Manual for Morphological Annotation

Instructions for Annotators

Jiří Hana, Hana Hanová

In cooperation with
Jan Hajič, Barbora Hladká, Emil Jeřábek

CKL MFF UK Praha

Contents

PREFACE

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 introduced them a morphological analyzer for processing isolated words we use (as a preprocessing step), 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-analytical annotations. This way we replaced the absence of annotation guidelines by sequential elimination of discrepancies across both the morphological and syntactic-analytical 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-analytical 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 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.

We are grateful to Petr Pajas – this document “as it is” would not appear without his XML and LaTex skills.

Typographical conventions

Vertical bar on the outer side of the page is used to highlight comments we make or suggestions we propose.

Gray is used to highlight something what should be checked.

Chapter 1

Introduction

Sometimes, the writer uses the word incorrectly – e.g. a name of a woman as a name of a man, surname as a first name, etc. it is necessary to annotate the real usage not the should-be usage.

Maybe it should be somehow marked, if we encounter it.

To get an idea what a foreign name, etc. mean it is useful to try to find using an internet portal, in an encyclopedia, on a map, etc. During annotation, we have found the following internet links useful:

Portals

`http://www.seznam.cz` – for Czech products, companies
`http://search.seznam.cz/search.cgi?mod=f&hlp=y` – for Czech companies
`http://www.google.com`
`http://www.altavista.com` (shop section for various searching products)

Encyclopedias

`http://www.britannica.com`
`http://www.encyclopedia.com`
`http://www.encarta.msn.com`

Dictionaries

`http://dictionary.oed.com/entrance.dtl` – Oxford English Dictionary
`http://slovnik.seznam.cz` – various dictionaries

Maps

`http://mapy.atlas.cz` – Czechia
`http://www.mapquest.com/maps` – U.S.A and the world

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.

Note: There is a convention that if lemmas use numbers to distinguish lexical items with the same base form, they all have to use them- i.e. instead of sets of lemmas {X, X-1, X- 2} or {X, X-2, X-3}, there should be a set {X-1, X-2, X-3}

Note: The lemmas having different semantic suffixes should have different numbers.

In this manual we behave as the annotator. We try to mark such improper numbers by roman font (other part of the lemma is in italics). For example *stop* in *akce Stop million* will be marked as *stop*-1*_;m* and not *stop-1_;m*).

Table 2.1: Examples

Whole lemma	**Lemma proper**	**Second part**
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 Derivational Information

The morphological component used in PDT 1.0, handles only inflection, not derivations – it means lemmas are rather shallow. However, sometimes the lemma contains information about lemmas it is derived. 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.

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*

Other examples:

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

2.1.2 Semantic Information

Some lemmas (esp. names) contain suffixes expressing semantic information about their use, etc.:

G – geographical name: *Praha, Ústí nad Labem*
Y – given (first) name, formely used as default: *Petr, John*
S – surname (last name): *Dvořák, Zelený, Agassi, Bush*
E – name of a nationality: *Čech, Kolumbijec*
R – name of a product: *Tatra* (the car),
K – name of a company: *Tatra* (the company)
m – default – names of mines, stadiums, guerilla bases, etc; also used for functional words in names.

2.2 Tag Structure

2.2.1 Positional tags

A positional tag is a string of 15 characters. Every position 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
6	PossGender	Possessor’s gender

continued on next page

continued from previous page

Position	Name	Description
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 no value (e.g. tense for nouns).

Not all combinations of tag values are possible. There is about 4K tags [1].

Examples:

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

2.2.1.1 1 – Part of speech

Value	Description
A	Adjective
C	Numeral
D	Adverb
I	Interjection
J	Conjunction
N	Noun

continued on next page

[1] See also: `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/docc0pos.pdf`, for quick reference: `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html`.

continued from previous page

Value	Description
P	Pronoun
V	Verb
R	Preposition
T	Particle
X	Unknown, Not Determined, Unclassifiable
Z	Punctuation (also used for the *Sentence Boundary* token)

2.2.1.2 2 – Detailed part of speech

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

Table 2.4: SUBPOS

Value	Description	POS
#	Sentence boundary	
	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 2.4 SUBPOS #)	
=	Number written using digits	C – numeral
?	Numeral *kolik* (lit. *how many/how much*)	C – numeral
@	Unrecognized word form	X – unknown
^	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
6	Reflexive pronoun *se* in long forms (*sebe, sobě, sebou*, lit. *myself / yourself / herself / himself* in various cases; *se* is personless)	P – pronoun

continued on next page

continued from previous page

Value	Description	POS
7	Reflexive pronouns *se* (2.8 CASE = 4), *si* (2.8 CASE = 3), plus the same two forms with contracted *-s*: *ses, sis* (distinguished by 2.11 PERSON = 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 2.16 VAR (for *-ž*) and 2.11 PERSON (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

continued on next page

continued from previous page

Value	**Description**	**POS**
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	
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 2.14 NEGATION as well as the 2.13 GRADE 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

continued on next page

continued from previous page

Value	Description	POS
f	Verb, infinitive	V – verb
g	Adverb, forming negation (2.14 NEGATION set to A/N) and degrees of comparison 2.13 GRADE 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* (2.16 VAR = 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

continued on next page

continued from previous page

Value	Description	POS
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

Table 2.5: 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

2.2.1.3 3 – 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

2.2.1.4 4 – 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

2.2.1.5 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*
6	Locative, e.g. *ženě*
7	Instrumental, e.g. *ženou*
X	Any

2.2.1.6 6 – 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*

2.2.1.7 7 – Possessor's Number

Value	Description
P	Plural, e.g. *náš*
S	Singular, e.g. *můj*

2.2.1.8 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

2.2.1.9 9 – Tense

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

2.2.1.10 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*

2.2.1.11 11 – Negation

Table 2.14: NEGATION

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

2.2.1.12 12 – Voice

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

2.2.1.13 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 [2]

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^-------------`

[2] `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/compact_tags.pdf`

`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

Proper names (either directly or the lemmas they consist of) have suffixes marking the category of that name:

G – geographical name: *Praha, Ústí nad Labem*
Y – given (first) name, formely used as default: *Petr, John*
S – surname (last name): *Dvořák, Zelený, Agassi, Bush*
E – name of a nationality: *Čech, Kolumbijec*
R – name of a product: *Tatra* (the car),
K – name of a company: *Tatra* (the company)
m – default – names of mines, stadiums, guerilla bases, etc; also used for functional words in names.

