Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank

Pavel Pecina, Pavel Straňák
Jan Hajič, Martin Holub, Marie Hučínová
Martin Pavlík, Pavel Šidák

Center for Computational Linguistics
Charles University, Prague, Czech Republic

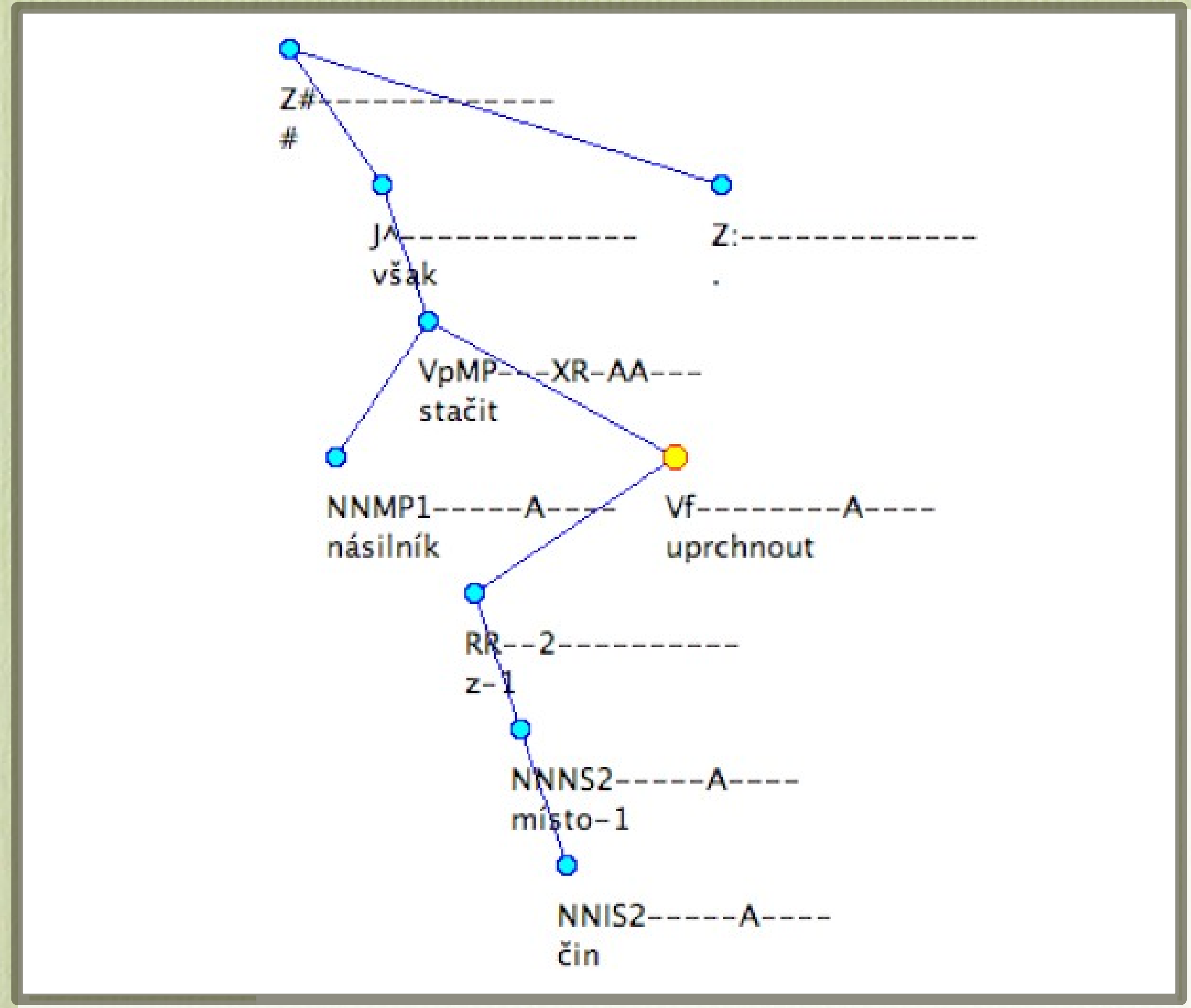
Outline

1. Prague Dependency Treebank
2. Motivation and project goals
3. Czech WordNet
4. Annotation process
5. Results and Statistics
6. CWN improvement
7. Conclusion and future work

