



Prague Dependency Treebank: Morphological Annotation

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz

Basic terms



- ***wordform / word form / form***

~ every string of letters that forms a "word" of a language

e.g.: *pencil, pencils, where, writes, written;*
ženou, píšícím

Basic terms



- **wordform / word form / form**

~ every string of letters that forms a "word" of a language

e.g.: *pencil, pencils, where, writes, written;*
ženou, píšícím

- **(morphological) lemma**

~ base form: **infinitive** for verbs

nom. sg. for nouns, numerals

nom. sg. masc. for adjectives

? pronouns

mně → *já*; *ona* → *ona* | ?*on*; *se* → *se*;

jeho → *on* | *jeho*; *jejich* → ?*jeho*; *svého* → *svůj*;

ta → *ta* | ?*ten*; *týmž* → *týž*; *koho* → *kdo*;

kdečím → *kdeco*



Basic terms

- **wordform / word form / form**
~ every string of letters that forms a "word" of a language
e.g.: *pencil, pencils, where, writes, written;*
ženou, píšícím
- **(morphological) lemma**
~ base form: **infinitive** for verbs
nom. sg. for nouns, numerals
nom. sg. masc. for adjectives
? pronouns
- **paradigm**
~ a set of forms created by means of **inflection** from a base form
e.g.: *psát* → {*psát, píšu, píši, píšeš, píše, píšeme, píšem, píšete, píšou, píší, psal, psala, psalo, psali, psaly, piš, pišme, pište, píšíc, píšíce, nepsat, nepíšu, ...*}



Basic terms

- **wordform / word form / form**

~ every string of letters that forms a "word" of a language

e.g.: *pencil, pencils, where, writes, written;*
ženou, píšícím

entry of a
morphological
lexicon

- **(morphological) lemma**

~ base form: **infinitive** for verbs

nom. sg. for nouns, numerals

nom. sg. masc. for adjectives

? pronouns

- **paradigm**

~ a set of forms created by means of **inflection** from a base form

e.g.: *psát* → {*psát, píšu, píši, píšeš, píše, píšeme, píšem, píšete, píšou, píší, psal, psala,*
psalo, psali, psaly, piš, pišme, pište, píšíc, píšíce, nepsat, nepíšu, ...}

Basic terms (cont.)



- **lexical unit** ... cz: (základní) lexikální jednotka, lexie
~ an abstract unit associating the paradigm (represented by the lemma) with a single meaning;
i.e., '**a given word in a given sense**'

- lemma: *write*
- paradigm: {*write, writes, writing, written, wrote*}

- gloss: *to make a record using letters*
- syntax: *sb writes st for sb*
- semantics: *agens creates a text for a receiver*

lexical unit

Basic terms (cont.)



- **lexeme**

~ set of (semantically related) lexical units that share the same paradigm

entry of a syntactic
/ valency lexicon

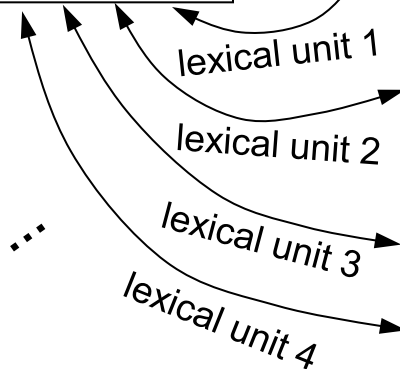
- lemma: *write*
- paradigm: {*write, writes, writing, written, wrote*}

- gloss: *to make a record using letters (for sb)*
- syntax: *sb writes st for sb*
- semantics: *agens creates a text for a receiver*

- gloss: *to send a message (to sb) via a letter*
- syntax: *sb writes to sb about st*
- semantics: *agens sends a letter to a receiver*

...

...



