

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.



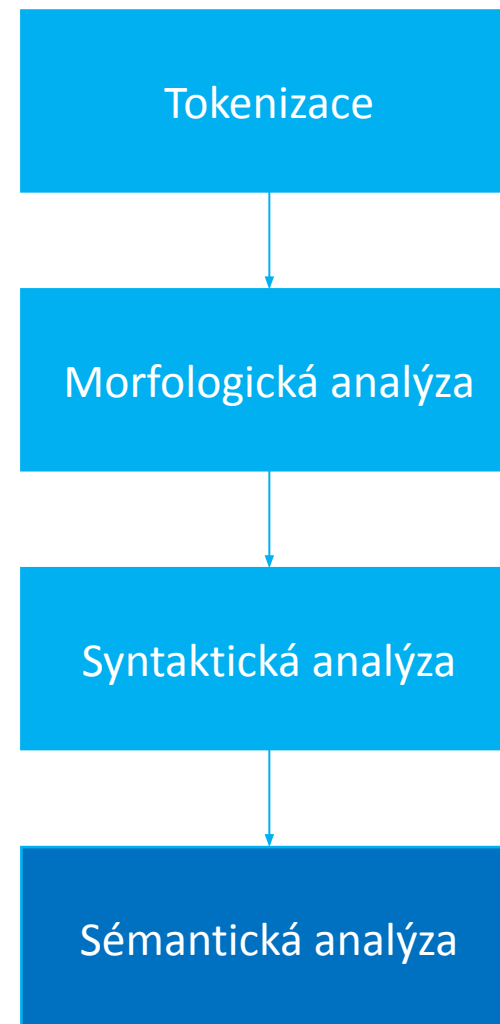
Termíny

- 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 UFAL 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>

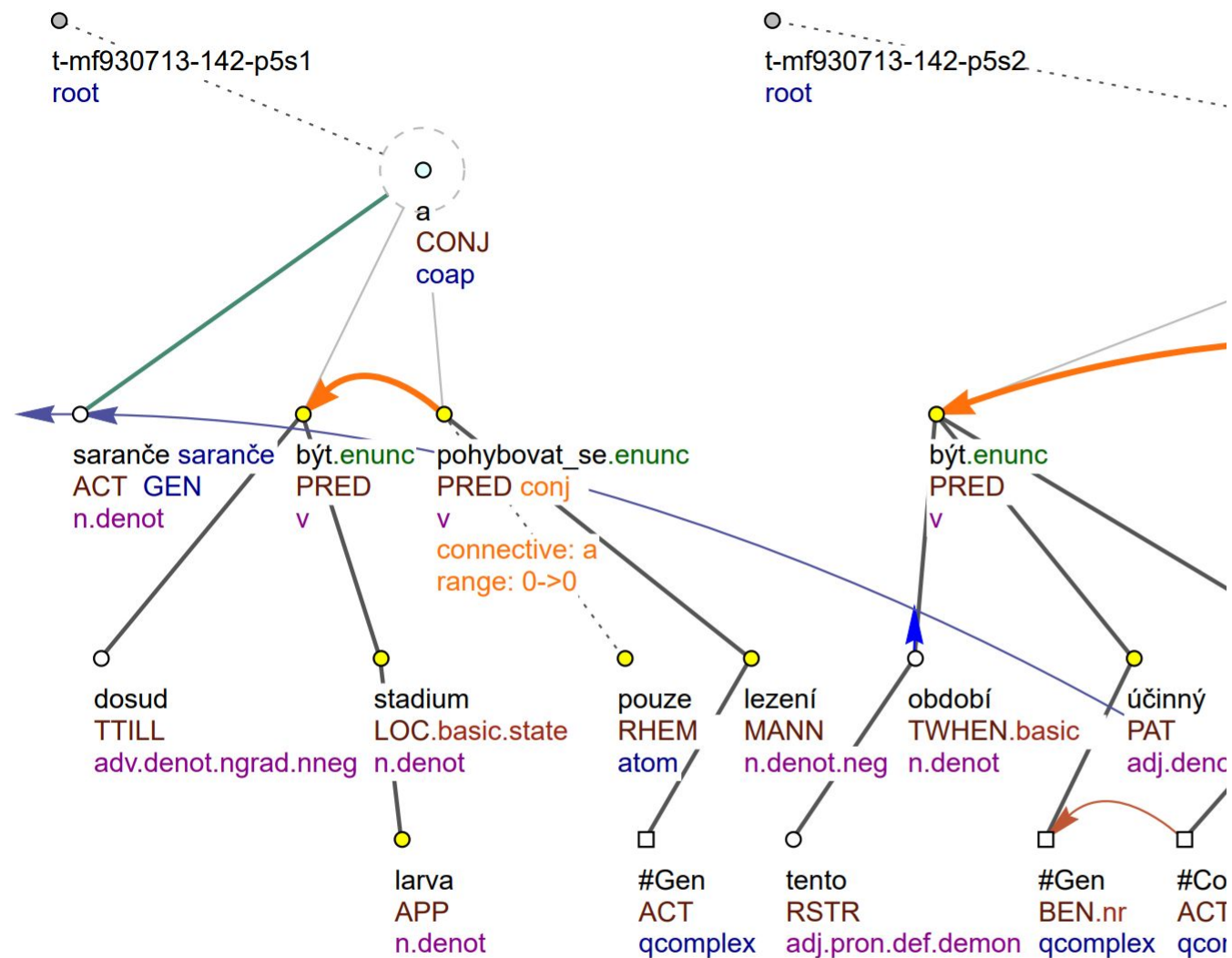
Pan Ing. Jan Novák z obce Chýně navrhl, aby se od 15. září 2014 městská rada (dále jen MR) scházela 4 km od památné lípy v centru obce. Svůj návrh podpořil citací § 6 zákona 121/2001 Sb. a svérázným tanečkem.

