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

entry of a morphological lexicon

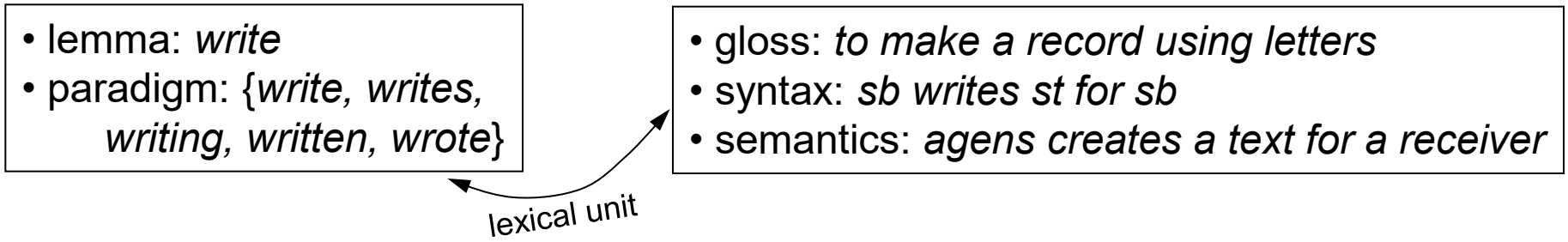
- ***paradigm***

  ~ 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, …

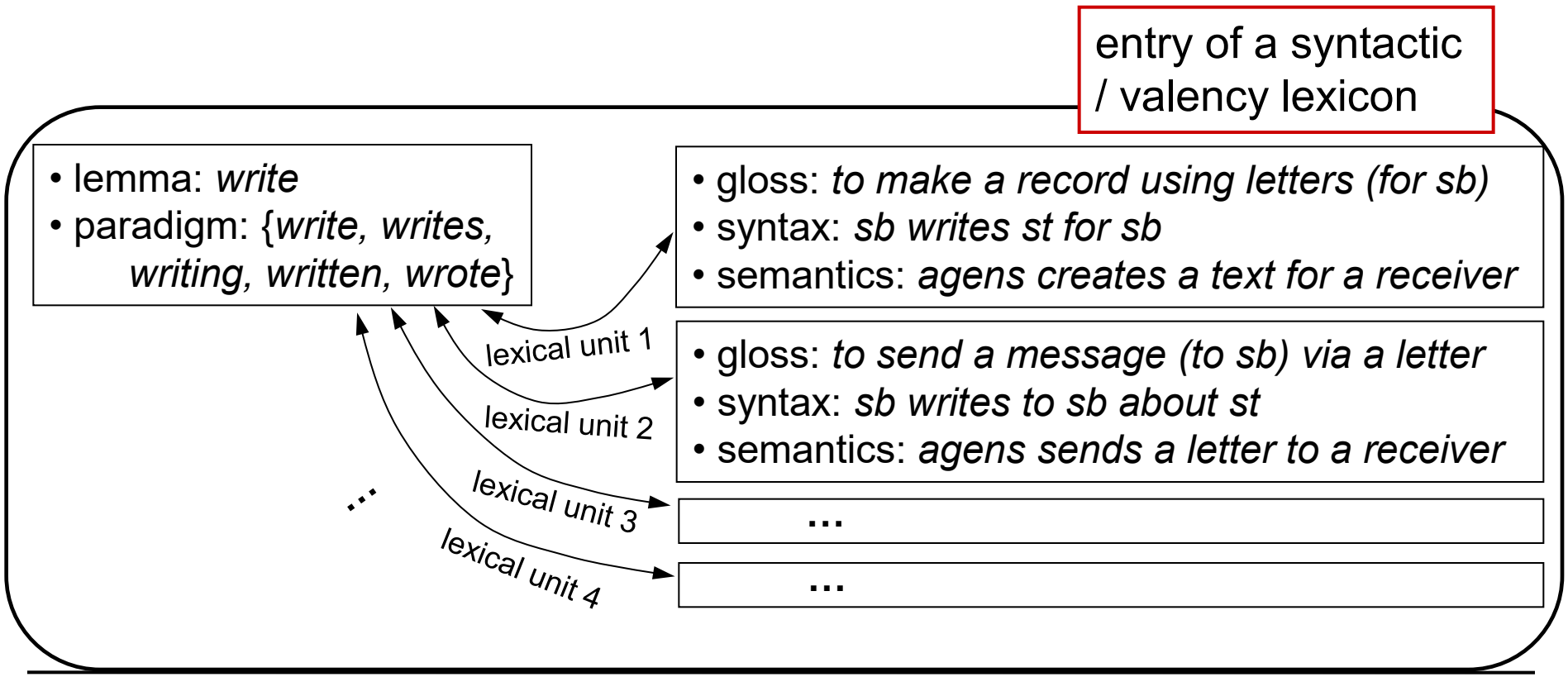
Basic terms (cont.)