'Golden rule' of morphology

lemma A \longrightarrow forms a_1, \dots, a_n

lemma B \longrightarrow forms b_1, \dots, b_m

different words with different wordform(s)



lemma + tag ... together should uniquely identify the word form

'Golden rule' of morphology

lemma A \longrightarrow forms $a_1, \dots a_n$
lemma B \longrightarrow forms $b_1, \dots b_m$

different words with different wordform(s)



lemma + tag ... together should uniquely identify the word form

lemma A
lemma B $\searrow \swarrow$ forms $c_1 \dots c_n$



different words with one or more shared form(s) ... ***homographs***

lemma C $\searrow \swarrow$ forms $c_1, \dots x, \dots c_n$
forms $c_1, \dots y, \dots c_n$



one lemma with different paradigms ... ***variants***

Variants



- those wordforms that
 - belong to the same lexeme and
 - values of all their morphological categories are identical

e.g.: *colour* / *color*;
okénko / *okýnko* / *vokýnko*;

lemmas as representatives
of whole paradigms
! affect ***the whole paradigm*** !



global variants

≠

lemma variants

got / *gotten* (as past participle);
lesu / *lese* (as locative singular)

wordforms of the same lemma,
with the same morph. properties
! affect only ***some wordform(s)*** !

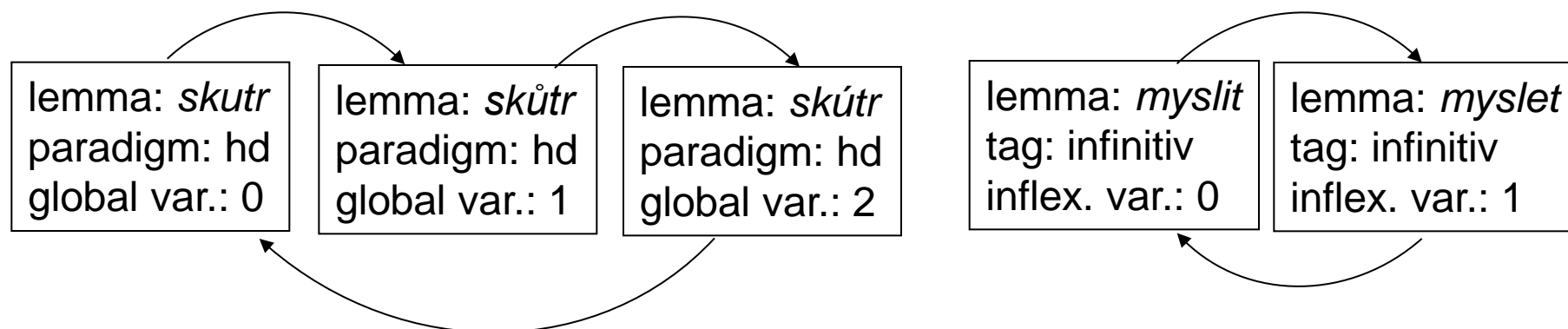


inflectional variants



Variants (cont.)

- different wordforms ... have to be distinguished
 - either by their lemma
 - or by their morphological tag \Rightarrow standard solution position for variants
- **BUT** lemma variants imply two (unrelated) entries in a lexicon
 \Rightarrow ? possible solution ... linking of lemma variants



Corpus query [lemma="skutr"] \longrightarrow all forms for all three lemmas {*skutr*, *skůtr*, *skútr*}

Homographs



- those wordforms that
 - have identical orthographic lettering, i.e. the identical strings of letters (regardless of their phonetic forms)
 - meanings of which are (substantially) different and cannot be connected
- e.g.: *pen* ~ writing instrument
~ enclosure
~ swan
- bank* ~ bench
~ riverside
~ financial institution



Inflectional homographs

~ homography affects only **particular wordforms**

+ **at most one** homographic word form is a **lemma**

(1) **syncretism** ~ wordforms with

- the same lemma and
- different morphological tags

<i>stopped</i>	<ul style="list-style-type: none">• past tense• past participle
----------------	--

<i>hradu</i> [castle]	<ul style="list-style-type: none">• genitive singular• dative singular
--------------------------	---

