# Morphology within the Multi-Layered Annotation Scenario of the Prague Dependency Treebank

Magda Ševčíková

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

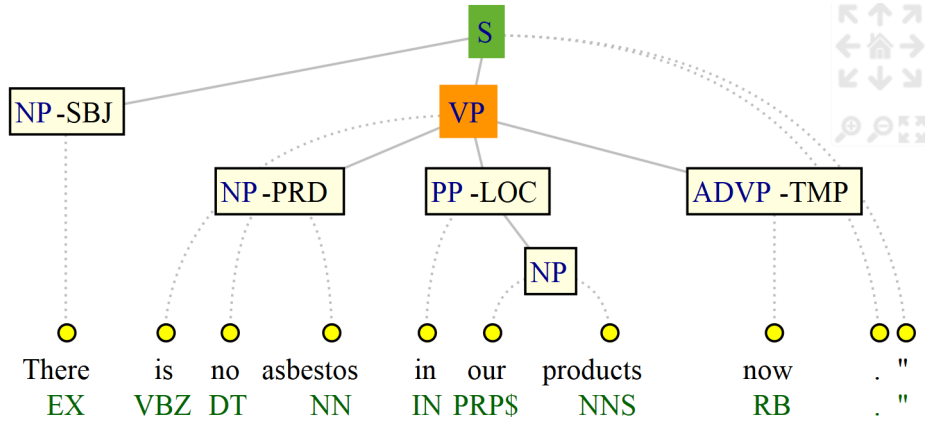
SFCM 2015, September 16–17, 2015

# Outline

1. Introduction
2. Morphology in Prague Dependency Treebank
   - PDT in a nutshell
   - Morphological layer
   - Tectogrammatical layer
3. Praguian morphology in NLP of Czech
   - Developing taggers
   - Named entity recognition
   - Derivational morphology
4. Conclusions

# Introduction: Treebanks without morphology?

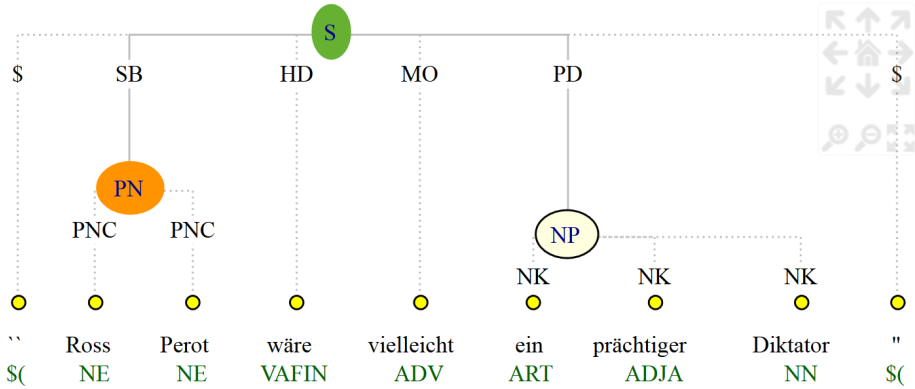
- 83 treebanks for 51 languages (Zeman 2015)
- from coarse-grained part-of-speech information to detailed description of morphological categories
- according to the theoretical approach (and morphological richness of the language)




Penn Treebank
`https://lindat.mff.cuni.cz/services/pmltq/`

# Introduction: Treebanks without morphology?

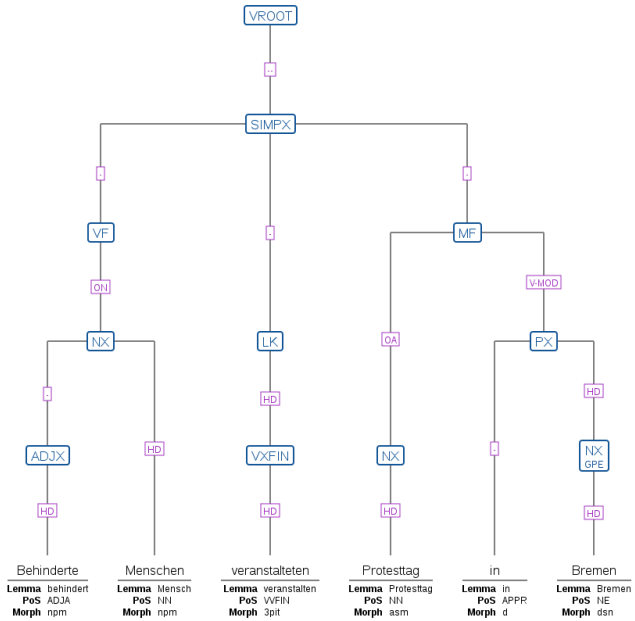
- 83 treebanks for 51 languages (Zeman 2015)
- from coarse-grained part-of-speech information to detailed description of morphological categories
- according to the theoretical approach (and morphological richness of the language)




TIGER treebank
`https://lindat.mff.cuni.cz/services/pmltq/`

# Introduction: Treebanks without morphology?

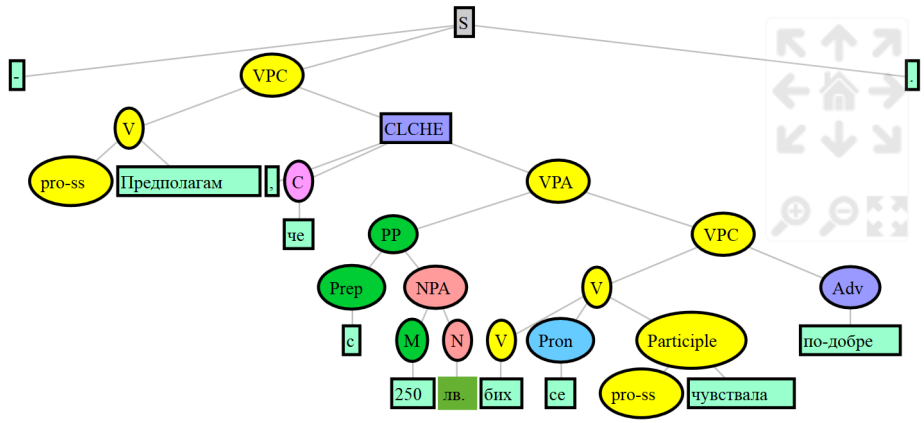
- 83 treebanks for 51 languages (Zeman 2015)
- from coarse-grained part-of-speech information to detailed description of morphological categories
- according to the theoretical approach (and morphological richness of the language)




TüBa-D/Z
https://weblicht.sfs.uni-tuebingen.de/Tundra/

# Introduction: Treebanks without morphology?

- 83 treebanks for 51 languages (Zeman 2015)
- from coarse-grained part-of-speech information to detailed description of morphological categories
- according to the theoretical approach (and morphological richness of the language)




BulTreeBank
`https://lindat.mff.cuni.cz/services/pmltq/`

# Introduction: Morphology in recent treebanking projects

- HamleDT (HArmonized Multi-LanguagE Dependency Treebank)
  - `http://ufal.mff.cuni.cz/hamledt`
  - 42 treebanks for 36 languages in version 3.0 (August 18, 2015)
  - surface-syntactic annotation based on Stanford Dependencies (de Marneffe et al. 2014)
  - Interset interlingua for morphological features (Zeman 2008)
- Universal Dependencies
  - `http://universaldependencies.github.io/docs/`
  - 34 languages in version 1.1 (May 15, 2015)
  - Universal Dependencies standard based on Stanford Dep.
  - "interlingua" based on Zeman's Interset and Google universal part-of-speech tags (Petrov et al. 2012)

# Introduction: Interset interlingua for morphological tagsets

- converting tagsets into interlingua (and/or into other tagsets)
- comparing tagsets (`http://quest.ms.mff.cuni.cz/cgi-bin/interset/index.pl`)
  - Penn treebank tagset: 48 tags for English
  - SynTagRus tagset: 376 tags for Russian
  - Hajič's tagset for Czech (PDT): 4,294 tags
    - vs. 846 tags for Czech assigned by the *ajka* tagger

**Penn**

NNPS
VB

**Interset**

pos="noun", subpos="prop", number="plu"
pos="verb", verbform="inf"

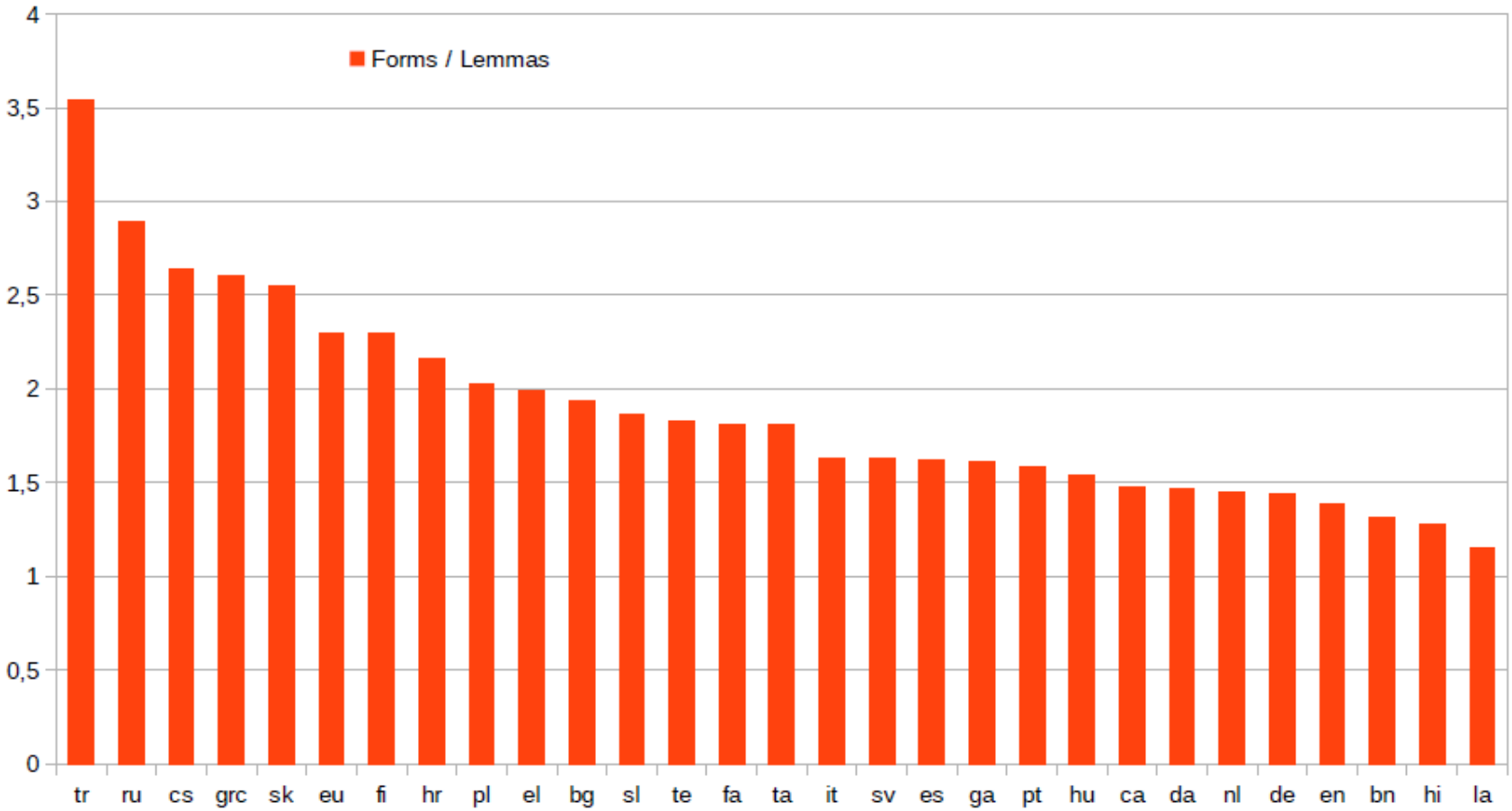
**PDT**

NNFP1-----A----
VB-P---3P-AA---

**Interset**

pos="noun", negativeness="pos", gender="fem", number="plu", case="nom"
pos="verb", negativeness="pos", number="plu", person="3", verbform="fin", mood="ind", tense="pres", voice="act"

