Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation

Markéta Lopatková, Ondřej Bojar,
Jiří Semecký, Václava Benešová, Zdeněk Žabokrtský
ÚFAL, Charles University, Prague

September 13, 2005

Overview

- VALLEX structure
- VALEVAL: lexical sampling experiment
- Inter-annotator agreement
- WFD: Word-frame disambiguation experiments
- Conclusion

VALLEX = Valency Lexicon of Czech Verbs

- Closely related to Prague Dependency Treebank (PDT)
  However VALLEX ≠ PDT-VALLEX (**?**)

	Verbs	Valency Frames	Released	
VALLEX 1.0	1400	4000	autumn 2003	
VALLEX 1.5	2500	6000	spring 2005	← (internal release)

VALLEX 1.5 coverage:

- high thanks to the primary focus on frequent lemmas
- around 90% verb tokens from Czech National Corpus (CNC), also counting auxiliary *být* (In PDT, we observe that 42% tokens of *být* are auxiliary.)

`http://ckl.ms.mff.cuni.cz/~zabokrtsky/vallex/1.0/`

VALLEX Structure

Key components: Frames, functors, obligatoriness, morphemic form(s)

Word entry

Frame entry

odpovídat (imperfective)

1 odpovídat$_1$ $\sim$ odvětit [answer; respond]

- frame: $\mathrm{ACT}^{obl}{}_1\ \mathrm{ADDR}^{obl}{}_3\ \mathrm{PAT}^{opt}{}_{na+4,4}\ \mathrm{EFF}^{obl}{}_{4,aby,a\check{t},zda,\check{z}e}\ \mathrm{MANN}^{typ}$
- example: *odpovídal mu na jeho dotaz pravdu / že ...* [he responded to his question truthfully / that . . .]
- asp.counterpart: odpovědět$_1$ pf.
- class: communication

2 odpovídat$_2$ $\sim$ reagovat [react]

- frame: $\mathrm{ACT}^{obl}{}_1\ \mathrm{PAT}^{obl}{}_{na+4}\ \mathrm{MEANS}^{typ}{}_7$
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_2$ pf.

. . .

odpovídat se (imperfective)

1 odpovídat se$_1$ $\sim$ být zodpovědný [be responsible]

- frame: $\mathrm{ACT}^{obl}{}_1\mathrm{ADDR}^{obl}{}_3\mathrm{PAT}^{obl}{}_{z+2}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

An abbreviated example for the base lemma "odpovídat".

VALEVAL: Task Description

VALEVAL = lexical sampling experiment with VALLEX 1.0
. . . annotate sample verb occurrences with VALLEX frame entries

Goals:

- Check quality of VALLEX entries
- Estimate inter-annotator agreement
- Prepare data for experimenting with automatic word-frame disambiguation

Base lemmas selected randomly but conforming the following criteria:

- Cover both easy and difficult verbs
- Cover all aspectual counterparts of the verbs

For each selected base lemma, up to 100 random examples from CNC.

Overall Annotation Statistics

Lemmas annotated	109
Sentences annotated	10256
Parallel annotators	3
Total annotations	30765 (100%)
Uncertain annotations	1045 (3.4%)
Ambiguous annotations	703 (2.3%)
Marked as invalid example	172 (0.6%)
Annotator got confused	90 (0.3%)
Marked as missing frame	1673 (5.4%)
Sentences where all were sure	9280 (90.5%)
Sentences where all were sure that the frame is missing	125 (1.2%)

Inter-Annotator Agreement and κ

	Average Pairwise Match			
	IAA [%]		κ	
	wØ	Ø	wØ	Ø
Exact	70.8	74.8	0.54	0.54
Ignoring Uncertainty	74.8	77.7	0.60	0.59
Where All Were Sure	76.7	80.9	0.61	0.64

Ø: Average over tested lemmas, wØ: Weighted by frequency in CNC.

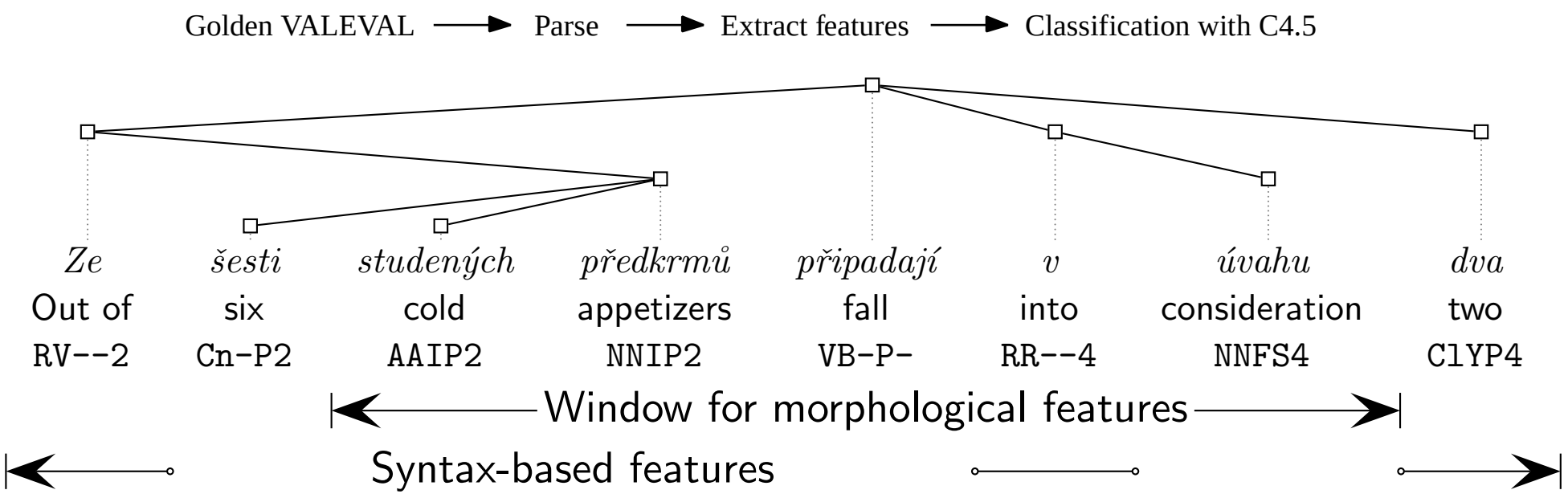
- κ values indicate $moderate\ agreement$, comparable to **?** (pairwise IAA for French verbs 60–65%, κ 0.41)
- Pairwise IAA higher than annotating verbs with CzechWordNet senses (**?**: 45–64%)
- Our 3-annotator IAA wØ: 61–68%

"Golden VALEVAL": Data for WFD

- Useful as a corpus for word-frame disambiguation (WFD $\sim$ WSD = word-sense disambiguation).
- Sentences with full agreement or post-editing: 8066 sentences for 108 lemmas.

	wØ	Ø
Entropy	1.54	1.28
VALLEX frames per lemma	12.46	7.61
Seen frames per lemma	5.85	4.85
10-fold Baseline WSD Accuracy	59.79%	66.19%

First Experiments with WFD



Morphological features: `AAIP2 NNIP2 VB-P- RR--4 NNFS4`

Syntax-based features: *ze+2*, *v+4*, *4*

Boolean features describe the presence or absence of types of the verb's modifications.

Word-Frame Disambiguation Results

	wØ	Ø	
choose commonest frame → Baseline	63.3%	67.9%	
Morphological	67.1%	73.9%	← morph. features from a 5-word window around the verb
Syntax-based	70.8%	78.5%	← morph. features of children of the verb

- Reasonable improvement over the baseline
- Difference between Ø and wØ caused by difficulties with *mít* [to have]
- Still room for improvement ⇒ further experiments with idiomatic expressions, WordNet classes and animacy
  (Some of the experiments described in **?**)

Note: Baseline different, because only 6666 sentences were successfully parsed.

Correcting VALLEX Errors

- **frame entries** (75 corrections in total, 39 "not serious" – missing idioms)
- **mistakes within frame entries** (32)
  - **mistakes in functors** (16)
  - **mistakes in morphemic realization** (12)
  - **mistakes in obligatoriness** (4)
- mistakes in the gloss or example (30)

Example: Inappropriate functor:

*Zůstal bez peněz.*PAT. [He remains out of pocket.]
ACT(1;obl) ACMP(bez+2;obl)
⇒ should be changed to ACT(1;obl) PAT(bez+2;obl)

See **?** for more examples.

Conclusion and Further Research

- Described VALLEX structure
- Evaluated inter-annotator agreement $\Rightarrow$ VALLEX potential applicability for WSD.
- Prepared golden data for verb frame (sense) disambiguation.
- Described first experiments with automatic WFD
  . . . to our knowledge the first automatic WSD of this scale for Czech
- Provided VALLEX with feedback on missing or errorneous entries
  $\Rightarrow$ fixed in VALLEX 1.5
- Continue the development of VALLEX:
  qualitative aspects: describe verb alternations and types of reflexivity
  quantitative aspects: add more verbs

References