(2) identical wordforms with

- different lemmas

<i>smaž</i> imp.	<ul style="list-style-type: none">• <i>smazat</i> [to erase]• <i>smažit</i> [to fry]
------------------	---

<i>ženu</i>	<ul style="list-style-type: none">• acc sg. <i>žena</i> [woman]• 1. pers. sg. pres. <i>hnát</i> [to rush]
-------------	--



Inflectional homographs

~ homography affects only **particular wordforms**

+ **at most one** homographic word form is a **lemma**

(1) **syncretism** ~ wordforms with

- the same lemma and
- different morphological tags

⇒ homographic wordforms
belong to one lexeme

(2) identical wordforms with

- different lemmas

⇒ two different lexemes

'Golden Rule of Morphology':

<lemma, morphological tag> = unique wordform

Global homographs



~ homography affects **all wordforms of a paradigm**

⇒ the **same lemma** represents two / more **different lexemes**

<i>flower</i>	<ul style="list-style-type: none">• noun• verb
---------------	---

<i>nakupovat</i>	<ul style="list-style-type: none">• [to buy]• [to heap]
------------------	--

<i>žít</i>	<ul style="list-style-type: none">• [to live]• [to mow]
------------	--

**two wordforms with
the same lemmas and
morph. properties**

(1) either their paradigms differ

<i>flower</i>	• <i>flowers</i>
<i>flower</i>	• <i>flowered</i>

<i>žít</i> [to live]	• <i>žil</i> for past tense
<i>žít</i> [to mow]	• <i>žal</i> for past tense

(2) or they are derived from different words

<i>odrolovat</i> [to roll away]	• <i>od-rol-ovat</i>
<i>odrolovat</i> [to crumble]	• <i>o-drol-ovat</i>

Global homographs (cont.)



Standard solution:

- no morphological category can distinguish them
➡ necessary to ***distinguish lemmas***

žít-1 [to live]
žít-2 [to mow]

nakupovat-1 [to buy]
nakupovat-2 [to heap]

flower-1 as a noun
flower-2 as a verb

stát-1 [the state]
stát-2 [to stand]



Homography vs. polysemy

- **homography** ~ wordforms with identical orthographic lettering with **(substantially) different meanings**



it concerns **separate lexemes**

- **polysemy** ~ a single word having **two / more related meanings**



usually treated within **a single lexeme**

! No clear cut between polysemy and homography !

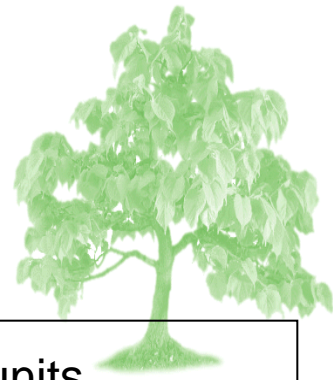
hradit [to fence]

• one polysemic lexeme with two lexical units (SSJČ)

hradit [to reimburse]

• homographic lemma, i.e. two lexemes (SSČ)

Homography vs. polysemy



hradit [to fence]
hradit [to reimburse]

- one polysemic lexeme with two lexical units

žít-1 [to live]
žít-2 [to mow]

- two lexemes represented by lemmas *žít-1*, *žít-2*

odpovídat [to answer]
odpovídat [to react]
odpovídat [to be responsible]
odpovídat [to correspond]

- one polysemic lexeme with four lexical units

stát-1 [the state]
stát-2 [to stand], [to cost]
stát-3 (se) [to happen]
stát-4 [to melt]