Prague Dependency Treebank

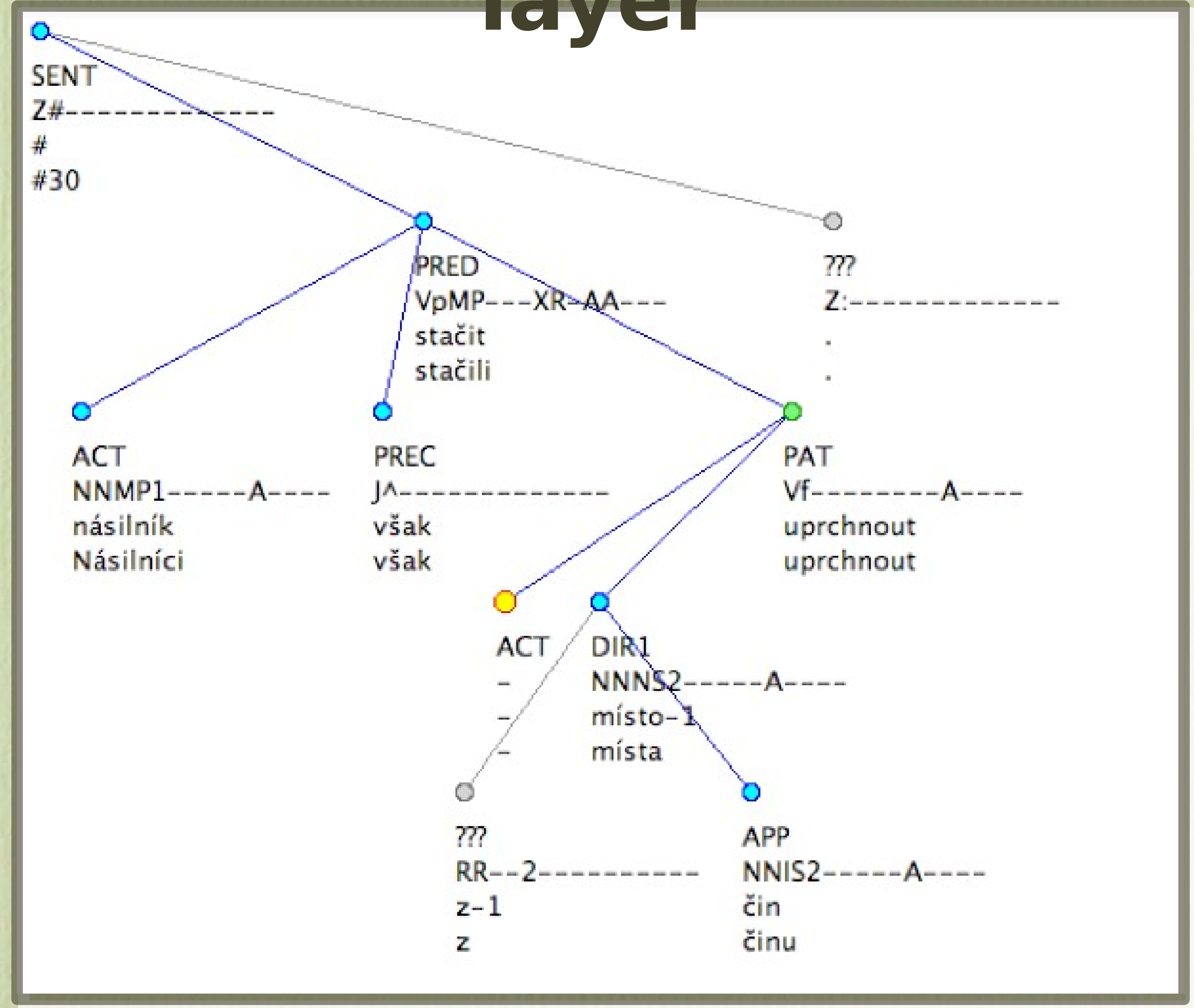
- **Subcollection of *the Czech National Corpus:***
  - 1.8 mil. tokens; 100,000 sentences; 1,500 docs
- **Three-layer annotation scheme:**
  - morphemic **`<lemma, tag>`**
  - analytical (*surface syntax*) **`<head pointer,analytical function>`**
  - tectogrammatical (*deep syntax*) **`<head pointer, functor>`**

PDT example: analytical layer



“Criminals, however, managed to escape from the scene of the crime.”

PDT example: tectogrammatical layer



"Criminals, however, managed to escape from the scene of the crime."

Motivation: lexico-semantic disambiguation

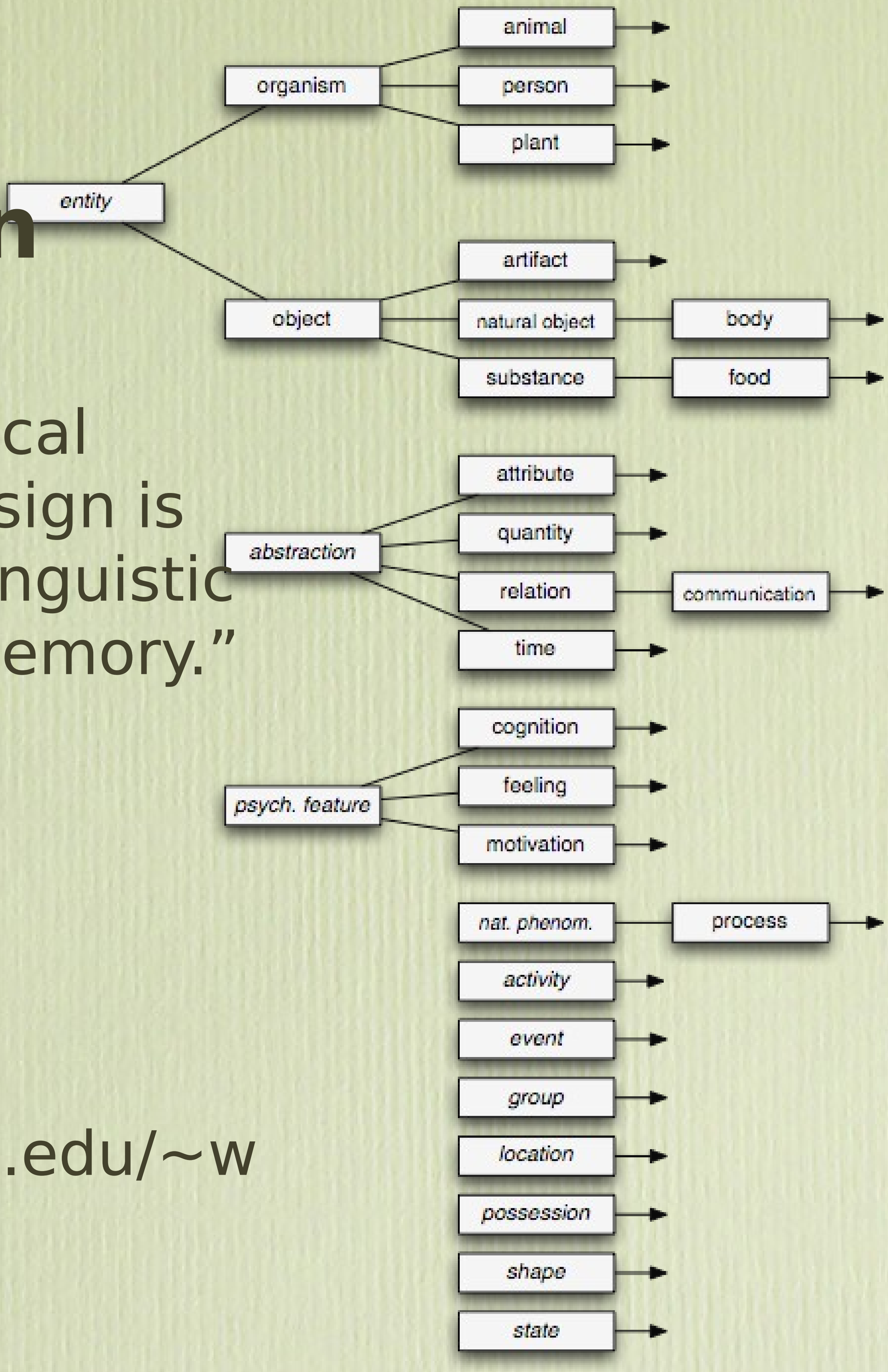
- **Task:**
    - “Automatic identification of word senses in a raw text.”
- **Requirements:**
    - *A semantic lexicon* — set of all possible meanings (labels/tags) for each word.
    - *A method/procedure* that assigns a semantic tag to each occurrence of a word.
        - Supervised methods -> need for **training data**