The lemma should start with upper case if the word is always in upper-case in names (*Tatra* is always in uppercase, but *banka* not).

Keeping this categorization in the same level as lemmas is quite unsustainable and very unsuitable.

1. In theory every word can occur in any category. For example, *new* in *New York* (G), *New Jersey Devils* (sport club – K), *New Jersey Devils cards* (product – R), etc. Because it would explode the lexicon, usually common words (besides new and alike, all functional words) have only two lemmas: one for common words and one for all names (using default category m). But such approach is highly unsystematic and works only for a small corpus.
2. But even then, the system is not used consistently – some functional words even do not have the above mentioned two versions (normal and m). For example, *nad* in *Ústí nad Labem*, should be nad-2_;m, but there is no such a lemma. Similarly *a* in *a.s.* should be annotated differently when part of a name of some company and when not, but it is not.
3. Moreover lemmas having different categories are formally not connected, e.g. if you see Martin-3_;K, you do not know if it is derived from Martin-1_;Y (*Martin + Martin, s.r.o*) or Martin-2_;G (*DS Martin, a.s.*).

Proposed solution:

- The G, K, etc. categories should be independent of the morphology, and should be assigned to phrases on a different level. This would also require some enhancement to the annotation tool DA.
- Only the words that always (i.e. >90% or so) belong to some category would marked with that category (e.g. *new* has no special suffix, *England* has G).
- Personal name would be annotated as separate lemmas *Petr Pánvička* – Pánvička_;S not (*pánvička*)$_S$.

- Names containing name of a person, where the original link is not perceived (usually geographical names that do not contain possessive construction) have separate entries. N.B. that current guidelines require all the following words to be annotated with lemmas containing G (incl. *ostrov*, and *úžina*!!)

  Columbus (town in Ohio) – *Columbus-2_;G* not (*Columbus-1_;S*)$_G$
  Martin (town in Slovakia)- *Martin-2_;G* not (*Martin-1_;S*)$_G$
  Beringova úžina – (*Beringův_;S_^(*2) úžina*)$_G$ not *Beringův-2_;G úžina-2_;G*
  Ostrov Sergeje Kirova – (ostrov *Sergej_;Y Kirov-1_;S*)$_G$
  Kirov (town in Russia) – *Kirov-2_;G* not (*Kirov-1_;S*)$_G$

3.1 Personal names

Some names are sometimes declined, sometimes not (*Bill* – *o Bill Clintonovi*, *o Billu Clintonovi*, *o Billovi*). The tag for nondeclined form is `NgXXA`.

3.1.1 von, van, etc.

For names (e.g *Ludwig van Beethoven*) the *van*, etc. phrase is perceived as a surname – annotate it that way. For other it is still perceived as geographical name (e.g *Kryštof Harant z Polžic a Bezdružic*). Of course the borderline is fuzzy.

Examples:

Ludwig van Beethoven – *Ludwig_;Y van-2_,t_^(v_hol._jménech) Beethoven_;S*
Vincent van Gogh – *Vincent_;Y van-2_,t_^(v_hol._jménech) Gogh_;S*
Kryštof Harant z Polžic a Bezdružic – *Kryštof_;Y Harant_;S z-1 Polžice_;G a-1 Bezdružice_;G*
Brigida z Háje – *Brigida_;Y z-1 Háj_;G*

3.1.2 Chinese 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).

> That was the original recommendation, but probably annotating them as S would be better, because they are often used that way (You can say *Clinton for Bill Clinton*, but you cannot say *Po for Po Li*).

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). Some can be widely used, some can be unique. Often it is impossible 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):

Cai, Ceng-Zeng, Chen, Chen-Shen, Deng, Gao, Guo, He, Hu, Huang, Li, Liang, Lin, Lü, Ma, She, Sun, Tang, Wang, Wu, Xie, Xu, Yang, Ye, Zhang, Zhao, Zheng, Zhu

Links

`http://www.wlu.edu/ hhill/names.html` – Chinese names explained
`http://www.geocities.com/Tokyo/3919/atoz.html` – Alphabetical Index of Chinese Surnames (incl. Pinyin, Anglicized and other versions)

3.1.3 Korean names

Korean names behave similarly as Chinese names. Surname precedes given name. Given name of most Koreans consists of two parts, in Latin alphabet often connected with a dash. Most common Korean surnames are (45% of the population): *Kim*, *Lee* (often spelled as *Rhee, Yi* or *Li*), *Park*.

Examples:

Yang Sung-jin – S: Yang, Y: Sung, jin
Yang Sungjin – S: Yang, Y: Sungjin
Kim Il-Sung (former dictator of North Korea) – *S: Kim, Y: Ir, Sen*
Kim Ir-Sen (= Kim Il-Sung) - *S: Kim, Y: Il, Sung*
He Wung – S: He, Y: Wung

3.1.4 Foreignized Czech names

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

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 three different tags (gender M, F, X [1])

  Peter Janda – *Janda_;S* + `NNMXX-----A----` or `NNMS1-----A----`
  Jane Janda – *Janda_;S* + `NNFXX-----A----`
  Jane a Peter Janda – *Janda_;S* + `NNXXX-----A----`

- a name that has the same spelling as in Czech, should use the Czech lemma *Jane Janda* – *Janda_;S* + `NNFXX-----A----`

- a name with altered spelling has its own lemma (with ,t suffix) *Judy Sedivy* – *Sedivy_;S_,t* + `NNFXX-----A----`

[1] If {M,F} gender is introduced, the tag `NN{M,F}XX-----A----` should be used.

3.2 Compound names (names consisting of other names)

All lemmas of autosemantic words in compound names must have the category determined by the whole name (e.g. K, R). The lemmas of functional words contain default type category (m).

The problem is that a name of one type can occur as part of a name of a different type:

New England – G
New England Association of Chemistry Teachers – K
New England Association of Chemistry Teachers Journal – R

England is G noun in the first, K adjective in the second and R adj. in the third name.

