Modelování slovotvorných vztahů ve slovní zásobě češtiny

Magda Ševčíková

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Seminář formální lingvistiky
29. 5. 2017

Obsah

Úvod: Slovotvorba v teoretickém popisu češtiny

- tvoření slov v češtině
  - odvozování slov afixy
    - *říd-i-t*→*řed-i-tel*, *dát*→*vy-dat*, *drz-ý*→*při-drz-lý*
  - skládání slov
    - $hor\text{-}a_N + lezec_N \rightarrow hor\text{-}o\text{-}lezec_N$
    - $nov\text{-}ý_A + doba_N \rightarrow nov\text{-}o\text{-}dob\text{-}ý_A$
    - $log\text{-}os + \$pedie \rightarrow log\text{-}o\text{-}pedie$
- slovotvorba součástí mluvnic češtiny, nikoli ale gramatického systému
  - v *Mluvnici češtiny* (Komárek et al. 1986) „přechodovou oblastí mezi morfologií a slovníkem“, slovotvorné prostředky jako „prostředky nemorfologické, stejně jako např. prostředky lexikální, slovosledné, intonační a jiné“ (Bednaříková 2009:24)
  - derivace nezahrnována do morfologie, morfologie omezena na flexi
  - minimální pozornost k souvislostem mezi derivací a flexí
  - vs. popis angličtiny: morfologie flektivní vs. derivační

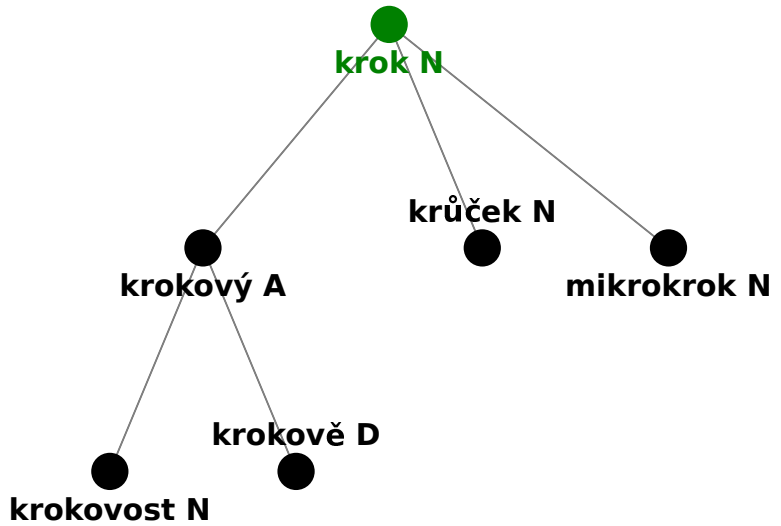
Úvod: Derivace v NLP

- derivační morfologii zatím věnována zásadně menší pozornost než morfologii flektivní
- pro češtinu
  - Derivancze („a new derivational analyser for Czech“; Pala – Šmerk 2015)
  - Morfio (Cvrček – Vondřička 2013), Deriv (Osolsobě et al. 2009)
  - MorfFlex CZ (Hajič – Hlaváčová 2013); tektogramatická anotace PDT; Czech Wordnet (Pala – Hlaváčková 2007)
- zvýšená pozornost k derivaci jiných jazyků relativně nedávno
  - CELEX (en, de, nl; Baayen et al. 1995), DerivBase (de; Zeller et al. 2013), DerivBase.Hr (Šnajder et al. 2014), jazykově nezávislý přístup (Baranes – Sagot 2014), Démonette (fr; Hathout – Namer 2014), Word Formation Latin (Litta et al. 2016)
  - dva workshopy při konferenci Societas Linguistica Europaea 2016, dva samostatné workshopy ve 2017 (Toulouse, Milán)

DeriNet: Zdroj jazykových dat specializovaný na derivaci

http://ufal.mff.cuni.cz/derinet
- vývoj dat: Z. Žabokrtský, J. Vidra, A. Kalužová, N. Mediankin, M. Straka
- teoretická témata: prof. J. Panevová, prof. P. Pognan, J. Hlaváčová

- pouze derivace
  - max. jeden rodič pro každý uzel
  - orientovaná hrana
  - derivační strom: kořenem nemotivované slovo, odpovídá slovotvorné čeledi
- pouze substantiva, adjektiva, adverbia, slovesa
  - deriváty jiných slovních druhů zatím jako neodvozené:
    dvojče, třetina, zčásti