# Introduction: Morphological richness (HamleDT)




[Zeman 2015]

# Introduction: How rich is Czech?

- rich inflectional and derivational morphology in Czech

**_agent_ 'agent'**

*agent* (nom.sg.)
*agenta* (gen.sg.|acc.sg.)
*agentu* (dat.sg.|loc.sg.)
*agentovi* (dat.sg.|loc.sg.)
*agente* (voc.sg.)
*agentem* (instr.sg.)
*agenti* (nom.pl.|voc.pl.)
*agentové* (nom.pl.|voc.pl.)
*agentů* (gen.pl.)
*agentům* (dat.pl.)
*agenty* (acc.pl.|instr.pl.)
*agentech* (loc.pl.)

# Introduction: How rich is Czech?

- rich inflectional and derivational morphology in Czech

**_agent_ 'agent'**

_agent_ (nom.sg.)
_agenta_ (gen.sg.|acc.sg.)
_agentu_ (dat.sg.|loc.sg.)
_agentovi_ (dat.sg.|loc.sg.)
_agente_ (voc.sg.)
_agentem_ (instr.sg.)
_agenti_ (nom.pl.|voc.pl.)
_agentové_ (nom.pl.|voc.pl.)
_agentů_ (gen.pl.)
_agentům_ (dat.pl.)
_agenty_ (acc.pl.|instr.pl.)
_agentech_ (loc.pl.)

**_agent_ 'agent'**

> _agentův_ 'agent's'
> _agentka_ 'female agent'
> _agentský_ 'agency'
> _superagent_ 'superagent'

...

## *zvát* 'to invite'

- ind.pres.act.:
  *zvu, zveš, zve; zveme, zvete, zvou*
- ind.pret.act.:
  *zval(a) jsem, zval(a) jsi, zval(a); zvali/y jsme, zvali/y jste, zvali/y*
- ind.fut.act.:
  *budu zvát, budeš zvát, bude zvát; budeme zvát, budete zvát, budou zvát*
- ind.pres.pass.:
  *jsem zván(a), jsi zván(a), je zván(a); jsme zváni/y, jste zváni/y, jsou zváni/y*
- ind.pret.pass.:
  *byl(a) jsem zván(a), byl(a) jsi zván(a), byl(a) zván(a); byli/y jsme zváni/y, ...*
- ind.fut.pass.:
  *budu zván(a), budeš zván(a), bude zván(a); budeme zváni/y, ...*
- cond.pres.act.:
  *zval(a) bych, zval(a) bys, zval(a) by; zvali/y bychom, ...*
- cond.pres.pass.:
  *byl(a) bych zván(a), byl(a) bys zván(a), byl(a) zván(a); byli/y by zváni/y, ...*
- ...

# Morphology in Prague Dependency Treebank: Form and meaning

- multiple annotation layers
  - morphology as a separate layer of annotation
    - lemma and positional (POS+) tag (Hajič 2004)

**_agentu_ '(to an) agent'**

agent NNMS3-----A---1

**_byli jste zváni_ '(you) were invited'**

být VpMP---XR-AA---
být VB-P---2P-AA---
zvát VsMP---XX-AP---

# Morphology in Prague Dependency Treebank: Form and meaning

- multiple annotation layers
  - morphology as a separate layer of annotation
    - lemma and positional (POS+) tag (Hajič 2004)
  - meanings expressed by morphological categories captured at the tectogrammatical layer
    - grammateme attributes

**_agentu_ '(to an) agent'**

agent NNMS3-----A---1

**_agentu_ '(to an) agent'**

one entity

**_byli jste zváni_ '(you) were invited'**

být VpMP---XR-AA---
být VB-P---2P-AA---
zvát VsMP---XX-AP---

**_byli jste zváni_ '(you) were invited'**

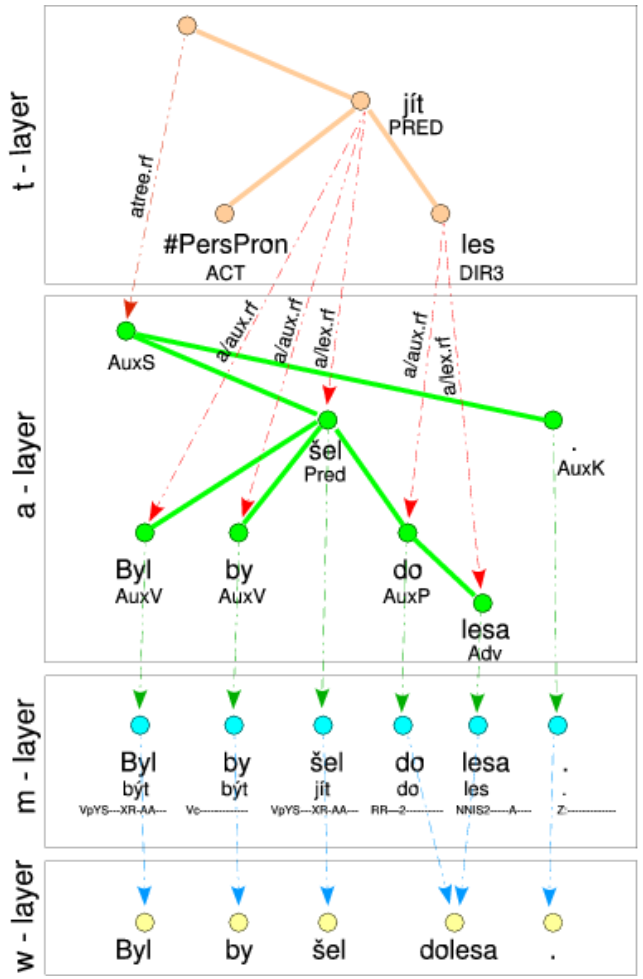
past event

# Prague Dependency Treebank – a short history

- theoretically rooted in Functional Generative Description (Sgall 1967, Sgall et al. 1986)
  - language system decomposed in multiple layers
  - relation of form and function between neighboring layers
  - **unambiguity and self-containedness of the sentence representation at each layer**
- annotation of Prague Dependency Treebank
  - started in the late 1990s
  - PDT 1.0 (2001): morphological and analytical annotation
  - PDT 2.0 (2006): plus tectogrammatical annotation
  - PDT 2.5 (2011)
  - PDT 3.0 (2013)

# Annotation layers in PDT

- one non-annotation (word) layer
- three layers of annotation
  - morphological layer
    - 1,960k tokens in 116k sent. in PDT 2.0
  - analytical layer
    - 88k sentences with 1,503k tokens
  - tectogrammatical layer
    - 49k sentences with 830k tokens
- cross-layer references between nodes of neighboring layers




*lit.:* 'Was would gone to-forest.'
'He would have gone to the forest.'

# Annotation at the morphological layer of PDT

- automatic morphological analysis
  - MorfFlex dictionary with 350k+ manual entries (Hajič – Hlaváčová 1990)
  - recognizer of about 12M Czech word forms
- manual disambiguation
  - each file annotated by two annotators in parallel
  - instances of disagreement decided by a third annotator
- each token
  - two-component lemma (lemma proper and technical suffix)
  - positional tag (15 positions)

| *agentu* '(to an) agent' | |
|---|---|
| agent | NNMS3-----A---1 |

| *Hrbkovu* '(to) Hrbek's' | |
|---|---|
| Hrbkův_;S_^(*3ek) | AUMS3M-------- |

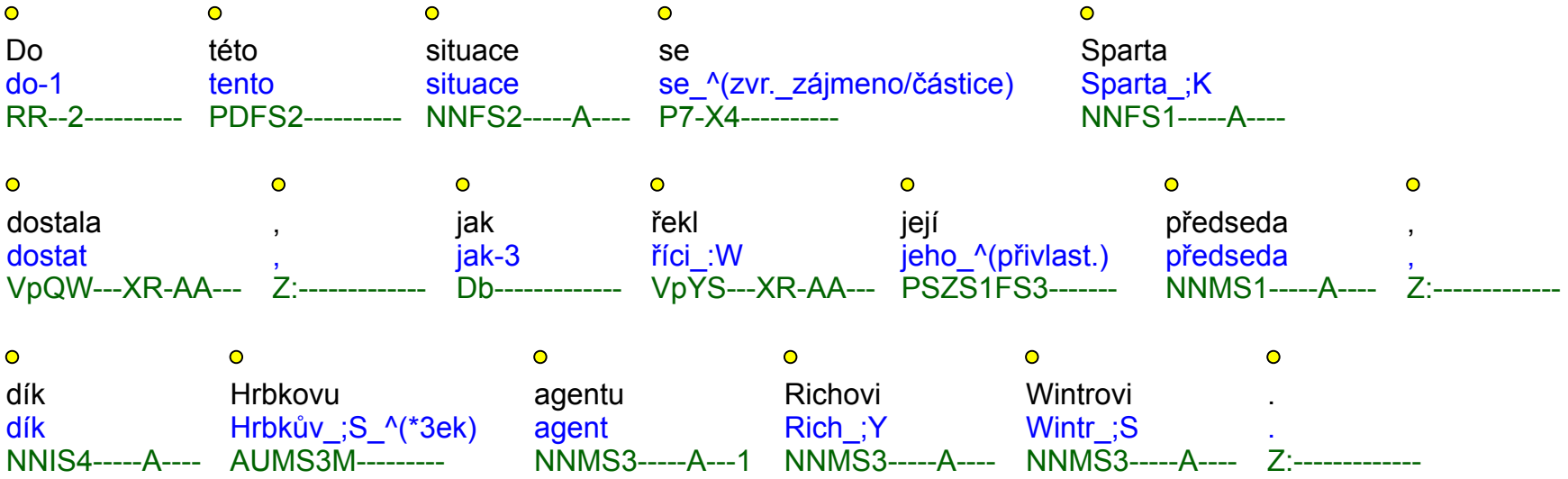
*Hrbkovu*

| | |
|---|---|
| Hrbkův_;S_^(*3ek) | AUMS3M--------- |

| Lemma part | Explanation |
|---|---|
| Hrbkův | lemma proper |
| _;S | technical suffix<br>*named entity type: surname* |
| _^(*3ek) | technical suffix<br>*derivation: substitute 3 last characters with "ek"* |

| | Position | In example |
|---|---|---|
| 1 | part of speech | A: adjective |
| 2 | detailed POS | U: possessive |
| 3 | gender | M: masc.anim. |
| 4 | number | S: singular |
| 5 | morph. case | 3: dative |
| 6 | possessor's gender | M: masc.anim. |
| 7 | possessor's number | |
| 8 | person | |
| 9 | tense | |
| 10 | degree of comp. | |
| 11 | negation | |
| 12 | verbal voice | |
| 13 | unused | |
| 14 | unused | |
| 15 | variant, register | |