If the lemma of the category you need does not exist and you have to insert a new one, do not care about numbering of lemmas, somebody else will do it (it would impossible to ensure that the numbers were unique across all annotators). That means, if there is other lemma having just different category (e.g. there is England_;G available, but you need England_;R), just change the category label.

> Using the above-proposed separation [2] of morphology and name categorization, the *New England* example would be annotated quite easily (only *England* is marked by a category (G) by the morphological analyzer, the rest is done by some other kind of tool):
>
> (new England_;G)$_G$
> ((new England_;G)$_G$ association of chemistry teacher)$_K$
> (((new England_;G)$_G$ association of chemistry teacher)$_K$ journal)$_R$
>
> If the annotator did not recognize the components of the name (e.g. it is in Burmese), (s)he would annotate just the highest level.

The categorization is sometimes quite tricky – you do not know, whether to consider a phrase a name or a name plus normal word:

Nobelova nadace – Nobelův_;K nadace_;K [3]
Nobelův stůl (e.g. in a museum) – Nobelův_;S stůl
Nobelova cena – hard to say (m vs. normal), decided: Nobelův_;S *cena*.

Examples:

Brownův pohyb – Brownův_;S
Cena J. Debrau – Debrau_;S *cena*
Mérieuxův ústav – Mérieuxův_;K *ústav* (Should be ústav_;K but is not)
Divadlo J. Grossmana – divadlo_;K J-4_:B_;K Grossman_;K
příloha Kolumbus (in Lidové noviny) – Kolumbus_;m
v *Dobrovského ulici nejezdí ...* – Dobrovský_;G
v *Dobrovského nejezdí ...* – Dobrovský_;G
poliklinika Dobrovského (unofficial, it is located in D. Street) – Dobrovský_;G

Using the separation of morphology and name categorization, this is quite easy:

Nobelova nadace – (Nobelův_;S nadace)$_K$
Nobelův stůl (e.g. in a museum) – Nobelův_;S *stůl*
Nobelova cena – easy to say: (Nobelův_;S cena)$_m$.

[2] See the beginning of 3.
[3] The lemmas have different numbers (e.g. *Nobelův-1_;S, Nobelův-2_;K*).

Examples:

Brownův pohyb – *Brownův_;S pohyb*
Cena J. Debrau – (*Debrau_;S cena*)m
Mérieuxův ústav – (*Mérieuxův_;S ústav*)$_K$
Divadlo J. Grossmana – (*divadlo J-0_:B_;Y Grossman_;S*)$_K$
příloha Kolumbus (in Lidové noviny) – (*příloha Columbus_;S*)$_m$
Dobrovského ulice – (*Dobrovský_;S ulice*)$_G$
v Dobrovského – (*Dobrovský_;S*)$_G$
poliklinika Dobrovského(unofficial, it is located in D. street) – (*poliklinika* (*Dobrovský_;S*)$_G$)$_K$

3.3 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*), and quite often you do not know if it is female or male (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.

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

> In our opinion, if the Y category were independent of the lemma, the horse name should be annotated correctly.

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.

3.4 Sport clubs, etc.

Name of the town in the club name: if only the town is noted, it is annotated as a geographic name (G), if the whole name of the club is noted, it is annotated as an institution (K). It is analogous to countries. (*Česko vs. Německo* are annotated as G)

Examples:

Cheb vs. Plzeň – *Cheb_;G Plzeň_;G*
SKP Union Cheb vs. Plzeň – *SKP_:B_;K Union_;K*[4] *Cheb_;K Plzeň_;G*

Of course, it can be a problem to know it with foreign clubs. If you do not know, annotate it as an institution (K).

Examples:

Chelsea – part of London, UK
Chelsea – *Chelsea_;G*
Chelsea FC – *Chelsea_;K FC-1_:B_;K_;w_^(...)*
Ferencvaros – part of Budapest, Hungary
Ferencvaros – *Ferencvaros_;G*

[4] In PDT 1.0, the lemma is *Union*, but it should *Union_;K*

Ferencvaros TC – Ferencvaros_;K TC-6_:B_;K
Sparta – Sparta-2_;K
Sparta Praha – Sparta-2_;K Praha_;K
Viktorie Žižkov – Viktoria-2_;K_^(jméno_sport.klubu) Žižkov_;K
Udinese – Udinese_;K_,t + `NNNXX-----A----`

It is the adjective of Udine (town in NE Italy), the official name of the football club is *Udinese Calcio* (*calcio = football*). However in Czech, the name is perceived as a noun and as the name of that club, therefore it is probably better to use it in that way:

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`) and also find the club (e.g. `http://www.soccerage.com`).

Using the above-proposed [5] separation of morphology and name categorization, this looks much more consistent:

Cheb vs. Plzeň – $(\text{Cheb\_;G})_K$ $(\text{Plzeň\_;G})_K$
SKP Union Cheb vs. Plzeň – $(\text{SKP\_:B\_;K Union\_;K Cheb\_;G})_K$ $(\text{Plzeň\_;G})_K$
Ferencvaros – $(\text{Ferencvaros\_;G})_K$
Ferencvaros TC – $(\text{Ferencvaros\_;G TC-6\_:B\_;K})_K$
Chelsea – $(\text{Chelsea\_;G})_K$
Chelsea FC – $(\text{Chelsea\_;G FC-1\_:B\_;K\_;w\_\^{}(...)})_K$
Viktorie Žižkov – $(\text{Viktoria-2\_;K\_\^{}(jméno\_sport.klubu) Žižkov\_;G})_K$
Sparta – Sparta-2_;K
Sparta Praha – $(\text{Sparta-2\_;K Praha\_;G})_K$
Udinese – $(\text{Udinese\_;G\_,t})_K$
Udinese Calcio – $(\text{Udinese\_;G\_,t calcio\_,t})_K$

The name of the sport club often contains some abbreviation. 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, entering them, suffixing the lemma by _:B_;K_;w and using `NNNXX-----A---8` as tag,

3.5 Other

Insisting on inclusion of name categories (K, R, etc.), implies explosion of number of lemmas. We follow each examples section by analogous examples using the above- proposed separation of morphology and name categorization (see 3.2).

