pochybovat, až nakonec uvěří, že ..*
- *?? Bylo mi 24, a byl jsem plný touhy se pomstít. Až jsem se ocitl před člověkem, který*
- *dostal zabrat víc než já.*
- **až-3 + Db**

If omitted, the sentence stays grammatical. It is often possible to replace it by teprve.

- *Dostanete až 250 mil zdarma.*
- *kam až: Kam až půjdeš?*
- *Až on me přesvědčil, že tomu tak bude.*

Modifies functional word (should be probably TT)

- až + conj: Je geolog a až pak filozof
- až + prep: z Brna až do Prahy (Cf. až-1)

8.2 jak

- *jak-1_;L_ˆ(živočich) + NNMnc——A—– Obvious.*
- *jak-2 + J,*

1. Meaning že ()

   - Jak řekl M. Zeman, bude třeba ..
   - Jak ukazuje vývoj poslednich let, je to ..
   - Jak známo, ...
   - Skutečnost, jak už to býva, byla trochu jiná.

   However, rarely it can be Db - depending on the interpretation

   - Viděl, jak upadla.
   - Meaning Viděl, že upadla. - J,
   - Meaning Viděl, jakým způsobem upadla. - Db
   - Kamera zabírá poslance, jak otvírají krabici

2. - Time, meaning když, až, jakmile

- Přijdu, (hned) jak budu hotov^ssč^.
- Hned jak budu moct, zavolám.

3.
- In comparison, meaning než, jako:
- Byl větší jak on^ssč^
- rychlý jak vítr^ssč^

4.
- Condition (coll.), having the meaning jestliže, když
- Jak budeš zlobit, nepůjdeš nikam^ssč^

Asi to sem patří, ale do které kategorie?

- *Japonskému turistovi upadla lžička, jak chtěl zmáčknout spoušť foťáku.*
- *Poslední šancí, jak se probojovat do finále, bude ...*
- *Stát to měl spravovat zvláštním ministerstvem (jak je tomu např. v Rakousku)*

- **jak-2 + Jˆ**
- In the phrase jak ... tak ... , having the meaning of i...i . However cf. jak-3 2.
- Byli tam jak odborníci, tak amatéři.
- **jak-3 + Db**
- Pronominal adverb

1.
- Interrogative - manner or extend (expr. jak pak).
- *Jak se jmenuješ?*
- *Jak je to možné?*
- Sometimes expressing large extend (often in exclamations).
- *Jak ten čas letí^ssč^*
- *Jak (pak) by ne^ssč^. Japa by ne.*
- *Líbí se ti to? - A jak!.*

2.
- Relative - marks subordinative adverbial clause (mostly manner expressing comparison, often with tak - however cf. jak-2 + Jˆ)
- *Jak řekli, tak udělali^ssč^*
- *tak dlouho, jak je možné (tak .., jak ..)*
- *Jak si kdo ustele, tak si lehne*

3.
- Relative (coll.) - meaning co, který
- ten člověk, jak jsem ti o něm říkal^ssč^

4.
- Indefinite
- buď jak buď (the verb is repeated)
- jak kdo, jak kde, jak kdy, etc. -

Kam s tím, je to asi Db, ale proč?

- *Jak se kůže sama obnovuje, postupně vylučuje ..*
- *?? Jak jsem chodil o berlích, tak jsem si zničil i druhé koleno.*

8.3 málo

- Similar to moc.
- **málo-1_ˆ(málo_+_2._p.,_málo_peněz) + Ca–c———-**

It has to be modified (in the shallow syntax) by a noun in genitive. Has only two forms:

- málo and mála (only in genitive).
- *Máme málo zájemců.*
- *bez mála peněz*
- *před málo lety^ssč^*
- *Je jen o málo důslednější.* - but *Je málo důsledný.* is *málo-3 (Dg)*

  - *Udělal to jako jeden z mála odborníků, ..*
  - *Udělal to jako jeden z mála. - ?? not modified by anything*
  - Udělal to jako jeden z mála, co přišli.

- **málo-2_ˆ(př._to_málo_co_měl) + NNNnc——A—-**
- *vystačit s málem*^{ssč}
- *vařit z mála*^{ssč}
- *Děkuji. - Za málo.* ^{ssč}
- **málo-3_ˆ(málo_+_příd._jm.,_př._byl_málo_důsledný) + Dg——-dA—-**
- *Málo mluví, hodně dělá.*^{ssč}
- *Je málo důsledný.*
- *Ve srovnání s loňskou sezónou je to velmi málo. - you can say méně.*
- *Zdržím se jen málo*^{ssč}.