*Do této situace se Sparta dostala, jak řekl její předseda, dík Hrbkovu agentu Richovi Wintrovi.*
lit.: Into this situation REFL Sparta got, as said her chairman, thanks Hrbek's agent Rich Winter.
'Sparta found itself in this situation, as its chairman said, thanks to Hrbek's agent Rich Winter.'



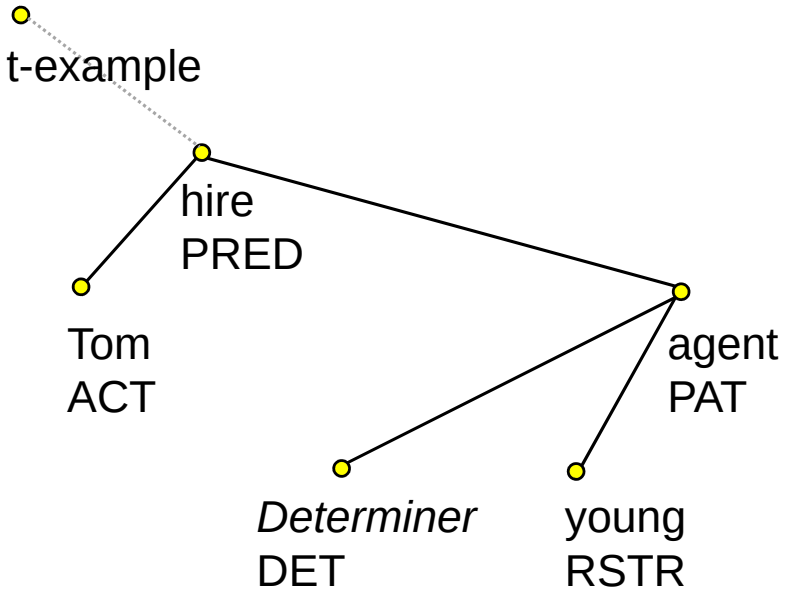
# Morphological annotation: an overview

- 1,960,000 tokens at the morphological layer of the PDT 3.0
  - 1,574 different positional tags (vs. 4k possible tags)
  - 71,503 different morphological lemmas

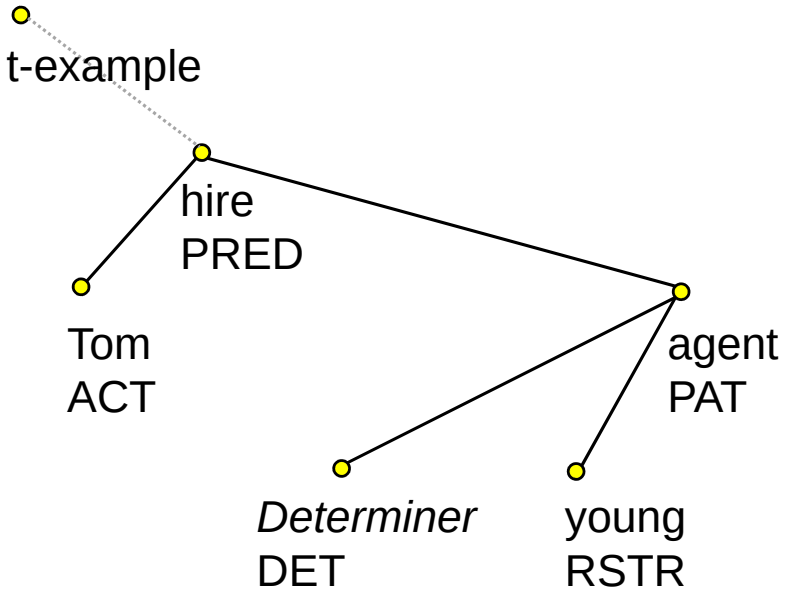
# Meanings expressed by morphological categories: Grammateme attributes at the tectogrammatical layer

- (a type of) node attributes in the tectogrammatical tree
- represent morphological meanings that participate in creating the meaning of the sentence, e.g.
  - number with nouns
  - degree of comparison with adjectives
  - tense with verbs
- **no** grammatemes for categories imposed by government or agreement
  - case with nouns
  - number and gender with adjectives
  - person, gender and number with verbs

# Grammatemes: Disambiguating meaning of the sentence

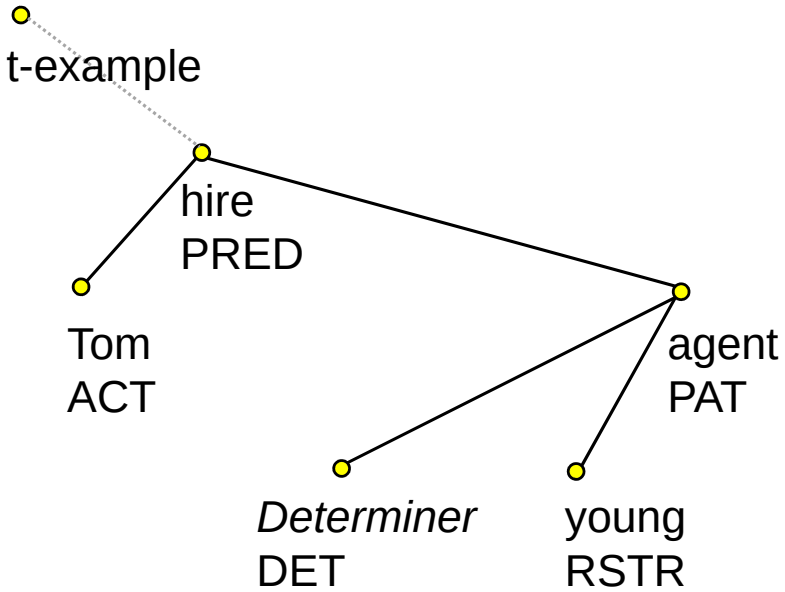


# Grammatemes: Disambiguating meaning of the sentence



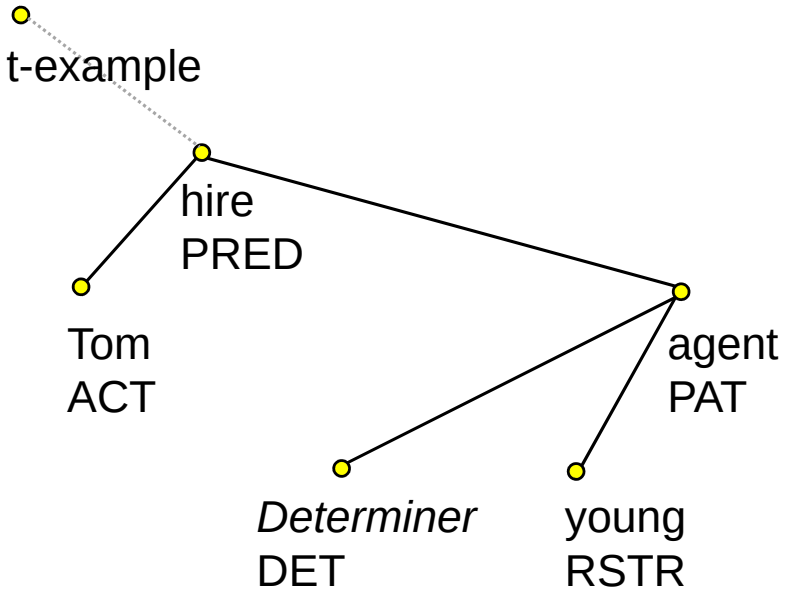

1. *Tom hired a young agent.*

# Grammatemes: Disambiguating meaning of the sentence



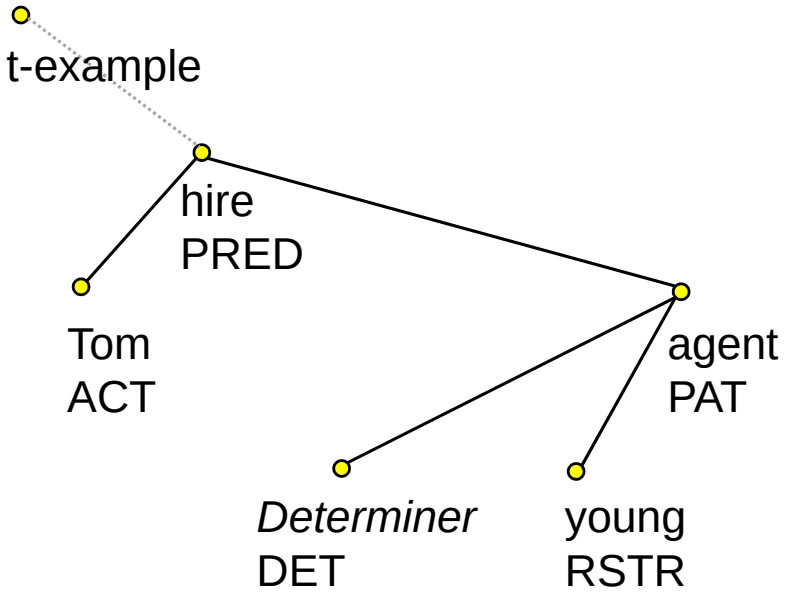

1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*

# Grammatemes: Disambiguating meaning of the sentence



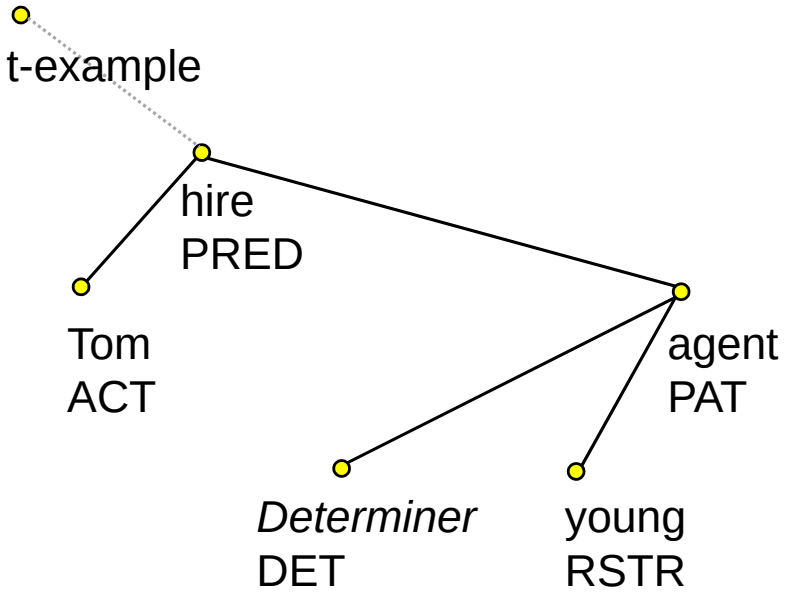

1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*
3. *Tom is hiring the youngest agent.*