3.5.1 Geographical names

Streets

We suppose that the word *ulice*, etc. is always present, even if elided on the surface.

Examples:

Dlouhá – dlouhý_;G+ `AAFS1----1A----`
Dlouhá ulice – dlouhý_;G+ `AAFS1----1A----` *ulice* + `NNFS1-----A----`
Palackého, *Dobrovského*, etc. – Palacký_;G, Dobrovský_;G+ `NNMS2-----A----`

[5] See the beginning of the 3.

Examples:

Dlouhá – $(dlouhý)_{G}$ + `AAFS1----1A----`
Dlouhá ulice – $(dlouhý\ ulice)_{G}$ or $(dlouhý)_{G}$ *ulice* + `AAFS1----1A---- NNFS1-----A-` `---`
Palackého, Dobrovského, etc. – $(Palacký\_;S)_{G}$, $(Dobrovský\_;S)_{G}$ + `NNMS2-----A----`

Towns

Words in one-word names consisting that were originally adjectives are annotated as nouns.

Examples:

Hluboká – *Hluboká_;G* + `NFS1`
Dobrá Voda – *dobrá_;G* + `AFS1` *Voda_;G^(součást_názvu_Odolena_Voda)* + `NFS1`
Ohrada u Hluboké – *Ohrada_;G* + `NFS1` *u_;m* + `RR2` *Hluboká_;G* + `NFS2`

Examples:

Hluboká – *Hluboká_;G* [6] + `NFS1`
Dobrá Voda – $(dobrá\ voda)_{G}$ + `AFS1 NFS1`
Ohrada u Hluboké – $(ohrada\ u\ Hluboká\_;G)_{G}$ + `NFS1 RR2 NFS2`

3.5.2 Initials

A separate character for aggregate gender {`M,F`} would be good (for initials following a letter in newspaper, an initial before a foreign last name, foreign names, etc.).

3.5.3 Institutions, companies

This category contains for example companies, foundations, shops, clubs, sport clubs, restaurants, etc. All autosemantic words in names of restaurants have lemmas with K. The exceptions are functional words that are annotated as default type (m)

Examples:

Porcela Plus: *Plus-3* + `TT`

Restaurants

Examples:

Bar Viola – *bar-2_;K, Viola-2_;K*
U Medvídků – *u-2_;m, medvídek-2_;K*
La cambusa – *Le-1_;m_,t_^(franc._člen_jako_souč._jmen_a_názvů)* [7]*, cambusa_;K_,t*
Restaurant HaPi – *restaurant-2_;K HaPi_;K*

[6] Frequent names of towns and names when POS changes, have separate entries. Therefore not $(hluboká)_{G}$

[7] In the current morphological lexicon, the m is missing.

Čínská restaurace Jin Jiang – čínský-2_;K, restaurace-2_;K, jin-2_;K, jiang-2_;K_,t
restaurace Jin Jiang – restaurace-1, jin-2_;K, jiang-2_;K_,t
Francouzská restaurace v Obecním domě – francouzský-2_;K, restaurace-2_;K, v-2_;m obecní-2_;K dům-2_;K
Hospůdka U vylitýho mrože – hospůdka-2_;K u-2_;m vylitý-2_;K mrož-2_;K

Examples:

Bar Viola – (bar, Viola_;Y)$_K$ or (bar, viola)$_K$ (select anyone, if you do not know the orig.)
U Medvídků – (u-1, medvídek)$_K$
La cambusa – (le-1_,t_^(franc._člen), cambusa_,t)$_K$
Restaurant HaPi – (restaurant, HaPi_;K)$_K$
Čínská restaurace Jin Jiang – (čínský, restaurace, jin, jiang_,t)$_K$
restaurace Jin Jiang – (restaurace, jin, jiang_,t)$_K$
Francouzská restaurace v Obecním domě – (francouzský restaurace, v-1 (obecní dům)$_K$)$_K$
Hospůdka U vylitýho mrože – (hospůdka, u-1, vylitý, mrož)$_K$

3.5.4 Sporting and other events

All events should receive special lemmas with m. However, if it is registered as a company and used in that meaning, then it should be K. If not certain use m.

Examples: [8]

Paris Indoor – Paris-2_;m_,t Indoor_;m_,t + `NNNXX-----A----`
US Open – US-3_:B_;m_,t + `AAXXX----1A---8` Open-1_,t_;m `AAXXX----1A----` [9]
akce Stop milión – stop-1_;m milión‘1000000_;m$_m$

Pohár mistrů – pohár_;m mistr_;m
Mistrovství světa – mistrovství_;m svět_;m

Examples:

Paris Indoor – (Paris-2_;G_,t Indoor_,t_;m)$_m$
US Open – (US-2_:B_,t_^(americký) Open-1_,t_;m)$_m$
akce Stop milion – (stop-1 milión‘1000000)$_m$
Pohár mistrů – (pohár mistr)$_m$
Mistrovství světa – (mistrovství svět)$_m$

[8] Many of these entries are not in the lexicon, therefore the actual numbers can be different once it is there. See note in 2.1, e.g. *mistrovství*: mistrovství-1, mistrovství-2_;m, mistrovství-3_;R, etc.
[9] We think, it is perceived as noun, probably inanimate, in Czech.

3.5.5 Televisions

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

Examples:

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

3.5.6 News, Magazines

All autosemantic word in names of news or magazines have lemmas with R. Currently, some of the newspapers are in the lexicon as institutions (e.g. *Sme*), this is not correct. Foreign names are often used as in plural, even if in the original there are in singular.

Examples:

Sme – Sme_;R_^(noviny) + `NNXX`
Zeitung – Zeitung-1_;R_,t_^(souč._názvu_něm._novin) + `NFPX` or `NISX`

3.5.7 Song names, etc.

Names of songs, TV programs etc. are annotated as normal words. The only reason is practical – it would cause explosion of the lexicon. If the categories and morphology are separated (see beginning of 3), these items can be annotated as R or m.

