

## Preparing VerbaLex Printed Edition

Dana Hlaváčková   Aleš Horák   Karel Pala

Natural Language Processing Centre  
Faculty of Informatics, Masaryk University  
Botanická 68a, CZ-602 00 Brno, Czech Republic  
E-mail: {hlavack, hales, pala}@fi.muni.cz

RASLAN 2013

# Outline

History

VerbaLex Valency Lexicon

VerbaLex in Print

# History

until 2005:

- ▶ Pala, Ševeček: Valence českých sloves (Valencies of Czech Verbs), 1997. Syntactic valency frames (called **BRIEF**), 15 000 verbs:  
`opustit <v>hPTc4,hPTc4-hPTc3r{kvůli},hPTc4-hPTc4r{pro}`
- ▶ Balkanet EU project, 2002. **Czech WordNet** valency frames, 3 000 verbs:  
 Synonyms: `opustit:6(opouštět), nechat:9(-)`  
 Valency: `kdo1*AG(person:1)=koho4*PAT((person:1)|(animal:1))  
 ?(kvůli komu3|pro koho4)*CAUSE(person:1)`
- ▶ Prague **Vallex** 1.0, 1 000 verbs:  
`~ impf: opouštět pf: opustit  
 + ACT(1;obl) PAT(4;obl) CAUS(kvůli+3,pro+4;typ)`

**Problems** – small coverage, machine processing features not unified

# History

2006–2013:

- ▶ development of **VerbaLex** valency lexicon
- ▶ start – inspiration from the Vallex development procedure
- ▶ all tools newly developed, HTML layout reused
- ▶ edit – formatted plain text in **VIM**
- ▶ transform tool to **XML**
- ▶ exports to **HTML** and **LaTeX** (**PDF**)
- ▶ main editing work by 5–10 linguists

## VIM VerbaLex Editing

verbalex-NLP.txt = (/nlp/verbalex/data/verbalex-NLP) - GVIM

Soubor Úpravy Nástroje Syntaxe Buffery Okna VerbaLex Nápověda

```

* NAVÁZAT:1(NAVAZOVAT), SPOJIT:2(SPOJOVAT)
~ dok: navázat:1 ned: navazovat
+ AG(kdo1;<person:1>;obl1)+++VERB+++OBJ(co4;<ob
  -synon: dok: spojit:2 ned: spojovat
  -example: dok: navázal provaz
  -example: dok: navázal jeden konec na druh
  -use: prim
  -class:
  -occurs: verbalex
~ dok: spojit:2 ned: spojovat
+ AG(kdo1;<person:1>;obl1)+++VERB+++OBJ(co4;<ob
  -synon: dok: navázat:1 ned: navazovat
  -example: dok: spojil televizi s videem
  -use: prim
  -class:
  -occurs: verbalex

* NAVÁZAT:3(NAVAZOVAT)
~ dok: navázat:3 ned: navazovat
+ AG(kdo1;<person:1>;obl1)+++VERB+++SOC(co4;<relationship:3>;obl1)+++SOC(s kým7;<
  -synon:
  -example: dok: navázal přátelství s novými lidmi
  -use: prim
  -class:
  -occurs: verbalex

* NAVINOUT:1(NAVINOVAT), NAVÍT:1(NAVÍJET), NATOČIT:6(NATAČET)
~ dok: navinout:1 ned: navinovat dok: navít:1 ned: navíjet dok: natočit:6
+ AG(kdo1;<person:1>;obl1)+++VERB+++ART(co4;<line:18>;obl1)+++OBJ(na co4|kolem če
  -synon:
  -example: dok: navinula přízi kolem cívky
  -use: prim
-- INSERT --

```

6452,15 43%

vložit Sémantickou roli  
Nové sloveso  
Nový ráMec  
Změna velikosti písmen  
otevřít Soubor pro doplňování  
zavřít soubor pro Doplňování  
Zobrazit jen Lemmata zM  
Zobrazit i Rámce  
Zobrazit i valenCe  
Zobrazit vše zR

ABS-INFO  
INS-TIME  
INS  
INT  
ISUB  
KNOW  
LOC  
MAN  
MEANS  
OBJ  
PART  
PAT  
PHEN  
POS  
REAS  
REC  
REL  
SOC  
STATE  
SUB  
SUBS  
TIME