- ***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'***



Basic terms (cont.)

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



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 interpunction

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

FGD: morphematics

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

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

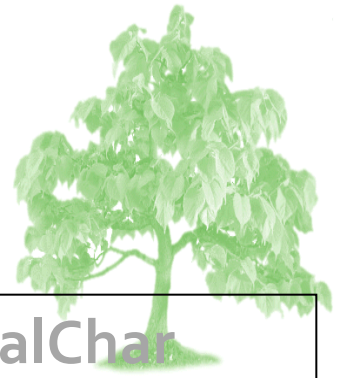
PDT: lemma structure

- 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
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)

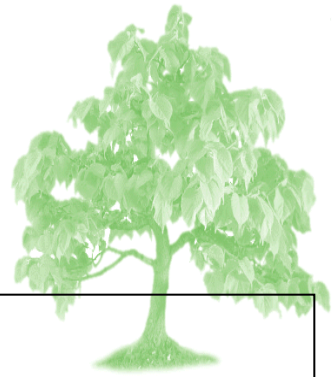
Lemma proper and base form



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



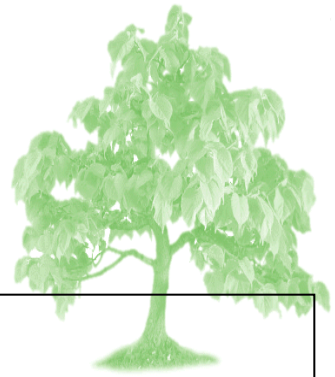
AddInfo ::= Reference Category Term Style Comment

- **Reference ::= <empty> | `LemmaProper**

  for explaning 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

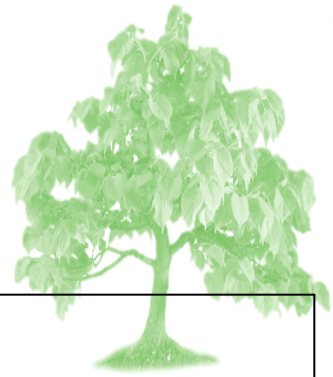
e.g.: *běhat*_:T, *říci*_:W, *analyzovat*_:T_:W

_:B for abbreviation

_:X 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

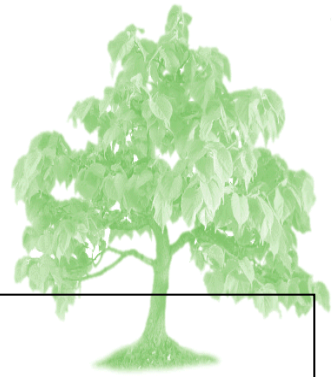
letter

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

letter

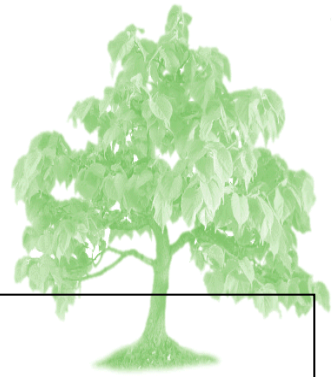
standard lemmas … no stylistic flag

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***

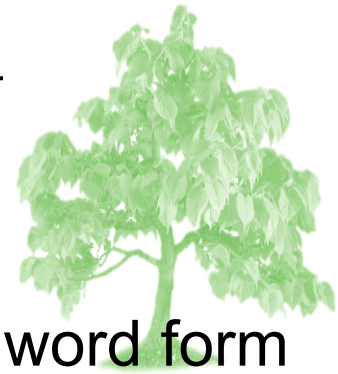
e.g.: kardinálův_^(*2)
… remove two letters: kardinál
Karlův_;Y_^(*3el)
přijetí-2_^(např._návrh)_(*5mout-2)
podání_^(něco_[někomu]_[někam])_(*3at)
protiprávnost_^(*3ý)

Morphological features

PennTreebank: Tag Set (36 values)

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

PDT: tag structure



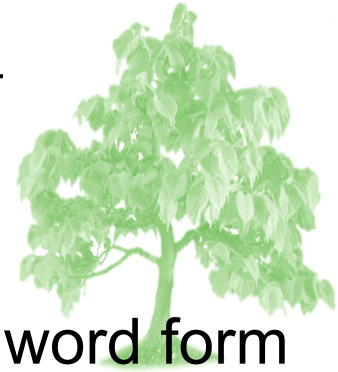
- lemma + tag … together should uniquely identify the word form
- positional tags … 15/16 characters
- every position ~ one morphological category (one character)