Project Goals

- **Primary:**
  - To obtain a training data for automatic lexico-semantic tagging
- **Secondary:**
  - To find the flaws of the system of semantic tags and get information for its improvement

WordNet: our semantic lexicon

- “WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory.”
- Electronic Lexical Database
- George A. Miller, Christiane Fellbaum, Randee Tengi
- http://www.cogsci.princeton.edu/~wn/



Structure of WordNet

- Only autosemantic words – nouns, adjectives, verbs and adverbs
- The basic semantic relation in WordNet is synonymy.
- Sets of synonyms are called *synsets*.
- Other relations: *meronymy* (“is a part of”), *antonymy*, *hyponymy* (“is a kind of”), *hypernymy* ...

EuroWordNet

- New wordnets:
  - EWN1: Dutch, English, Italian, Spanish
  - EWN2: **Czech**, Estonian, French, German
- Interlingual Index (ILI)
- Interlingual Relations (ILR)
- Top ontology (63 top concepts), 1053 basic concepts

Czech WordNet

- Developed at The Masaryk University, Brno
- Originally in EuroWordNet 2, continuing development within the Balkanet project
- Mapped *directly to the Princeton WordNet 2.0*
- XML format
- 17,000 nouns; 2,000 verbs; 4,000 adjectives and adverbs

Czech Wordnet: “a driver” example

```
• <SYNSET>
    <ID>ENG171-08137652-n</ID>
    <POS>n</POS>
    <SYNONYM>
        <LITERAL>
            šofér
            <SENSE>1</SENSE>
        </LITERAL>
        <LITERAL>
            řidič
            <SENSE>1</SENSE>
        </LITERAL>
    </SYNONYM>
    <ILR>
        <TYPE>hypernym</TYPE>
        ENG171-08506030-n
    </ILR>
</SYNSET>
```

Annotation Process

- *Data preprocessing*
  - For each word to be annotated (its lemma exists in the CWN) get a list of all its synsets: **uniliteral synsets, multiliteral synsets, exceptions**
- *Annotation itself*
  - Performed independently by two people with linguistic education (1 doc ~ 50 sentences ~ 100-300 words ~ 1 hr)
  - *Instructions*: always assign one tag, prefer uniliteral synsets, only the very last option is the „missing synset“ exception.

