Sémantická analýza, valenční slovníky

Počítačové nástroje pro češtinu
jaro 2021

Markéta Audy Masopustová
audy.masopustova@phil.muni.cz

Sémantická analýza

Termíny

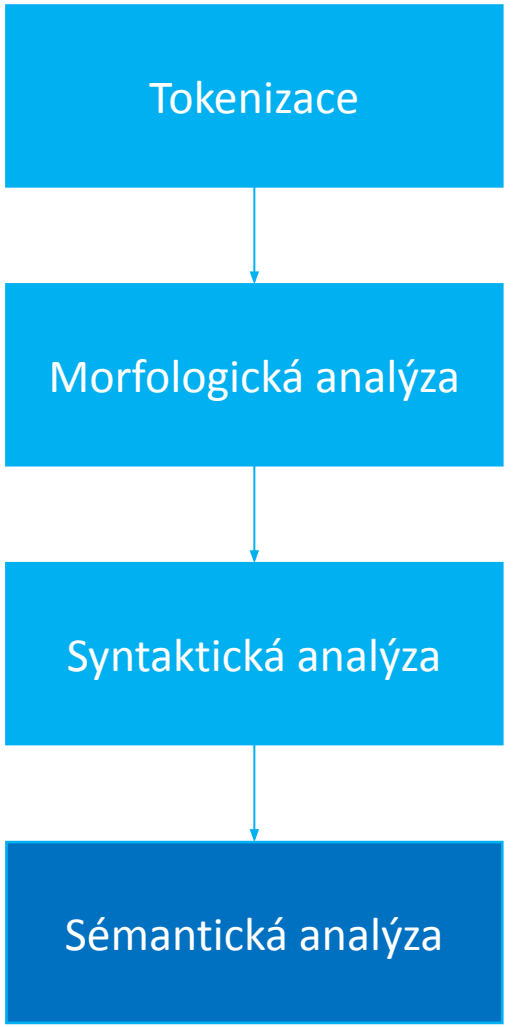
- Sémantika;
- Hyponyma;
- Hyperonyma;
- Kohyponyma;
- Meronyma;
- Synset.

Termíny

- Sémantika – nauka o významu;
- Hyponyma – podřazená slova;
- Hyperonyma – nadřazená slova;
- Kohyponyma – významově shodná slova;
- Meronyma – označuje část celku;
- Synset – synonymická řada.

Sémantická analýza

- Snaží se o formální popis významu – rozdělit slova do skupin a dát jim nějakou nálepku.
- Snaží se o zobecnění světa.
- Měla by být jazykově nezávislá.
- Počátky můžeme najít v ontologiích (v inf. explicitní a formalizovaný popis určité problematiky).
- Většinou se jedná spíše o nějaký slovník, který uchovává znalosti z určité problematiky.
- V ČR se tím zabývají především v rámci CZPJ FI MU a na ÚFAL MFF CUNI.