Position	Name
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

** not in PDT*

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)

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

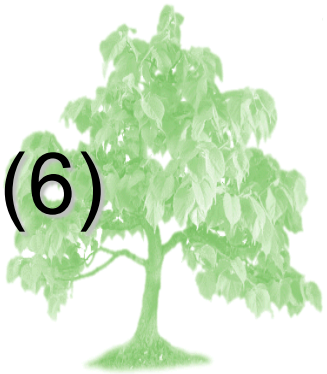
PDT: tag structure – Number (4)

Value	Description
D	Dual , e.g. *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 (*děláno*)
X	Any

PDT: tag structure – Case (5)

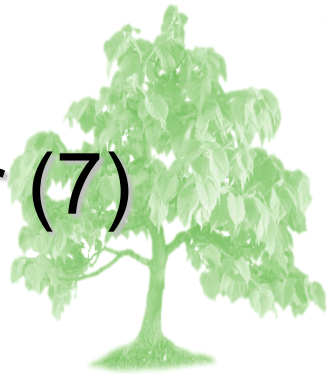
Value	Description
1	Nominative, e.g. *žena*
2	Genitive, e.g. *ženy,*
3	Dative, e.g. *ženě*
4	Accusative, e.g. *ženu*
5	Vocative, e.g. *ženo*
6	Locative, e.g. *ženě*
7	Instrumental, e.g. *ženou*
X	Any

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



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*

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



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

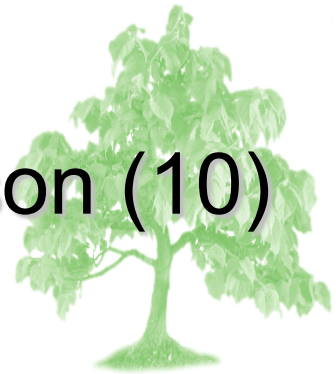
PDT: tag structure – Person (8)

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

PDT: tag structure – Tense (9)

Value	Description
F	Future, e.g. *pojede*
H	{R, P} - Past or Present
P	Present
R	Past
X	Any, e.g. *chráněn, vyhrazen, uloženi*

PDT: tag structure – Degree of Comparison (10)



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

PDT: tag structure – Negation (11)

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

PDT: tag structure – Voice (12)

Value	Description
A	Active, e.g. *píše, jsem, sílila*
P	Passive, e.g. *udělán, napsán, varování, dovoleno*

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 – Acpect (16)

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

Not in PDT !!

Addition – Variants and Homographs

'Golden rule' of morphology

lemma *A* → forms $a_1, \ldots a_n$
lemma *B* → forms $b_1, \ldots b_m$

different words with different wordform(s)

lemma + tag … together should uniquely identify the word form

---

lemma *A*
lemma *B* → forms ***$c_1 \ldots c_n$***

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

---

lemma *C* → forms $c_1, \ldots$ ***x***, $\ldots c_n$
lemma *C* → forms $c_1, \ldots$ ***y***, $\ldots c_n$

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

stopped	• past tense • past participle

hradu [castle]	• genitive singular • dative singular

(2) identical wordforms with

- different lemmas

smaž imp.	• *smazat* [to erase] • *smažit* [to fry]

ženu	• acc sg. *žena* [woman] • 1. pers. sg. pres. *hnát* [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***

flower	• noun • verb

nakupovat	• [to buy] • [to heap]

žít	• [to live] • [to mow]

two wordforms with the same lemmas and morph. properties

(1) either their paradigms differ

flower	• *flowers*
flower	• *flowered*

žít	[to live]	• *žil* for past tense
žít	[to mow]	• *žal* for past tense

(2) or they are derived from different words

odrolovat	[to roll away]	• *od-rol-ovat*
odrolovat	[to crumble]	• *o-drol-ovat*

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***

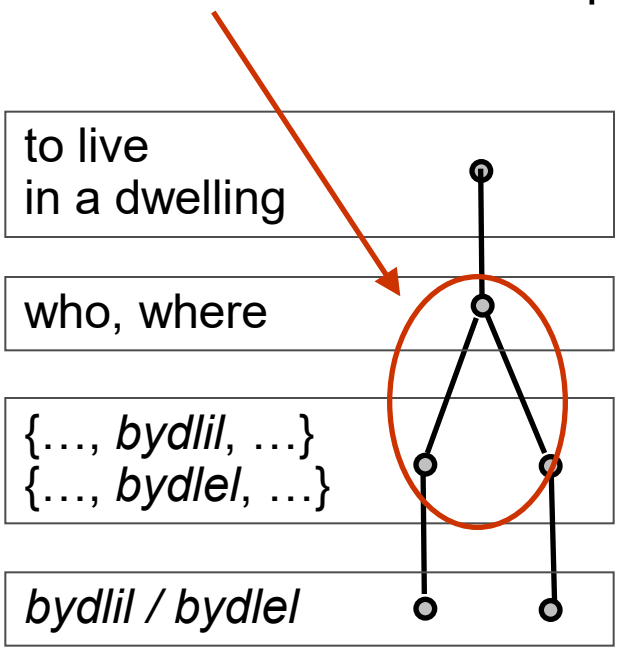
! No clear cut between polysemy and homography !