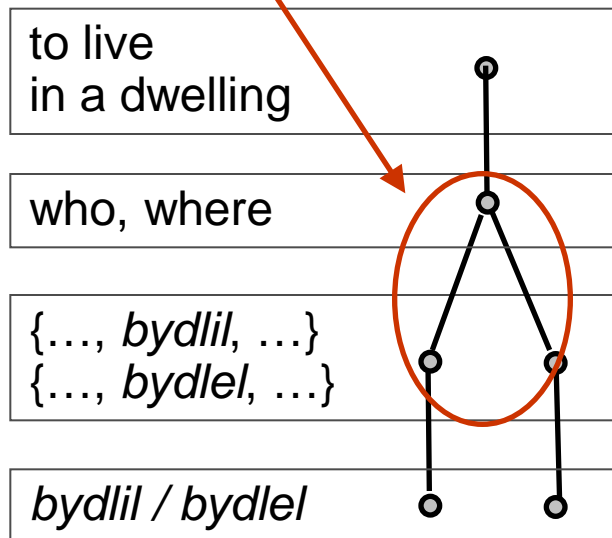
- four lexemes with four different paradigms

Duality of variants and homographs

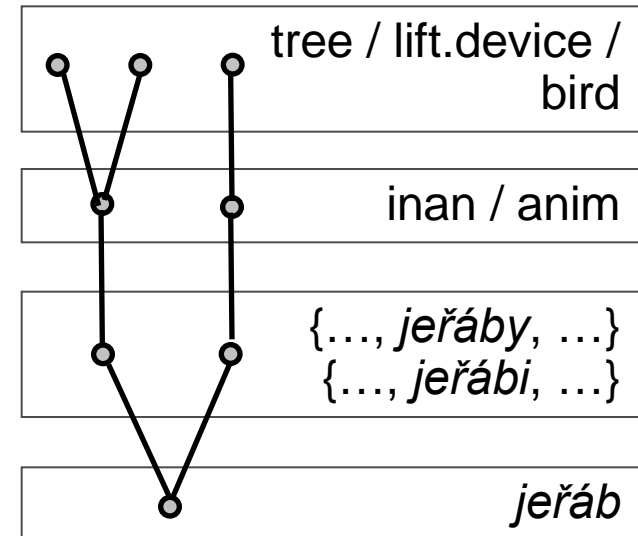


Schema of

variants for the example *bydlit / bydlet* homographs for the word *jeřáb*



meaning
syntactic /
semantic features
paradigms
(set of wordforms)
lemmas
(orthographic variants
of lemma)

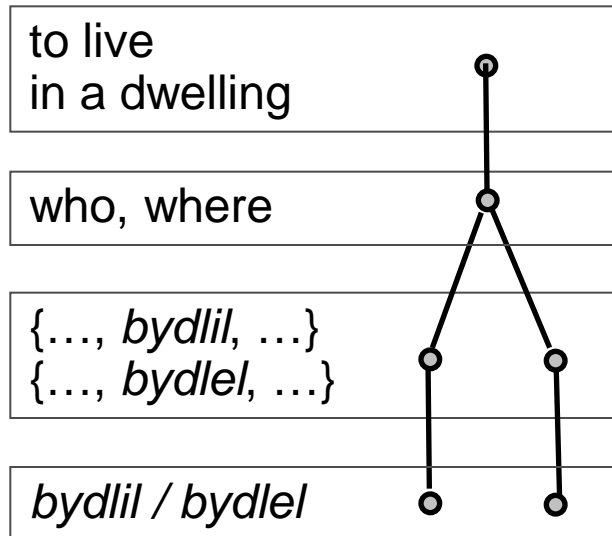


Duality of variants and homographs

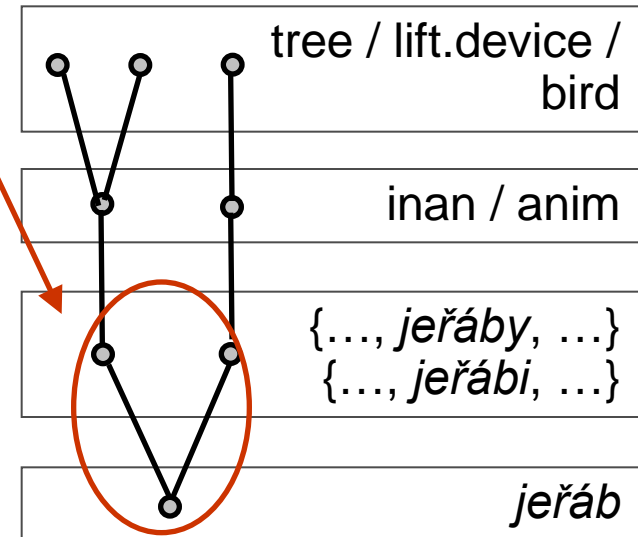


Schema of

variants for the example *bydlit / bydlet* **homographs** for the word *jeřáb*



meaning
syntactic /
semantic features
paradigms
(set of wordforms)
lemmas
(orthographic variants
of lemma)