8.4 moc

- Similar to málo.
- **moc-1_ˆ(nad_někým;_politická,_vojenská;_plná,...)**
- Obvious.
- *převzít moc*
- *moc proletariátu*
- *udělám, co je v mé moci*
- *mermo mocí*
- **moc-2_ˆ(mnoho_něčeho_[se_subst._v_gen.]) + `Ca--X----------`**
- Cannot be replaced by velmi. Can mean příliš, but is more colloquial. It has to be modified (in the shallow syntax) by a noun in genitive.
- *Má moc peněz.*
- *Všeho moc škodí.*
- **moc-3_ˆ(velmi,_ve_spojení_s_adj.,_př._moc_hezká) + Db**
- Can be replaced by velmi (except ellipses). Modifies an adjective, adverb or verb.
- *Je moc hezká.*
- *Vím to moc dobře.*
- *Moc se snažil.*
- Ve srovnání s loňskem je to moc. - ellipse.

8.5 proto

- **proto-1_ˆ(proto;_a_proto,_ale_proto,...) + Jˆ**
- Coordinative conjunction expressing consequence (implication). Structure: reason → consequence. Replaceable by tedy. Usually a proto or a ... proto
- *Nesplnil úkol, (a) proto nedostal odměnu.*
- *Každé proč má své proto.*
- *Německo se začalo dusit, a rozhodlo se proto omezit ...*
- Na začátku vět, bez a (to je tam implicitní)
- **proto-2_ˆ(dal_mu_co_proto,_tak_proto!) + Db**
- Pronominal adverb. Refers to the subordinative clause Structure: what → reason
- *proto, že: Udělal to proto, že musel.*
- *Udělal to proto, aby/že mu pomohl.*
- *co proto: dát někomu co proto; dostat co proto*
- no proto: Říkal, že tam přece jen půjde - No proto! (Sometimes classified as a modal particle)

8.6 svůj

- **svůj-1_ˆ(přivlast.) + P8gnc———v**
- Obvious.
- **svůj-2_ˆ(být_svůj) + AOgn———-v**
-
- *Vzít za své.*
- *Víme své. Víme svoje.*

8.7 tak

In general:

- replaceable by a proto ⇒ Jˆ

- replaceable by tím způsobem, stejně, zrovna ⇒ Db

  tak-2 + Jˆ

  Coordinative conjunction. If one of the clause is subordinative, tak has the meaning of an adverb: (Cf. Bál se, tak si pískal. - Jˆ vs. Kdyby se bál, tak si pískal - Db)

  1. - meaning (a) proto, tedy
     - *Bál se, (a) tak si pískal.*$^{\text{ssč}}$
     - *Neudělali..., příspěvek tak budou muset vrátit.*
     - *Byly zakázané, a tak přitahovaly*
     - *Zmizí bariéry, a tak bude možné využívat ..*
     - *Zpozdila se, a tak musela běžet.*
     - *Jsou profíci, tak ať se podle toho zařídí/*
     - *Počítá se s tím, že některé se sloučí, i tak bude třeba ..*

  2. in jak - tak

tak-3 + Db

1. - Refering to something known, to other sentence, etc.
   - *tak - jak: Bylo to tak, jak jsem myslel.*$^{\text{ssč}}$
   - *jak - tak: Jak řekli, tak udělali.*
   - *Přesně tak.*
   - *tak zvaný*
   - *Ať je to tak nebo tak ...*$^{\text{ssč}}$
   - *jen tak: Udělal to jen tak.*
   - *tak tak: Stihl to (jen) tak tak.*
   - *> to: Stalo se tak při ..*
   - *Tak se tehdy žilo*$^{\text{ssč}}$
   - Sub-Clause, tak Main-clause:
   - *Když - tak: Když jsem počítal já, tak mi vyšlo velké číslo.*
   - *Pokud - tak: Pokud to není diskriminace, tak nevidím důvod ..*
   - *Dokud se člověk raduje, tak je život pěkný.*
   - *Kdyby - tak: Kdyby/Pokud by se bál, tak by si pískal.*
   - (Cf. Bál se, tak si pískal. - Jˆ)