Exception List

1. Incorrect Reflexivity *l is reflexive but CWN knows only its non-reflexive form or vice versa.*

2. Missing Positive Sense *l is positive, but CWN includes only its negative form.*

3. Missing Negative Sense *l is negative, but CWN includes only its positive form.*

4. Incorrect Lemma *The lemma l assigned to the word is incorrect (therefore the synsets proposed are incorrect too).*

5. Figurative Use *The word is used in a metaphorical or other figurative way.*

6. Proper Name *Assigned to proper names not included in the CWN.*

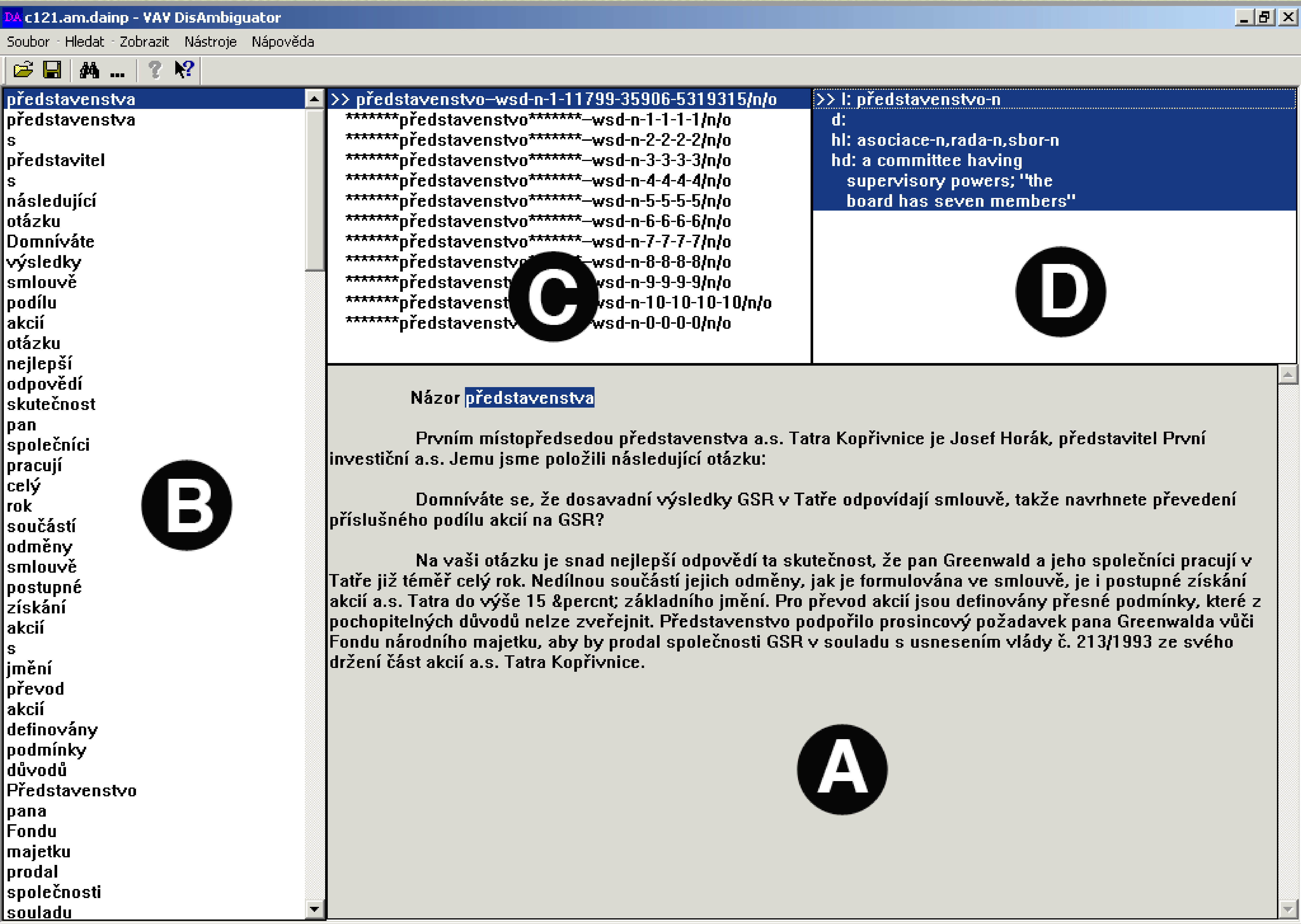
7. Unclear Word Meaning in Text *The meaning of l is unclear (therefore no synset can be assigned).*

8. Unclear CWN Sense *The meaning of a synset is unclear and no other proposed synset can be used.*

9. Missing More General Sense *At least one of the proposed synsets corresponds to the meaning of l, but is too specific and so expressing only part of l.*

10. Missing Sense *None of the synsets proposed expresses the meaning of l and more specific exceptions can not be used.*

0. Other Problem *Assigned if no other category can be used.*



```
předstawenstva
představenstva
s
představitel
s
následující
otázku
Domníváte
výsledky
smlouvě
podílu
akcií
otázku
nejlepší
odpovědí
skutečnost
pan
společníci
pracují
celý
rok
součástí
odměny
smlouvě
postupné
získání
akcií
s
jmění
převod
akcií
definovány
podmínky
důvodů
Představenstvo
pana
Fondu
majetku
prodal
společnosti
souladu
```

```
>> představenstvo–wsd-n-1-11799-35906-5319315/n/o
   *******představenstvo*******–wsd-n-1-1-1-1/n/o
   *******představenstvo*******–wsd-n-2-2-2-2/n/o
   *******představenstvo*******–wsd-n-3-3-3-3/n/o
   *******představenstvo*******–wsd-n-4-4-4-4/n/o
   *******představenstvo*******–wsd-n-5-5-5-5/n/o
   *******představenstvo*******–wsd-n-6-6-6-6/n/o
   *******představenstvo*******–wsd-n-7-7-7-7/n/o
   *******představenstv        –wsd-n-8-8-8-8/n/o
   *******představenst          sd-n-9-9-9-9/n/o
   *******představenst          sd-n-10-10-10-10/n/o
   *******představenstv          wsd-n-0-0-0-0/n/o
```

```
>> I: představenstvo-n
   d:
   hl: asociace-n,rada-n,sbor-n
   hd: a committee having
     supervisory powers; "the
     board has seven members"
```

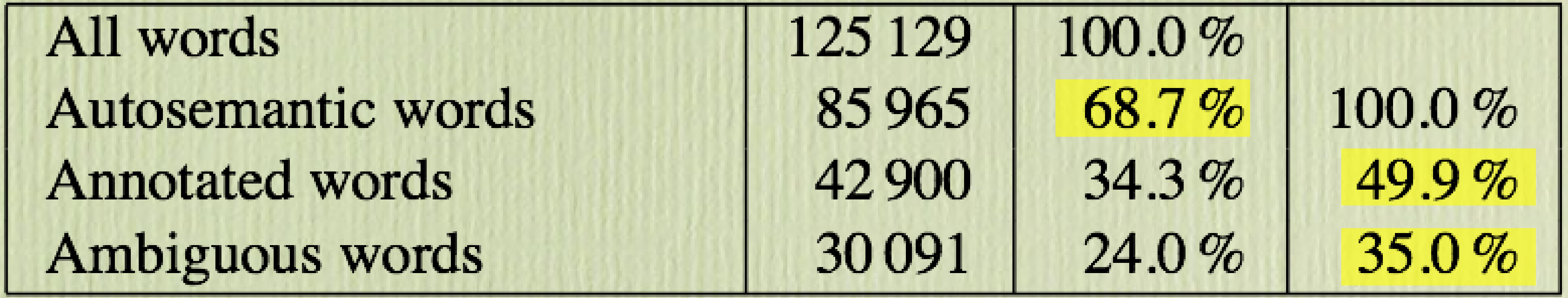
Názor představenstva

Prvním místopředsedou představenstva a.s. Tatra Kopřivnice je Josef Horák, představitel První investiční a.s. Jemu jsme položili následující otázku:

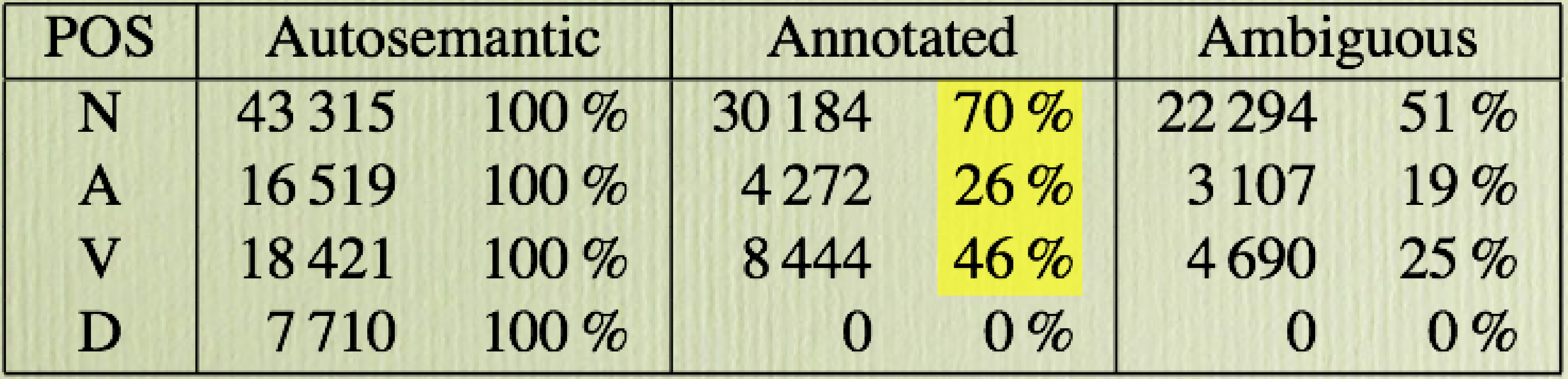
Domníváte se, že dosavadní výsledky GSR v Tatře odpovídají smlouvě, takže navrhnete převedení příslušného podílu akcií na GSR?

Na vaši otázku je snad nejlepší odpovědí ta skutečnost, že pan Greenwald a jeho společníci pracují v Tatře již téměř celý rok. Nedílnou součástí jejich odměny, jak je formulována ve smlouvě, je i postupné získání akcií a.s. Tatra do výše 15 &percnt; základního jmění. Pro převod akcií jsou definovány přesné podmínky, které z pochopitelných důvodů nelze zveřejnit. Představenstvo podpořilo prosincový požadavek pana Greenwalda vůči Fondu národního majetku, aby by prodal společnosti GSR v souladu s usnesením vlády č. 213/1993 ze svého držení část akcií a.s. Tatra Kopřivnice.

Statistics: annotated text

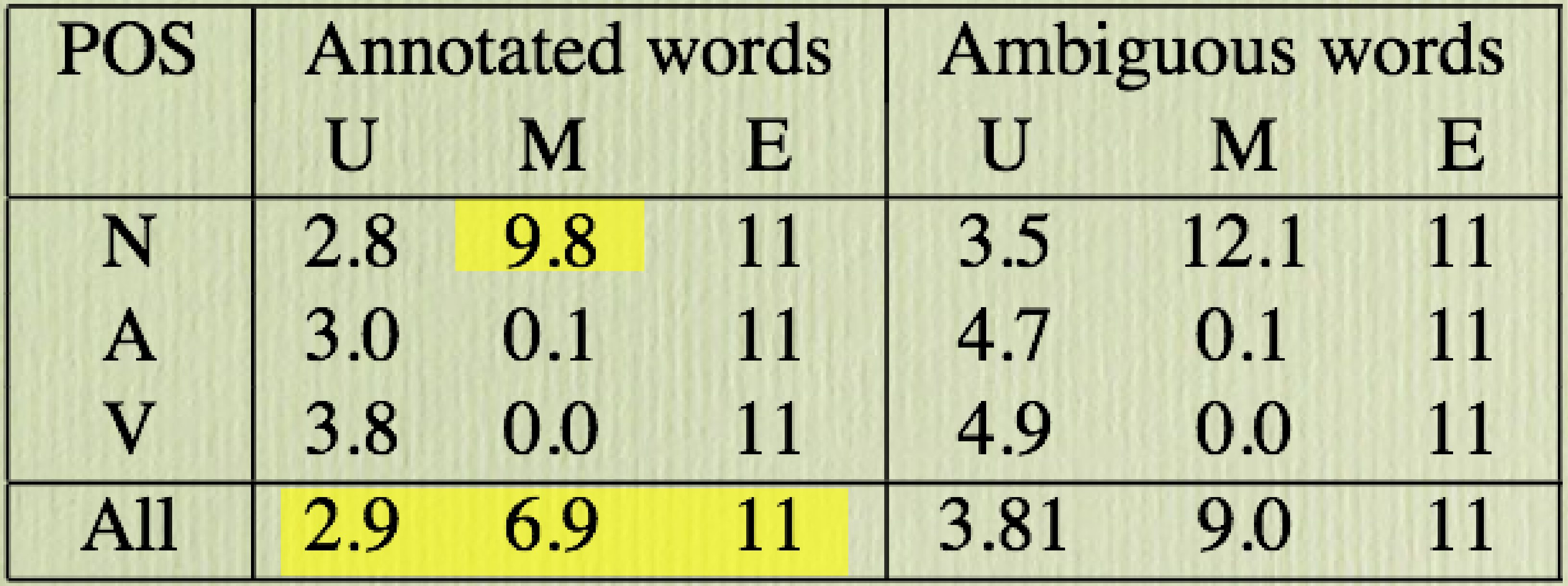


All words	125 129	100.0 %	
Autosemantic words	85 965	68.7 %	100.0 %
Annotated words	42 900	34.3 %	49.9 %
Ambiguous words	30 091	24.0 %	35.0 %



