

# **VALEVAL: Anotace VALLEXem**

Ondřej Bojar, obo@cuni.cz

26. leden, 2004

---

# Obsah

- Proč VALEVAL
- Co je VALEVAL, postup anotace
- Meziannotátorská shoda
- Korekce VALLEXu díky VALEVALu
- Golden VALEVAL
- Poznámky ke způsobu práce
- Co s VALEVALEM dál

## Cíle VALEVALu

- Ověřit použitelnost hesel
- Testovat konzistenci a úplnost VALLEXu
- Připravit data pro automatické metody identifikace a generování valenčních rámců

Pro slovesa: sense  $\sim$  frame

$\Rightarrow$  word frame disambiguation (WFD)

## Anotovaná data

- 2 x cca 50 náhodně vybraných sloves
  - napříč slovesy všech složitostí
  - napříč sémantickými třídami
  - vidové dvojice vybírány jako celek
  - reflexivní varianty vybírány jako celek
- 100 náhodně vybraných příkladů ke každému slovesu z ČNK, plus kontext tří předchozích vět.
- 3 paralelní anotace (VB-KCH-KP, VB-MR-obo)

# Ukázka

- Jak se anotuje: `nechat.htm`
- Jak se prohlížejí výsledky:  
`wfa_results.htm`, `pick_examples.htm`, `prozlacovani.htm`

## Výsledky anotace

Průměrná doba anotace: 34 minut  
26 minut (Magda Razímová) až 47 (Klára Chvátalová).

Celkem anotovaných vět: 10256  
Celkem anotací: 30765  
- z toho chybějící rámeček 1758 5,7 %

Anotovaných lemmat: 109  
Anotovaných sloves: 197 (když počítáme reflexivní varianty zvlášť)

Průměrně možných rámců na sloveso:  
ve VALLEXu celkem 3.92  
u vybraných sloves 3.77

## Mezianotátorská shoda

	Celkem vět	10256
	Přesná shoda	67.3 %
	Neprázdný průnik	70.4 %
Ignorujeme-li nejistotu	Přesná shoda	70.5 %
	Neprázdný průnik	74.2 %
	Celkem vět	9280
Jen věty, kde všichni jisti	Přesná shoda	74.4%
	Neprázdný průnik	77.6%
	VB KCH	75.0 %
	KCH KaPa	77.1 %
Po dvojicích	VB KaPa	77.5 %
	VB obo	78.2 %
	magda obo	78.9 %
	VB magda	83.4 %

# VALEVAL VALLEX<sub>u</sub>

Data vychází jen z dávky 1, tj. 1/2.

Celkem 68 “závažných” chyb.

- chyby v rámcích (75 chyb)
  - chybějící rámce (57 chyb, z toho 39 idiomy)
  - nesprávně spojené nebo oddělené rámce (6 chyb)
  - nadbytečné rámce (12)
- chyby ve funktorech, povrchových realizacích a typech doplnění (32)
- chybné glosy a příklady (30 corrections)

## Golden VALEVAL

= věty, kde se všichni shodli a volili smysluplnou odpověď  
(nebo to autoritativně Vendula rozhodla)

	VALEVAL	PDT
Celkem výskytů sloves:	8066	65057
Výskytů s 1 možností:	864	17811
Přesnost při volbě nejčtetnějšího:	67.06 %	74.12 %
Přesnost při náhodném výběru:	32.09 %	47.18 %
Unikátních sloves:	108	4740
Sloves s 1 možností:	13	3431
Průměrně možností na sloveso:	4.85	1.63

## Poznámky ke způsobu práce

- osvědčilo se primitivní anotování v textovém editoru
- velmi se osvědčil kontrolní skript pro ověření konzistence
- velmi se osvědčil výstup v HTML, souhrnné tabulky shody, odkazy
- problém je s omezením XSLT – vše musí být v paměti, takže 80 MB valeval.xml je již obtížné zpracovat na notebooku s 512 MB RAM

## Další plány s VALEVALem

- Publikovat:
  - Nepřijato na VerbWorkshop
  - Studentská sekce ACL (já, Vendula)
  - PBML (já, Vendula, Jirka Semecký)
  - TSD (Markéta, Zdeněk + informace o VALEVALu)
  
- Zařadit VALEVAL mezi SENSEVAL
  
- Jirka Semecký: automatická identifikace rámce
  
- Eda Bejček: ? automatické generování rámců ?

## Shrnutí

- golden VALEVAL = 8 000 vět anotovaných VALLEXem, k experimentům s WFD
- není jasný vztah mezi počtem rámců a shodou anotátorů, záleží na kvalitě hesel
- není jasný vztah mezi četností a shodou anotátorů, záleží na kvalitě hesel
- WFD baseline: 67 % (nový golden VALEVAL)