Chapter 4

Abbreviations

For discussion about inserting abbreviation not present in the morphological lexicon, see 10

4.1 Gender

Abbreviations can be used with different genders (e.g *ODS* – feminine (*strana*) or neuter). Any abbreviation can have neuter gender. If the gender cannot be disambiguated by the context, use the gender used elsewhere in article. If the author mixes genders or there are no disambiguating contexts, use the gender inherent gender of the abbreviation. In Czech, is usually easy to determine – it is the gender of the head of unabbreviated equivalent (e.g. *ODS* – *strana* → f). With foreign abbreviations it is much more problematic, different people use different genders (e.g. because of different translation). If you are not certain which of the gender is most widely used, use the default neutrum.

Examples:

UK – F (*univerzita*)
FBI (Federal Bureau of Investigation) – I (*úřad*), N (default or *byro*), F (probably á la *CIA*), *MP* (pl., referring to the members of the FBI)
CIA (Central Intelligence Agency) – F (*agentura*)

4.2 Normal abbreviations

Normal abbreviations have sometimes as a lemma the abbreviation (and sometimes the original unabbreviated word. Usually the former method is used for abbreviation that are more common then the unabbreviated word (and for abbreviation of multi word expressions). But it is not always true.

For discussion about determining the gender of an abbreviation, see 4.1

Examples:

např.: *například_:B* + `Db-------------8`
P.S.:
post-2_:B_,t_^(lat.,_po,_např._P.S.) + `RR--X----------8`
scriptum_:B_,t_^(př._P.S.) + `NNNXX-----A---8`
n.L.: *nad-1_:B* [1]+ `RR--7----------8`, *Labe_:B_;G* + `NNNS7-----A---8`

[1]Should be *nad-2_:B_;m*, but is not.

r. 1998: *rok_:B* + `NNIXX-----A---8`
r.: *režie_:B* + `NNFXX-----A---8`
rež.: *režie_:B*+ `NNFXX-----A---8`

4.3 Isolated letters

Note: The following is still not official.

Isolated letters (e.g. *A-konto*) are handled as abbreviations. The only exception is if they are not in the name (*zápas skupiny B*). Many of the annotations suggested bellow are still not offered by the morphological analyzer. Moreover, sometimes the morphological analyzer is constrained to offer appropriate lemma and tag only if the letter is followed by a dot. Should be repaired.

You have to select (or insert) the lemma according to the semantic category:

K-0_:B_;Y – first (and most middle) 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 – last (and some middle) names
k-8_:B_^(ost._zkratka) – other abbreviations (not names)
k-3_^(označení_pomocí_písmene) – other letters (not abbreviations, not in names)

Frequent abbreviations have their own lemmas, for example V – *V-1'volt_:B* or k: *ABC k.s.* – *komanditní_:B_^(jen_komanditní_společnost)*.

Tag selection (or insertion):

- noun: gender is known: NNgXX—–A—8 (g ∈ {`MFIN`})
- noun: gender is unknown: `NNXXX-----A---8`
- adjective: `AAXXX----1A---8` or AAgXX—-1A—8
- others: `X@-----------1` (variant of `X@-------------` for one letter words)

Examples:

A: A-mužstvo – *a-3_^(označení_pomocí_písmene)* + `AAXXX----1A----`
d: odst. 1 písm. d) – *d-3_^(označení_pomocí_písmene)* + `NNNXX-----A----`
A: 16 A – *A-1'ampér_:B* + `NNIXX-----A---8`
A: A konto (or *A-konto*) – *A-6_:B_;R* + `AAXXX----1A----`
a: ABC a.s. – *akciový_:B_^(jen_akciová_společnost)* + `AAXXX----1A---8`
s: na s. 128 – *strana-4_:B_^(v_knize,_rukopise,...)* + `NNFXX-----A---8`

1. It is hard to decide, whether an isolated letter is an abbreviation or a label using a letter (e.g. a-3). For example, *B* in *B-konto* can be from *bežný*, but *A* in *A-konto* probably means better than *B*. Maybe not, maybe yes, who knows. What is important, we mostly annotate texts written by people that do not know. Therefore it would be reasonable to merge these two possibilities together. Maybe annotate all single letters as abbreviation, the possible exception could be labels of paragraphs and cases (*odst. 1 písm. d*) or *za a)*.

2. Letters in similar configuration as nouns in noun cluster should be treated as nouns – they should be annotated as nouns. See also 6.1.1.
3. The category of a name (K, G, R, etc) and lemma selection should be orthogonal – see also 3.2.

   Examples:

   A: *A-mužstvo* – *a-8_^(ost._zkratka_nebo_označení)* + `NNNXX-----A---8`
   d: odst. 1 písm. d) – *d-3_^(př._odst._a,_za_a)* + `NNNXX-----A---8`
   A: 16 A – *A-1'ampér_:B* + `NNIXX-----A---8`
   A: A konto (or *A-konto*) – (*a-8_^(ost._zkratka_nebo_označení)...*)$_R$ + `NNXXX-----A---8`
   a: ABC a.s. – (*... akciový_:B_^(jen_akciová_společnost)*)$_K$ + `AAXXX----1A---8`
   s: na s. 128 – *strana-4_:B_^(v_knize,_rukopise,...)* + `NNFXX-----A---8`

4.4 RM-systém, samopal SA-58

An abbreviation preceding a noun is an adjective, an abbreviation following a noun is a noun. We would suggest to annotate them all as nouns (see 6.1.1). Does it mean that *HIV* in *HIV virus* and *virus HIV* have different POS.

Examples:

RM-systém – *RM_:B_;K* + `AAXXX----1A---8` `NNXXX-----A---8`
samopal SA-58 – *SA-2_:B_;R* + `NNXXX-----A---8`
virus AH 3 B -*AH-1_:B* + `NNXXX-----A---8`
virus HIV- *HIV_:B_;L_;U_^(lidský_virus_způsobující_AIDS)* + `NNXXX-----A---8`

4.5 Units of measurements

Units called after some males person (*V – volt*, *A – ampér*, etc.), have inanimate gender. However, units using degrees (°*C*, °*F*) have masculine animate gender, because the word *stupeň* is always present (even if omitted in the written text). Absolute temperature uses as the unit called *Kelvin* (K) not *degree of Kelvin*. Therefore the unit has inanimate masculine gender. However, if the author uses it errorneously as degree, the tag as to be masculine animate.