PDT: m-layer





PDT: m-layer

- the sequence of tokens divided into sentences
- annotation ~ attaching a set attributes to each token
 - **lemma** ... base wordform
 - **tag** ... set of morphological categories
 - **id** ... PDT unique identifier
 - **w.rt** ... reference to w-layer
 - **form** ... (corrected) wordform
 - attributes identifying type of corrections
- PDT 2.0: Manual for Morphological Annotation
<http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html>
- Morphological Analysis of Czech Word Forms (Hajič)
<http://ufal.mff.cuni.cz/pdt2.0/tools/machine-annotation/morphology/>
DEMO: <http://quest.ms.mff.cuni.cz/morph/>



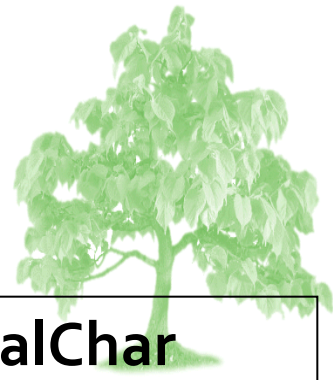
PDT: lemma structure

- lemma proper
 - a unique identifier ~ entry of the morphological lexicon
 - basic wordform (+ number for homographs)
 - no lemma is allowed to occur with two different POS
- additional information
 - e.g. semantic or derivational information

Lemma ::= LemmaProper | LemmaProper AddInfo

lemma	LemmaProper	AddInfo
<i>Chemik</i>	<i>chemik</i>	
<i>maso</i> ^(jídlo_apod.)	<i>maso</i>	^(jídlo_apod.)
<i>Bonn</i> _;G	<i>Bonn</i>	_;G
<i>vazba-1</i> ^(obviněného)	<i>vazba-1</i>	^(obviněného)
<i>vazba-2</i> ^(spojení)	<i>vazba-2</i>	^(spojení)
<i>Martinův-1</i> _;Y_^(*4-1)	<i>Martinův-1</i>	_;Y_^(*4-1)

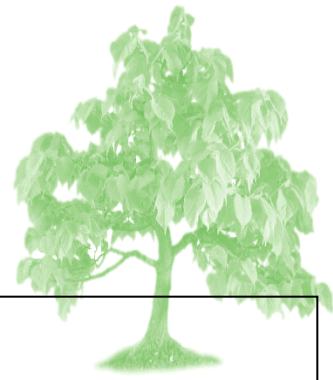
Lemma proper and base form




LemmaProper ::= Word | Word-Number | Number | SpecialChar

- **Word** ... base form of the respective paradigm
(case sensitive)
- **Number** ... to distinguish several senses of a homographic base form
(‘arbitrary’, some conventions for human readers)
- **SpecialChar** ::= ! | " | # | \$ | % | & | ' | (|) | * | + | , | - | . | / | : | ; | < | = | > | ? | @
| [| \ |] | ^ | _ | ` | { | | | } | ~ | \$ | °

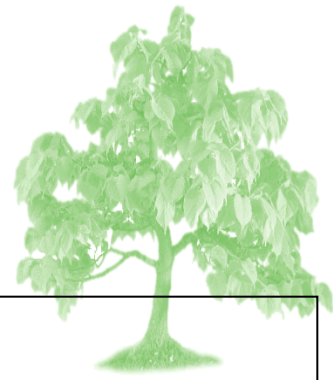
Additional information



AddInfo ::= Reference Category Term Style Comment

- **Reference ::= <empty> |  Lemma Proper**
for explaining the meaning of course lemma
e.g.: *kWh`kilowatthodina, jeden`1, oba`2*

