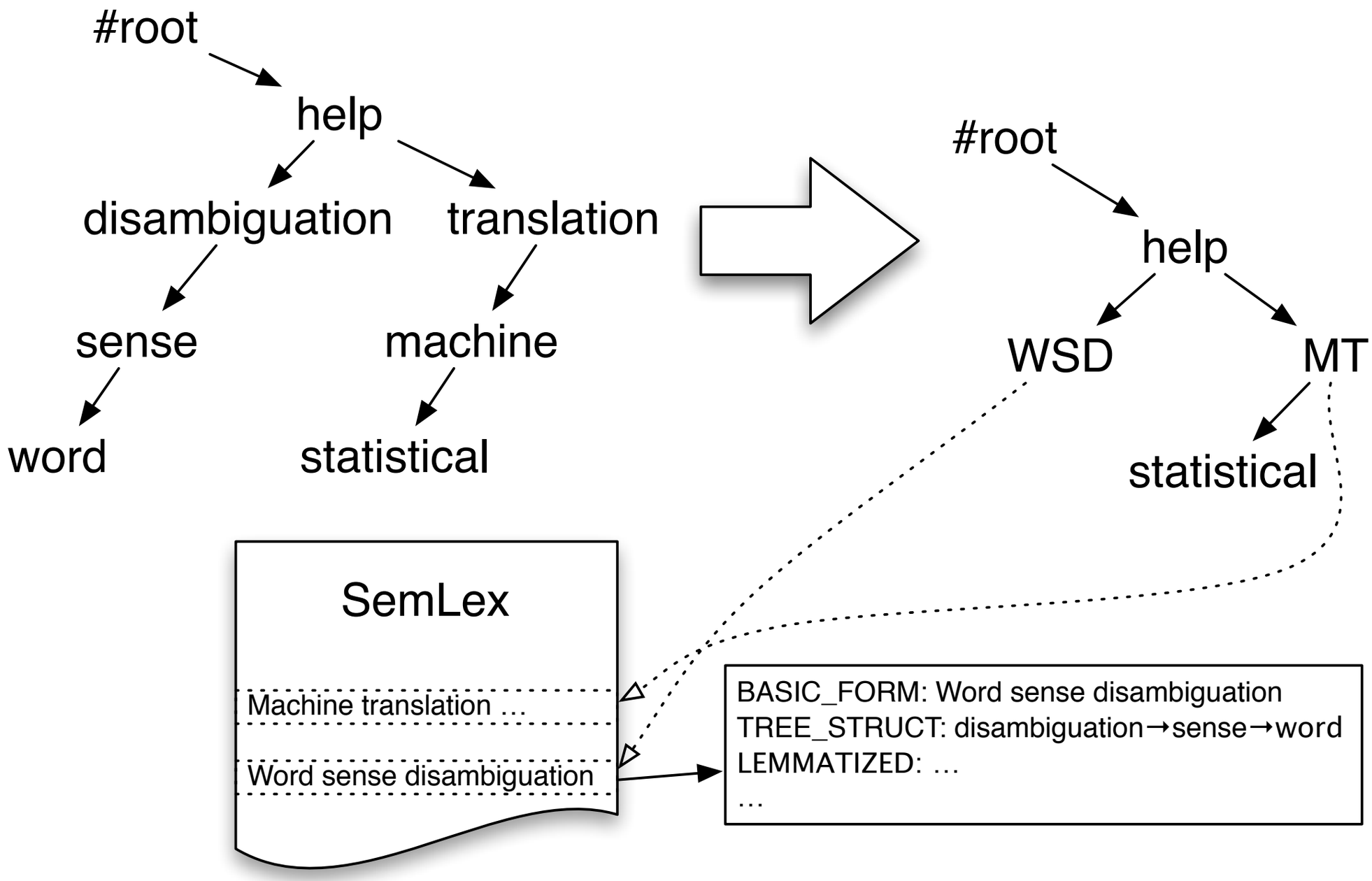
Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures

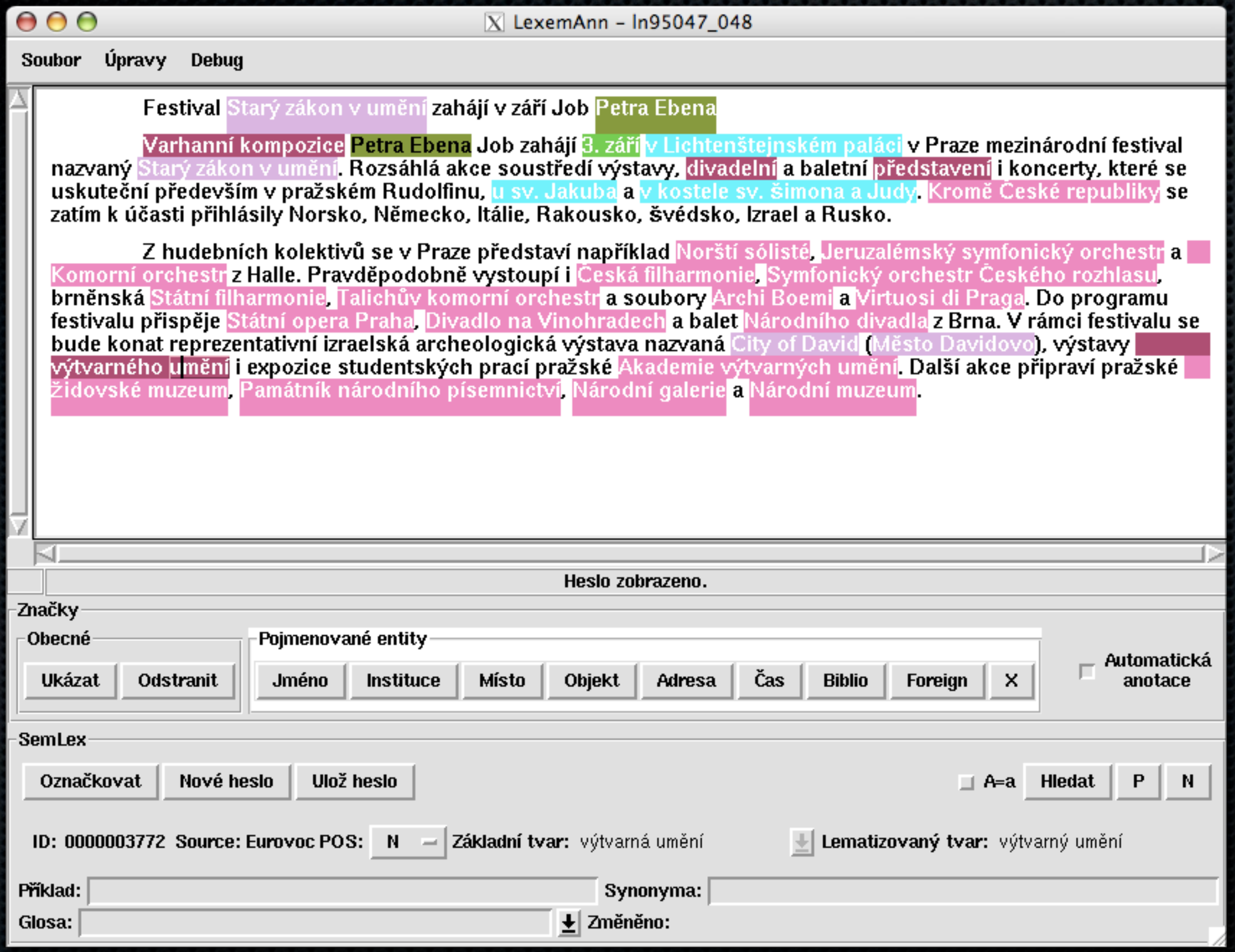
Eduard Bejček, Pavel Straňák, Pavel Pecina