POS	Autosemantic		Annotated		Ambiguous	
N	43 315	100 %	30 184	70 %	22 294	51 %
A	16 519	100 %	4 272	26 %	3 107	19 %
V	18 421	100 %	8 444	46 %	4 690	25 %
D	7 710	100 %	0	0 %	0	0 %

Statistics: Was the annotation difficult?



POS	Annotated words			Ambiguous words		
	U	M	E	U	M	E
N	2.8	9.8	11	3.5	12.1	11
A	3.0	0.1	11	4.7	0.1	11
V	3.8	0.0	11	4.9	0.0	11
All	2.9	6.9	11	3.81	9.0	11

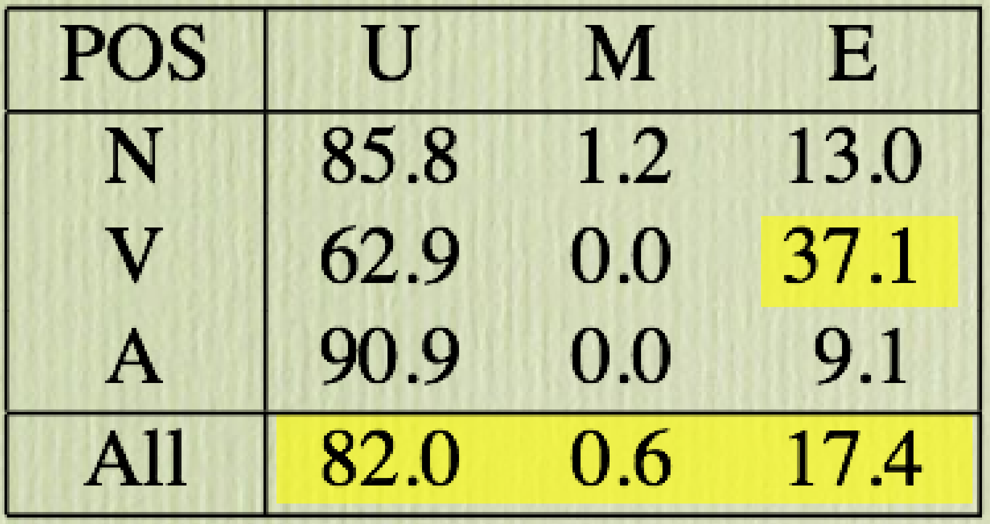
An average list of possible tags for a word consists of 3 uniliteral synset, 7 multiliteral synsets and 11 exceptions.

U – uniliteral synsets

M – multiliteral synsets

E – exceptions

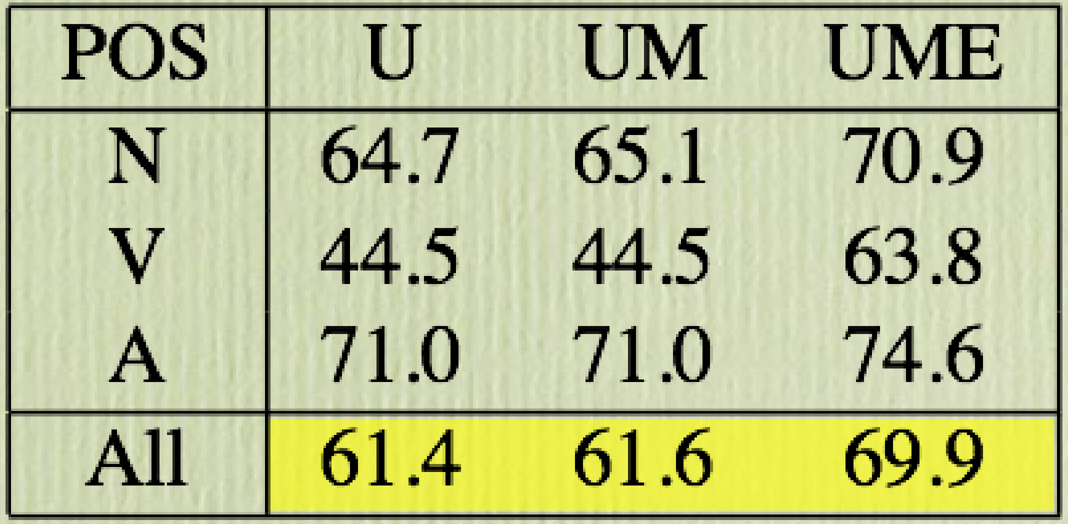
Statistics: average tag types usage



POS	U	M	E
N	85.8	1.2	13.0
V	62.9	0.0	37.1
A	90.9	0.0	9.1
All	82.0	0.6	17.4

- ***Exceptions were used in 17.4 % of cases***
- ***37.1% were assigned an exception***