Additional information



AddInfo ::= Reference **Category** Term Style Comment

• Category ::= <empty> | **⌊**Category₁ | **⌋**: Category₁ Category

letter

⌊:**T** and **⌊**:**W** for verbal aspect

e.g.: *běhat*:T, *říci*:W, *analyzovat*:T:W

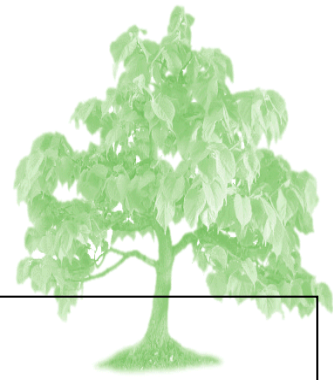
⌊:**B** for abbreviation

for part of speech (rarely used)

e.g.: *vedle-1*:D, *vedle-2*:P

(also possible: *vedle-1*^(*je_z_toho_vedle*), *vedle-2*^(*vedle_něčeho*))

Additional information



AddInfo ::= Reference Category **Term Style Comment**

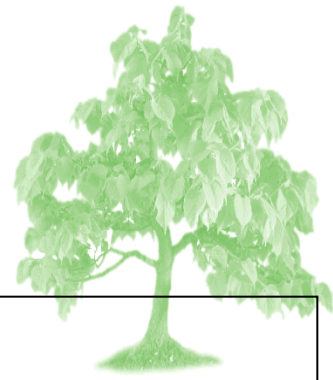
• **Term ::= <empty> | _ ; Term1 | _ ; Term1** **Term**

letter

named entities (mandatory) and
scientific/professional terms

e.g.: Y *John_*;Y ... given name
S *Agassi_*;S ... family name
E *Čech_*;E ... member of a particular nation
G *Praha_*;G ... geographic name
R *Tatra_*;R ... product
j ... justice
c ... computers and electronics
g ... technology
z ... ecology, environment

Additional information



AddInfo ::= Reference Category Term *Style* Comment

- **Style ::= <empty> | _, Style1 | _, Style1 *Style***

standard lemmas ... no stylistic flag

t ... foreign

n ... dialect

a ... archaic

s ... bookish

h ... colloquial

e ... expressive

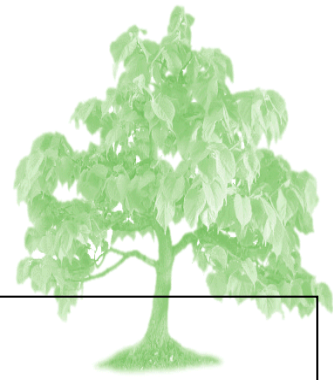
l ... slang, argot

v ... vulgar

x ... outdated spelling or misspelling

stylistic flag for a lemma vs. stylistic flag for a particular wordform

Additional information



AddInfo ::= Reference Category Term Style **Comment**

• **Comment ::= <empty> | _ ^ Comment₁**

**Comment₁ ::= (Explanation) | (Derivation) |
(Explanation) _ (Derivation)**

***string of letters, digits
and spec. characters***

(without spaces and parentheses;
in Czech)

**** Number Word | * Word***

e.g.: kardinálův _ ^ (*2)

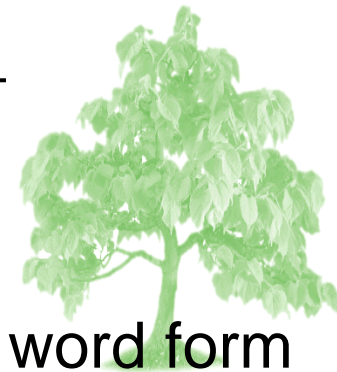
... remove two letters: kardinál

Karlův _ ; Y _ ^ (*3el)

přijetí-2 _ ^ (např. _ návrh) _ (*5mout-2)

podání _ ^ (něco _ [někomu] _ [někam]) _ (*3at)

protiprávnost _ ^ (*3ý)