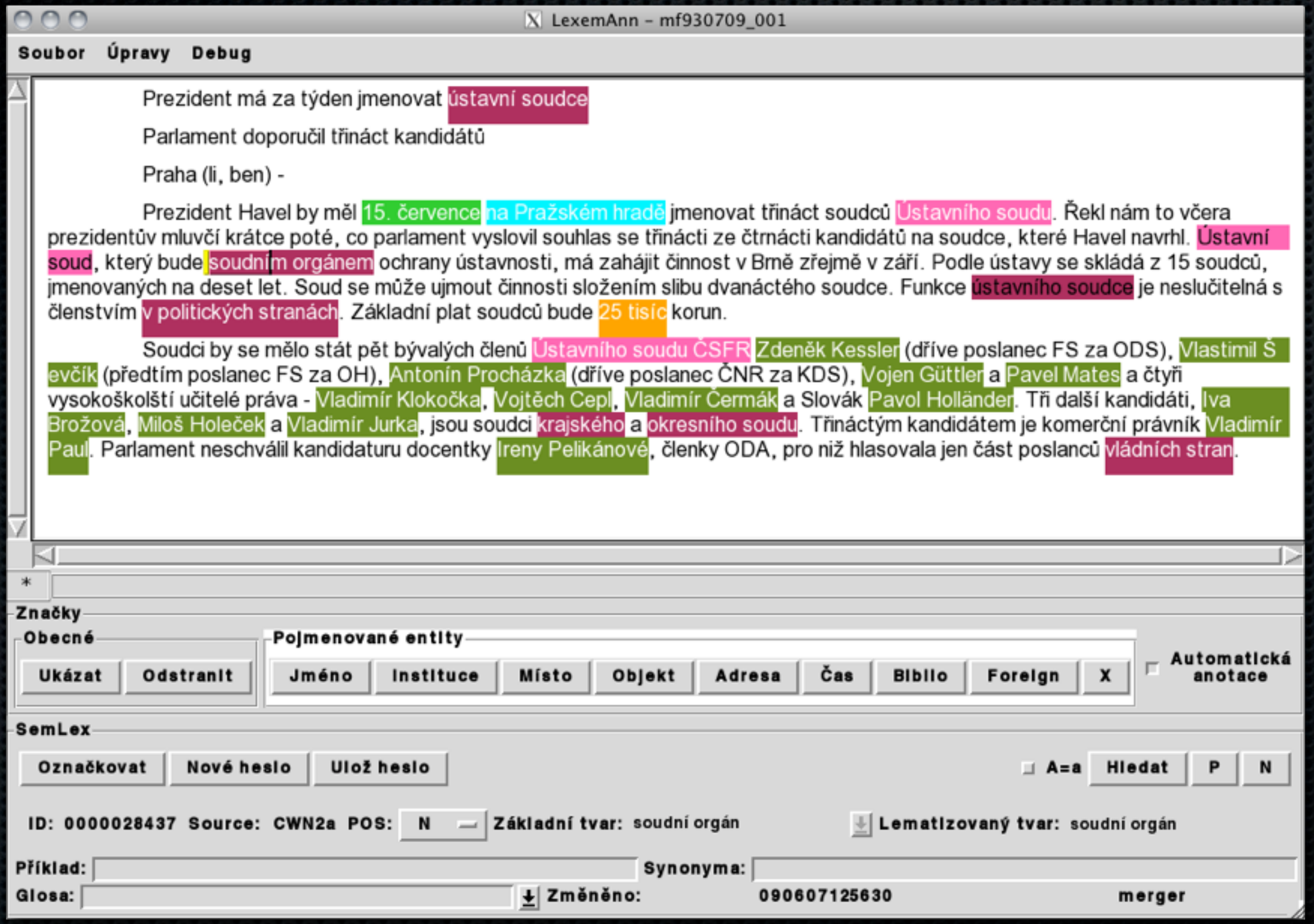
Objective

- We have **SemLex** – lexicon of all MWEs in PDT 2.5
  - basic (quotation) forms
  - lemmatised forms
  - dependency structures
- We have **PDT 2.5** with MWE annotation (instances)
- **How to find MWEs from Semlex in new texts?**

Can word sense disambiguation help statistical machine translation?




LexemAnn - ln95047_048
Soubor
Úpravy
Debug
Festival Starý zákon v umění zahájí v září Job Petra Ebena
Varhanní kompozice Petra Ebena Job zahájí 3. září v Lichtenštejnském paláci v Praze mezinárodní festival nazvaný Starý zákon v umění. Rozsáhlá akce soustředí výstavy, divadelní a baletní představení i koncerty, které se uskuteční především v pražském Rudolfinu, u sv. Jakuba a v kostele sv. šimona a Judy. Kromě České republiky se zatím k účasti přihlásily Norsko, Německo, Itálie, Rakousko, švédsko, Izrael a Rusko.
Z hudebních kolektivů se v Praze představí například Norští sólisté, Jeruzalémský symfonický orchestr a Komorní orchestr z Halle. Pravděpodobně vystoupí i Česká filharmonie, Symfonický orchestr Českého rozhlasu, brněnská Státní filharmonie, Talichův komorní orchestr a soubory Archi Boemi a Virtuosi di Praga. Do programu festivalu přispěje Státní opera Praha, Divadlo na Vinohradech a balet Národního divadla z Brna. V rámci festivalu se bude konat reprezentativní izraelská archeologická výstava nazvaná City of David (Město Davidovo), výstavy výtvarného umění i expozice studentských prací pražské Akademie výtvarných umění. Další akce připraví pražské židovské muzeum, Památník národního písemnictví, Národní galerie a Národní muzeum.
Heslo zobrazeno.
Značky
Obecné
Ukázat
Odstranit
Pojmenované entity
Jméno
Instituce
Místo
Objekt
Adresa
Čas
Biblio
Foreign
X
Automatická anotace
SemLex
Označkovat
Nové heslo
Ulož heslo
A=a
Hledat
P
N
ID: 0000003772 Source: Eurovoc POS: N
Základní tvar: výtvarná umění
Lematizovaný tvar: výtvarný umění
Příklad:
Synonyma:
Glosa:
Změněno:


LexemAnn - mf930709_001
Soubor Úpravy Debug
Prezident má za týden jmenovat ústavní soudce
Parlament doporučil třináct kandidátů
Praha (li, ben) -
Prezident Havel by měl 15. července na Pražském hradě jmenovat třináct soudců Ústavního soudu. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třináct ze čtrnácti kandidátů na soudce, které Havel navrhl. Ústavní soud, který bude soudním orgánem ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce ústavního soudce je neslučitelná s členstvím v politických stranách. Základní plat soudců bude 25 tisíc korun.
Soudci by se mělo stát pět bývalých členů Ústavního soudu ČSFR Zdeněk Kessler (dříve poslanec FS za ODS), Vlastimil Ševčík (předtím poslanec FS za OH), Antonín Procházka (dříve poslanec ČNR za KDS), Vojen Güttler a Pavel Mates a čtyři vysokoškolští učitelé práva - Vladimír Klokočka, Vojtěch Cepl, Vladimír Čermák a Slovák Pavol Holländer. Tři další kandidáti, Iva Brožová, Miloš Holeček a Vladimír Jurka, jsou soudci krajského a okresního soudu. Třináctým kandidátem je komerční právník Vladimír Paul. Parlament neschválil kandidaturu docentky Ireny Pelikánové, členky ODA, pro niž hlasovala jen část poslanců vládních stran.
*
Značky
Obecné
Ukázat Odstranit
Pojmenované entity
Jméno Instituce Místo Objekt Adresa Čas Biblio Foreign X
Automatická anotace
SemLex
Označkovat Nové heslo Ulož heslo
A=a Hledat P N
ID: 0000028437 Source: CWN2a POS: N Základní tvar: soudní orgán
Lematizovaný tvar: soudní orgán
Příklad:
Synonyma:
Glosa:
Změněno: 090607125630 merger