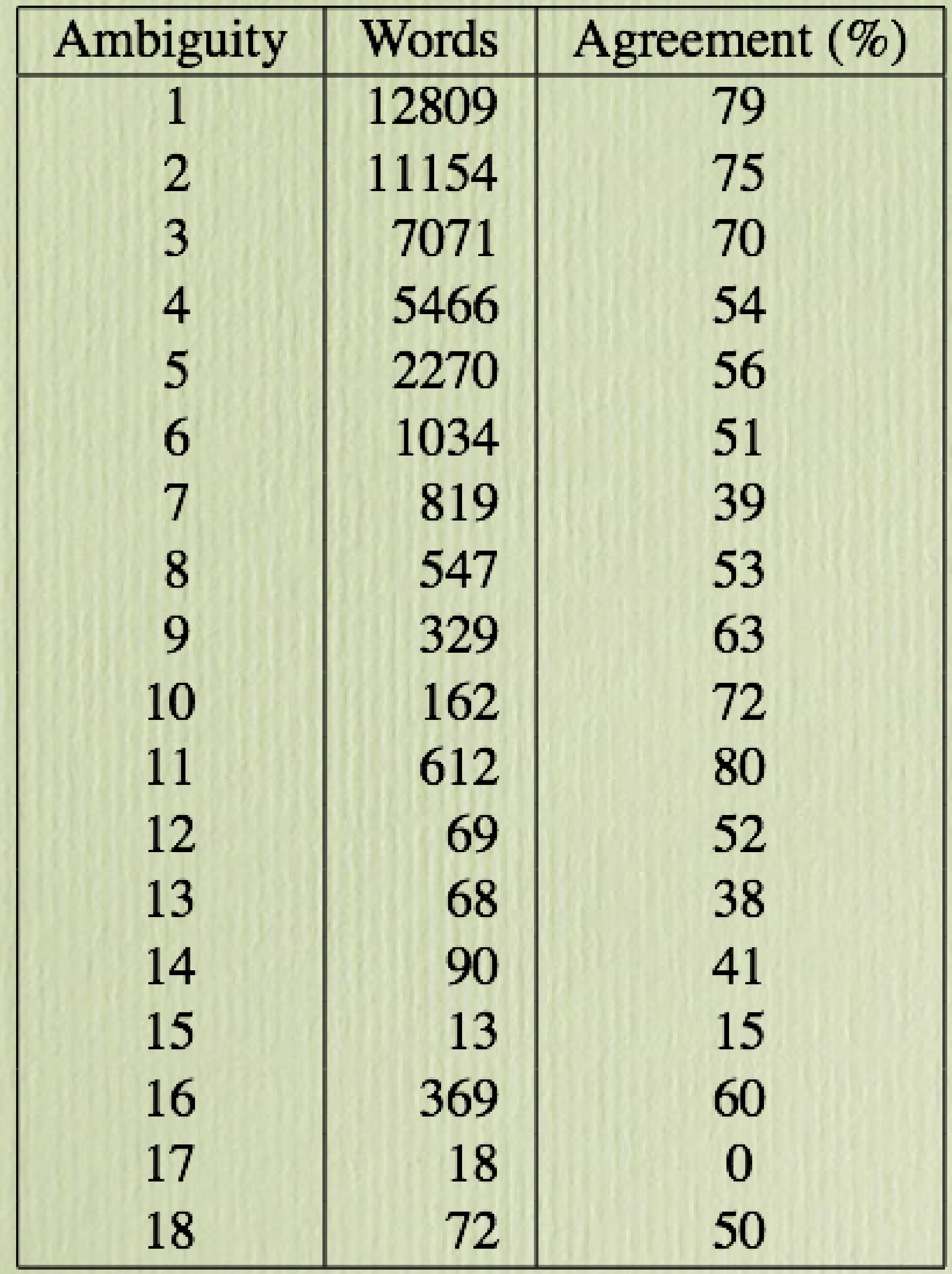
Statistics: interannotator agreement



POS	U	UM	UME
N	64.7	65.1	70.9
V	44.5	44.5	63.8
A	71.0	71.0	74.6
All	61.4	61.6	69.9

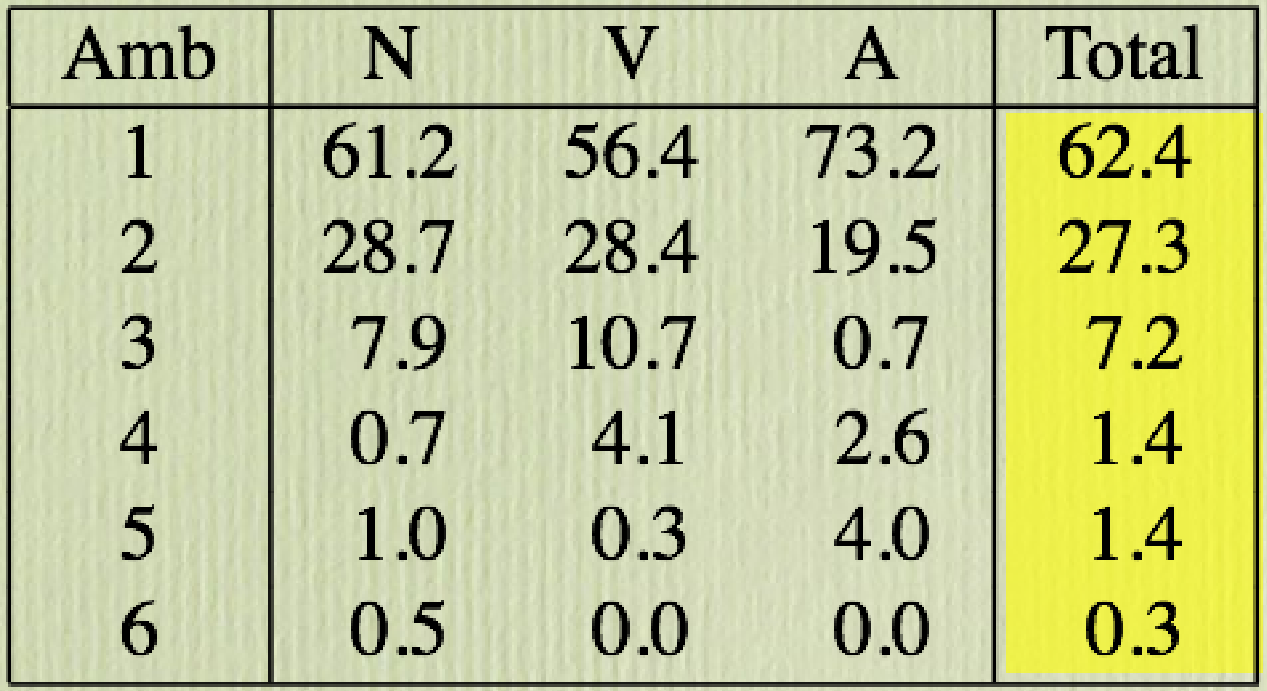
- ***Interannotator agreement on synset selection is 61.6 %***
- ***Over all interannotator agreement is 69.9 %***