PDT: tag structure

- lemma + tag ... together should uniquely identify the word form
- positional tags ... 15 characters
- every position ~ one morphological category (one character)

Position	Name
1	POS
2	SubPOS
3	Gender
4	Number
5	Case
6	PossGender
7	PossNumber
8	Person

Position	Name
9	Tense
10	Grade
11	Negation
12	Voice
13	Reserve1
14	Reserve2
15	Variant, style
16*	Aspect

PDT: tag structure



Examples:

dash (-) ... not applicable (e.g., tense for nouns)

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

píšící: AGMS1-----A-----

adjective, adjective derived from present transgressive form of a verb, masculine
animate, singular, nominative, affirmative

PDT: tag structure – POS (1)



- 'traditional' part of speech ... lexical category
- 10 classes + unknown (X) + punctuation (Z)

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)

PDT: tag structure – SubPOS (2)



- POS can be derived from SubPOS (67 classes)

e.g., for verbs (POS ... V)

- B ... present or future form
- c ... conditional of the verb *být* (*by, bych, bys, bychom, byste*, lit. would)
- e ... transgressive present (endings -e/-ě, -íc, -íce)
- f ... infinitive
- i ... imperative
- m ... past transgressive; also archaic pr. transgressive of pf verbs *udělav, udělaje*
- p ... past participle, active (*dělal, dělala, dělalo, dělali, dělaly, dělala*)
- q ... past participle, active, with the enclitic -ť (*bylť, bylať, bylotť, ...*)
- s ... past participle, passive (*dělán, dělána, děláno, dělání, dělány, dělána*)
- t ... present or future tense, with the enclitic -ť

PDT: tag structure – Gender (3)



- morphological property
for adjectives, pronouns, numerals and verbs
- lexical property ... nouns (→ no noun lemma have two different genders)

F	Feminine
H	{F, N} - Feminine or Neuter (<i>uběhnuvši</i>)
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 (<i>dělána</i>)
T	Masculine inanimate or Feminine (plural only); used only with participles and nominal forms of adjectives (<i>ležely</i>)
X	Any (<i>štěkajíce</i>)
Y	{M, I} - Masculine (either animate or inanimate) (<i>utíkáje</i>)
Z	{M, I, N} - Not feminine (i.e., Masculine animate/inanimate or Neuter); only for (some) pronoun forms and certain numerals

PDT: tag structure – Number (4)



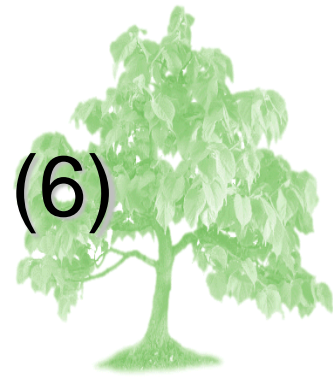
Value	Description
D	Dual , e.g. <i>nohama</i>
P	Plural, e.g. <i>nohami</i>
S	Singular, e.g. <i>noha</i>
W	Singular for feminine gender, plural with neuter; can only appear in participle or nominal adjective form with gender value Q (<i>dělána</i>)
X	Any

PDT: tag structure – Case (5)



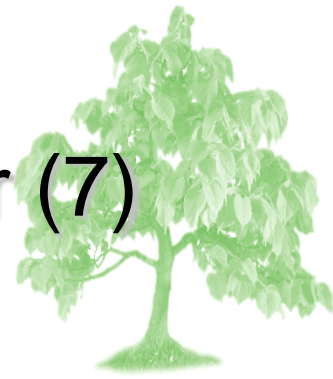
Value	Description
1	Nominative, e.g. <i>žena</i>
2	Genitive, e.g. <i>ženy</i> ,
3	Dative, e.g. <i>ženě</i>
4	Accusative, e.g. <i>ženu</i>
5	Vocative, e.g. <i>ženo</i>
6	Locative, e.g. <i>ženě</i>
7	Instrumental, e.g. <i>ženou</i>
X	Any