Prezident má za týden jmenovat ústavní soudce

Parlament doporučil třináct kandidátů

Praha (li, ben) -

Prezident Havel by měl 15. července na Pražském hradě jmenovat třináct soudců Ústavního soudu. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. Ústavní soud, který bude soudním orgánem ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce ústavního soudce je neslučitelná s členstvím v politických stranách. Základní plat soudců bude 25 tisíc korun.

Soudci by se mělo stát pět bývalých členů Ústavního soudu ČSFR Zdeněk Kessler (dříve poslanec FS za ODS), Vlastimil Ševčík (předtím poslanec FS za OH), Antonín Procházka (dříve poslanec ČNR za KDS), Vojen Güttler a Pavel Mates a čtyři vysokoškolští učitelé práva - Vladimír Klokočka, Vojtěch Cepl, Vladimír Čermák a Slovák Pavol Holländer. Tři další kandidáti, Iva Brožová, Miloš Holeček a Vladimír Jurka, jsou soudci krajského a okresního soudu. Třináctým kandidátem je komerční právník Vladimír Paul. Parlament neschválil kandidaturu docentky Ireny Pelikánové, členky ODA, pro niž hlasovala jen část poslanců vládních stran.

(Prezident|t-mf930709-001-p1s1w1) (má|t-mf930709-001-p1s1w5) (za|t-mf930709-001-p1s1w4) (týden|t-mf930709-001-p1s1w4) (jmenovat|t-mf930709-001-p1s1w5) (ústavní|t-mf930709-001-p1s1w6) (soudce|t-mf930709-001-p1s1w7)

(Parlament|t-mf930709-001-p2s1w1) (doporučil|t-mf930709-001-p2s1w2) (třináct|t-mf930709-001-p2s1w3) (kandidátů|t-mf930709-001-p2s1w4)

(Praha|t-mf930709-001-p3s1Aw1) ((li|t-mf930709-001-p3s1Aw3)(,|t-mf930709-001-p3s1Aw4) (ben|t-mf930709-001-p3s1Aw5)) -

(Prezident|t-mf930709-001-p3s1Bw1) (Havel|t-mf930709-001-p3s1Bw2) (by|t-mf930709-001-p3s1Bw11) (měl|t-mf930709-001-p3s1Bw11) (15|t-mf930709-001-p3s1Bw5). (července|t-mf930709-001-p3s1Bw7) (na|t-mf930709-001-p3s1Bw10) (Pražském|t-mf930709-001-p3s1Bw9) (hradě|t-mf930709-001-p3s1Bw10) (jmenovat|t-mf930709-001-p3s1Bw11) (třináct|t-mf930709-001-p3s1Bw12) (soudců|t-mf930709-001-p3s1Bw13) (Ústavního|t-mf930709-001-p3s1Bw14) (soudu|t-mf930709-001-p3s1Bw15). (Řekl|t-mf930709-001-p3s2w1)

Prezident má za týden jmenovat ústavní soudce

Parlament doporučil třináct kandidátů

Praha (li, ben) -

Prezident Havel by měl 15. července na Pražském hradě jmenovat třináct soudců Ústavního soudu. Řekl nám to včera prezidentův mluvčí krátce poté, co parlament vyslovil souhlas se třinácti ze čtrnácti kandidátů na soudce, které Havel navrhl. Ústavn soud, který bude soudním orgánem ochrany ústavnosti, má zahájit činnost v Brně zřejmě v září. Podle ústavy se skládá z 15 soudců, jmenovaných na deset let. Soud se může ujmout činnosti složením slibu dvanáctého soudce. Funkce ústavního soudce je neslučitelná členstvím v politických stranách. Základní plat soudců bude 25 tisíc korun.

Soudci by se mělo stát pět bývalých členů Ústavního soudu ČSFR Zdeněk Kessler (dříve poslanec FS za ODS), Vlastimil evčík (předtím poslanec FS za OH), Antonín Procházka (dříve poslanec ČNR za KDS), Vojen Güttler a Pavel Mates a čtyři vysokoškolští učitelé práva - Vladimír Klokočka, Vojtěch Cepl, Vladimír Čermák a Slovák Pavol Holländer. Tři další kandidáti, Iva Brožová, Miloš Holeček a Vladimír Jurka, jsou soudci krajského a okresního soudu. Třináctým kandidátem je komerční právník Vladim Paul. Parlament neschválil kandidaturu docentky Ireny Pelikánové, členky ODA, pro niž hlasovala jen část poslanců vládních stran.