2. - Expressing amount (usually large) of a property, etc.
   - *Kam tak rychle?*$^{\text{ssč}}$
   - *tak jako: Je tak velký jako já.*
   - *Zmizel z povědomí tak jako jeho pomnik;*
   - *Nabízí se tak okatě, až je to hanba.*
   - *To je ale tak daleko .*
   - *tak vysoká; tak oslaben, že ...*
   - *Buďte tak laskav.*$^{\text{ssč}}$
   - *ani tak o ..., jako o ...: Nejde ani tak o mzdu, jako o ...*
   - *> přibližně: Dostane se na burzu asi tak třetí den od ..*
   - *hned tak: Hned tak nepřijde. (koneckoců)*
   - *odmítá to, stejně tak jako ...*
   - *.. a zrovna tak hyzdit;*
   - *tak jako tak*

Chapter 9

Selected words

strana. *na jedné straně ..., na druhé straně ...*: `druhý-1_^(jiný) strana-1_^(v_prostoru)`

nerespektované ze strany Israele: `strana-3_^(u_soudu,_na_úřadě,_smluvní_strany;_na_něčí_straně)`

stát. *stane se ministrem*: `stát-2_^(něco_se_přihodilo)`

s=to. *být sto něco udělat*: `sto-3_^(být_sto) / TT-------------`

vážit. *vážit cestu*: `vážit-1_:T_^(na_váze)` (similar to *zvažovat něco*; besides that, the only other possibility would be `vážit-2_:T_^(ctít_si_někoho)` but that verb is reflexive.

vedení. One of the lemma groups for which the morphological analyzer currently violates the rule that each lemma should be numbered. There are two variants, one unnumbered, and the other `vedení-1_^(*7ést-1)`. The unnumbered lemma is used only for *elektrické vedení* and similar uses. Otherwise the numbered variant should be assigned, including but not limited to: *pod vedením kamarádky, vedení podniku, čínské vedení*.

Chapter 10

Date and time

- *v* + a day: accusative (4) (*v sobotu, v neděli*)
- *v* + a month: locative (6) (*v lednu, v září*)
- *v* + an hour: accusative (4) (*ve 4 hodiny, v 6 hodin*)
- *ve dne*: locative (6) - `NNIS6-----A---9` - special kind of locative that occurs only in this context (*v noci* is also in locative)
- month in a date: genitive (2) (*25. září, 2. října*)

Chapter 11

Numbers, numerals and quantifiers

An adjective modifying a quantified expression agrees in case with the noun, not the numeral.

Example 11.0.1: Case agreement in counted phrases

- *za* (gen)
- *těch* (gen)
- *mizerných* (acc)
- *deset* (gen)
- *korun* (gen)

1x. Lemma equal to the form, e.g. `1x`. Tag `Cv-------------`.
4x5. It should be tokenized into three tokens, e.g. `4`, `x-5_^(náhr._symbolu_krát)`, and `5`.
tři stovky, dvacet tisíc lidí, necelých 9000

- *sto* and *pětiset* in *sto-, pětiset- a tisícikoruny*

  Not solved. The closest existing tag is the one of first parts of hyphenated adjectives (`A2--------A----`). But a lemma of a numeral should not have an adjectival tag.

- *Domníváme se, že* ***poslední*** *půl miliardy let udržuje...*

  What case should *poslední* get? Does it agree with *půl* (accusative), or with *miliardy* (genitive)? Solution: genitive should be preferred.

  za těch patnáct let: *patnáct* = accusative, *těch* = genitive.

- *Výsledkem bylo zase jen pár marek.* *pár* can be a numeral (`C...[2367]`) or a noun (`N...[14]`). But in this particular context, it should be `C` due to agreement with the predicate and `N` due to the nominative case. Solution: use `ClXP1----------`, the morphological analyzer must be adjusted.

Chapter 12

Hyphenated composites

If the hyphenated word ends with -o, and by a replacement of that -o by an adjective ending we obtain an adjective (normal or possesive), the lemma for the word is that adjective (e.g. *česko-německý - česko → český, Karlo-Ferdinanova - Karlo → Karlův*). Some words cannot be viewed as derived from adjectives, but rather from nouns (e.g. rap- jazzová - rap → rap vs. rapovo-jazzová - rapovo → rapový).