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 →

Location of cat

en my lap →
en a bed →
en the windowsill →
en a chair →
en a table →
en a vet →
en the barn →
en the floor →
en your way →
en the backyard →
en bag →
en someone's home →
en the rug →
en an alley →
en a back yard →
en a cat box →
en a closet →
en a house →
en the roof →

cat is capable of...

en hunt mice →
en catch a mouse →
en drink water →
en climb up a tree →
en corner a mouse →
en look at a king →
en kill birds →
en mother her kittens →
en catch a bird →
en cleaning itself →
en drink milk →
en scratch →
en scratch furniture →
en sleep →
en wash its paws →
en eat cat food →
en eye a mouse →
en hide under the bed →
en meow →

WordNet

- G. A. Miller (*Princeton University*) – psycholog, psycholinguista, 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
→

▲ Základní

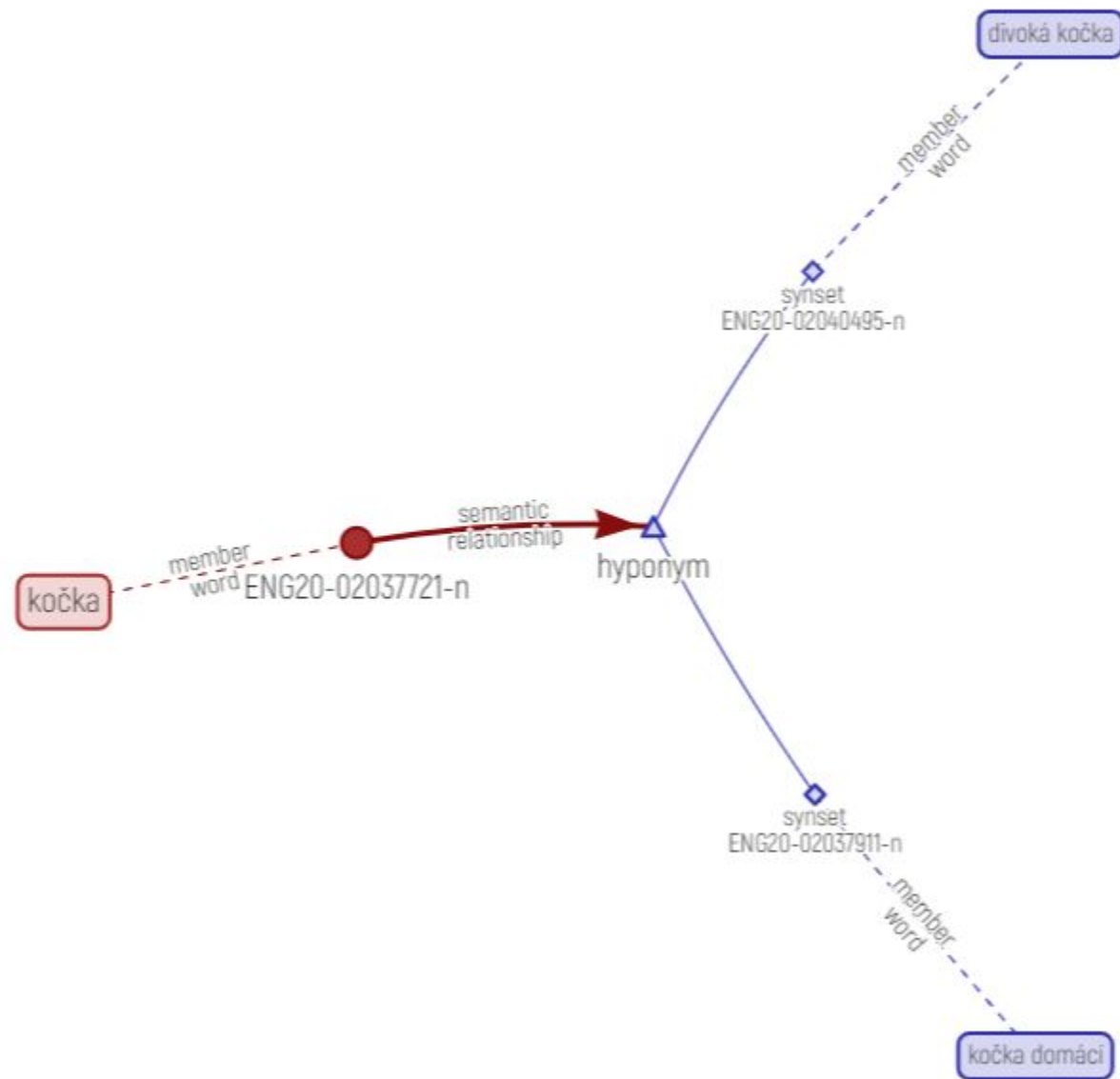
Doslova „Síť slov“. Lexikální databáze, ve které jsou slova a slovní spojení (tzv. *literály*) seskupena do synonymických řad neboli *synsetů* (z angl. *synonymical set*) a jednotlivé synsety jsou propojeny sémantickými vztahy. Každý sémantický vztah je spojnicí mezi dvěma synsety, díky čemuž tvoří W. 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, sportřák} tvoří synset, jehož hyperonymem je synset {auto, vůz}, který je spojen vztahem hyperonymie s (jednoduchým) 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 okouzljuje.