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*

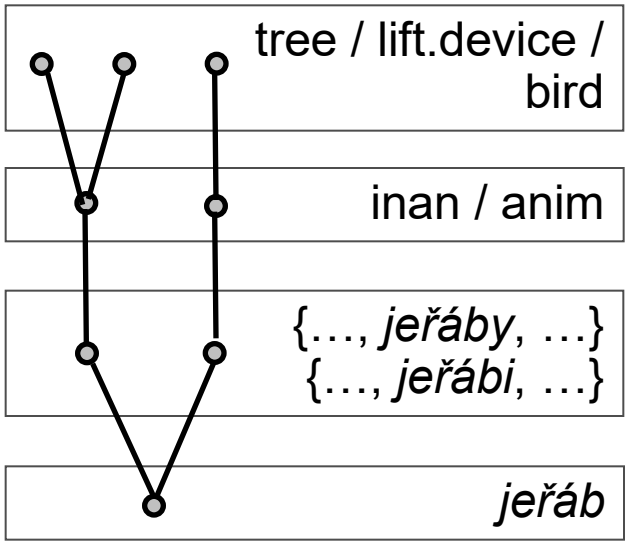


meaning

syntactic /
semantic features

paradigms
(set of wordforms)

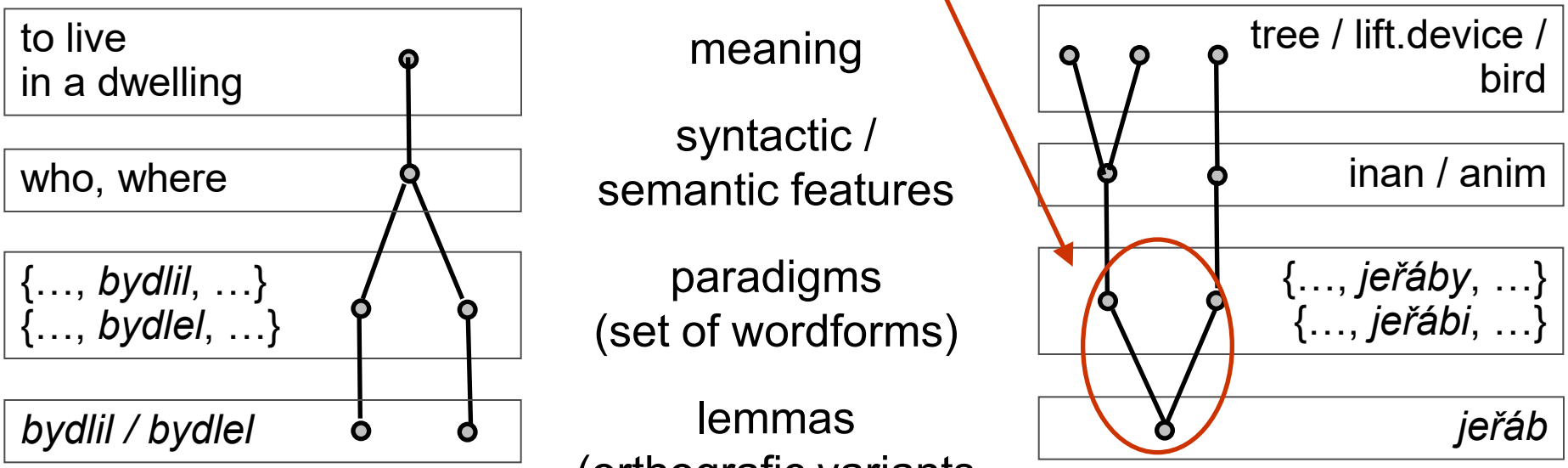
lemmas
(orthografic variants
of lemma)



Duality of variants and homographs

Schema of

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



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 Univeristy Press, Prague.

  https://wiki.korpus.cz/doku.php/seznamy:tagy

  http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptaggr.html

  DEMO: http://quest.ms.mff.cuni.cz/morph/
- 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