(na|t-mf930709-001-p3s1Bw10) (Pražském|t-mf930709-001-p3s1Bw9) (hradě|t-mf930709-001-p3s1Bw10)
(jmenovat|t-mf930709-001-p3s1Bw11) (třináct|t-mf930709-001-p3s1Bw12) (soudců|t-mf930709-001-p3s1Bw13)
(Ústavního|t-mf930709-001-p3s1Bw14) (soudu|t-mf930709-001-p3s1Bw15). (Řekl|t-mf930709-001-p3s2w1)
(nám|t-mf930709-001-p3s2w2) (to|t-mf930709-001-p3s2w3) (včera|t-mf930709-001-p3s2w4)
(prezidentův|t-mf930709-001-p3s2w5) (mluvčí|t-mf930709-001-p3s2w6) (krátce|t-mf930709-001-p3s2w7)
(poté|t-mf930709-001-p3s2w12), (co|t-mf930709-001-p3s2w12) (parlament|t-mf930709-001-p3s2w11)
(vyslovil|t-mf930709-001-p3s2w12) (souhlas|t-mf930709-001-p3s2w13) (se|t-mf930709-001-p3s2w15)
(třinácti|t-mf930709-001-p3s2w15) (ze|t-mf930709-001-p3s2w18) (čtrnácti|t-mf930709-001-p3s2w17)
(kandidátů|t-mf930709-001-p3s2w18) (na|t-mf930709-001-p3s2w20) (soudce|t-mf930709-001-p3s2w20),
(které|t-mf930709-001-p3s2w22) (Havel|t-mf930709-001-p3s2w23) (navrhl|t-mf930709-001-p3s2w24).
(Ústavní|t-mf930709-001-p3s3w1) (soud|t-mf930709-001-p3s3w2), (který|t-mf930709-001-p3s3w4)
(bude|t-mf930709-001-p3s3w5) (soudním|t-mf930709-001-p3s3w6) (orgánem|t-mf930709-001-p3s3w7)
(ochrany|t-mf930709-001-p3s3w8) (ústavnosti|t-mf930709-001-p3s3w9), (má|t-mf930709-001-p3s3w12)
(zahájit|t-mf930709-001-p3s3w12) (činnost|t-mf930709-001-p3s3w13) (v|t-mf930709-001-p3s3w15)
(Brně|t-mf930709-001-p3s3w15) (zřejmě|t-mf930709-001-p3s3w16) (v|t-mf930709-001-p3s3w18)
(září|t-mf930709-001-p3s3w18). (Podle|t-mf930709-001-p3s4w2) (ústavy|t-mf930709-001-p3s4w2) (se|t-mf930709-001-p3s4w
(skládá|t-mf930709-001-p3s4w4) (z|t-mf930709-001-p3s4w7) (15|t-mf930709-001-p3s4w6) (soudců|t-mf930709-001-p3s4w7),
(jmenovaných|t-mf930709-001-p3s4w9) (na|t-mf930709-001-p3s4w12) (deset|t-mf930709-001-p3s4w11)

MWEs in PDT 2.5

- Full annotation of t-layer
- Continuous text, but annotating t-trees
- Storing the lexicon
  - Pre-annotating known MWEs using trees
- NEs also annotated – not discussed hereafter
- 16,3% of content words

Semlex

- All MWEs in PDT (t-layer)
- basic (quotation) forms
- lemmatized forms
- dependency structures

```
!!perl/hash:SemLex_heslo
BASIC_FORM: deficit státního rozpočtu
CREATED: '110914162730'
EXAMPLE: ~
GLOSS: ~
ID: '0000032344'
LEMMATIZED: deficit státní rozpočet
MODIFIED: ~
MODIFIER: bejcek
MORPHO_TAGS: ''
ORIGID: ~
PDT25_FREQ: 2
POS: 'N'
SOURCE: stastna
SYNONYMS: []

TREE_STRUCT:
  -
    - deficit
    - ~
  -
    - rozpočet
    - 0
  -
    - státní
    - 1
```

Datasets

- Prague Dependency Treebank 2.5
  - full manual annotation
    - morphology (m), surface syntax (a), deep syntax (t)
    - MWE
  - automatic analysis: (m), (a), (t)
- Czech National Corpus: SYN2006-PUB – automatic

ÚFAL

Automatic analysis of data

- Treex (http://ufal.mff.cuni.cz/treex)
  - scenario “Analysis of Czech”