“ – „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 spojení jako Slavnostně prohlašují, že ...; Místopřísežně prohlašují...; 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 rámeček, příklad a případně sémantická třída.
- Valenční pozice obsahují informaci o obligatornosti / fakultativnosti a číslo pádu.

VALLEX 2.7
the latest version is 3.0

alphabet class functors forms aspect control reflex. diat. alter. re

- A (14)
- B (32)
- C (11)
- Č (10)
- D (129)
- E (8)
- F (10)
- G (1)
- H (51)
- CH (22)
- I (17)
- J (13)
- K (73)
- L (37)
- M (53)
- N (133)
- O (220)
- P (529)
- R (104)
- Ř (12)
- S (225)
- Š (13)
- T (61)

- absolvovat
- absorbovat
- adresovat
- akceptovat
- aktivovat
- aktivovat se
- aktualizovat
- analyzovat
- angažovat
- angažovat se
- apelovat
- aplikovat
- argumentovat
- asistovat

absolvovat^{biasp}

1 ≈ zakončit, končit

-frame: **ACT**₁^{obl} **PAT**₄^{obl}

-example: absolvovat studium

-class: phase verb

deagent: studium se absolvovalo s vyzr

passive: DAAD ve svém rozhodnutí tak

-diat: absolvována, aby mohlo být stipendium

possres-sb: Kurzy s počtem do 30 žáků
mít absolvován kurz první pomoci.

2 ≈ zažít, zažívat

-frame: **ACT**₁^{obl} **PAT**₄^{obl}

-example: absolvovat operaci

deagent: poté, co se absolvuje operace

passive: Moje cesty služebním autem s

tímto autem absolvovány.

-diat: possres-sb: Nikol po vleklém zranění m

neměla absolvovány žádné závody a tír

mistrovství jet nemohla.