## Current VerbaLex Editor

Verbalex zvolit: 1, volit: 1

Head

lemma: zvolit sense: 1 aspect: pf + lemma - lemma

lemma: volit sense: 1 aspect: impf + lemma - lemma

WN ID ▾ ENG20-00652154-v

can be  reflexivity: syn use: prim

passive:

def: vybrat, vybírat z nabízených možností (v politice)

Frames

	left	right	right
+frame	form: kdo1	prep:	prep:
copy	SR: AG	form: koho4	form: kým7
-frame	constraint: person:1 institution:1	type: adj	type:
		SR: PAT	SR: PAT
		constraint: person:1 group:1	constraint: role:1
		opt: <input type="checkbox"/>	opt: <input checked="" type="checkbox"/>
		+ slot - slot	+ slot - slot

+ex -ex  pf  impf  biasp Členové skupiny zvolili Pavla svým předsedou.

```

<synset_entry id="id440051">
  <wnid resource="ENG20-00334609-v"/><wneq>get down:7, begin:1, get:34...
  <definition>přistoupit ke konání něčeho</definition>
  <synset_variants>
    <lemma_variants id="id440051_1" canbepassive="no" sense="1">
      <lemma aspect="pf"> <variant>dát se</variant> </lemma>
      <lemma aspect="impf"> <variant>dávat se</variant> </lemma>
    </lemma_variants> ... </synset_variants>
  <frame_entry>
    <example aspect="pf">stavební firma <example_verb>se</example_verb>
      <example_verb tag="k5eAaPmAgNnP" lemma="pustit">pustila</example_verb>
      do stavby domu </example>
  <frame_slots>
    <slot obl="obl">
      <functor role="AG"> <constraint wnclass="institution:1"/> </functor>
      <grammar type="direct_case" case="1" anim="false"/>
    </slot>
    <slot obl="obl">
      <functor role="ACT"> <constraint wnclass="act:2"/> </functor>
      <grammar type="prepos_case" case="2" anim="false" prepos_lemma="do"/>
    </slot> </frame_slots>
  </frame_entry> ...
</synset_entry>

```

alphabet	semantic role	sel. restriction	gram. structure	verb class	phraseme	aspect	complexity
misc.			↶	⊥	CS		
<b>Alphabet</b>	<b>Verbs starting with letter "d"</b>	<p>dát se<sup>pf</sup><sub>1</sub>    pustit se<sup>pf</sup><sub>1</sub>    začít<sup>pf</sup><sub>1</sub>    chopit</p> <p>dávat se<sup>impf</sup><sub>1</sub>    pouštět se<sup>impf</sup><sub>1</sub>    začínat<sup>impf</sup><sub>1</sub>    chápat</p>					
<ul style="list-style-type: none"> <li>• A (82)</li> <li>• B (183)</li> <li>• C (72)</li> <li>• Ć (73)</li> <li>• <b>D (523)</b></li> <li>• Ď (3)</li> <li>• E (16)</li> <li>• F (33)</li> <li>• G (9)</li> <li>• H (107)</li> <li>• CH (50)</li> <li>• I (19)</li> <li>• J (18)</li> <li>• K (418)</li> <li>• L (139)</li> <li>• M (220)</li> <li>• N (854)</li> <li>• Ń (2)</li> <li>• O (653)</li> <li>• P (2699)</li> <li>• R (690)</li> <li>• Ř (22)</li> <li>• S (556)</li> <li>• Š (47)</li> <li>• T (98)</li> <li>• Ť (4)</li> <li>• U (596)</li> <li>• V (1214)</li> <li>• X (1)</li> <li>• Z (1029)</li> <li>• Ž (19)</li> </ul>	<ul style="list-style-type: none"> <li>• dabovat</li> <li>• dálit</li> <li>• darebačit</li> <li>• darovat</li> <li>• dařit se</li> <li>• dát</li> <li>• datovat</li> <li>• datovat se</li> <li>• <b>dát se</b></li> <li>• dát si</li> <li>• dávat</li> <li>• dávat se</li> <li>• dávat si</li> <li>• dávit</li> <li>• dávit se</li> <li>• dávkovat</li> <li>• dbát</li> <li>• debatovat</li> <li>• debitovat</li> <li>• debutovat</li> <li>• decimovat</li> <li>• dedikovat</li> <li>• dědit</li> <li>• dedukovat</li> <li>• defilovat</li> <li>• definovat</li> <li>• deformovat</li> <li>• deformovat se</li> <li>• defraudovat</li> <li>• degenerovat</li> </ul>	<p><b>1</b> dát se<sub>1</sub>, dávat se<sub>1</sub>, pustit se<sub>1</sub>, pouštět se<sub>1</sub> ≈</p> <p>-frame: <b>AG</b> &lt;institution:1&gt;<sub>i1</sub> <b>VERB</b> obl <b>ACT</b> &lt;act:2&gt;<sub>obl do+i2</sub></p> <p>-example: <b>stavební firma se pustila do stavby domu (pf)</b></p> <p><b>2</b> začít<sub>1</sub>, začínat<sub>1</sub> ≈</p> <p><b>3</b> začít<sub>1</sub>, začínat<sub>1</sub> ≈</p> <p><b>4</b> chopit se<sub>3</sub>, chápat se<sub>3</sub>, ujmout se<sub>4</sub>, ujímat se<sub>4</sub> ≈</p> <p><b>5</b> začít<sub>1</sub>, začínat<sub>1</sub> ≈</p> <p><b>6</b> dát se<sub>1</sub>, dávat se<sub>1</sub>, pustit se<sub>1</sub>, pouštět se<sub>1</sub> ≈</p> <p>-frame: <b>AG</b> &lt;person:1&gt;<sub>obl a1</sub> <b>VERB</b> obl <b>ACT</b> &lt;act:2&gt;<sub>obl do+i2</sub></p> <p>-example: <b>pustil se do práce (pf)</b></p> <p><b>7</b> začít<sub>1</sub>, začínat<sub>1</sub> ≈</p>					



## Valency Frames and Semantic Roles in Czech

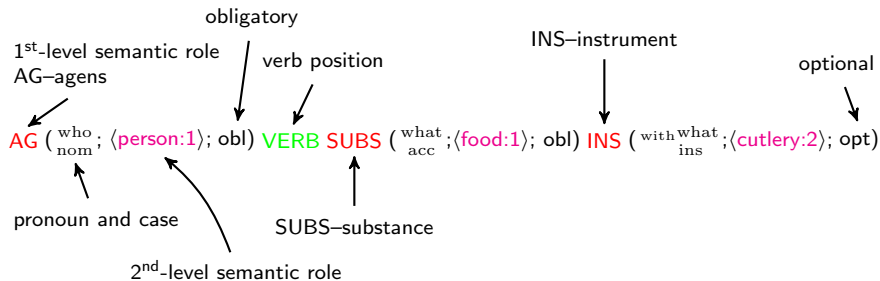
### VerbaLex:

- ▶ 10 500 Czech verb lemmata, 19 000 verb frames
- ▶ Surface and deep valencies
- ▶ Inventory of semantic roles
- ▶ Reasons for two-level notation – EWN Top Ontology, BCs – general labels
- ▶ Subcategorization features – literals from PWN 2
- ▶ Verb semantic classes
- ▶ Strong connection to WordNet – machine processing

### Vallex:

- ▶ 6 500 verbs
- ▶ different approaches – oriented to linguistic processing

## Surface and deep valencies



- ▶ **basic valency frames** – predicate-argument structure of a verb
- ▶ semantic roles – **subcategorization features** (or selectional restrictions) required by the meaning of the verb.

## Surface and deep valencies

**jíst:1 (impf), požit:2(pf), požívat:2(impf)** (eat:1)

definition: přijímat potravu (take in solid food)

class: eat-39.1

passive: yes

**jíst:1** (eat:1) ≈

-frame: **AG**(<sup>who</sup><sub>nom</sub>; ⟨**person:1**⟩;obl) **VERB SUBS**(<sup>what</sup><sub>acc</sub>; ⟨**food:1**⟩;obl)

**INS**(<sup>with what</sup><sub>ins</sub>; ⟨**cutlery:2**⟩;opt)

-example: synovec jedl zmrzlinu (impf) (the nephew ate an ice cream)

-example: dcera jí polévku lžící (impf) (the daughter eats a soup with a spoon)

-use: prim

-reflexivity: no

**complex valency frames** – verb sense, aspect, verb semantic classes, ability to form passive voice (transitive and intransitive verbs), reflexivity, behaviour of the verb in a context: primary, figurative, idiomatic usage

## VerbaLex Semantic roles

- ▶ 1<sup>st</sup>-level – 32 top PWN hypernyms, broad class
- ▶ 2<sup>nd</sup>-level – PWN literals, typical class

**Substance** – in VerbaLex a semantic role:

**1<sup>st</sup>-level – SUBS**

**2<sup>nd</sup>-level, PWN hypernym – substance:1**

**Two-layer semantic role – SUBS(substance:1)**

**Hyponymic lexical units as specifiers:**

SUBS⟨solid:1⟩, SUBS⟨liquid:3⟩, SUBS⟨gas:2⟩, SUBS⟨food:1⟩,  
SUBS⟨beverage:1⟩, ...

**Hyponymic subclass of particular examples:**

SUBS⟨beverage:1⟩ = milk:1, alcohol:1, chocolate:1, fruit juice:1, soft  
drink:1, coffee:1, tea:1, drinking water:1, ...

## VerbaLex and WordNet

- ▶ valency frames for (sub)[synsets](#) and verb senses, not verb lemma
- ▶ synsets are linked to their [English equivalents](#) in PWN
- ▶ 3 686 whole **new synsets**
- ▶ **15 %** of them – [no lexicalized equivalent](#) in English – perfective, reflexive or prefixed verbs verbs with expressive or metaphoric meaning
  - povyskočit* (“jump up a little”)
  - povyskakovat* (“jump out of [something] one after another”)
  - povyřizovat* (“finish doing things successively”)
  - dovyplnit* (“fill in an extra information”)

## Semantic Classes

- ▶ classification system of English verbs by **B. Levin** – 48 classes
- ▶ extended by **M. Palmer** VerbNet project – 83 classes
- ▶ **VerbaLex** – 109 classes

**withdraw-80** – abdikovat:1, odstoupit:2, vzdát se:1, couvnout:2

**contribute-13.2-1** – alokovat:1, přidělit:5, rozdělit:7

**correlate-84** – adaptovat se:1, aklimatizovat se:1, přizpůsobit se:1

## Preparation of VerbaLex Book

- ▶ 500–800 verbs
  - ▶ selection by frequency
- |        |             |
|--------|-------------|
| být    | 160 587 902 |
| mít    | 31 663 942  |
| mocet  | 16 108 948  |
| chtít  | 7 301 412   |
| stát   | 6 207 847   |
| jít    | 6 092 565   |
| dát    | 5 138 742   |
| vědět  | 4 846 575   |
| říct   | 4 494 682   |
| musit  | 4 394 941   |
| vidět  | 3 943 882   |
| dostat | 3 568 890   |
| začít  | 3 509 862   |

## VerbaLex Book Format

### ◆PŘIPUSTIT<sub>1</sub> (PŘIPOUŠTĚT<sub>1</sub>)

připustit<sub>1</sub>, připouštět<sub>1</sub>, **přiznat<sub>1</sub>**, **přiznávat<sub>1</sub>**

D: *po jistém zdráhání uznat něco za možné*

V: *člověk/instituce - připustí, připouští - stav*

PWN: admit:1, acknowledge:1

class: allow-63, approve-75, declare-29.4, conjecture-29.5-2

ORG(person:1)**kdo1** VERB STATE(state:4)**co4**

① připustit<sub>1</sub>, připouštět<sub>1</sub>, přiznat<sub>1</sub>, přiznávat<sub>1</sub>

*připouštím tuto možnost; připustili nebezpečí; vědec připustil existenci výjimek*

use: *prim*, reflexivity: *no*

GROUP(institution:1)**kdo1** VERB STATE(state:4)**co4**

② připustit<sub>1</sub>, připouštět<sub>1</sub>, přiznat<sub>1</sub>, přiznávat<sub>1</sub>

*ruská vláda tento stav připouští velmi neochotně; parlament připustil velký vliv ministerstva vnitra*

use: *prim*, reflexivity: *no*