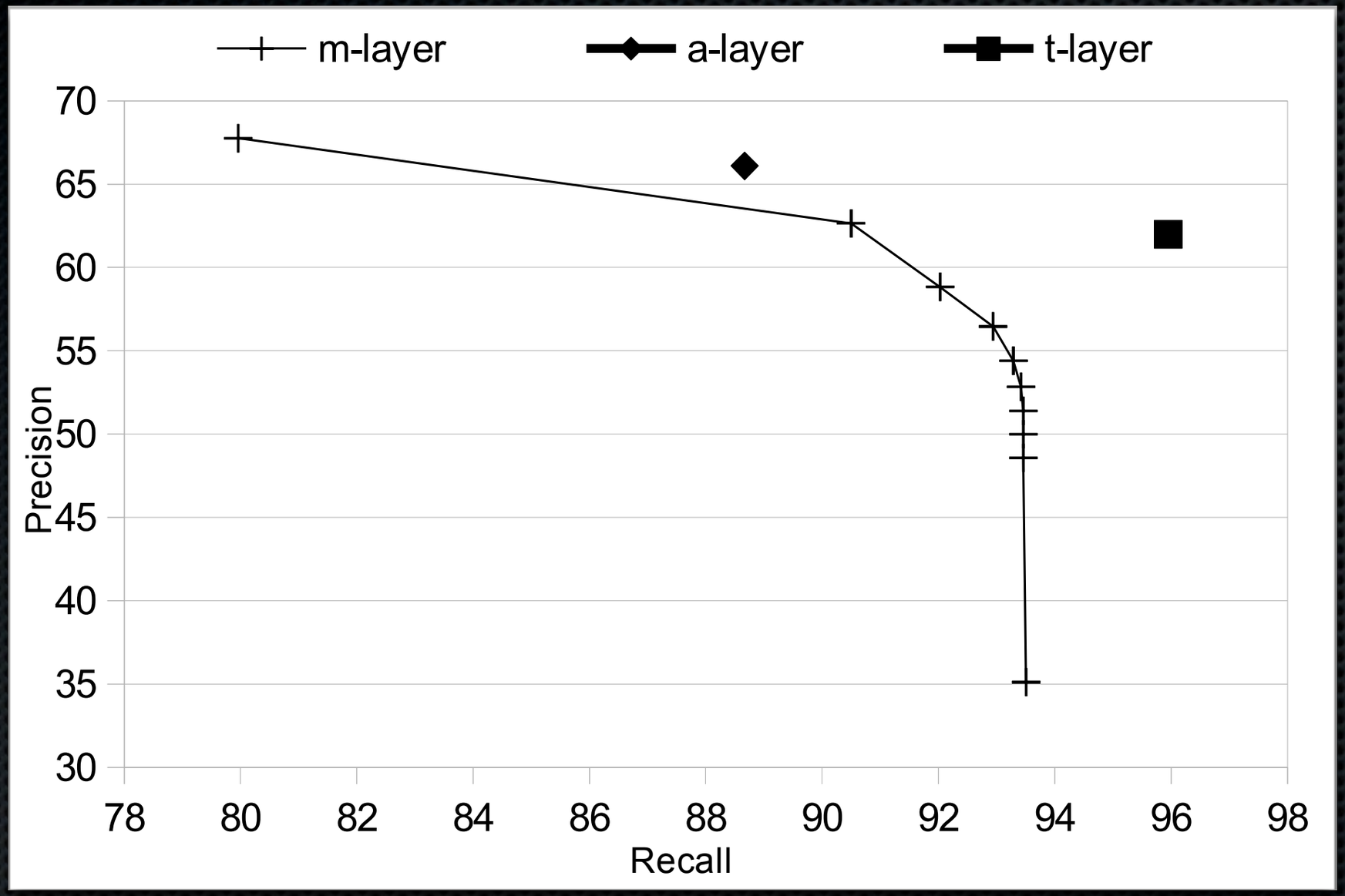
1. rule-based segmentation and tokenisation
2. morphology and tagger (Hajič + Featurama)→ m-layer
3. parser (MST with improved features) → a-layer
4. t-tree transformation (rule-based) → t-layer

Experiments

- Purely syntactic. No semantics.
- t-layer: matching t-trees
- a-layer:
  - a-tree structures, including auxiliaries, m-lemmas
  - from annotated data; t-tree:a-trees – 1:N
- m-layer: surface collocations of lemmas in a window