# Grammatemes: Disambiguating meaning of the sentence



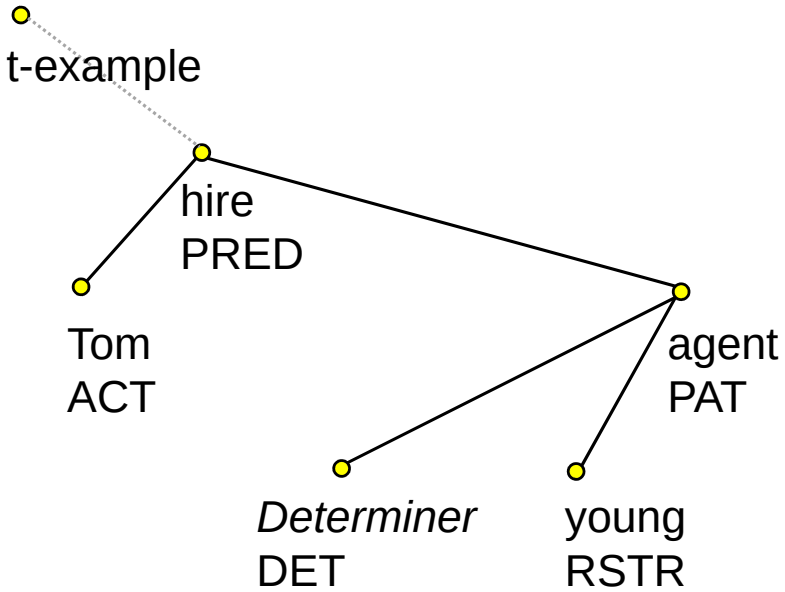

1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*
3. *Tom is hiring the youngest agent.*
4. *Tom will hire younger agents.*

# Grammatemes: Disambiguating meaning of the sentence



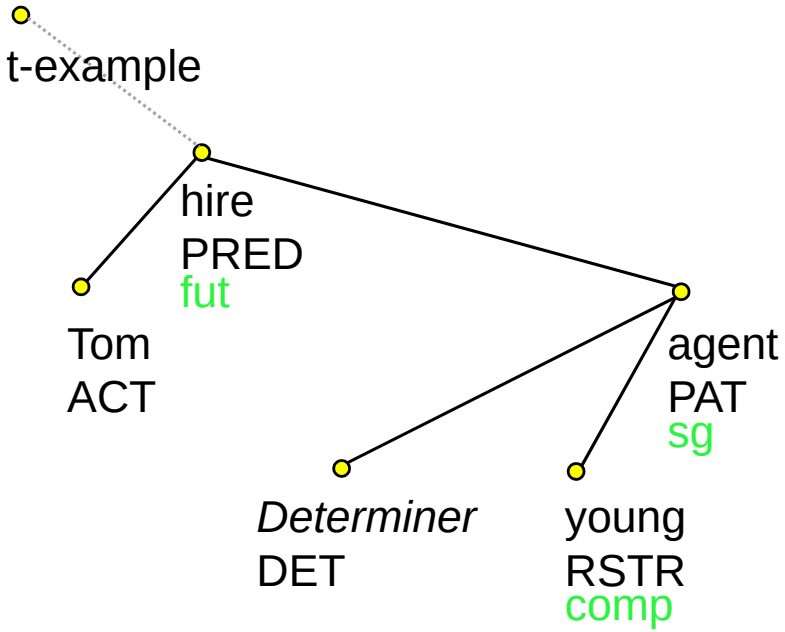

1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*
3. *Tom is hiring the youngest agent.*
4. *Tom will hire younger agents.*
5. *Tom hired young agents.*

# Grammatemes: Disambiguating meaning of the sentence




1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*
3. *Tom is hiring the youngest agent.*
4. *Tom will hire younger agents.*
5. *Tom hired young agents.*
6. . . .

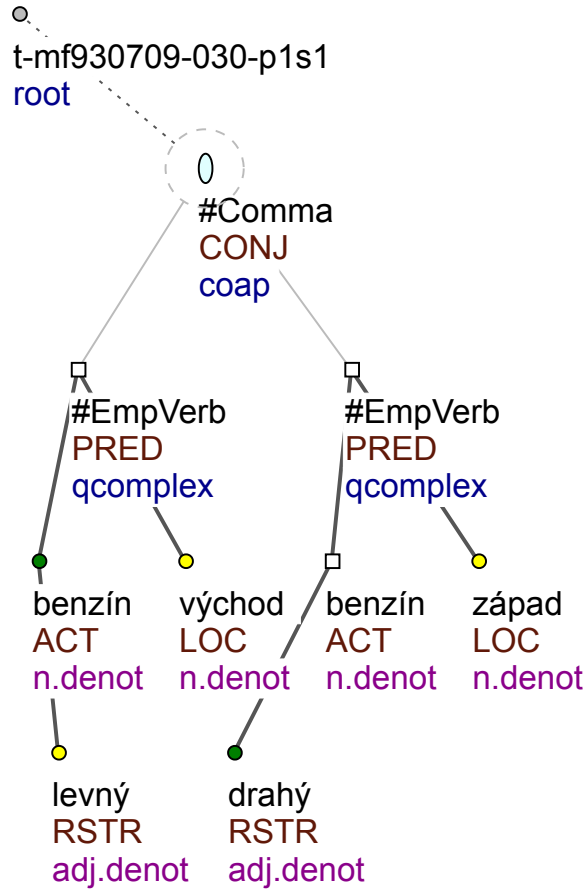
# Grammatemes: Disambiguating meaning of the sentence




1. *Tom hired a young agent.*
2. *Tom will hire a younger agent.*
3. *Tom is hiring the youngest agent.*
4. *Tom will hire younger agents.*
5. *Tom hired young agents.*
6. . . .

# Two-level typing of tectogrammatical nodes

1. 8 types of nodes
   - nodetype attribute
   - grammatemes relevant for complex nodes only
2. 4 semantic parts of speech
   - sempos attribute
   - semantic nouns, adjectives, adverbs, and verbs
   - 19 subgroups
   - the sempos value delimits the set of relevant gramamtemes




*Levnější benzín na Východě, dražší na Západě*
'Cheaper gasoline in the East, more expensive one in the West'

# 15 grammatemes in PDT 3.0
## Semantic nouns, adjectives, and adverbs

1. number: number of entities which a noun refers to
2. typgroup: plural forms of nouns denoting pairs/groups
3. gender: grammatical gender of nouns
4. person: with pronouns (speaker vs. hearer vs. nonparticipant)
5. politeness: polite usage of 2nd person pronouns
6. degcmp: degree of comparision with adjectives and adverbs
7. negation: negated nouns etc. represented by positive counterparts
8. indeftype: pronominals reduced to a small set of lemmas
9. numertype: numerals reduced to cardinals

# 15 grammatemes in PDT 3.0
## Semantic nouns, adjectives, and adverbs

1. number: number of entities which a noun refers to
2. typgroup: plural forms of nouns denoting pairs/groups
3. gender: grammatical gender of nouns
4. person: with pronouns (speaker vs. hearer vs. nonparticipant)
5. politeness: polite usage of 2nd person pronouns
6. degcmp: degree of comparision with adjectives and adverbs
7. negation: negated nouns etc. represented by positive counterparts
8. indeftype: pronominals reduced to a small set of lemmas
9. numertype: numerals reduced to cardinals

# 15 grammatemes in PDT 3.0
## Semantic nouns, adjectives, and adverbs

1. number: number of entities which a noun refers to
2. typgroup: plural forms of nouns denoting pairs/groups
3. gender: grammatical gender of nouns
4. person: with pronouns (speaker vs. hearer vs. nonparticipant)
5. politeness: polite usage of 2nd person pronouns
6. degcmp: degree of comparision with adjectives and adverbs
7. negation: negated nouns etc. represented by positive counterparts
8. indeftype: pronominals reduced to a small set of lemmas
9. numertype: numerals reduced to cardinals