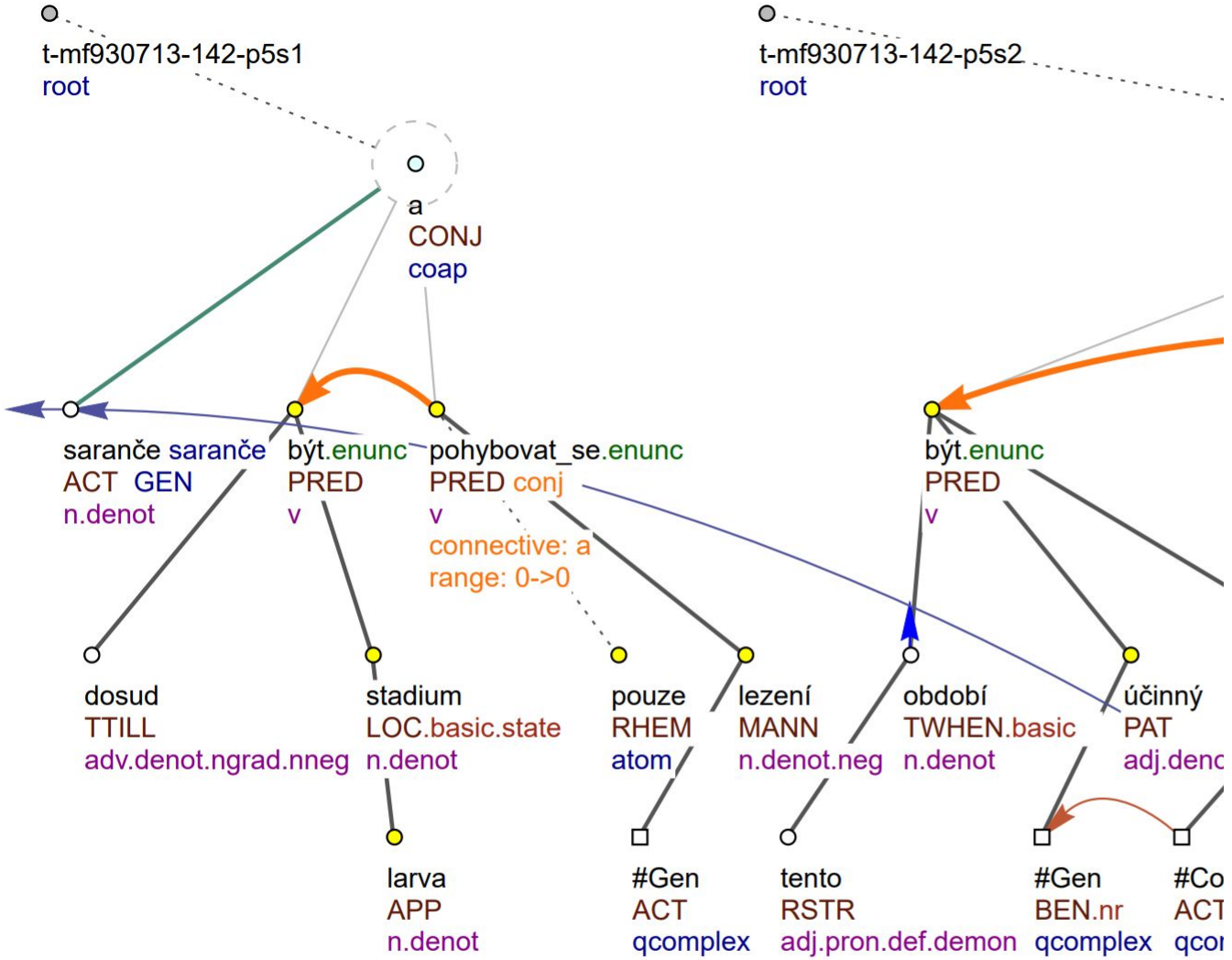
NER

- Rozpoznání pojmenovaných entit (Named Entity Recognition).
- Cílem je najít předem definované kategorie v nestrukturovaném textu.
- Poměrně složité, musí se předem definovat, co je jmenná entita.
- Jmenná entita může být např. jméno, město, datum, značka, …
- https://nlp.fi.muni.cz/projekty/ner/v2/
- http://ufal.mff.cuni.cz/cnec/cnec2.0

Pan Ing. Jan Novák z obce Chýně navrhl, aby se od 15. září 2014 městská rada (dále jen MR) scházela 4 km od památné lípy v centru obce. Svůj návrh podpořil citací § 6 zákona 121/2001 Sb. a svérázným tanečkem.

V rámci ÚFALu

- PDT:
  - Prague Dependency Treebank;
  - České texty doplněné o morfologickou a syntaktickou informaci.
  - Vyznačený není význam, ale o významová roli ve větě (agens, patiens, …).
- SEANCe – projekt ke značkování sentimentu (emocí v textu).

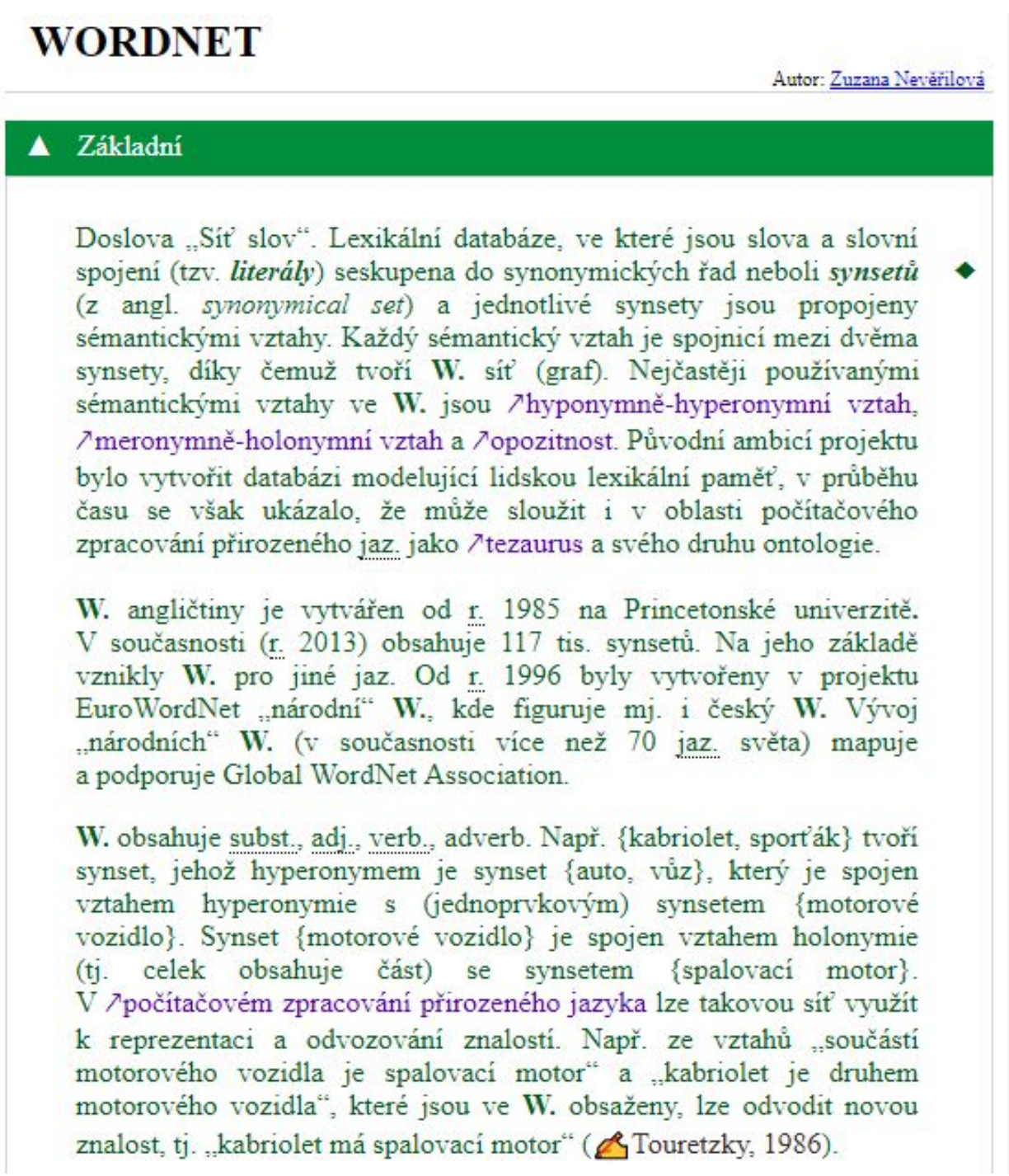


Sémantické sítě (anglické)

- FrameNet
  - lexikální síť čitelná pro člověka i stroj
  - https://framenet.icsi.berkeley.edu/fndrupal/
- VerbNet
  - slovník sloves
  - https://verbs.colorado.edu/~mpalmer/projects/verbnet.html
- ConceptNet
  - sémantická síť vytvořená pro lepší porozumění významu pro stroj
  - http://conceptnet.io/
  - obrázek z ConceptNetu →