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.

vallex 4.0

DATA | GRAMMAR | GUIDE | THEORY |

frames | reflexivity & reciprocity ^{new!} | control | alternation | class | MWE | lexemes

a 14
b 31
c 10
č 11
d 132
e 8
f 10
g 1
h 52
ch 23
i 17
j 13
k 77
l 37
m 53
n 140
o 222
p 537
r 105
ř 12
s 230
š 14
t 61
u 179
v 380
z 393
ž 10

search (2772 lexemes)



řadit, řadívát
řádit, řádívát
řadit se
řešit, řešívát
řezat, řezávát
řezat se
řít
řít se
řít (si), říci (si)/řít (si),
říkávát (si)
říkat si, říkávát si
řítit se, řítívát se
řvát, řvávát
sahat, sáhnout
sázet, sadit
sázet se
sbalovat, sbalit
sbírat, sebrat, sbírávát
sbírat se, sebrat se,
sbírávát se
sblížovat, sblížit
sblížovat se, sblížit se
sdělovát, sdělit
sdělovát si, sdělit si
sdílet
sdružovat, sdružit
sdružovat se, sdružit se
sečítat/sčítat,
sečíst/sčíst,
sečítávát/sčítávát

řít ^{impf}

① určovat směr pohybu; vést; ovládat

frame	ACT ₁ ^{obl}	PAT ₄ ^{obl}
example	znamenitě řítit koně; řítit auto; řítit orchestr / p	
recipr	ACT-PAT	Tyto systémy dokážou samostatně p se řítit.
reflex	ACT-PAT	Stát neumí řítit ani sám sebe ani vyk
diat	deagent	letadlo se řítit pomocí několika pák
	passive	Při třetím pokusu byla operace řízena
	poss-result _{both}	Ceny máme centrálně řízeny z
	poss-result _{conv}	Model má rádiem řízeny všech

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.

alphabet	semantic role	sel. restriction	gram. structure	verb class	phraseme
aspect	complexity	patterns	misc.		↵ ⊥ CS
Alphabet • A (82) • B (183) • C (72) • Č (73) • D (523) • ě (3) • E (16) • F (33) • G (9) • H (107) • CH (50) • I (19) • J (18) • K (418) • L (139) • M (220) • N (854) • ň (2) • O (653) • P (2699) • R (690) • Ř (22) • S (556) • Š (47) • T (98) • ě (4) • U (506)	Verbs starting with letter "a" • abdikovat • abonovat • absentovat • absolvovat • absorbovat • abstrahovat • adaptovat • adaptovat se • adjustovat • administrovat • adoptovat • adorovat • adresovat • agitovat • akcelarovat • akcentovat • akceptovat • aklimatizovat • aklimatizovat se • akolytovat • akomodovat • akomodovat se • akreditovat • aktivizovat • aktivizovat se • aktivovat			prožit ^{pf} ₁ prožívat ^{impf} ₁ absolvovat ^{biasp} ₂ žít ^{impf} ₆	
				1 žít ₆ ≈ 2 absolvovat ₂ , prožit ₁ , prožívat ₁ ≈ -frame: AG <person:1> obl _{a1} VERB obl EVEN <experience:3> obl _{i4} LOC <location:1> opt _{na+i6} -example: <i>absolvoval vyšetření na psychiatrické klinice (biasp)</i>	
				3 absolvovat ₂ , prožit ₁ , prožívat ₁ ≈ -frame: AG <person:1> obl _{a1} VERB obl EVEN <experience:3> obl _{i4} OBJ <vehicle:1> opt _{na+i6, i7} -example: <i>absolvoval raketou cestu do vesmíru (biasp)</i> -example: <i>absolvoval výlet na kolech (biasp)</i>	

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.

alphabet	semantic role	sel. restriction	gram. structure	verb class	phraseme
aspect	complexity	patterns	misc.		↙ ⊥ CS
Alphabet	Verbs starting with letter "a"				
<ul style="list-style-type: none"> • A (82) • B (183) • C (72) • Č (73) • D (523) • ě (3) • E (16) • F (33) • G (9) • H (107) • CH (50) • I (19) • J (18) • K (418) • L (139) • M (220) • N (854) • ň (2) • O (653) • P (2699) • R (690) • Ř (22) • S (556) • Š (47) • T (98) • ě (4) • U (506) 	<ul style="list-style-type: none"> • abdikovat • abonovat • absentovat • absolvovat • absorbovat • abstrahovat • adaptovat • adaptovat se • adjustovat • administrovat • adoptovat • adorat • adresovat • agitovat • akcelarovat • akcentovat • akceptovat • aklimatizovat • aklimatizovat se • akolytovat • akomodovat • akomodovat se • akreditovat • aktivizovat • aktivizovat se • aktivovat 			<p>prožít^{pf}₁ absolvovat^{biasp}₂ žít^{impf}₆</p> <p>prožívat^{impf}₁</p>	
					<p>1 žít₆ ≈</p>
					<p>2 absolvovat₂, prožít₁, prožívat₁ ≈</p> <p>-frame: AG <person:1>_{a1} VERB obl EVEN</p> <p><experience:3>_{i4} LOC <location:1>_{opt na+i6}</p> <p>-example: <i>absolvoval vyšetření na psychiatrické klinice (biasp)</i></p>
					<p>3 absolvovat₂, prožít₁, prožívat₁ ≈</p> <p>-frame: AG <person:1>_{a1} VERB obl EVEN</p> <p><experience:3>_{i4} OBJ <vehicle:1>_{opt na+i6, i7}</p> <p>-example: <i>absolvoval raketou cestu do vesmíru (biasp)</i></p> <p>-example: <i>absolvoval výlet na kolech (biasp)</i></p>