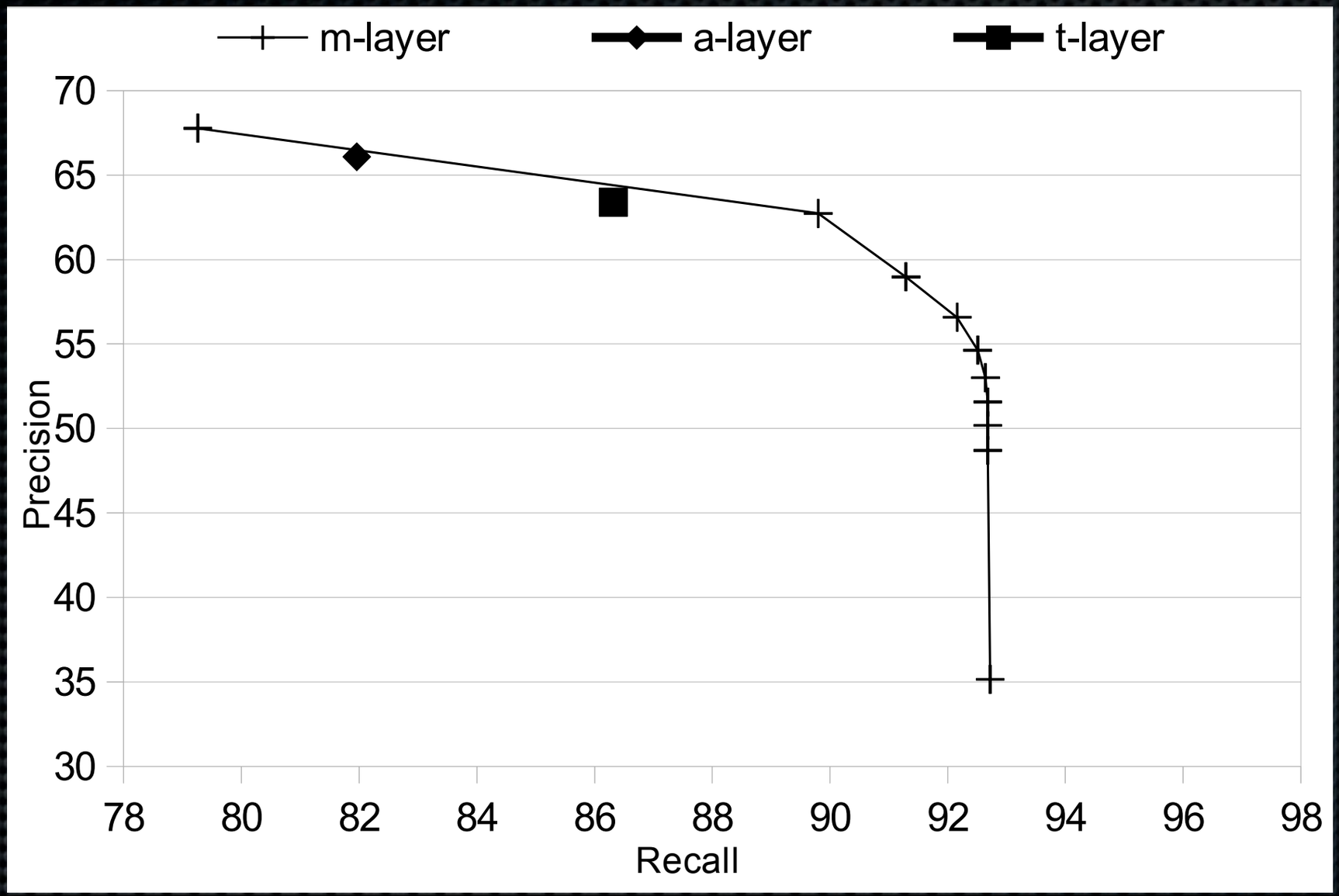
layer/span	PDT/man	PDT/auto	CNC/auto
tecto	61.99 / 95.95 / 75.32	63.40 / 86.32 / 73.11	44.44 / 58.00 / 50.33
analytical	66.11 / 88.67 / 75.75	66.09 / 81.96 / 73.18	45.22 / 60.00 / 51.58
morpho / 2	67.76 / 79.96 / 73.36	67.77 / 79.26 / 73.07	51.85 / 56.00 / 53.85
3	62.65 / 90.50 / 74.05	62.73 / 89.80 / 73.86	46.99 / 60.00 / 52.70
4	58.84 / 92.03 / 71.78	58.97 / 91.29 / 71.65	42.83 / 61.33 / 50.48
5	56.46 / 92.94 / 70.25	56.59 / 92.16 / 70.12	40.09 / 61.33 / 48.49
6	54.40 / 93.29 / 68.81	54.64 / 92.51 / 68.70	38.27 / 61.33 / 47.13
7	52.85 / 93.42 / 67.51	53.01 / 92.64 / 67.43	36.99 / 61.33 / 46.15
8	51.39 / 93.46 / 66.32	51.57 / 92.68 / 66.27	35.59 / 61.33 / 45.04
9	50.00 / 93.46 / 65.15	50.18 / 92.68 / 65.11	34.67 / 61.33 / 44.30
10	48.57 / 93.46 / 63.92	48.71 / 92.68 / 63.86	33.84 / 61.33 / 43.64
∞	35.12 / 93.51 / 51.06	35.16 / 92.72 / 50.99	22.70 / 62.00 / 33.24
	P / R / F	P / R / F	P / R / F