WordNet

- G. A. Miller (*Princeton University*) – psycholog, psycholingvista, psycholexikolog.
- Základním je Princeton WordNet (1985), postupně vytvářeny národní Wordnety.
- http://wordnet.princeton.edu
- http://globalwordnet.org/resources/wordnets-in-the-world/
- podrobnosti viz NESČ a obrázek →



WORDNET

Autor: Zuzana Nevěřilová

▲ Základní

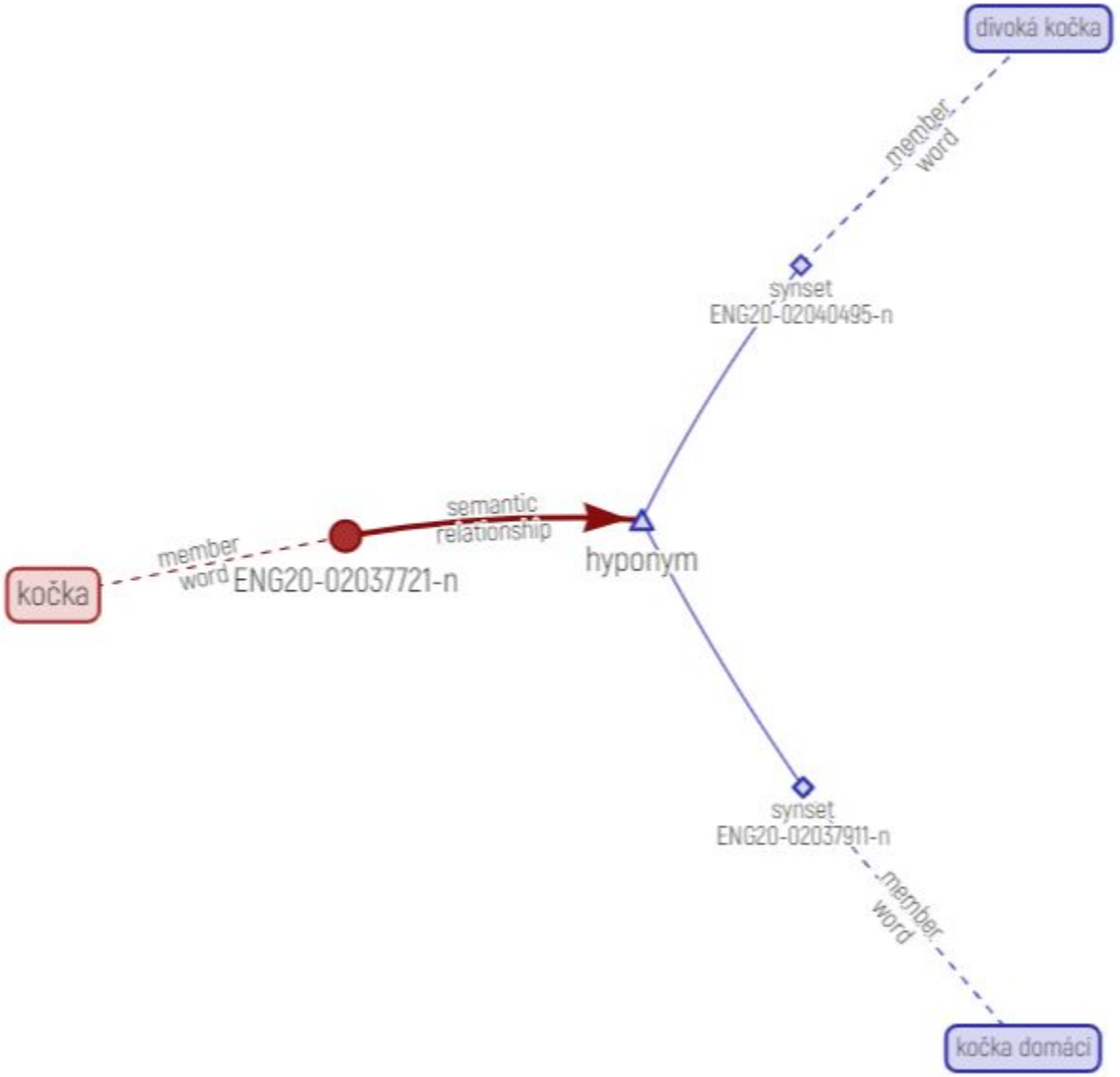
Doslova „Síť slov“. Lexikální databáze, ve které jsou slova a slovní spojení (tzv. ***literály***) seskupena do synonymických řad neboli ***synsetů*** (z angl. *synonymical set*) a jednotlivé synsety jsou propojeny sémantickými vztahy. Každý sémantický vztah je spojnicí mezi dvěma synsety, díky čemuž tvoří **W.** síť (graf). Nejčastěji používanými sémantickými vztahy ve **W.** jsou ↗hyponymně-hyperonymní vztah, ↗meronymně-holonymní vztah a ↗opozitnost. Původní ambicí projektu bylo vytvořit databázi modelující lidskou lexikální paměť, v průběhu času se však ukázalo, že může sloužit i v oblasti počítačového zpracování přirozeného jaz. jako ↗tezaurus a svého druhu ontologie.

W. angličtiny je vytvářen od r. 1985 na Princetonské univerzitě. V současnosti (r. 2013) obsahuje 117 tis. synsetů. Na jeho základě vznikly **W.** pro jiné jaz. Od r. 1996 byly vytvořeny v projektu EuroWordNet „národní“ **W.**, kde figuruje mj. i český **W.** Vývoj „národních“ **W.** (v současnosti více než 70 jaz. světa) mapuje a podporuje Global WordNet Association.

W. obsahuje subst., adj., verb., adverb. Např. {kabriolet, sporťák} tvoří synset, jehož hyperonymem je synset {auto, vůz}, který je spojen vztahem hyperonymie s (jednoprvkovým) synsetem {motorové vozidlo}. Synset {motorové vozidlo} je spojen vztahem holonymie (tj. celek obsahuje část) se synsetem {spalovací motor}. V ↗počítačovém zpracování přirozeného jazyka lze takovou síť využít k reprezentaci a odvozování znalostí. Např. ze vztahů „součástí motorového vozidla je spalovací motor“ a „kabriolet je druhem motorového vozidla“, které jsou ve **W.** obsaženy, lze odvodit novou znalost, tj. „kabriolet má spalovací motor“ (✍Touretzky, 1986).

WordNet prakticky

- Přístup: demo/demo; read/read.
- DebVisDic 2:
  - https://deb.fi.muni.cz/proj_debvisdic-cs.php
- RAW viewer:
  - https://deb.fi.muni.cz/raw-viewer/rawviewer.html



Valenční slovníky

Termíny

- Valence;
- Verbum finitum;
- Verbum infinitum;
- Synset;
- Funktor.

Termíny

- Valence – schopnost vázat na sebe syntaktické pozice, substantiva, adverbia, verba;
- Verbum finitum – sloveso v určitém tvaru;
- Verbum infinitum – sloveso v neurčitém tvaru;
- Synset – synonymická řada;
- Funktor – typ syntakticko-sémantického vztahu mezi slovesem a jeho doplněním.

Valenční slovníky

- není jich mnoho
- snaha vytvořit rozsáhlou elektronickou databázi českých slovesných valenčních rámců
- valence – významem determinovaná schopnost slovesa vázat na sebe další slova
- snaha zachytit valenci sloves na syntaktické a sémantické úrovni a doplnit je o další relevantní informace o chování v přirozeném kontextu

Slovníky:

- Slovesa pro praxi: valenční slovník nejčastějších českých sloves;
- BRIEF;
- VALLEX;
- VerbaLex.

Slovesa pro praxi

- N. Svozilová, 1997.
- První a dlouho jediný tištěný valenční slovník pro češtinu.
- Zdrojem lístkový lexikální archiv ÚJČ AV ČR a *Frekvenční slovník češtiny* (Jelínek, 1961).
- Celkem 767 valenčně analyzovaných sloves.
- Zápis obsahuje heslové slovo, informaci o vidu, stylovém zařazení, výklad významu, větný vzorec a příklady.
- Navazuje na něj *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (Svozilová, Prouzová, Jirsová, 2005).

PROHLÁSIT dok. – ***PROHLAŠOVAT*** ned.

I. ‚důrazně říci / říkat, oznámit / oznamovat'

Val 1 – VF – Val 2

někdo – prohlásí/prohlašuje – že.../„...“/něco

Val 1: S nom [hum > instit] v [zool antropomorf]
Val 2: *že* SENT / „SENT“ // S acc [inform]
Nový velitel prohlásil, že přejímá plnou odpovědnost. – Rakouský spolkový vicekancléř prohlásil před novináři, že rakouská vláda zaujala k vyhlášení palestinského státu kladný postoj. – Jeden západní průmyslník prohlásil: „Váš vztah ke koním mne okouzluje.“ – „To je nesmysl,“ prohlásila rozčileně Alice. – Pán podstrčil Václavovi pod nos bílou legitimaci a prohlásil něco portugalsky. – To není pomluva, Kamila to o sobě prohlásila sama. – Gazette prohlásila, že jednoho dne Poláci budou hrdi na svého Chopina. – Kozel Bobeš radostně prohlásil, že k tomu cirkusu za Mikešem pojede. – – Mluvčí městské policie prohlašuje, že za tento okrsek nenesou zodpovědnost. – Často ve spojeních jako *Slavnostně prohlašuji, že ...; Místopřísežně prohlašuji...; Prohlašujeme závazně...; Všichni prohlašují s plným vědomím...*

BRIEF

- K. Pala a P. Ševeček, 1997.
- Elektronický slovník na FI MU.
- Obsahuje 15 000 sloves a přes 50 000 valenčních rámců.
- Zdrojem knižně vydané slovníky (*SSJČ, SSČ, Slovník českých synonym*).
- Pouze pravostranné valence, u slovesa uvedeny přímé a předložkové pády.
- Základem pro *Český syntaktický slovník* (Skoumalová, 2001).
- Nejsou přístupné pro veřejnost.

Struktura zápisu formátu *Brief*:

bít <v>hPTc4, hPTc4-hTc7, hPc6r{po}, hTc4r{na}, hTc7-hTc4r{o}, hTc7-hTc2r{do}

bít se <v>hPTc7r{s}, hPTc7r{s}-hTc3r{kvůli}, hPc7r{s}, hTc4r{o}

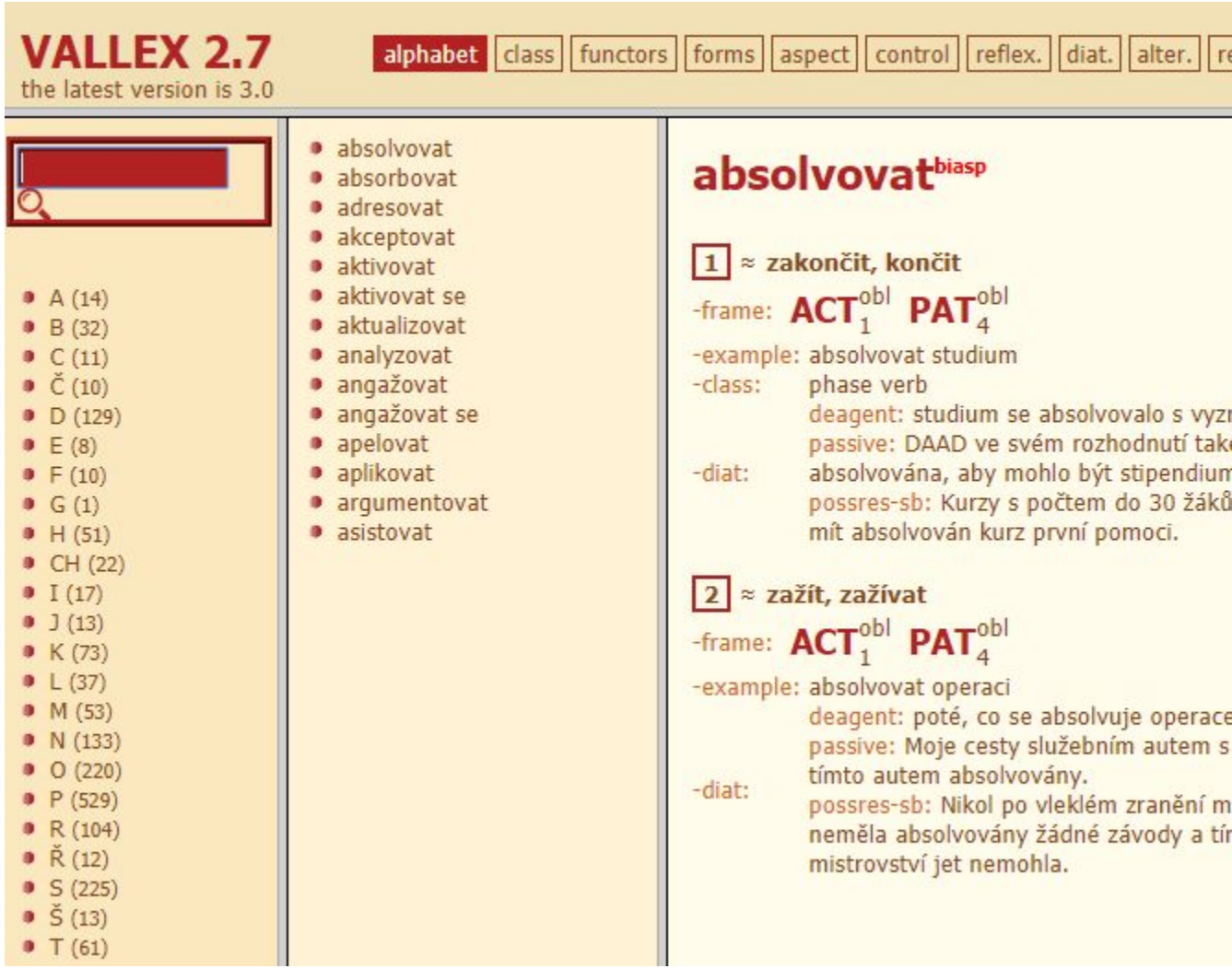
Struktura zápisu formátu *Verbose*:

bít
= koho|co
= koho|co & čím
= po kom
= na co
= čím & do čeho
= čím & o co

bít se
= za koho|co
= s kým|čím
= s kým|čím & kvůli čemu
= s kým
= s kým & o co
= o co

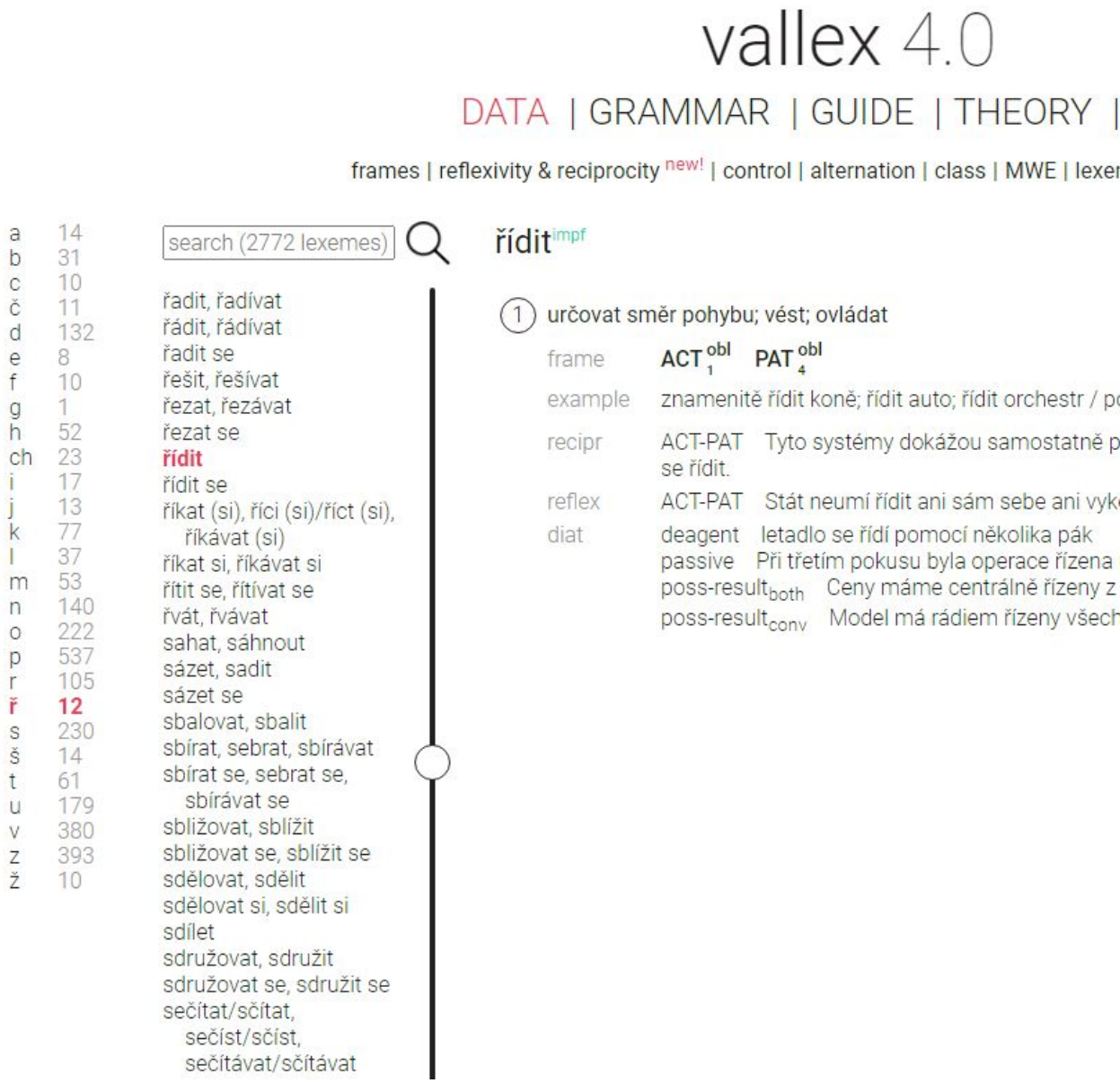
VALLEX

- M. Lopatková, V. Kettnerová, Z. Žabokrtský; vzniká od roku 2001.
- Několik verzí, v roce 2008 vyšla první knižní verze.
- Formální popis valenčních rámců; využívá sémantické role (funktory).
- Vychází z funkčního generativního popisu sloves.
- Zápis obsahuje sloveso v základním tvaru, informaci o vidu, jednotlivé významy, upřesnění pomocí synonymických výrazů, valenční rámec, příklad a případně sémantická třída.
- Valenční pozice obsahují informaci o obligatornosti / fakultativnosti a číslo pádu.



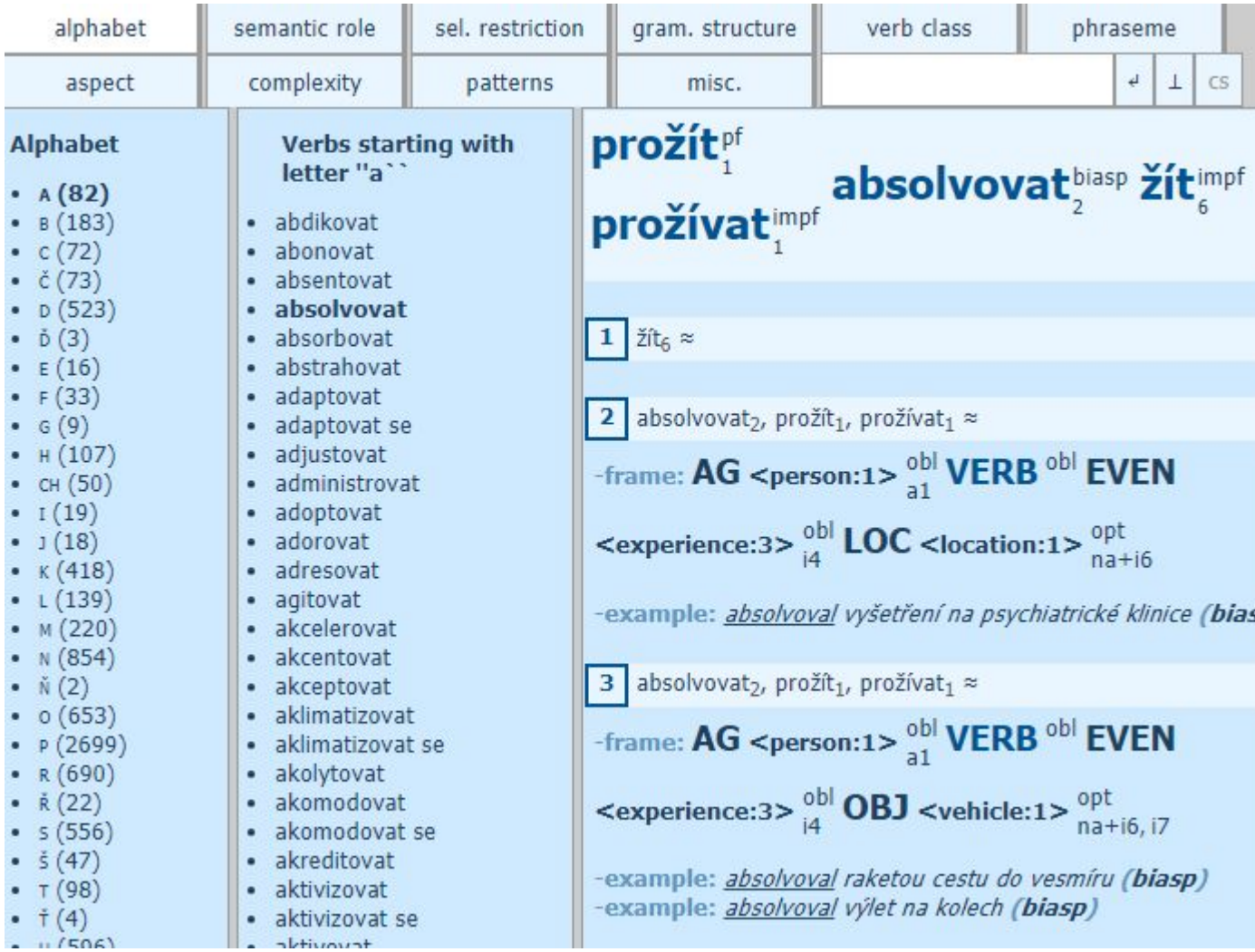
VALLEX 4.0

- http://ufal.mff.cuni.cz/vallex/4.0/
- 4 659 českých sloves, která odpovídají 11 030 lexikálním jednotkám.
- Přímé propojení s PDT.
- Zdrojem BRIEF, *SSČ*, *SSJČ*, *Slovesa pro praxi*, korpusy ČNK řady SYN, PDT.



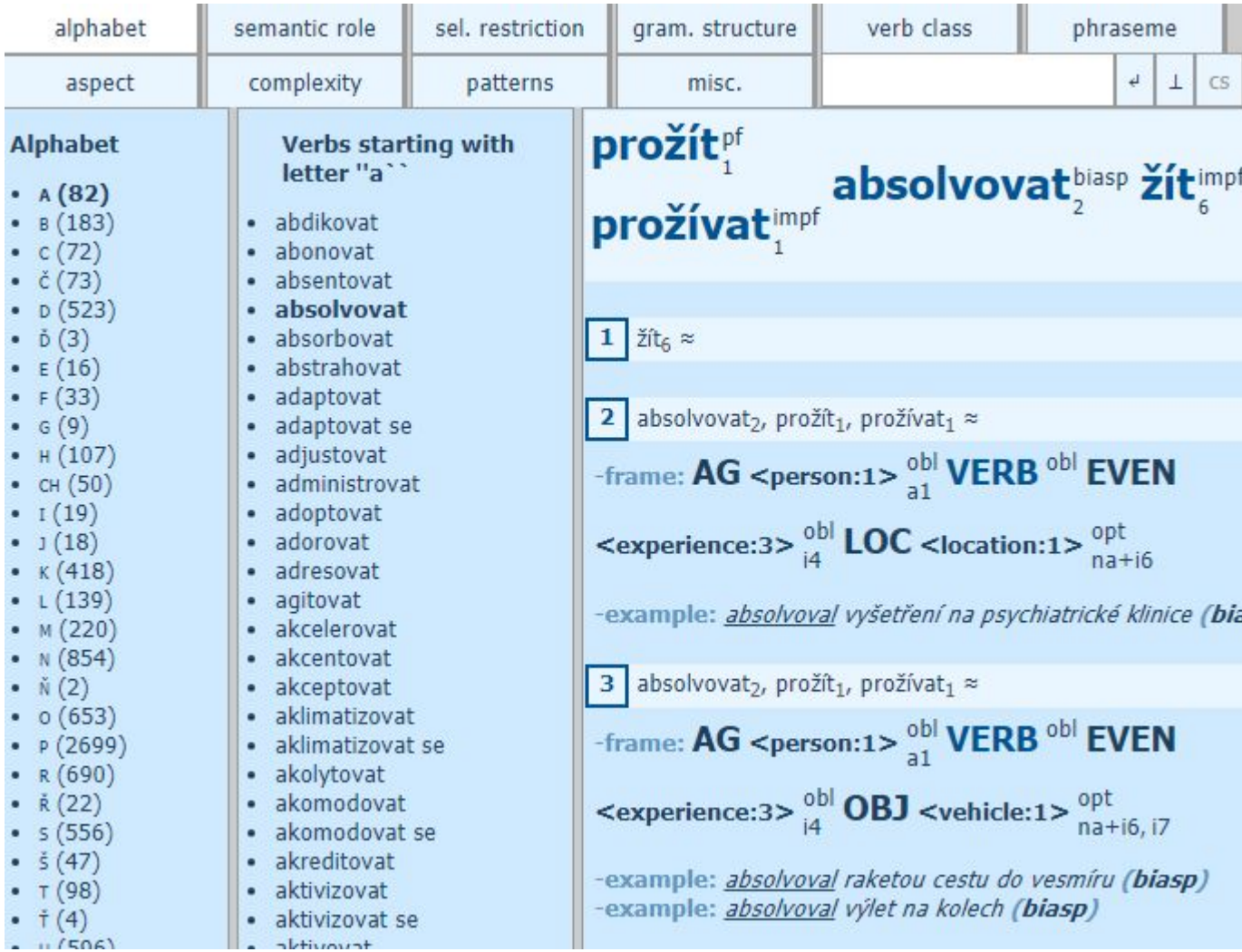
VerbaLex

- D. Hlaváčková, A. Horák; vzniká od roku 2005.
- Inspirace ve VALLEXu.
- Zdrojem BRIEF, VALLEX a český WordNet.
- Systém synonymických řad převzatý z WordNetu (odlišnost od ostatních slovníků).
- Dvě úrovně sémantických rolí:
  - První úroveň – sémantická role podle EuroWordNetu; celkem 38 rolí.
  - Druhá úroveň – hyperonymum; přímý odkaz na Princeton WordNet; otevřená množina.



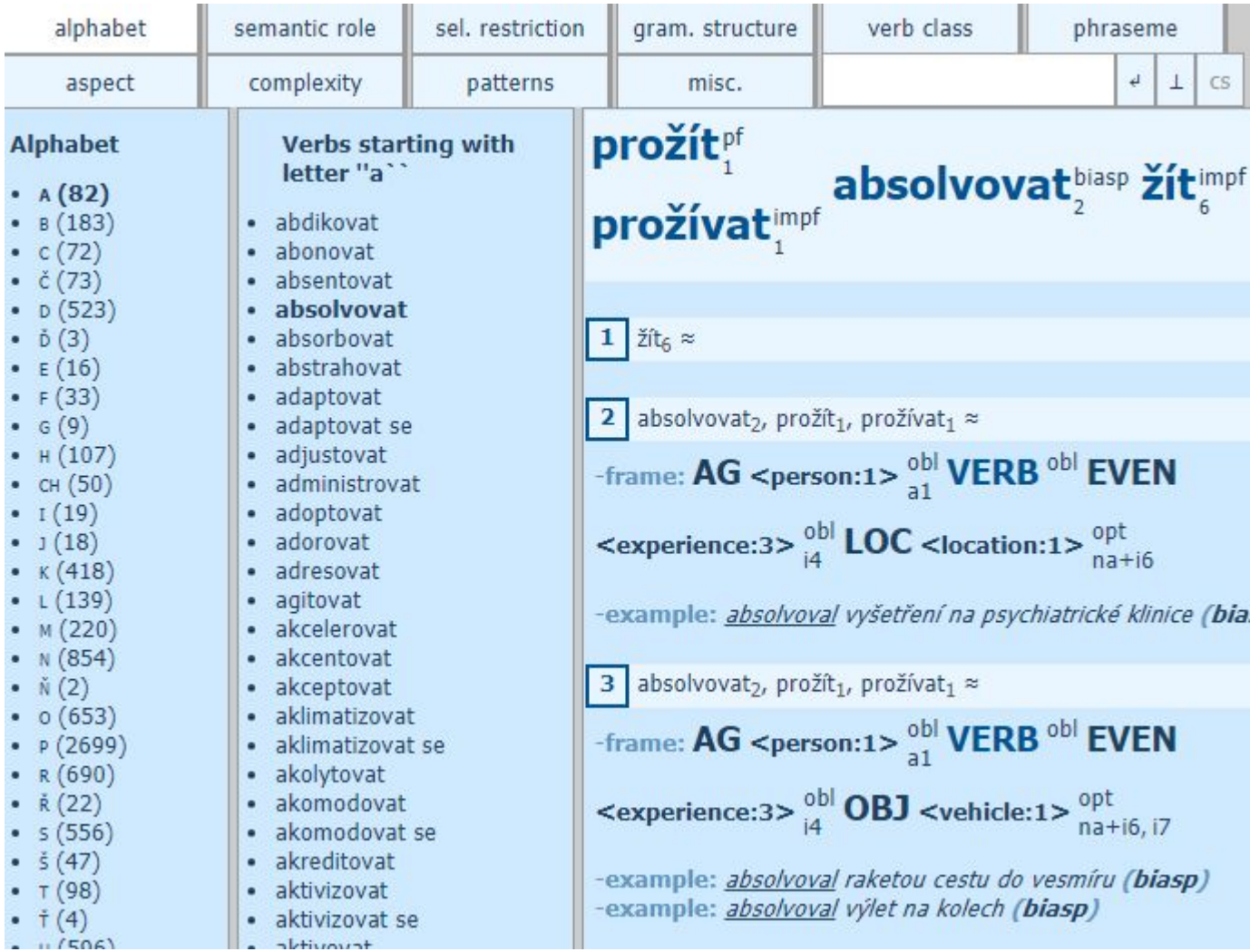
VerbaLex

- Obsahuje 10 469 sloves (slovesných lemmat) a 19 247 valenčních rámců.
- Zápis obsahuje synset (slovesa mají uvedenou vidovou variantu, číslem je označeno pořadí), seznam jednotlivých sloves (čísla v rámečku) s jejich valenčními rámci (obě úrovně), informaci o pádu a příklad.



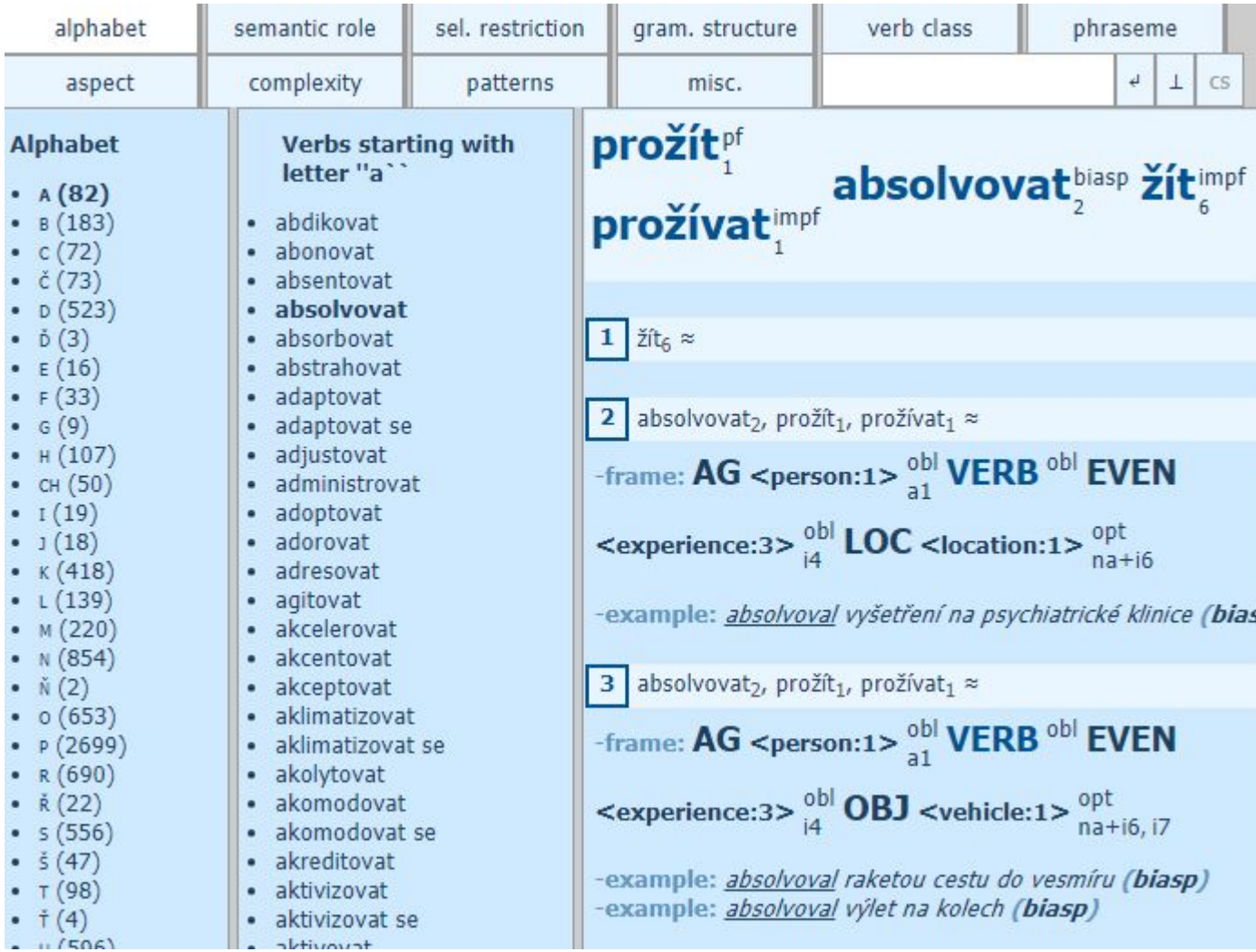
VerbaLex – sémantické třídy sloves

- Motivace v sémantické klasifikaci predikátů (Daneš, Grepl, Karlík).
- Východiskem:
  - B. Levin: *English Verb Classes and Alternations* (48 základních sémantických tříd);
  - M. Palmer: *VerbNet* (82 základních sémantických tříd, celkem 395 podtříd).
- České sémantické třídy:
  - modifikovaný překlad, doplnění o další synonyma, vidové protějšky, prefigovaná slovesa;
  - 82 základních sémantických tříd, celkem 258 podtříd, aktuálně seznam zahrnuje 11 241 sloves, z toho 6 393 různých lemmat.



VerbaLex

- https://nlp.fi.muni.cz/verbalex/html3/index.php
  - Přístup: plin/plin.



Děkuji za pozornost.