Examples:

C: Ráno byly 3 °*C.* – *Celsius_:B* – `NNMXX-----A---8`°
C: Ráno byly 3 C. (read as *Ráno byly tři stupně Celsia*) – *Celsius_:B* – `NNMXX-----A---8`
K: Teplota 5000 K. – *Celsius_:B* – `NNMXX-----A---8`
K: Teplota 5000 °*K.*- *Celsius_:B* – `NNMXX-----A---8`°

If the *C* character is preceded by some character trying to look as the degree symbol ° (eg. -*C*, *o C*, *O C*), then you should mark it as an error – as lemma insert the degree[2] symbol ° and as tag `X@-----------1`. It should be converted into a punctuation mark.

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

4.6 Authors abbreviations

The author's name abbreviations used in newspapers (e.g. *Ber, mas, jst, ...*) have lemma as the form + *-99_:B_;S* and tag `NNXXX-----A---8`. There is X for gender because usually we do not know it. If the {M,F} gender is introduced, it should be used here. These abbreviations are not present in the lexicon, therefore you have to insert them.

Examples:

ač: PRAHA (ČTK, ač) Problém Gabčíkova ... – ač-99_:B_;S – `NNXXX-----A---8`
gap: DUKOVANY (gap) Na základě posudku ... – gap-99_:B_;S`NNXXX-----A---8`

4.7 Academic titles

Titles distinguish genders – there has to be one lemma for men, and one lemma for women (*JUDr-1_:B_^(doktor_práv)* vs. *JUDr-2_:B_^(doktorka_práv)*); to keep it consistent the masculine has number 1, the feminine has number 2. We think, the titles should have the same form for women and men. Just the tag should be different, with possibility to have X if the gender is not known (e.g. a letter subscribed as Dr. A. B.)

Chapter 5

Colloquial Czech

If an official alternative to the colloquial form exist, then the the colloquial form has the same tag except a different variant ('5', '6', '7', ev. '3' – see 2.2.1.13).

Examples:

které: stavení, které – `P4NP4---------5`
Novákovic: Novákovic pes – Novákův_;S_^(*2) -`AUXXXM--------6` [1]
takovejhlema: takovýhle – `AAFP7----1A---6`
hovadinama: hovadina – `NNFP7-----A---6`
naší: pro naší atletiku (officially short: naši) – můj_^(přivlast.) – `PSFS4-P1------6`

5.1 Cos, jaks, kdys ...

We tagged these words as if they were without -s and added -9 at the end.

> In our opinion it would be better to divide such an expression in two words (e.g. cos → co + být, analogous to *abych* → *aby* + *být*) and tag them like two normal words, just with some variant recognize it.

5.2 Suffix -é in plural of neuter

Should not be treated as misspelling, but annotated as (colloquial) variant of official -á forms (variant '5').

Examples:

které: stavení, které – `P4NP4---------5`

[1] In PDT 1.0, this is sometimes obsoletely annotated as `AUMS1M--------6` or `NNXXX-----A---6`

Chapter 6

Foreign words and phrases

General rule

1. For a longer phrase (or citations) in a foreign language, use morphology of that language (but distinguish genders M and I ??) (Hence citation use).
2. For a single word or shorter phrase use Czech morphology. (Hence word use) The borderline is fuzzy, of course.

6.1 Part of speech

Many foreign words used in Czech sentence can have different part of speech than in their original language. Usually the hint is how it behaves in different context, if it is declined as a noun, if agrees with its head, etc.

All foreign lemmas have _,t suffix.

It would be good to somehow distinguish foreign words in word or citation use.

6.1.1 English noun clusters

All nouns in attributive use are annotated as adjectives.

> That's quite problematic:
>
> 1. Virtually all English nouns can be used as attributes of other nouns
> 2. It is imported to Czech: *Staropramen Extraliga*, *Český Telecom Cup*, etc.
>
> We think, it should be annotated as two nouns.

6.1.2 Examples

V kostele XY zpívala Musica Bohemica.
Bohemica annotated as a noun; in Latin it is an adjective.
Reason: When the phrase is declined, Bohemica is declined as a noun (*žena*): *pozvali Musicu Bohemicu*, **pozvali Musicu Bohemicou*
Annotation: *Musica_,t_;K* + `NFS1A`, *Bohemica_,t_;K* `NFS1A`
To je trochu ad hoc.
hoc is annotated as a noun; in Latin it is an adverb.
Annotation: *ad_,t* `RRX`, *hoc_,t* `NXXXA`

In the following, the section headers refer to the categories of the foreign language.

6.2 Articles

English *an* should be a form of *a*.

Articles merged with a preposition (fra *du*, ita *della*, deu *im*, *aufs*, *zur*) are treated as prepositions (?Split into two words?)

Arabic short words (##)(?articles, ?prepositions) are treated as articles.

6.2.1 Citation use

Same as single words

Should distinguish gender, number, and/or case Therefore: `TTgnc` or `AAgnc` ??

6.2.2 Word use

Tag: `TT-------------`
Lemma: Usually the same as the form

Originally, we wanted to treat articles as adjectives. Forms having different gender, number and/or case, would have the same lemma (*der* for forms *der, die, das, des, dem, den*). The problem is that Czech does not respect the original categories (*nebezpečný La Manche* – *la* in French F, in Czech the phrase is I; *Los Angeles* – in Spanish pl, in Czech sg.)

6.2.3 Examples

l' *l-5_,t_^(př._l'Arc,_stažený_tvar_fr._členu)* + `TT`
L' *L-10_^(př._L'Aqua,_stažený_tvar_fr._členu)* + `TT` (should be m [1])
la *la-2_,t* + `TT`
il *il_,t_^(it.__len)* + `TT`
as *as_,t* + `TT` (Arabic)
al *al_,t* + `TT` (Arabic)
el *el_,t* + `TT` (Arabic)
della *della_,t* + `RRX` (sometimes incorrectly annotated as `AAXXX`)
am *am_,t* + `RRX`

6.3 Nouns

6.3.1 Citation use

6.3.1.1 English

To keep it simple, number of English nouns is annotated in the same way in citation use as in word use. That means X is used instead S for nouns in singular. The difference is in X is used instead S for nouns in singular. The difference is in cases – in citation use, the case is always X, but word use it can sometimes be declined.