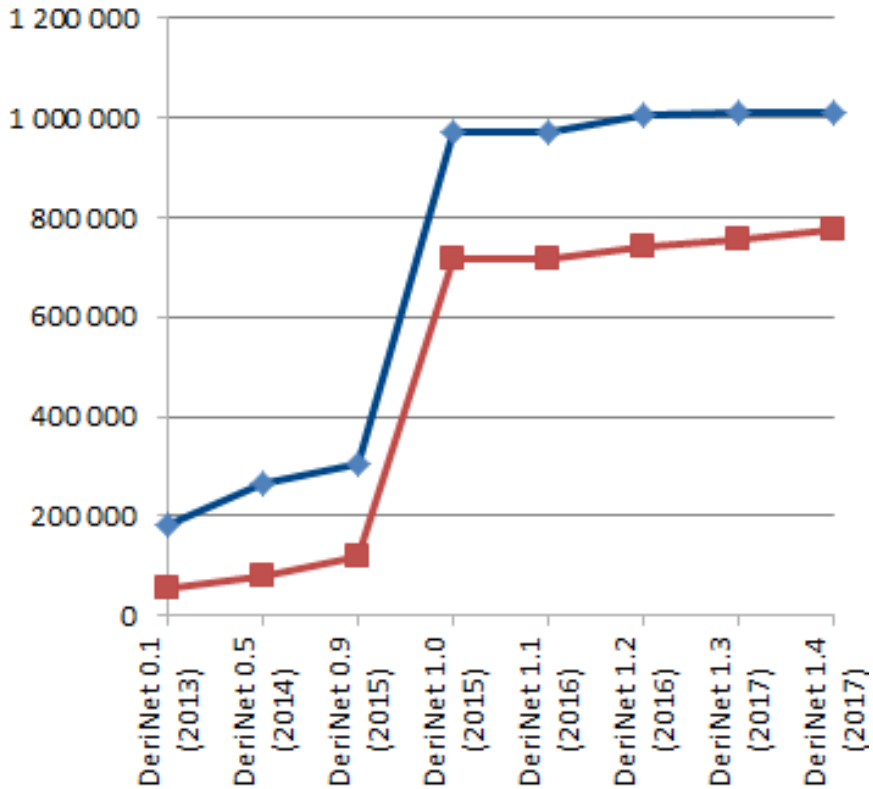


DeriNet Search (Jonáš Vidra)
http://ufal.mff.cuni.cz/derinet/search

DeriNet Viewer (Milan Straka)
http://ufal.mff.cuni.cz/derinet/viewer

Verze sítě DeriNet, grantová podpora

- 2012/13 založení sítě v rámci postdoc projektu o deadjektivní derivaci
  - 180K lexémů (z korpusu), 56k derivačních vztahů (DeriNet 0.1)
  - z toho 18k základových adj., 26k derivátů (AdjDeriNet, 2014)
- 2015 rozšíření sady lexémů, aby odpovídala slovníku MorfFlex CZ
  - 970K lexémů, 716K deriv. vztahů (DeriNet 1.0)
  - → 1.012K lexémů, 773K deriv. vztahů (DeriNet 1.4)



- grantová podpora
  - postdoc GAČR (2012–14) – LINDAT (2015) – GAČR (2016–18) – ?

Hláskové alternace

- „pravidelné střídání vybraných fonémů, popř. skupin fonémů, na němž je založena alomorfie“ (Osolsobě 2002)
  - *lét-o→let-ní*, *černo-och→černo-oš-ka*
  - vs. resufixace (*tanečn-ík→tanečn-ice*) a náhrada koncovky sufixem (*brank-a→brank-ář*)
- zdrojem systémové hláskové změny ve vývoji češtiny
- v současné češtině nesystémový až arbitrární charakter
- nedostatečný popis v odborné literatuře (Ziková 2015, 2016)
- zásadní problém pro poloautomatické metody hledání derivačních vztahů

Masivní přítomnost alternací v české slovní zásobě

- do alternací vstupují všechny vokály, kromě p, b, f, v, m a l všechny konsonanty
- cca 80 párů při zohlednění směru alternace: $a{>}á$ vs. $á{>}a$
- obousměrné vs. jednosměrné: $a{>}á$, $á{>}a$ vs. $h{>}z$
- poměr 1:více a více:1: $a{>}á/e/ě/o$ vs. $á/e/é/í{>}a$
- možné ve všech morfémech (prefix, kořen, sufix)

alter.	prefix	kořen	sufix
$a{>}á$	$na\text{-}lézt{\rightarrow}ná\text{-}lez$	$vrat\text{-}a{\rightarrow}vrát\text{-}ka/vrát\text{-}ný$	$stoj\text{-}an{\rightarrow}stoj\text{-}án\text{-}ek$
	vs. $na\text{-}bídnout{\rightarrow}na\text{-}bídka$	vs. $pat\text{-}a{\rightarrow}pat\text{-}ka$	
$a{>}e$		$úřad{\rightarrow}úřed\text{-}ník$	
$a{>}ě$		$šťast\text{-}ný{\rightarrow}štěst\text{-}í$	
$a{>}o$		$hrab\text{-}a\text{-}t{\rightarrow}hrob$	
$á{>}a$		$vrát\text{-}i\text{-}t{\rightarrow}vrat\text{-}ný$	

Masivní přítomnost alternací v české slovní zásobě

- do alternací vstupují všechny vokály, kromě p, b, f, v, m a l všechny konsonanty
- cca 80 párů při zohlednění směru alternace: $a{>}á$ vs. $á{>}a$
- obousměrné vs. jednosměrné: $a{>}á$, $á{>}a$ vs. $h{>}z$
- poměr 1:více a více:1: $a{>}á/e/ě/o$ vs. $á/e/é/í{>}a$
- možné ve všech morfémech (prefix, kořen, sufix)

alter.	prefix	kořen	sufix
$a{>}á$	$na\text{-}lézt{\rightarrow}ná\text{-}lez$	$vrat\text{-}a{\rightarrow}vrát\text{-}ka/vrát\text{-}ný$	$stoj\text{-}an{\rightarrow}stoj\text{-}án\text{-}ek$
	vs. $na\text{-}bídnout{\rightarrow}na\text{-}bídka$	vs. $pat\text{-}a{\rightarrow}pat\text{-}ka$	
$a{>}e$		$úřad{\rightarrow}úřed\text{-}ník$	
$a{>}ě$		$šťast\text{-}ný{\rightarrow}štěst\text{-}í$	
$a{>}o$		$hrab\text{-}a\text{-}t{\rightarrow}hrob$	
$á{>}a$		$vrát\text{-}i\text{-}t{\rightarrow}vrat\text{-}ný$	

Modelování alternací v síti DeriNet (i/ii)

1/ alternace (zvl. ve finální hlásce kořene/kmene) zakomponována do pravidel pro nahrazování koncových řetězců

- automaticky vyvozená, ručně sepsaná
  N>N-ka: *učitel→učitel-ka*
  N-c>N-čka: *herec→hereč-ka*
  N>A-ový: *achát→achát-ový*
  N-ec>A-cový: *konec→konc-ový*

2/ alternace povoleny plošně při aplikaci substitučních pravidel

- V-it>N-a: *otráv-i-t→otrav-a*, *osvít-i-t→osvět-a*
  N>A-ový: *cela→*cíl-ový*, *kuře→*kouřový*

3/ alternace součástí technických přípon lemmatu ve slovníku MorfFlex CZ

- polepšení_^(*3it): *polepš-i-t→polepš-e-ní*

Modelování alternací v síti DeriNet (ii/ii)

4/ experiment zaměřený na alternace v kořeni

- alternace při derivaci často totožné s alternacemi ve flektivním paradigmatu základového slovesa:
  - derivace: $bůh \rightarrow bož\text{-}í$
  - flexe: $bůh - boh\text{-}a_{gen.sg} - bož\text{-}e_{voc.sg} - boz\text{-}i_{nom.pl}$
- pro jednotlivé lexémy extrahovány z MorfFlexu řetězce s alternacemi, technická definice alternace (Milan Straka):
  - $bůh \quad boh \quad bož \quad boz$
- řetězce s alternacemi použity jako vstup pro substituční pravidla:
  - N>A-í: $čert \rightarrow čertí,\ sob \rightarrow sobí,\ pstruh \rightarrow pstruží$
  - $bůh\ /\ bož \rightarrow boží$