Results



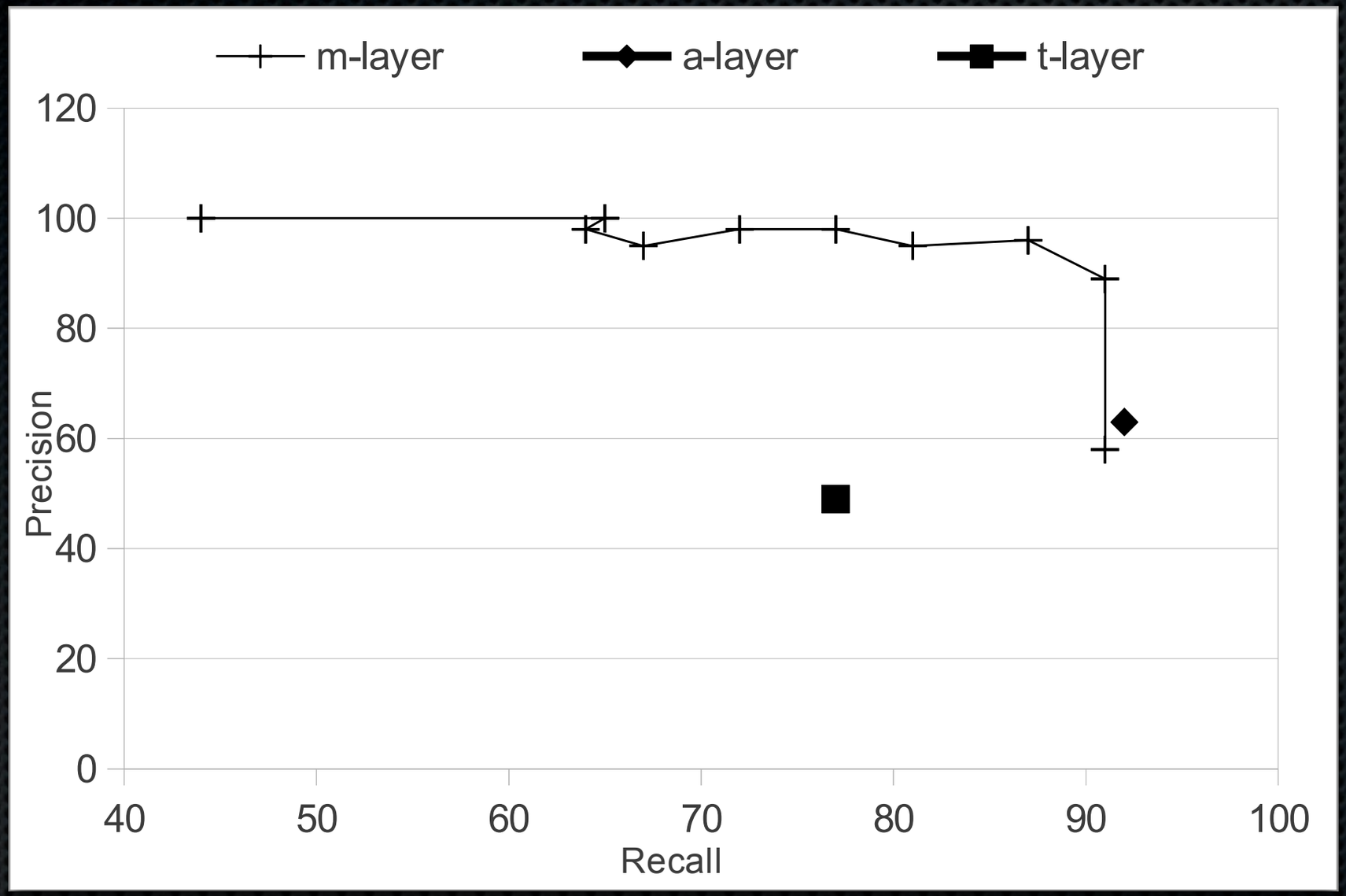
Results²

manual analysis – gold data



Results[3]

automatic analysis of gold data (PDT)



Results[4]

automatic analysis of CNC

Problem: t-lemma variability

- All of these variations share the basic lexical meaning
- They should be specified by attributes explicating the relationship to the basic lemma (meaning) – **someday**
  - Diminutives: dům, domek, domeček (1st, 2nd degree)
  - Gender opposites: ředitelka (female director)
  - Lemma variants: občanský zákoník (citizen law codex)
- 771 of 8816 Semlex entries with >1 tree-structure

ÚFAL

More Problems

Quality of t-layer parsing

- empty nodes
  - especially in coordinations (red [wine] and white wine)
- Diminutives: dům, domek, domeček (1st, 2nd degree)

Semlex

- Too simple tree structures
  - Aux representation needed for MWE semantics
- Insufficient coverage (cf. PDT-auto vs. CNC)

ÚFAL

Conclusions

- On new data, surface (morphologic) method wins
  - With current syntactic methods
- Syntactic method with t-analysis has potential
  - Improve Semlex (auxiliaries, lemma variants)
  - Improve t-analysis (overall quality, generated nodes)

Data and Tools Used

- PDT 2.5 – http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8
- ČNK (shuffled) – coming soon!
- Semlex – http://ufal.mff.cuni.cz/lexemann/mwe/
- Morphology – http://hdl.handle.net/11858/00-097C-0000-0015-A780-9
- Tagger – http://hdl.handle.net/11858/00-097C-0000-0001-4904-2
- Treex – http://ufal.mff.cuni.cz/treex

ÚFAL