PDT: tag structure – Possessor's gender (6)



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

PDT: tag structure – Possessor's number (7)



Value	Description
P	Plural, e.g. <i>náš</i>
S	Singular, e.g. <i>můj</i>
X	Any, e.g. <i>your</i>

PDT: tag structure – Person (8)



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

PDT: tag structure – Tense (9)

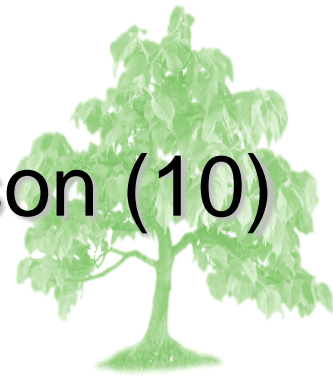


Value	Description
F	Future, e.g. <i>pojede</i>
H	{R, P} - Past or Present (???)
P	Present
R	Past
X	Any, e.g. <i>chráněn, vyhrazen, uloženi</i>

ČNK: Vs[FN]---2H-AP---[PI] errors!

bombardována-s (prep.), Jatas (NE), Klenos (NE), Kutas (NE, příjm.), litas, Litos (NE), **manipulováno-s** (prep.), Minutos (NE, příjm.), mytos, Oblitas (NE, příjm.), Pitas (NE, příjm.), Plutos, **počítáno-s** (prep.), probitas (lat.), **propuštěna-s** (prep.), Rytas (NE), Setas (NE), **spojena-s** (prep.), Vitas (NE), **vzdálenos** (-t)

PDT: tag structure – Degree of Comparison (10)



Value	Description
1	Positive, e.g. <i>velký</i>
2	Comparative, e.g. <i>větší</i>
3	Superlative, e.g. <i>největší</i>

PDT: tag structure – Negation (11)



Value	Description
A	Affirmative (not negated), e.g. <i>možný, kniha, neštěstí, utíká, udělaný</i>
N	Negated, e.g. <i>nemožný, nešťastný</i>

PDT: tag structure – Voice (12)



Value	Description
A	Active, e.g. <i>píše, jsem, sílila</i>
P	Passive, e.g. <i>udělán, napsán, varování, dovoleno</i>

PDT: tag structure – Variant (15)



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.

PDT: tag structure – Aspect (16)



Value	Description
P	perfective, e.g. <i>napsal, soustředěna, přijde</i>
I	imperfective, e.g. <i>píše, vlastnila</i>
B	biaspectual, e.g. <i>fascinovalo, jsem, defiovat</i>

Not in PDT !!

PennTreebank: Tag Set



CC	Coordinating conjunction	PP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential <i>there</i>	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	<i>to</i>
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VCN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PP	Personal pronoun	WRB	Wh-adverb



References

- Hajič, J. (2004) *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.
- Matthews, H. (1997) *The Concise Oxford Dictionary of Linguistics*. Oxford University Press, Oxford
- Filipec, J. (1994) Lexicology and Lexicography: Development and State of the Research. In Luelsdorff, P.A. (ed.) *The Prague School of Structural and Functional Linguistics*, Amsterdam-Philadelphia, John Benjamins, p.163–183
- Spoustová J., Hajič J., Raab J., Spousta M. (2009) Semi-Supervised Training for the Averaged Perceptron POS Tagger. In: *Proceedings of the EACL 2009*, pp. 763-771
- PDT documentation: Manual for morphological annotation
<http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html>
- Morphological Analysis of Czech Word Forms (Hajič, J.)
<http://ufal.mff.cuni.cz/pdt2.0/tools/machine-annotation/morphology/>
- DEMO: <http://quest.ms.mff.cuni.cz/morph/>
- Morfologický analyzátor češtiny ajka (Laboratoř NLP, Masarykova univerzita, Brno)
<http://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm>