



# Prague Dependency Treebank: Morphological Annotation

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

[lopatkova@ufal.mff.cuni.cz](mailto: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*

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, ...*}  
write → {write, wrote, written, writing, has been writing, ...}



---

## 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

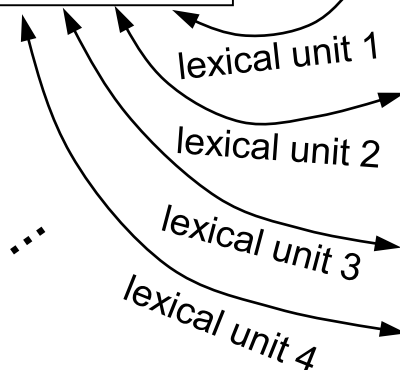
• lemma: *write*  
• paradigm: {*write, writes, writing, written, wrote*}

• 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*

...

...



---

# Morphological information in LRs



- *word form*  
~ every string of letters that forms a "word" of a language
- *(morphological) lemma*  
~ base form of a "word"
- *(morphological) tag*  
~ set of features specifying morphological categories
- *token*  
~ either a wordform or an punctuation

---

# Morphology in LRs vs. FGD theory



## **PDT: m-layer**

- *word form*  
~ every string of letters that forms a "word" of a language
- *(morphological) lemma*  
~ base form of a "word"
- *(morphological) tag*  
~ set of features specifying morphological categories

annotated text divided into sentences & morphemes grouped into words

## **FGD: morphemematics**

- words divided into morphemes
- lexical morphemes (roots and derivational m.)  
vs. grammatical morphemes (semas)
- formemes

---

# 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>
- MorphoDiTa  
<http://ufal.mff.cuni.cz/morphodita>



# PDT: lemma structure

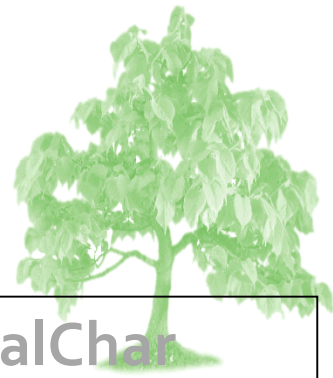
- **lemma proper**
  - a unique identifier ~ entry of the morphological lexicon
  - basic wordform (+ number for homographs)
  - no lemma is allowed to occur with two different POS
- additional information
  - e.g. semantic or derivational information

**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

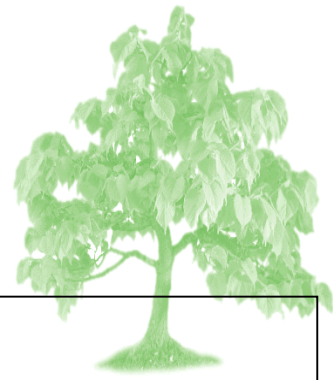


LemmaProper ::= Word | Word-Number | Number | SpecialChar


- **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

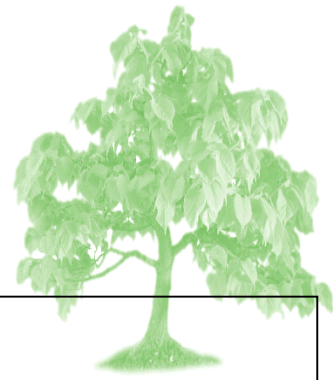


AddInfo ::= **Reference** Category Term Style Comment

- Reference ::= <empty> |  LemmaProper  
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> | **⓪**:Category<sub>1</sub> | **\_**:Category<sub>1</sub> Category

**letter**

**\_:T** and **\_:W** for verbal aspect

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

**\_:B** for abbreviation

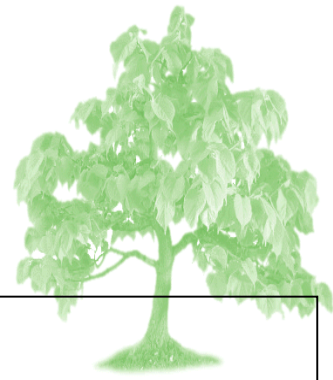
**\_:X** for part of speech (rarely used)

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

(also possible: *vedle-1*\_<sup>^</sup>(*je\_z\_toho\_vedle*), *vedle-2*\_<sup>^</sup>(*vedle\_něčeho*) )

---

# Additional information



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

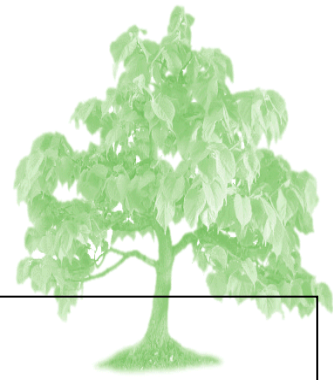
- Term ::= <empty> | **⓪**; Term1 | **letter**; Term1 Term

named entities (mandatory) and  
scientific/professional terms

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

---

# Additional information



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

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

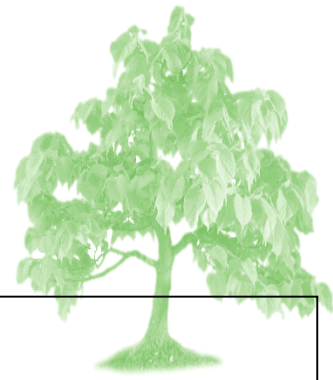
standard lemmas ... no stylistic flag

t ... foreign	
n ... dialect	e ... expressive
a ... archaic	l ... slang, argot
s ... bookish	v ... vulgar
h ... colloquial	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> | **\_** ^ Comment<sub>1</sub>

Comment<sub>1</sub> ::= ( 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)



---

# Morphological features

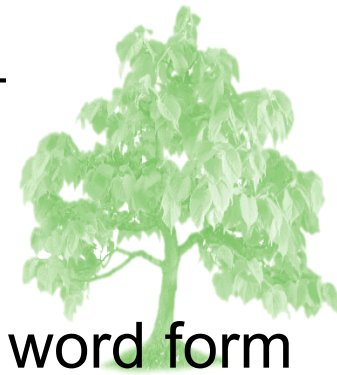


---

# PennTreebank: Tag Set (36 values)



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	VBN	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



# PDT: tag structure

- lemma + tag ... together should uniquely identify the word form
- positional tags ... 15/16 characters
- **every position ~ one morphological category**  
(one character)

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

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

---

# PDT: tag structure



- lemma + tag ... together should uniquely identify the word form
- positional tags ... 15/16 characters
- every position ~ one morphological category (one character)
  - ➔ about 2 600 values (linguistically adequate)  
1 454 values in PDT 3.0 (train and dtest )

# CZECH POSITIONAL MORPHOLOGICAL TAGS

## 1. PART OF SPEECH

**A** Adjectives

**C** Numerals

**D** Adverbs

**I** Interjection

**J** Conjunction

**N** Noun

**P** Pronoun

**V** Verb

**R** Preposition

**T** Particle

**X** Unknown, Not Determined, Unclassifiable

**Z** Punctuation (also used for the Sentence Boundary Token)

## 2. SUB PART OF SPEECH

**#** Sentence boundary

**%** Author's signature, e.g. haš-99\_B\_S

**\*** Word krát (lit.: "times")

**,** Conjunction subordinate (incl. "aby", "kdyby" in all forms)

**}** Numeral, written using Roman numerals (XIV)

**:** Punctuation (except for the virtual sentence boundary word ###, which uses "Sub part of speech" #)

**=** Number written using digits

**?** Numeral "kolik" (lit. "how many"/"how much")

**@** Unrecognized word form

**^** Conjunction (connecting main clauses, not subordinate)

**4** Relative/interrogative pronoun with adjectival declension of both types (soft and hard) ("jaký", "který", "čí", ..., lit. "what", "which", "whose", ...)

**5** The pronoun he in forms requested after any preposition (with prefix n-: "něj", "něho", ..., lit. "him" in various cases)

**6** Reflexive pronoun "se" in long forms ("sebe", "sobě", "sebou", lit.

"myself" / "yourself" / "herself" / "himself" in various cases; "se" is personless)

**7** Reflexive pronouns "se" ("Case" = 4), "si" ("Case" = 3), plus the same two forms with contracted -s: "ses", "sis" (distinguished by "Person" = 2; also number is singular only) This should be done somehow more consistently, virtually any word can have this contracted -s ("cos", "polívkus", ...)

**8** Possessive reflexive pronoun "svůj" (lit. "my"/"your"/"her"/"his" when the possessor is the subject of the sentence)

**9** Relative pronoun "jenž", "již", ... after a preposition (n-: "něhož", "niž", ..., lit. "who")

**A** Adjective, general

**B** Verb, present or future form

**C** Adjective, nominal (short, participial) form "rád", "schopen", ...

**D** Pronoun, demonstrative ("ten", "onen", ..., lit. "this", "that", "that", ... "over there", ...)

**E** Relative pronoun "což" (corresponding to English which in subordinate clauses referring to a part of the preceding text)

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

**G** Adjective derived from present transgressive form of a verb

**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

**I** Interjections

**J** Relative pronoun "jenž", "již", ... not after a preposition (lit. "who", "whom")

**K** Relative/interrogative pronoun "kdo" (lit. "who"), incl. forms with affixes -ž and -s (affixes are distinguished by the category "Variant" (for -ž) and "Person" (for -s))

**L** Pronoun, indefinite "všechn", "sám" (lit. "all", "alone")

**M** Adjective derived from verbal past transgressive form

**N** Noun (general)

**O** Pronoun "svůj", "nesvůj", "tentam" alone (lit. "own self", "not-in-mood", "gone")

**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

**Q** Pronoun relative/interrogative "co", "copak", "cožpak" (lit. "what",

"isn't-it-true-that")

**R** Preposition (general, without vocalization)

**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

**T** Particle

**U** Adjective possessive (with the masculine ending -ův as well as feminine -in)

**V** Preposition (with vocalization -e or -u): ("ve", "pode", "ku", ..., lit. "in", "under", "to")

**W** Pronoun negative ("nic", "nikdo", "nijaký", "žádný", ..., lit. "nothing", "nobody", "not-worth-mentioning", "no"/"none")

**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")

**Z** Pronoun indefinite ("nějaký", "některý", "číkoli", "cosi", ..., lit. "some", "some", "anybody's", "something")

**a** Numeral, indefinite ("mnoho", "málo", "tolik", "několik", "kdovíkolik", ..., lit. "much"/"many", "little"/"few", "that much"/"many", "some" ("number of"), "who-knows-how-much/many")

**b** Adverb (without a possibility to form negation and degrees of comparison, e.g. "pozadu", "naplocho", ..., lit. "behind", "flatly"); i.e. both the "Negation" as well as the "Grade" attributes in the same tag are marked by – (Not applicable)

**c** Conditional (of the verb "být" (lit. "to be") only) ("by", "bych", "bys", "bychom", "byste", lit. "would")

**d** Numeral, generic with adjectival declension ("dvoji", "desaterý", ..., lit. "two-kinds"/"two", "ten-...")

**e** Verb, transgressive present (endings -e/-ě, -ic, -ice)

**f** Verb, infinitive

**g** Adverb (forming negation ("Negation" set to A/N) and degrees of comparison "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")

**i** Verb, imperative form

**j** Numeral, generic greater than or equal to 4 used as a syntactic noun ("čtvero", "desatero", ..., lit. "four-kinds"/"sorts-of", "ten-...")

- k** Numeral, generic greater than or equal to 4 used as a syntactic adjective, short form ("čtvery", ..., lit. "four-kinds"/"sorts-of")
- 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
- m** Verb, past transgressive; also archaic present transgressive of perfective verbs (ex.: "udělav", lit. "(he-)having-done"; arch. also "udělaje" ("Variant" = 4), lit. "(he-)having-done")
- n** Numeral, cardinal greater than or equal to 5
- o** Numeral, multiplicative indefinite ("-krát", lit. ("times"): "mnohokrát", "tolikrát", ..., lit. "many times", "that many times")
- p** Verb, past participle, active (including forms with the enclitic -s, lit. 're ("are"))
- q** Verb, past participle, active, with the enclitic -ř, lit. ("perhaps") - "could-you-imagine-that?" or "but-because-" (both archaic)
- r** Numeral, ordinal (adjective declension without degrees of comparison)
- s** Verb, past participle, passive (including forms with the enclitic -s, lit. 're ("are"))
- t** Verb, present or future tense, with the enclitic -ř, lit. ("perhaps") - "could-you-imagine-that?" or "but-because-" (both archaic)
- u** Numeral, interrogative "kolikrát", lit. "how many times?"
- v** Numeral, multiplicative, definite (-krát, lit. "times": "pětkrát", ..., lit. "five times")
- w** Numeral, indefinite, adjectival declension ("nejeden", "tolikátý", ..., lit. "not-only-one", "so-many-times-repeated")
- y** Numeral, fraction ending at -ina; used as a noun ("pětina", lit. "one-fifth")
- z** Numeral, interrogative "kolikátý", lit. "what" ("at-what-position-place-in-a-sequence")

### 3. GENDER

- 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 feminine (i.e., Masculine animate/inanimate or Neuter); only for (some) pronoun forms and certain numerals

### 4. NUMBER

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

### 5. CASE

- 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

### 6. POSSESSIVE GENDER

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

### 7. POSSESSIVE NUMBER

- P** Plural, e.g. "náš"
- S** Singular, e.g. "můj"
- X** Any, e.g. "your"

### 8. PERSON

- 1** 1st person, e.g. "píšu", "píšeme"
- 2** 2nd person, e.g. "píšeš", "píšete"
- 3** 3rd person, e.g. "píše", "píšou"

- X** Any person

### 9. TENSE

- F** Future
- H** {R, P} – Past or Present
- P** Present
- R** Past
- X** Any

### 10. GRADE

- 1** Positive, e.g. "velký"
- 2** Comparative, e.g. "větší"
- 3** Superlative, e.g. "největší"

### 11. NEGATION

- A** Affirmative (not negated), e.g. "možný"
- N** Negated, e.g. "nemožný"

### 12. VOICE

- A** Active, e.g. "píšíci"
- P** Passive, e.g. "psaní"

### 13., 14. RESERVE 1, RESERVE 2

- Not applicable

### 15. VARIANT

- 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 – 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, byloť, ...*)
- 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)

<b>F</b>	<b>Feminine</b>
H	{F, N} - Feminine or Neuter ( <i>uběhnuvši</i> )
<b>I</b>	<b>Masculine inanimate</b>
<b>M</b>	<b>Masculine animate</b>
<b>N</b>	<b>Neuter</b>
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>utíkaje</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

---

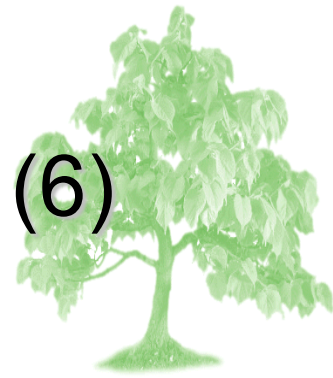
# PDT: tag structure – Case (5)



<b>Value</b>	<b>Description</b>
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

---

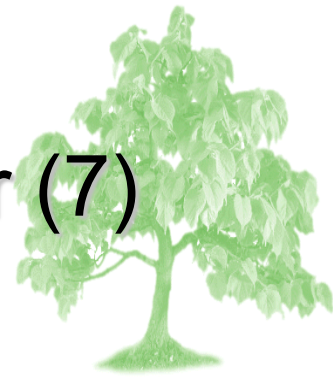
# PDT: tag structure – Possessor's gender (6)



Value	Description
F	Feminine, e.g. <i>matčin, 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)



<b>Value</b>	<b>Description</b>
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)



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

---

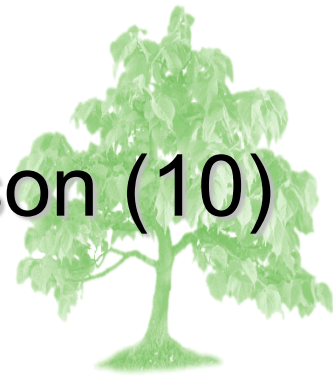
# PDT: tag structure – Tense (9)



<b>Value</b>	<b>Description</b>
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>

---

# PDT: tag structure – Degree of Comparison (10)



<b>Value</b>	<b>Description</b>
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)



<b>Value</b>	<b>Description</b>
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 !!**

---

# Addition – Variants and Homographs



---

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



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

---

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



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

---

# 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>	• past tense • past participle
----------------	-----------------------------------

<i>hradu</i> [castle]	• genitive singular • dative singular
--------------------------	--

(2) identical wordforms with

- different lemmas

<i>smaž</i> imp.	• <i>smazat</i> [to erase] • <i>smažit</i> [to fry]
------------------	--

<i>ženu</i>	• 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>	• noun • verb
---------------	------------------

<i>nakupovat</i>	• [to buy] • [to heap]
------------------	---------------------------

<i>žít</i>	• [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]
-----------------------------------

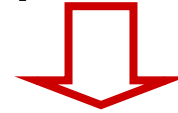
nakupovat-1 [to buy] nakupovat-2 [to heap]
---



---

# 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]

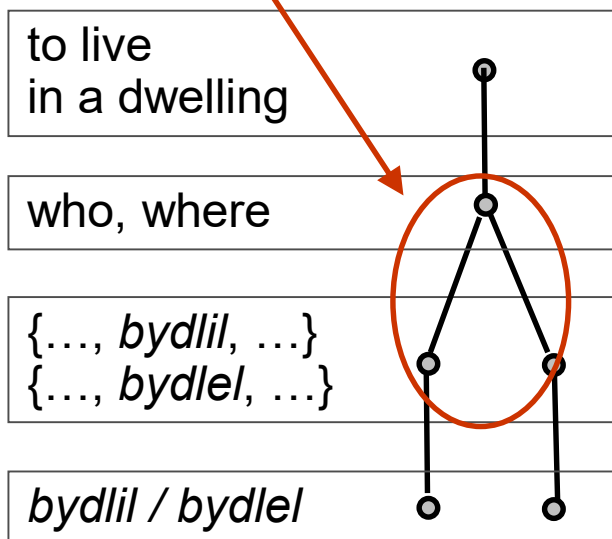
• homographic lemma, i.e. two lexemes (SSČ)



# Duality of variants and homographs

Schema of

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

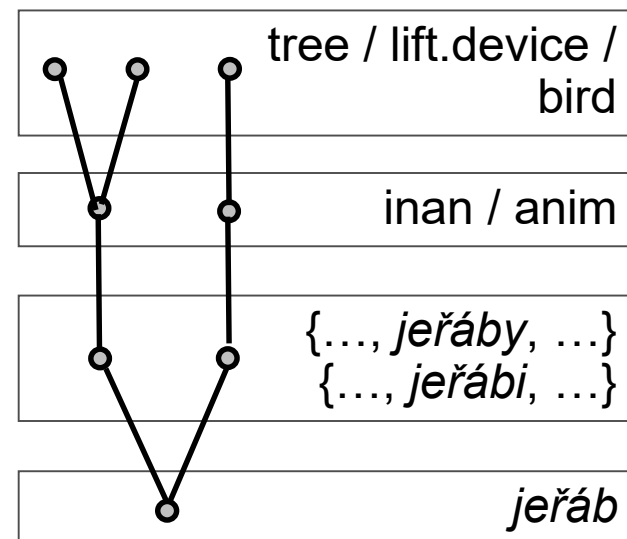


meaning

syntactic /  
semantic features

paradigms  
(set of wordforms)

lemmas  
(orthographic variants  
of lemma)

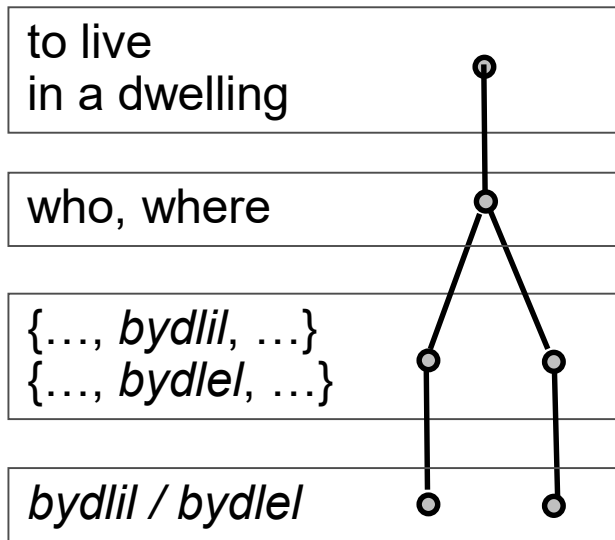




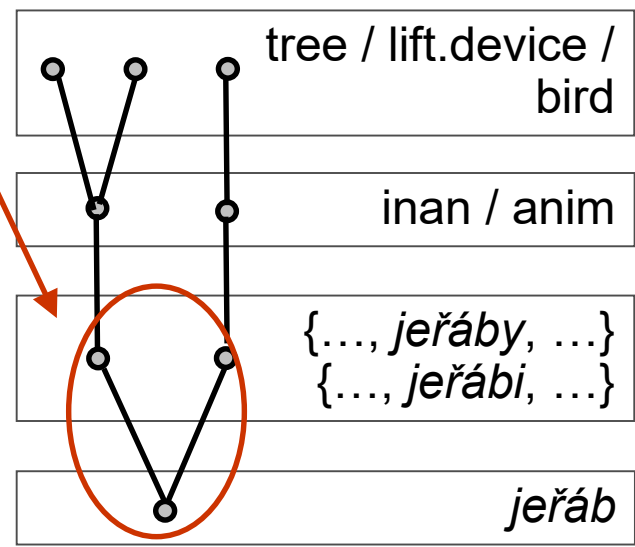
# Duality of variants and homographs

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)



---

# References



- 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
- Hajič, J. (2004) *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.
- <https://wiki.korpus.cz/doku.php/seznamy:tagy>
- [http://ufal.mff.cuni.cz/pdt/Morphology\\_and\\_Tagging/Doc/hmptagqr.html](http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html)
- Straková Jana, Straka Milan and Hajič Jan. (2014) Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13-18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- DEMO: <http://lindat.mff.cuni.cz/services/morphodita/>
- PDT documentation: Manual for morphological annotation <http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html>