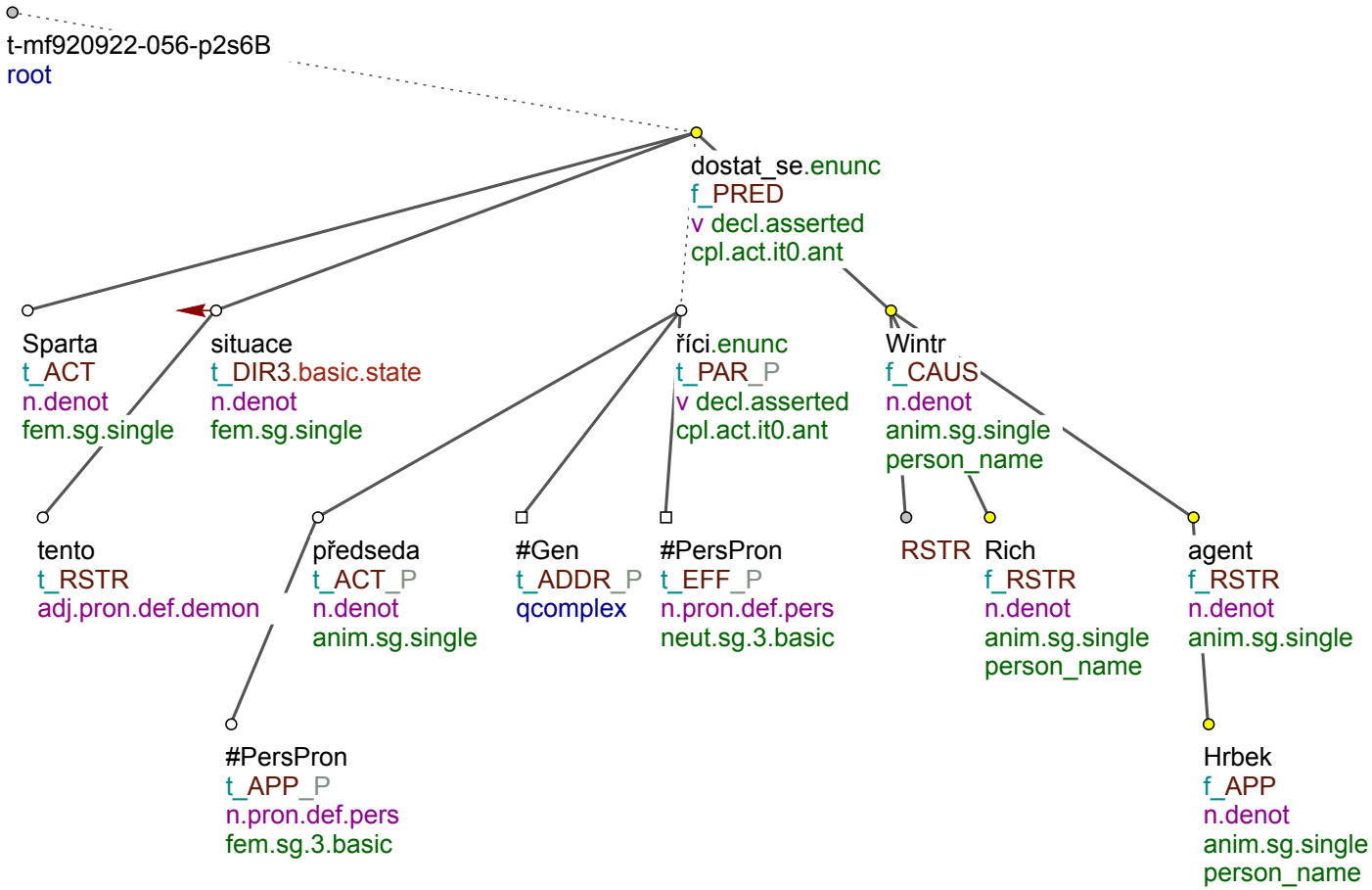
# 15 grammatemes in PDT 3.0
## Semantic nouns, adjectives, and adverbs

1. number: number of entities which a noun refers to
2. typgroup: plural forms of nouns denoting pairs/groups
3. gender: grammatical gender of nouns
4. person: with pronouns (speaker vs. hearer vs. nonparticipant)
5. politeness: polite usage of 2nd person pronouns
6. degcmp: degree of comparision with adjectives and adverbs
7. negation: negated nouns etc. represented by positive counterparts
8. indeftype: pronominals reduced to a small set of lemmas
9. numertype: numerals reduced to cardinals

# 15 grammatemes in PDT 3.0
## Semantic verbs

1. tense: past vs. present vs. future events
2. factmod: asserted vs. potential vs. irreal events
3. aspect: imperfective vs. perfective verbs
4. deontmod: modal verbs represented as auxiliaries
5. diatgram: gramaticalized diatheses of verbs
6. iterativeness: iterative verbs represented by non-iteratives

*Do této situace se Sparta dostala, jak řekl její předseda, dík Hrbkovu agentu Richovi Wintrovi.*
'Sparta found itself in this situation, as its chairman said, thanks to Hrbek's agent Rich Winter.'

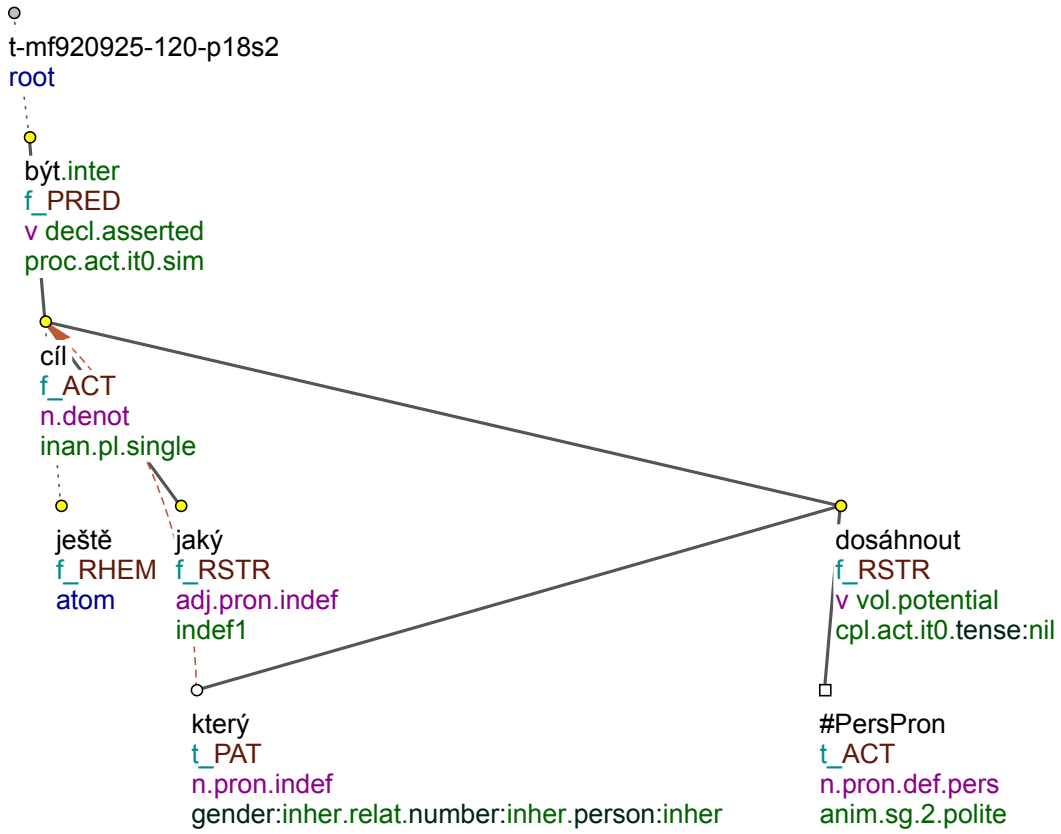


# Annotation of grammatemes

- the last task in the PDT 2.0 annotation procedure
- automatic assignment based on
  - morphological annotation
    - grammateme values cannot be mostly interpreted from the positional tag of a single word form
    - more complex structures including auxiliaries involved in the value assignment procedure
  - preceding tectogrammatical annotations
    - tree structure
    - semantic roles
    - coreference
  - lexical resources
    - special-purpose lists of pronouns, adverbs, verbs
- manual annotation of special problems
  - e.g. number with pluralia tantum

# Automatic assignment of grammatemes: using positional tags, tree structure, and lexical lists

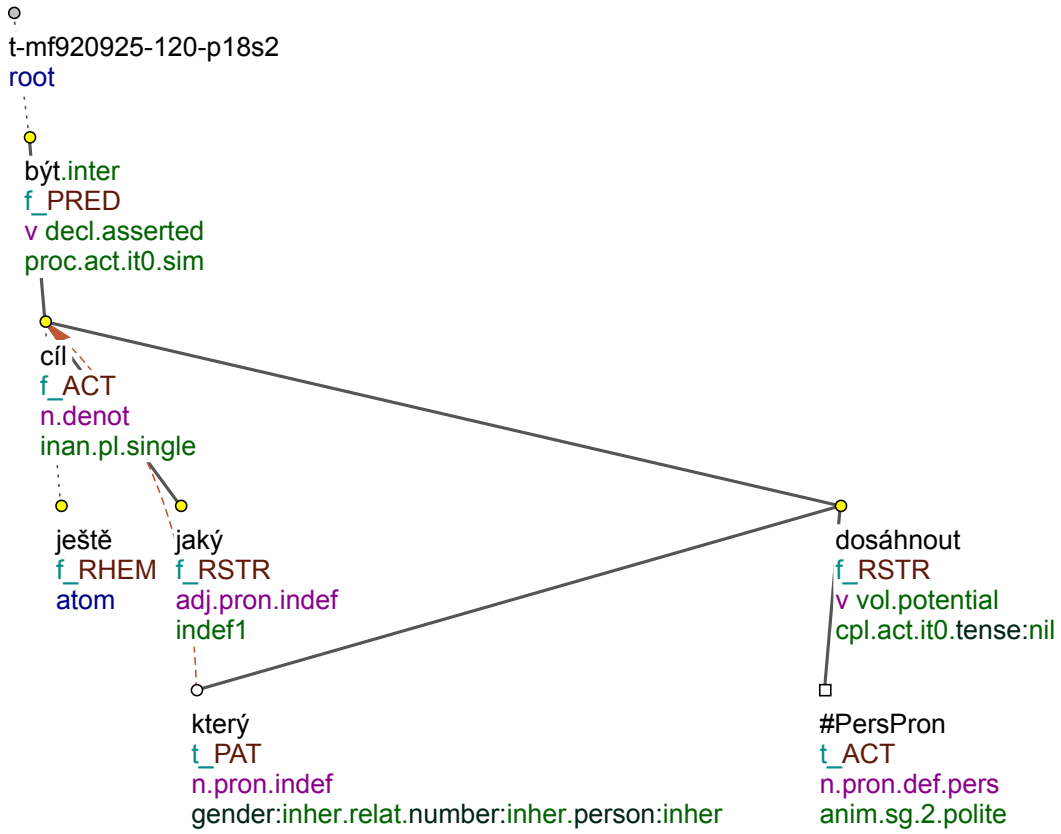
- number grammateme
  - from positional tags with most nouns
  - from verb forms with pro-drops
- factmod grammateme
  - from positional tags of (auxiliary) verb forms
- deontmod grammateme
  - *chtít* 'to want'
- indeftype grammateme
  - *nějaký* 'some' > *jaký*




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are there any goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using positional tags, tree structure, and lexical lists