Statistics: ambiguity vs. agreement



Ambiguity	Words	Agreement (%)
1	12809	79
2	11154	75
3	7071	70
4	5466	54
5	2270	56
6	1034	51
7	819	39
8	547	53
9	329	63
10	162	72
11	612	80
12	69	52
13	68	38
14	90	41
15	13	15
16	369	60
17	18	0
18	72	50

Statistics: ambiguity of annotated words

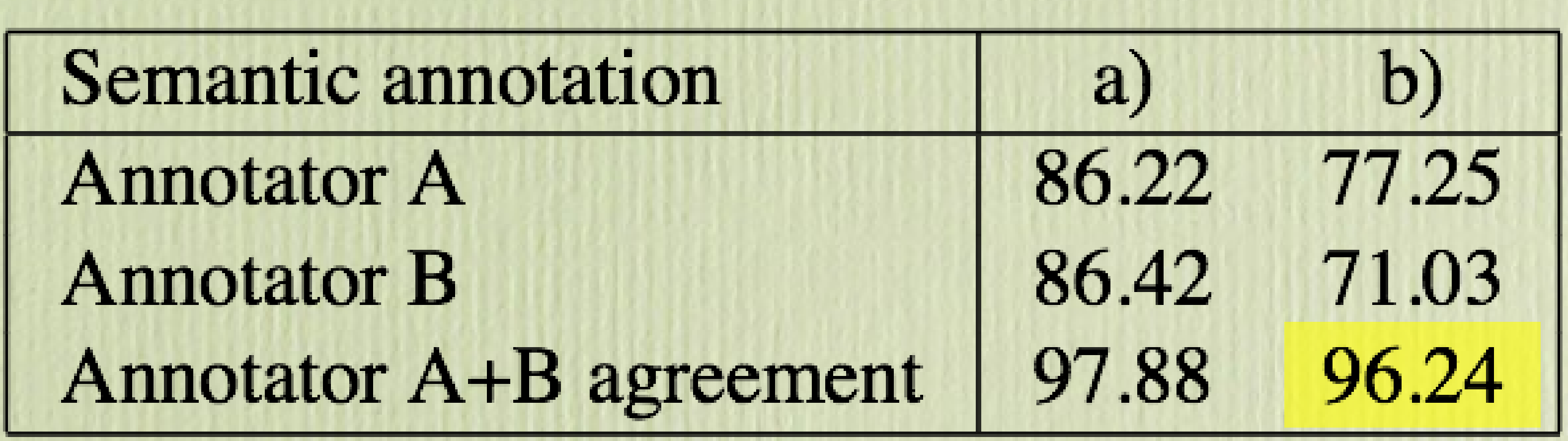


Amb	N	V	A	Total
1	61.2	56.4	73.2	62.4
2	28.7	28.4	19.5	27.3
3	7.9	10.7	0.7	7.2
4	0.7	4.1	2.6	1.4
5	1.0	0.3	4.0	1.4
6	0.5	0.0	0.0	0.3

Almost 2/3 of annotated words (types) were not ambiguous.

Statistics: “One sense per collocation”

Yarowsky (1995): *“All occurrences of a word in the same collocation have the same meaning.”*



Semantic annotation	a)	b)
Annotator A	86.22	77.25
Annotator B	86.42	71.03
Annotator A+B agreement	97.88	96.24

Manually extracted list of frequent collocations in the PDT

a) all

b) occurring at least twice in the annotated data

Czech WordNet: the facts and flaws

- *Less then 50%* of N,A,V in the annotated text appear in the CWN
- *Only 30%* of all N,A,V were successfully annotated with a CWN synset
- Some very common meanings of frequent words are not covered by the CWN
- *Only 12%* of all CWN synsets were assigned to a word.

- ***Uneven distribution of the CWN synsets***
- ***Insufficient word coverage***

Czech WordNet: the feedback

- Distribution of synset elements for individual synsets
  “this synonym is missing”
- Distribution of missing synsets / exceptions and their types
  “this synset is missing”
- Distribution of synsets for individual words
  “this word has this sense in this many cases”

Czech WordNet: the improvement

Ver.	**1.7**	**1.8**
N	17,000	21,000
A	2,000	2,000
V	4,000	5,000
D	0	200

Conclusions & Future Work

- **Achieved goals / work in progress**
  - Enrichment of the PDT by lexico-semantic tags
  - Validation of the CWN and stimulus for its improvement
- **Future work**
  - To employ a new version of the CWN
  - To improve the annotation methodology (tag lists, instructions) - *in order to increase the interannotator agreement.*
  - To perform the second annotation cycle.
  - Exploiting data for automatic WSD in Czech

Thank you.