[1] If the name categories and lemmas were independent, everything annotated as L-10 would become l-5 (see 3.2)

6.3.2 Word use

6.3.2.1 English

The nouns in singular in English have number annotated as X (English singular. is often used in Czech as plural). For nouns that are usually declined mark the case even if in base form, for nouns that are nondeclined mark it as X.

Examples:

flow: oba dva cash flow (oficiální i skutečný) ... – *flow_,t* – `NNIXX-----A----`
statement: v cash flow statementu ... – *statement_,t* – `NNIS6-----A----`
statement: Náš cash flow statement ... – *statement_,t* – `NNIS1-----A----`
flow: Náš cash flow ... – *flow_,t* – `NNIXX-----A----`
girl: Beatles: Girl – *girl_,t* – `NNFXX-----A----`
girls: A teď zahrajeme písničku Girls. – *girl_,t* – `NNFXX-----A----`

6.4 Verbs

6.4.1 Citation use

6.4.1.1 English verbs

The following tags are applied:

- Present non3sg: *go* `VB-X---XP-AA---`
- Present 3sg: *goes* `VB-S---3P-AA---`
- Imperative: *go* `Vi-X---X--A----`
- Infinitive: *go* `Vf--------A----`
- Past tense: *went* `Vp-X---XR-AA---`
- Passive participle *gone* `Vs-X---XX-AP---`

If it is hard to determine the base form usage, annotate it as infinitive. If it is hard to decide between past tense and passive participle, use past tense. In PDT 1.0, most of the verbs using base form were annotated using the default – infinitive.

Examples:

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

6.4.2 Word use

Usually the tags and lemmas are the same as in citation use.

6.5 Slovak language

If a Slovak word has the same form as corresponding Czech one (e.g. prepositions), you should annotate it as if it were Czech. Otherwise it has to be annotated as any other foreign language.

Chapter 7

Errors

The text can contain errors. It is reasonable to correct some of them, preserving the original form. However, only low-level errors – spelling and morphology should be corrected (We do not want to correct Engels' text into Heidegger's). Never correct a colloquial form by an official one (e.g. *zelené města* *$\rightarrow$ *zelená města*, *bez noh** *bez nohou*), even if the analyzer does not know the form [1].

The errors have to be just marked, do not edit the file. Try to insert lemma and tag as if the form is correct, and use the DA support for marking errors – it inserts the text "(*Chyba*)" at the end of lemma or tag. If the lemma is correct, insert it after the tag, otherwise insert it after the lemma, if you do not know just insert it somewhere. If you want to add some comment, write it before the closing parenthesis, preceded by a dash (e.g. (*Chyba-nad c by měla být čárka, ne háček*)). This convention makes it easy to find the errors automatically.

7.1 Characters

Sometimes, foreign characters had been be screwed (e.g. *Fran?oise*), and therefore the morphological analyzer did not recognize the whole word. Mark it as a lemma error (do not edit the file), it has to be corrected and run thru the analyzer once more. There is a problem with letters not contained in Latin 2, they should be replaced by corresponding characters without diacritics. In the future, Unicode (2 bytes or UTF) should be considered.

7.2 Separators, etc.

Sometimes, the text contains *o* or *I* as bullets or separators. They should be marked for deletion (Press L (de**l**ete) in the lemma or tag list).

[1] You have to insert a new lemma and/or tag – see 10 for more details.

Chapter 8

Hard to decide

8.1 až

až-1 + `J^`
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 replaceable 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_^(živočich) + `NNMnc-----A----`
Obvious.
jak-2 + `J,`

1. Meaning *že* (cannot be replaced by *jakpak*)

   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]

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

jak-2 + `J^`
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* + `J^`)
   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,

??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_^(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_^(př._to_málo_co_měl) + `NNNnc-----A----`
vystačit s málem[ssč]
vařit z mála[ssč]
Děkuji. – Za málo. [ssč]
málo-3_^(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_^(nad_někým;_politická,_vojenská;_plná,...)
Obvious.
převzít moc
moc proletariátu
udělám, co je v mé moci
mermo mocí
moc-2_^(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_^(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_^(proto;_a_proto,_ale_proto,...) + `J^`
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 ...
proto-2_^(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_^(přivlast.) + `P8gnc---------v`
Obvious.
svůj-2_^(být_svůj) + `AOgn----------v`
Problem with tags, analyzer probably needs update.
Vzít za své.
Víme své. Víme svoje.

8.7 tak

In general:

- replaceable by *a proto* ⇒ `J^`
- replaceable by *tím způsobem, stejně, zrovna* ⇒ `Db`

tak-2 + `J^`
Coordinative conjunction. If one of the clause is subordinative then *tak* has the meaning of an adverb: (Cf. *Bál se, tak si pískal.* – `J^` vs. *Kdyby se bál, tak si pískal* – `Db`)

1. A consequence — meaning *(a) proto, tedy*

   Bál se, (a) tak si pískal.[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. A conjunction — in *jak – tak*

tak-3 + `Db`

1. Refering to something known, to other sentence, etc.
   tak – jak: Bylo to tak, jak jsem myslel.[ssč]
   jak – tak: Jak řekli, tak udělali.
   Přesně tak.
   tak zvaný
   Ať je to tak nebo tak ...[ssč]
   jen tak: Udělal to jen tak.
   tak tak: Stihl to (jen) tak tak.
   to: Stalo se tak při ..
   Tak se tehdy žilo[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.* – `J^`)

2. Expressing amount (usually large) of a property, etc.
   Kam tak rychle?[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.[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

Sólokapři

Hradec Králové

Králová_;G_;S_^(Dvůr_Králové) + `NNFS2-----A----` It is *hradec* that belonged to *králová*

strana

na jedné straně ..., na druhé straně ...: *druhý-1* (jiný), *strana-1_^(v_prostoru)* *nerespektované ze strany Israele*: *strana-3_^(u_soudu, ..*

stát

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

s=to

být sto něco udělat lemma = *sto-3_^(být_sto)*, znacka `TT-------------`

tudíž

always J^ ##Why is there the other possibility

vážit

vážit cestu – vážit na váze? nebo *ctít někoho*

vedení

Everything except *elektrické vedení* type, is considered as form of *vedení-1*

Examples:

*pod vedenim kamarádky – vedení-1_^(*7ést-1)*
*vedení podniku – vedení-1_^(*7ést-1)*
*čínské vedení – vedení-1_^(*7ést-1)*
elektricke vedení – vedení

9.1 Date and time

v +

a day – accusative (4): *v sobotu, v neděli*
a month – locative (6): *v lednu, v září*
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*

9.2 Numbers, numerals and quantifiers

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

Examples:

za (gen) *těch* (gen) *mizerných* (acc) *deset* (gen) *korun*
Deset (nom) *nejlepších* (gen) *sportovců* (gen) *ukázalo*

1x

Lemma: as form. Tag `Cv-------------`

Example

1x – 1x + `Cv-------------`

4x5

Should be split into three parts. E.g. *4x5 → 4, x, 5*

9.3 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 word cannot be viewed as derived from adjectives, but rather from nouns (e.g. *rap- jazzová – rap → rap* vs. *rapovo-jazzová – rapovo → rapový*). However, the lemma of that noun cannot be used as a lemma for the hyphenated form, and a new lemma (having different number) has to be introduced.

> That is extremely inconvenient (padlý na hlavu) – virtually any noun can be used in such a context, therefore for every noun, there have to be two lemmas – one for normal usage and one for hyphenated usage. We strongly suggest to allow any noun to have a hyphenated form in several variants – at least the bare base form and form ending in -o (variant ’1’).

Examples:

srbsko-černohorská – *srbský* – `A2--------A----`
Univerzita Karlo-Ferdinandova – *Karlův_;K_^(*3el)* -`A2--------A----` [1]

[1] Better: *Karlův_;Y_^(*3el)* – `A2--------A----`. See 3.2

Univerzita Karel-Ferdinandova – *Karel-2_;K* – `A2--------A----`
rap-jazzová: *rap-2* – `A2--------A----`
Better: *rap* – `A2--------A----`
rapo-jazzová: *rap-2* – `A2--------A----`
Better: *rap* – `A2--------A---1`
rapovo-jazzová: *rapový* – `A2--------A----`

Chapter 10

Insertion

If the possibilities offered by 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.

10.1 Possessive adjectives

Lemmas of possessive adjectives show how the get the noun they are derived from (see also 2.1.1). 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*

Examples:

*premiérův_^(*2)*
*Sorosův_;S_^(*2)*
*chlapcův_^(*3ec)*
*Švehlův_;S_^(*2a)*
*Máchův_;S_^(*2a)*
*Hlinkův-1_;S_^(*4a-1)*
*Benderův-1_;S_^(*4-1)*

10.2 Words ending with -ismus, -izmus

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

Examples: [1]

mechanizmus: mechanizmus – `NNIS1-----A----`
mechanismus: mechanizmus – `NNIS1-----A---1`
exhibicionismus: exhibicionismus – `NNIS1-----A----`

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

exhibicionismus: exhibicionismus – `NNIS1-----A---1`
nacionalizmus: nacionalizmu – `NNIS1-----A----`
nacionalismus: nacionalizmus – `NNIS1-----A---1`

10.3 Strange and unique things

Transcription of pronunciation

Lemma: as the form, tag: `NNXXX`

Examples:

vyslovujeme "zpjev" – zpjev + `NNXXX`

Isolated morphemes

Lemma: as the form, tag: `NNXXX`

Examples:

...ve slovech končících na -ství píšeme...: ství + `NNXXX`

Geometry

We can meet an article about a geometric theme sometimes. It means, that there occur a lot of triangles ABC, abscissas (lines) PQ, RS, AB and so on in that article. It is necessary to create a new lemma ending 98 for every mentioned figure.

Chess codes

Lemma: The code + *-1_;w*. Tag `NNNXX-----A---8` (neuter because *pole* is neuter)

Example

Jh8 – *Jh8-1_;w* + `NNNXX-----A---8`

Crippled forms

Lemma: the same as the form + *_,t*
Tag: normal if possible, otherwise `NNXXX / AAXXX` according to the POS

Examples:

Waklaf Hafel – Waklaf_,t + `NNMS1`, Hafel_,t + `NNMS1`
Gaptschikowo – Gaptschikowo_,t + `NNNS1`
v Gaptschikowo – Gaptschikowo_,t + `NNNXX`

10.4 Other

This section contains especially examples of previously inserted lemmas/tag. Some of them are already in the dictionary, however they mainly serve as an inspiration, when inserting similar things.

ad hoc

ad-x_,t + `RRX`, *hoc-x_,t* + `N`

pele-mele

For example as a heading in a newspaper *pele* + `TT`, *mele* + `TT`

zprostředkovací vs. zprostředkovat

The morphology is rather shallow. It means, for example, that lemma for *zprostředkovací* is *zprostředkovací* (precisely *zprostředkovací_ ^(*2t)*), and *zprostředkovat* as it was in the past.

Chapter 11

Errors in PDT 1.0

- *HaDivadlo_;Y* → *HaDivadlo_;K*
- Theaters – most of the theaters do not have K category (search for *Divadlo* case sensitive): *Divadlo v Celetné, Divadlo Husa*, etc.
- S/NWS/1993/mf930701:105-p4s1 – *los* should be *los-2*
- *které* should be `P4NP4---------5` not corrected as an error. (look for *w <spell>které*)
- <s id="S/NWS/1992/lnd92251:095-p1s1">: *US* should be with m (U*S-3_:B_;m_,t*) not K.
- *Novákovic, Perotovic*, .. should be `AUMS1M--------6`, but it is either AUgncM———6 or `NNXXX-----A---6` (e.g. S/NWS/1992/lnd92254:051-p5s8)
- S/NWS/1992/lnd92258:077-p91s1

	cash	**flow**	**statement**
Is	*cash_,t* `AAXXX----1A----`	*flow* `NNFXX-----A----`	*statement* `NNIS1-----A----`
Should be	*cash-2_,t* `A2--------A----`	*flow-2_,t* `AAXXX----1A----`	*statement_,t*
Should be	*cash_,t* `NNIXX-----A----`	*flow_,t* `NNIXX-----A----`	