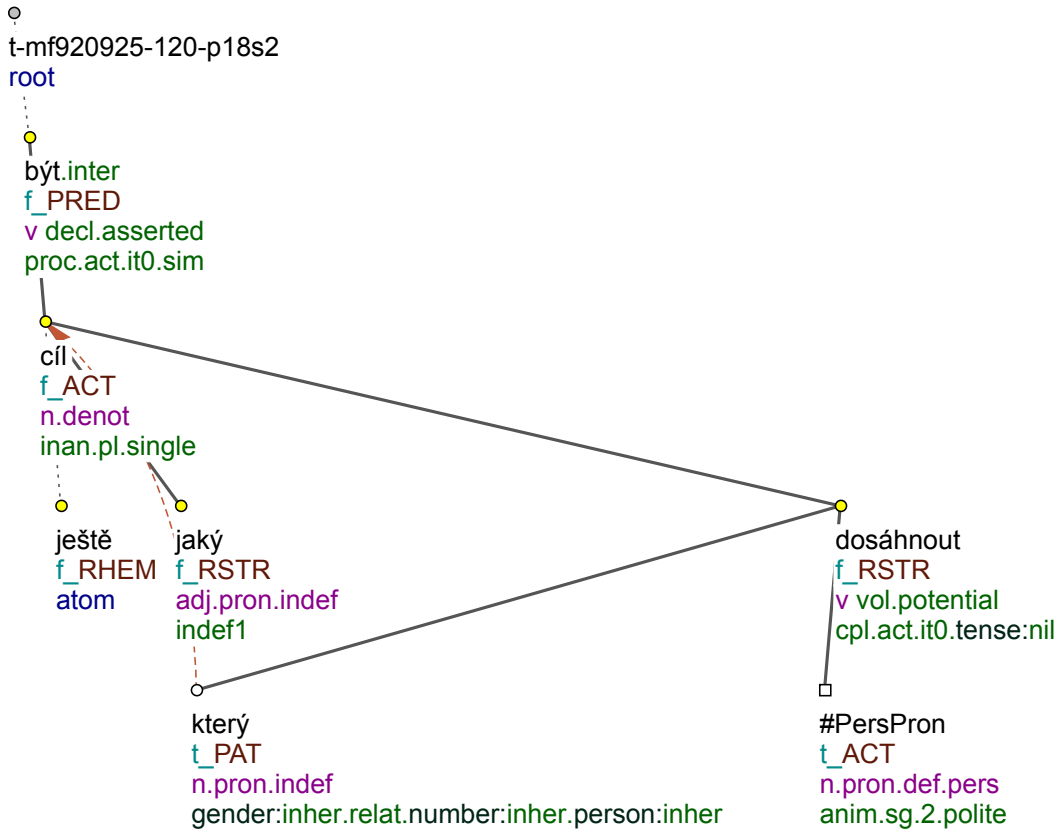
- number grammateme
  - from positional tags with most nouns
  - from verb forms with pro-drops
- factmod grammateme
  - from positional tags of (auxiliary) verb forms
- deontmod grammateme
  - *chtít* 'to want'
- indeftype grammateme
  - *nějaký* 'some' > *jaký*




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are there any goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using positional tags, tree structure, and lexical lists

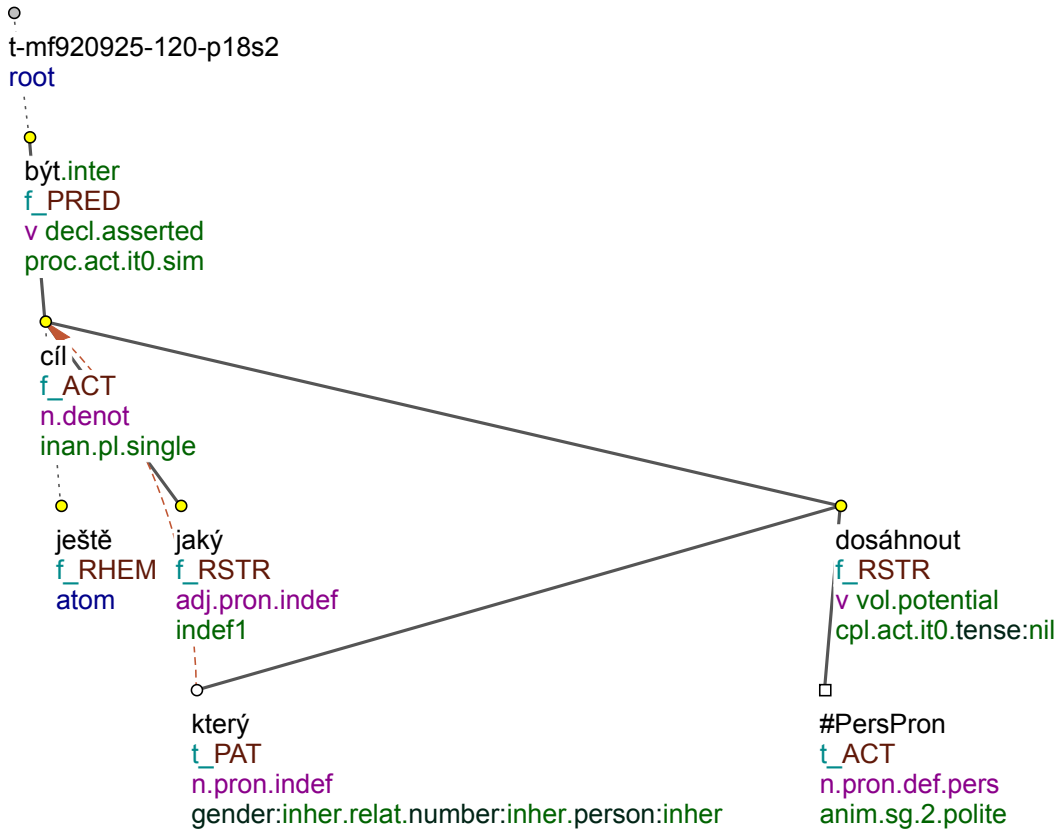
- number grammateme
  - from positional tags with most nouns
  - from verb forms with pro-drops
- factmod grammateme
  - from positional tags of (auxiliary) verb forms
- deontmod grammateme
  - *chtít* 'to want'
- indeftype grammateme
  - *nějaký* 'some' > *jaký*




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are there any goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using positional tags, tree structure, and lexical lists

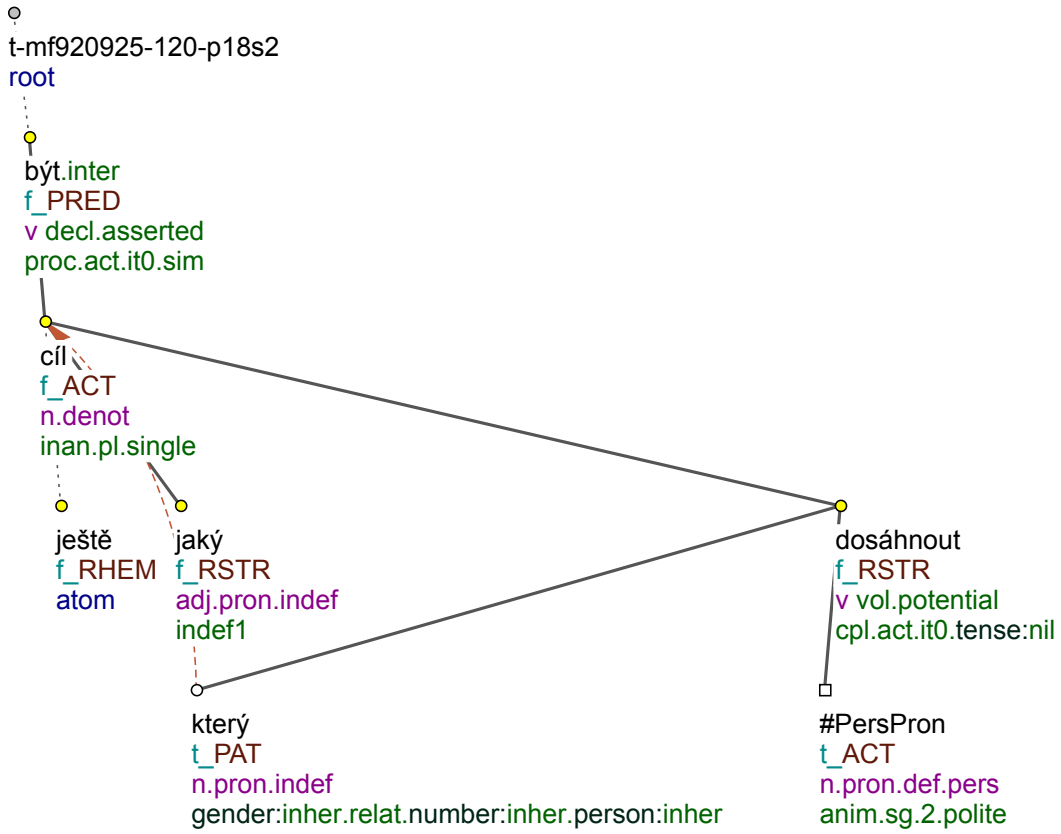
- number grammateme
  - from positional tags with most nouns
  - from verb forms with pro-drops
- factmod grammateme
  - from positional tags of (auxiliary) verb forms
- deontmod grammateme
  - *chtít* 'to want'
- indeftype grammateme
  - *nějaký* 'some' > *jaký*




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are there any goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using positional tags, tree structure, and lexical lists

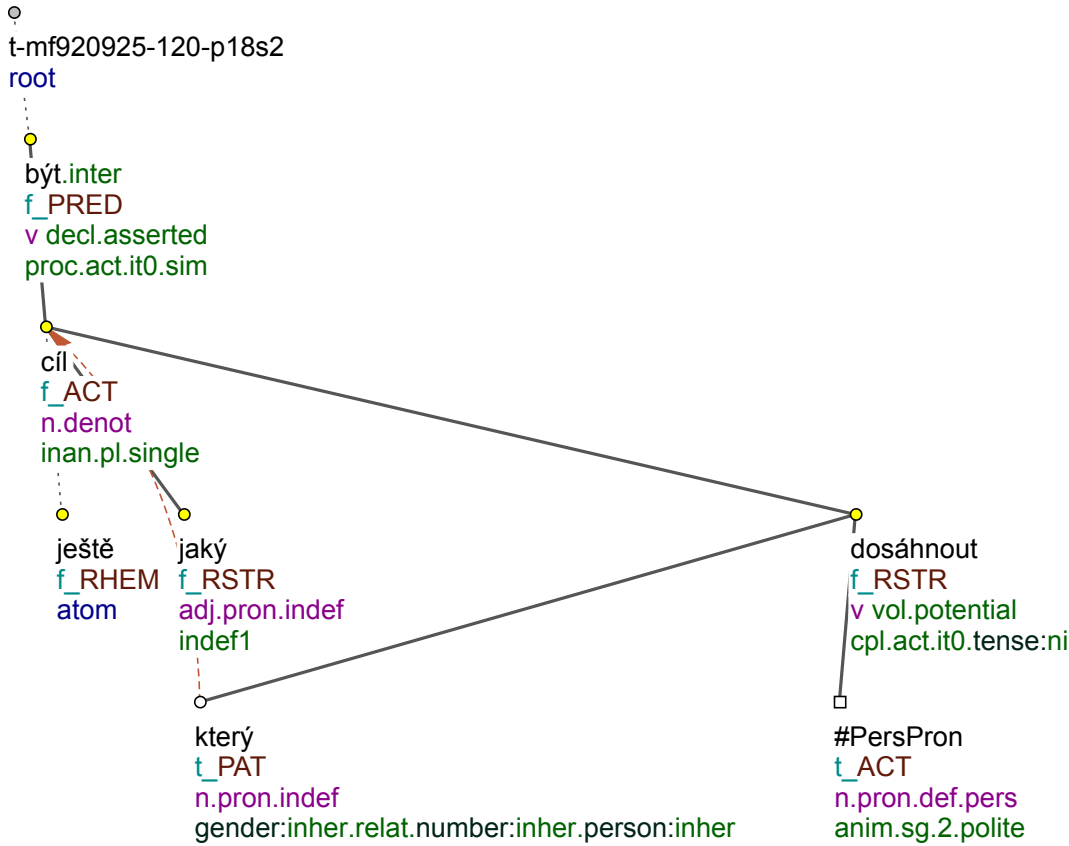
- number grammateme
  - from positional tags with most nouns
  - from verb forms with pro-drops
- factmod grammateme
  - from positional tags of (auxiliary) verb forms
- deontmod grammateme
  - *chtít* 'to want'
- indeftype grammateme
  - *nějaký* 'some' > *jaký*