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.

alphabet	semantic role	sel. restriction	gram. structure	verb class	phraseme
aspect	complexity	patterns	misc.		↵ ⊥ cs
Alphabet • A (82) • B (183) • C (72) • Č (73) • D (523) • ě (3) • E (16) • F (33) • G (9) • H (107) • CH (50) • I (19) • J (18) • K (418) • L (139) • M (220) • N (854) • ň (2) • O (653) • P (2699) • R (690) • Ř (22) • S (556) • Š (47) • T (98) • ě (4) • U (506)	Verbs starting with letter "a" • abdikovat • abonovat • absentovat • absolvovat • absorbovat • abstrahovat • adaptovat • adaptovat se • adjustovat • administrovat • adoptovat • adorat • adresovat • agitovat • akceleroovat • akcentovat • akceptovat • aklimatizovat • aklimatizovat se • akolytovat • akomodovat • akomodovat se • akreditovat • aktivizovat • aktivizovat se • aktivovat			prožít ^{pf} ₁ absolvovat ^{biasp} ₂ prožívat ^{impf} ₁ žít ^{impf} ₆	
				1 žít ₆ ≈ 2 absolvovat ₂ , prožít ₁ , prožívat ₁ ≈ -frame: AG <person:1> _{a1} VERB obl EVEN <experience:3> _{i4} LOC <location:1> _{opt na+i6} -example: <i>absolvoval vyšetření na psychiatrické klinice (biasp)</i>	
				3 absolvovat ₂ , prožít ₁ , prožívat ₁ ≈ -frame: AG <person:1> _{a1} VERB obl EVEN <experience:3> _{i4} OBJ <vehicle:1> _{opt na+i6, i7} -example: <i>absolvoval raketou cestu do vesmíru (biasp)</i> -example: <i>absolvoval výlet na kolech (biasp)</i>	

VerbaLex

- <https://nlp.fi.muni.cz/verballex/html3/index.php>
 - Přístup: plin/plin.

alphabet	semantic role	sel. restriction	gram. structure	verb class	phraseme
aspect	complexity	patterns	misc.		↵ ⊥ CS
Alphabet <ul style="list-style-type: none">• A (82)• B (183)• C (72)• Č (73)• D (523)• ě (3)• E (16)• F (33)• G (9)• H (107)• CH (50)• I (19)• J (18)• K (418)• L (139)• M (220)• N (854)• ň (2)• O (653)• P (2699)• R (690)• Ř (22)• S (556)• Š (47)• T (98)• ě (4)• U (506)	Verbs starting with letter "a" <ul style="list-style-type: none">• abdikovat• abonovat• absentovat• absolvovat• absorbovat• abstrahovat• adaptovat• adaptovat se• adjustovat• administrovat• adoptovat• adorovat• adresovat• agitovat• akcelarovat• akcentovat• akceptovat• aklimatizovat• aklimatizovat se• akolytovat• akomodovat• akomodovat se• akreditovat• aktivizovat• aktivizovat se• aktivovat	prožit ^{pf} ₁ prožívat ^{impf} ₁ absolvovat ^{biasp} ₂ žít ^{impf} ₆	1 žít ₆ ≈ 2 absolvovat ₂ , prožit ₁ , prožívat ₁ ≈ -frame: AG <person:1> ^{obl} _{a1} VERB ^{obl} EVEN <experience:3> ^{obl} _{i4} LOC <location:1> ^{opt} _{na+i6} -example: <i>absolvoval vyšetření na psychiatrické klinice (biasp)</i> 3 absolvovat ₂ , prožit ₁ , prožívat ₁ ≈ -frame: AG <person:1> ^{obl} _{a1} VERB ^{obl} EVEN <experience:3> ^{obl} _{i4} OBJ <vehicle:1> ^{opt} _{na+i6, i7} -example: <i>absolvoval raketou cestu do vesmíru (biasp)</i> -example: <i>absolvoval výlet na kolech (biasp)</i>		

Děkuji za pozornost.