Currently the only tag for first parts of hyphenated compounds is `A2--------A----`. The tag set has to be extended by a similar tag for nouns. Otherwise, we would have to introduce two lemmas for each noun, one tagged normally as noun, the other as an adjective before a hyphen. (One lemma must not occur with more than one part of speech.) Of course, that would be extremely inconvenient.

Example 12.0.2: Hyphenated composites

- *srbsko-černohorská*: `srbský / A2--------A----`
- *Univerzita Karlo-Ferdinandova*: `Karlův_;Y_^(*3el) / A2--------A----`
- *Univerzita Karel-Ferdinandova*: `Karel_;Y / A2--------A----`
- *rap-jazzová*: `rap-2 / A2--------A----`
- *rapo-jazzová*: `rap-2 / A2--------A----`
- *rapovo-jazzová*: `rapový / A2--------A----`

Chapter 13

Insertion

If the possibilities offered by the morphological analyzer are not suitable, you have to insert new lemma and/or tag. If you insert a new lemma, you have to ensure that the lemma (lemma proper) you insert is not already used. That usually means adding unique numbers to distinguish lexical items having the same base form.

13.1 Possessive adjectives

Lemmas of possessive adjectives show how the get the noun they are derived from (see also Section 2.1.7). For example:

- `kardinálův_^(*2)` - remove two letters: `kardinál`
- `Karlův_;Y_^(*3el)` - remove 3 characters, add "el": `Karel`
- `Martinův-1_;Y_^(*4-1)` - remove 4 characters, add "-1": `Martin-1`

13.2 Words ending with *-ismus, -izmus*

The base form should use -ismus ending, the form using -izmus is treated as variant '1'. Currently some entries still do not follow this convention.

The examples show the desired state, in the current version of morphological analyzer they are regarded as separate lexical items (they have different lemmas).

Example 13.2.1: *-ismus, -izmus*

- *mechanismus*: `mechanismus / NNIS1-----A----`
- *mechanizmus*: `mechanismus / NNIS1-----A---1`
- *exhibicionismus*: `exhibicionismus / NNIS1-----A----`
- *nacionalizmu*: `nacionalismus / NNIS2-----A---1`

13.3 Transcription of pronunciation

Example 13.3.1: Transcription of pronunciation

vyslovujeme "zpjev"
"měly" se čte "mněli"

The lemma should be equal to the word form, the tag should be `NNXXX-----A----` even if transcribing pronunciation of words that are not nouns: `mněli_^(přepis_výslovnosti) / NNXXX-----A----`

13.4 Crippled forms

Some crippled forms very closely resemble the pronunciation category. In *Gaptschikowo*, pronunciation is modeled using German spelling. In *"řada lidí chybuje a píše 'poměnka'"*, the author points out a spelling error other people do. However, the author's intention to use the wrong form should be clear, otherwise it is the author's error that should be corrected.

If possible, the crippled forms should be tagged as if they were spelled the standard way; otherwise, use `NNXXX-----A----` or `AAXXX----1A----` according to the part of speech.

Example 13.4.1: Crippled forms

- *Waklaf Hafel*: `Waklaf_;Y_,t / NNMS1-----A---- // Hafel_;S_,t / NNMS1-----A----`
- *Gaptschikowo*: `Gaptschikowo_;G_,t / NNNS1-----A----`
- *v Gaptschikowo*: `Gaptschikowo_;G_,t / NNNXX-----A----`

13.5 Isolated morphemes

The lemma should be equal to the form, the tag should be `NNXXX-----A----`

Example: *ve slovech končících na -ství píšeme...*: `ství / NNXXX-----A----`

13.6 Geometry

In documents on geometric subjects, lots of "triangles ABC", abscissas (lines) PQ, RS, AB etc. occur. The identifiers of the objects are not abbreviations! Instead, a new lemma numbered 98 must be created for each. As always, no program should rely on the number being 98 but the annotators should keep the rule for the sake of improving human readability.

Example: *trojúhelník ABC*: `ABC-98_^(označení_pomocí_písmene)`

13.7 Chess codes

Records of Chess games appear occasionally in the data. They contain move descriptions in the Chess notation. Currently there are errors in tokenization; whole move (figure, target column and target row) should be one token. The lemma should equal to the code + `-1_:B_;w_^(šachový_tah)`. The tag should be `NNNXX-----A---8` (the neuter gender corresponds to the gender of *pole* (field)).

Example: *Jh8*: `Jh8-1_:B_;w_^(šachový_tah) / NNNXX-----A---8`