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are there any goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using coreference

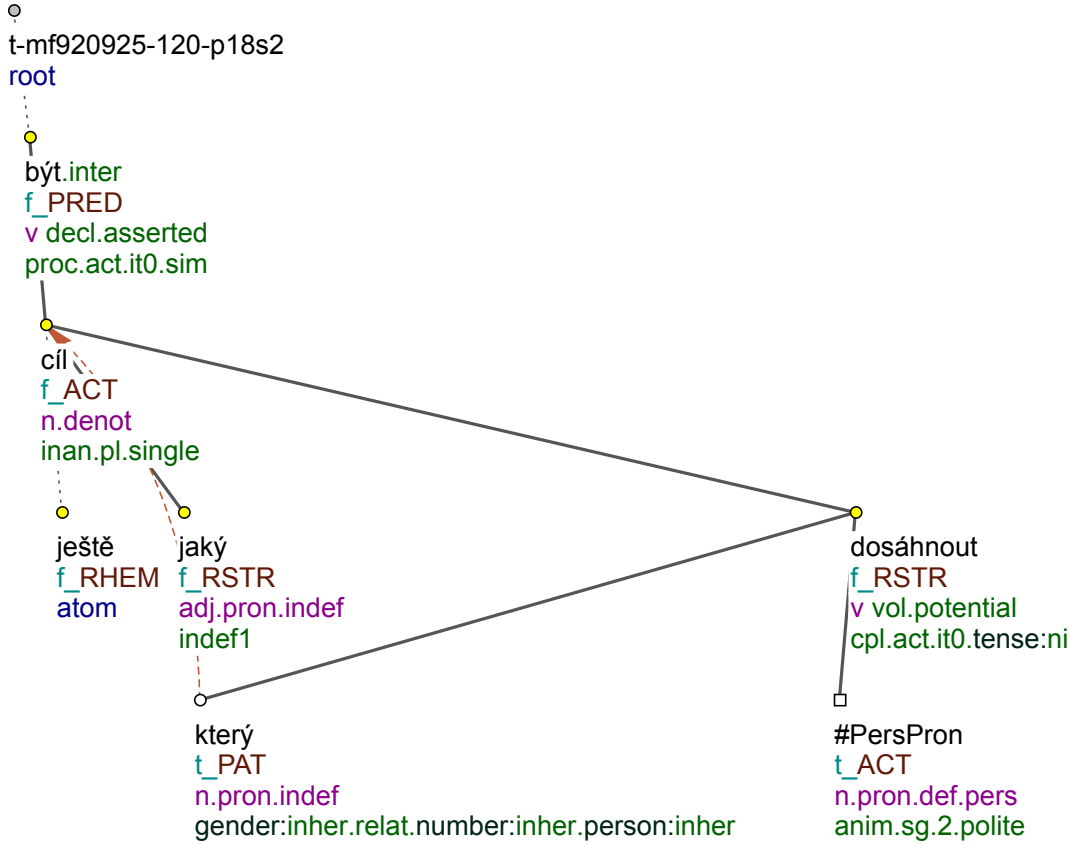
- relative pronouns
  - grammatical categories imposed by agreement
  - inherited from the antecedent
  - values underspecified (inher value)




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are thereany goals which (you) would like to achieve?'

# Automatic assignment of grammatemes: using coreference

- relative pronouns
  - grammatical categories imposed by agreement
  - inherited from the antecedent
  - values underspecified (inher value)




*Jsou ještě nějaké cíle, kterých byste chtěl dosáhnout?*
'Are thereany goals which (you) would like to achieve?'

# Automatic vs. manual annotation of grammatemes

- **1,600,000** grammateme values assigned to 550,000 complex nodes at the tectogrammatical layer of PDT 2.0
- **17,500** out of them assigned manually

# Manual annotation of grammatemes

- two annotators in parallel
  - inter-annotator agreement: 70–85 %
- simplified annotation environment
  - treebank positions extracted into simple HTML forms
- pluralia tantum
  - *Otevřel dveře*.sg *na terasu.* 'He opened the door to the terrace.'
  - vs. *několikery dveře*.pl 'several doors'
- polite usage of 2nd person pronouns
  - *Vy*.polite *jste se už přihlásil?* 'Have you logged in already?'
  - vs. *Vy*.basic *jste se už přihlásili?* 'Have you logged in already?'
- absolute usage of comparative forms of adjectives and adverbs
  - *starší*.acomp *žena* 'an elder woman'
  - vs. *jeho starší*.comp *bratr* 'his older brother'
- biaspectual verbs, pair/group meaning of plural forms, ...

# PDT data: developing taggers of Czech

- feature-based tagger (Hajič 2004)
  - part of the PDT 2.0 release
- HMM tagger (Krbec 2005)
- *Morče* tagger (Votrubec 2005)
  - averaged perceptron
- combined approach (Spoustová et al. 2007)
  - *Morče* tagger, feature-based tagger, HMM tagger, and a rule-based component
- *Morče* tagger semi-supervised (Spoustová et al. 2009)
- *MorphoDiTa* (Straková et al. 2014)
  - open-source tool for morphological analysis, tagging, lemmatization, tokenization, and morphological generation
  - available with trained linguistic models

# Accuracy of taggers

| Czech taggers (PDT 2.5) | Accuracy |
|---|---|
| *Morče* semi-supervised (Spoustová et al. 2009) | 95.89 % |
| *MorphoDiTa* (Straková et al. 2014) | 95.75 % |
| combination of taggers (Spoustová et al. 2007) | 95.70 % |
| *Morče* (Votrubec 2005) | 95.67 % |
| HMM (Krbec 2005) | 94.82 % |
| feature-based tagger (Hajič 2004) | 94.04 % |

# Accuracy of taggers

| Czech taggers (PDT 2.5) | Accuracy |
|---|---|
| *Morče* semi-supervised (Spoustová et al. 2009) | 95.89 % |
| *MorphoDiTa* (Straková et al. 2014) | 95.75 % |
| combination of taggers (Spoustová et al. 2007) | 95.70 % |
| *Morče* (Votrubec 2005) | 95.67 % |
| HMM (Krbec 2005) | 94.82 % |
| feature-based tagger (Hajič 2004) | 94.04 % |

| English taggers (PennTB/WSJ) | Accuracy |
|---|---|
| Shen et al. (2007) | 97.33 % |
| *MorphoDiTa* (Straková et al. 2014) | 97.27 % |
| *Morče* semi-supervised (Spoustová et al. 2009) | 97.23 % |

# Accuracy of taggers

| Czech taggers (PDT 2.5) | Accuracy |
|---|---|
| *Morče* semi-supervised (Spoustová et al. 2009) | 95.89 % |
| *MorphoDiTa* (Straková et al. 2014) | 95.75 % |
| combination of taggers (Spoustová et al. 2007) | 95.70 % |
| *Morče* (Votrubec 2005) | 95.67 % |
| HMM (Krbec 2005) | 94.82 % |
| feature-based tagger (Hajič 2004) | 94.04 % |

| English taggers (PennTB/WSJ) | Accuracy |
|---|---|
| Shen et al. (2007) | 97.33 % |
| *MorphoDiTa* (Straková et al. 2014) | 97.27 % |
| *Morče* semi-supervised (Spoustová et al. 2009) | 97.23 % |

*MorphoDiTa* (Czech, first 2 positions) 99.18 %

# Named entity classification and recognition in Czech

- pilot approach in 2007
- two-level classification
  - rough and detailed categories
  - embedding allowed

| g | *geographical names* |
|---|---|
| | gp *planets* |
| | gt *continents* |
| | gc *states* |
| | gu *towns* |
| | gs *streets, squares* |
| | gh *hydronyms* |
| | ... |

| p | *person names* |
|---|---|
| | pf *first names* |
| | ps *surnames* |
| | pm *second names* |
| | pd *(academic) titles* |
| | pc *inhabitant names* |
| | pp *religious/myth. persons* |
| | ... |

- 5 recognizers since 2007
  - trained on Czech Named Entity Corpus

# Czech Named Entity Corpus

http://ufal.mff.cuni.cz/cnec/

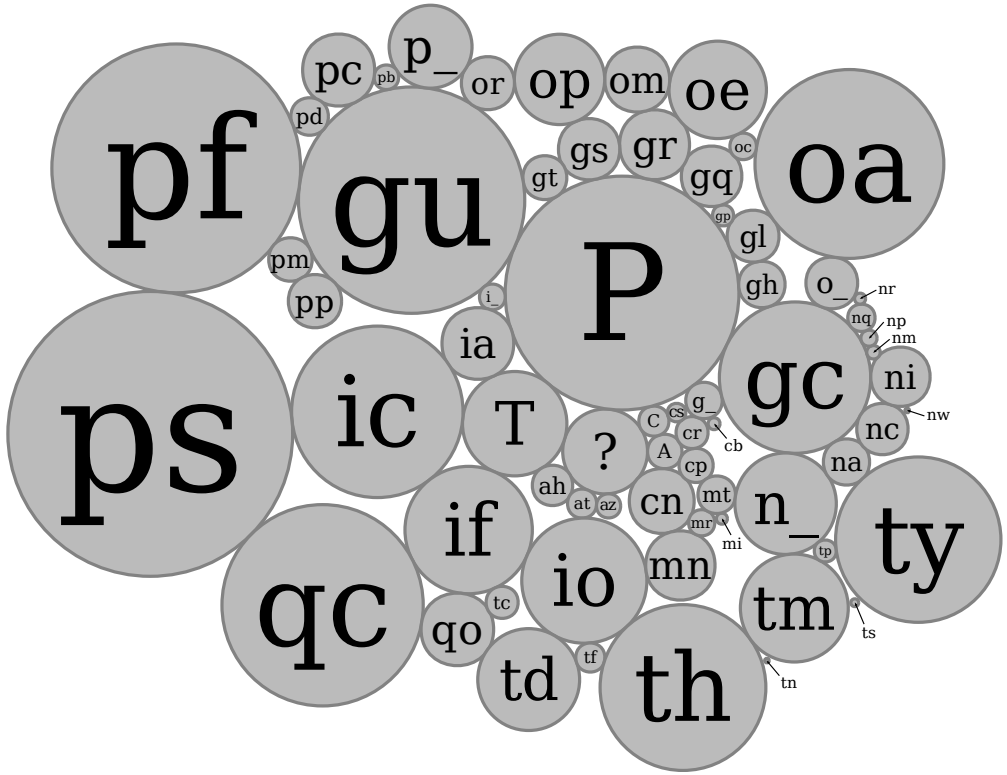
- data selection
  - random selection of isolated 6k sentences with 150k tokens
  - 33k NEs manually assigned by two annotators in parallel
- categories annotated
  - 7 rough categories in CNEC 1.1, 10 in CNEC 2.0
  - 42 detailed categories in CNEC 1.1, 62 in CNEC 2.0
- LINDAT/Clarin repository
  - CNEC 1.0 (2009)
  - CNEC 1.1 (2014)
  - CNEC 2.0 (2014)

# Czech Named Entity Corpus

http://ufal.mff.cuni.cz/cnec/

- data selection
  - random selection of isolated 6k sentences with 150k tokens
  - 33k NEs manually assigned by two annotators in parallel
- categories annotated
  - 7 rough categories in CNEC 1.1, 10 in CNEC 2.0
  - 42 detailed categories in CNEC 1.1, 62 in CNEC 2.0
- LINDAT/Clarin repository
  - CNEC 1.0 (2009)
  - CNEC 1.1 (2014)
  - CNEC 2.0 (2014)

```
1:  <P<pf Jan> <ps Stavěl>>
byl dlouho činným , zemřel
jako stařešina moravského
hasičstva krátce před
dovršením <qo 75 .>
narozenin v <tm únoru> <ty
1933> .
2:  " Začínala jsem v roce
<ty 1995> s osmi chovanci
místního ústavu , dnes jich
pracuje třináct , " uvedla
ke vzniku mimořádného
seskupení herečka <P<pf
Viera> <ps Dubačová>> .
3:  V současné době je v <i_
<s CECIMO>> tedy <qc 14>
členů .
4:  Vnitřní reforma <io
Unie> dosud neproběhla a
válka na <gl Balkáně>
odčerpá finanční prostředky
:  <io<s EU>> bude
investovat do poválečné
obnovy <gc Jugoslávie> .
```




[Straková et al. 2015]

# Named entity recognizers for Czech

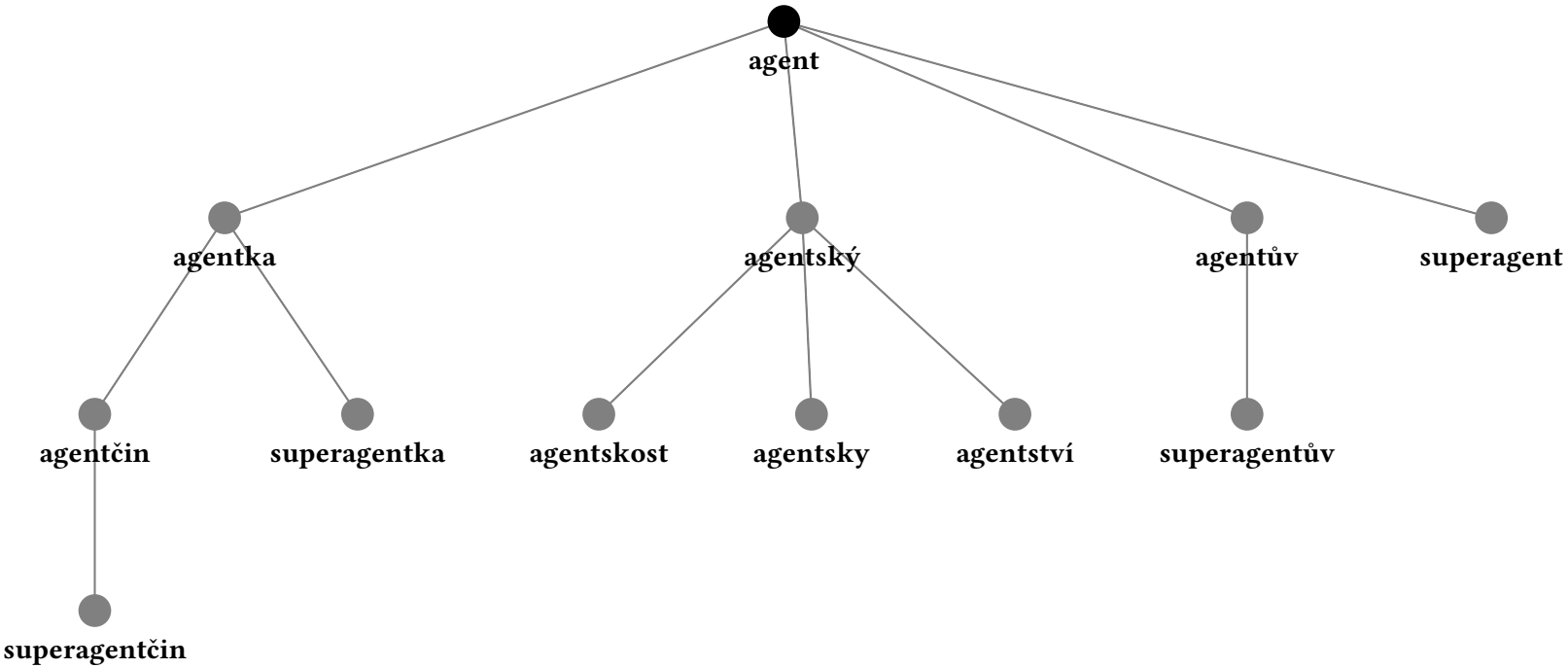
| System | F-measure (7 categories) | F-measure (42 categories) |
|---|---|---|
| Straková et al. (2013) | 82.82 | 79.23 |
| Straková et al. (NameTag; 2014) | 81.01 | 77.88 |
| Konkol – Konopík (2013) | 79.00 | na |
| Kravalová et al. (2009) | 71.00 | 68.00 |
| Ševčíková et al. (2007) | 68.00 | 62.00 |

# Derivational morphology of Czech

- derivational morphology underresourced in most languages
  - CELEX for English, German and Dutch (Baayen et al. 1995)
  - DerivBase for German (Zeller et al. 2013)
  - DerivBase.Hr for Croatian (Šnajder et al. 2014)
  - language-independent approach by Baranes – Sagot (2014)
  - Démonette network for French (Hathout – Namer 2014)
  - DeriNet for Czech (Ševčíková – Žabokrtský 2014)

# DeriNet: lexical resource of derivational relations in Czech

- 970k lemmas connected with 715k derivational relations
  - compatible with the MorfFlex dictionary



# DeriNet 1.0

| | DeriNet 1.0 |
|---|---|
| lemmas | **968,967** |
| unique lemmas | 965,535 |
| derivational links | **715,729** |
| derivational clusters | 253,238 |
| singleton clusters | 101,311 |
| maximum lemmas per cluster | 82 |
| maximum cluster depth | 8 |

[Vidra 2015]

# DeriNet 1.0

| | DeriNet 1.0 |
|---|---|
| lemmas | **968,967** |
| unique lemmas | 965,535 |
| derivational links | **715,729** |
| derivational clusters | 253,238 |
| singleton clusters | 101,311 |
| maximum lemmas per cluster | 82 |
| maximum cluster depth | 8 |

[Vidra 2015]

# Derivational information in dependency trees

- derivational information currently available in PDT
  - lemma suffix at the morphological layer
  - selected grammatemes and semantic roles at the tectogrammatical layer
- extending derivational annotation in tectogrammatical trees
  - most frequent semantic classes
  - derived words substituted by the lemma of the base word
  - the word-formation meaning stored in a deriveme attribute

# Dependencies and derivations: natural language processing

- machine translation: out-of-vocabulary words
  - English adverbs ending in *-ly*
  - Czech female profession names
  - Czech diminutives
- parsing
  - sublexical analysis helped to achieve state-of-the-art results in parsing the Turkish Treebank (Eryigit et al. in *Computational Linguistics* 2008)
- paraphrasing, ...

# Dependencies and derivations: linguistic research

- derivational morphemes vs. valency of verbs and nouns
  - Actor of *učit* 'to teach' incorporated in *učitel* 'teacher'
  - Patient of the verb *dát* 'to give' involved in *dárek* 'present'
- derivational morphology of Czech vs. other languages
  - *padnout – fallen – to fall*
  - *napadnout – auffallen – to stand out*
  - *vypadnout – ausfallen – to drop out*
- alignment at the level of morphemes
  - diminutive suffix *Karlík* vs. noun phrase *little Charles*

# Conclusions (i)

- Prague Dependency Treebank
  - morphology as a separate layer of annotation
  - grammateme attributes at the tectogrammatical layer
- universal part-of-speech tags
  - substituting language- and framework-specific tagsets
  - grammatemes not yet confronted
- lemmatization and tagging an essential prerequisite for most NLP tasks in Czech

# Conclusions (ii)

- analysing derivational morphology
- derivational analysis in dependency trees
  - substituting derived words with base words as an extended lemmatization
  - advantageous for NLP and linguistic research
  - beware of 'overloading' the data
- (automatic) morphemic analysis missing
  - semantic classification of affixes

# References


- Baayen, H. et al.: *The CELEX lexical database* (release 2). LDC 1995.
- Baranes, M. – Sagot, B.: A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. *LREC* 2014:2793–2799.
- Eryigit, G. et al.: Dependency Parsing of Turkish. *Computational Linguistics* 2008:34, 357–389.
- Hajič, J.: *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Prague 2004.
- Hajič, J. et al.: *Prague Dependency Treebank 2.0*. LDC 2006.
- Hathout, N. – Namer, F.: Démonette, a French Derivational Morpho-Semantic network. *LiLT* 2014:11, 125–168.
- Straková, J. et al.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. *ACL 2014: System Demonstrations*, 13–18.
- Ševčíková, M. – Žabokrtský, Z.: Word-Formation Network for Czech. *LREC* 2014: 1087–1093.
- Šnajder, J. et al.: DerivBase.Hr: A High-Coverage Derivational Morphology Resource for Croatian. *LREC* 2014: 3371–3377.
- Vidra, J.: *Extending the lexical nerwork DeriNet*. Bc Thesis, Charles University in Prague 2015.
- Zeller, B. et al.: DErivBase: Inducing and evaluating a derivational morphology resource for German. *ACL* 2013: 1201–1211.
- Zeman, D.: From the Jungle to a Park: Harmonizing Annotations across Languages. Key note at *SPMRL* 2015, Bilbao.
- ... *see References in the SFCM 2015 proceedings paper*