5/ alternace v rámci odvozování sloves od sloves řešeny zvlášť

Slovesný vid jako (lexikálně-)gramatická kategorie

- vid v teoretickém popisu češtiny
  - nejasný status: gramatická vs. lexikálně-gramatická kategorie
  - nejasná hranice mezi flektivním a derivačním paradigmatem slovesa
- vid vyjadřován derivačními afixy
  - sufixem: $dá\text{-}t_{pf} - dá\text{-}va\text{-}t_{impf}$
  - prefixem: $psát_{impf} - na\text{-}psat_{pf}$
- perfektivizační prefix u konkrétního slovesa
  - často jedním z mnoha prefixů kombinovatelných s daným slovesem:
    - $psát - napsat$ vs. $pře\text{-}psat$, $do\text{-}psat$, $vy\text{-}psat$, $nade\text{-}psat$, $pode\text{-}psat$, ...
  - u různých sloves plní tuto funkci různý prefix:
    - $psát - napsat$, $dělat - u\text{-}dělat$, $chválit - po\text{-}chválit$
  - identifikace čistě perfektivizačního prefixu:
    - $napsat - {}^{*}napisovat$ vs. $přepsat - přepisovat$

Slovesný vid a design derivačního stromu

- důsledky teoretického statutu pro zpracování v datech
  - vid lexikální kategorií: slovesa vidového páru ve vztahu derivace
  - vid gramatickou kategorií: vidový protějšek formou slovesa, do derivačního stromu nepatří
- v síti DeriNet:
  - vid vyjádřen derivačními afixy, vidové protějšky zachyceny jako deriváty
  - prefixální tvoření: nedokonavé sloveso→dokonavé sloveso
    $psát_{impf} \rightarrow na\text{-}psat_{pf}$
  - sufixální tvoření: podle délky sloves ve dvojici:
    - dokonavé sloveso→nedokonavé sloveso
      $dá\text{-}t_{pf} \rightarrow dá\text{-}va\text{-}t_{impf}$
      $koup\text{-}i\text{-}t_{pf} \rightarrow kup\text{-}ova\text{-}t_{impf}$
      $skoč\text{-}i\text{-}t_{pf} \rightarrow skák\text{-}a\text{-}t_{impf}$
    - nedokonavé sloveso→dokonavé sloveso
      $štěk\text{-}a\text{-}t_{impf} \rightarrow štěk\text{-}nou\text{-}t_{pf}$

Odvozování sloves od sloves

- deverbální derivace sloves specifickou subdoménou derivace
  - významná role prefixace (oproti tvoření jiných slovních druhů)
  - alternace v kořeni
  - afixy částečně vyjadřují kategorii vidu
  - uspořádání sloves do derivačního stromu nelze vyvodit z dat
  - několik ekvivalentních řešení

    $skoč\text{-}i\text{-}t_{pf} \rightarrow skák\text{-}a\text{-}t_{impf}$

    $skoč\text{-}i\text{-}t_{pf} \rightarrow vy\text{-}skoč\text{-}i\text{-}t_{pf}$

    $skák\text{-}a\text{-}t_{impf} \rightarrow vy\text{-}skák\text{-}a\text{-}t_{pf}$

    $skák\text{-}a\text{-}t_{impf} \rightarrow skák\text{-}áv\text{-}a\text{-}t_{pf}$
- řešeno jako samostatný problém (Adéla Kalužová)
  - vymezení slovotvorné čeledi daného slovesa
  - uspořádání členů slovotvorné čeledi do derivačního stromu

Odvozování sloves od sloves

- deverbální derivace sloves specifickou subdoménou derivace
  - významná role prefixace (oproti tvoření jiných slovních druhů)
  - alternace v kořeni
  - afixy částečně vyjadřují kategorii vidu
  - uspořádání sloves do derivačního stromu nelze vyvodit z dat
  - několik ekvivalentních řešení

    $skoč\text{-}i\text{-}t_{pf} \rightarrow skák\text{-}a\text{-}t_{impf}$

    $skoč\text{-}i\text{-}t_{pf} \rightarrow vy\text{-}skoč\text{-}i\text{-}t_{pf}$

    $skák\text{-}a\text{-}t_{impf} \rightarrow vy\text{-}skák\text{-}a\text{-}t_{pf}$

    $skák\text{-}a\text{-}t_{impf} \rightarrow skák\text{-}áv\text{-}a\text{-}t_{pf}$
- řešeno jako samostatný problém (Adéla Kalužová)
  - vymezení slovotvorné čeledi daného slovesa
  - uspořádání členů slovotvorné čeledi do derivačního stromu

Slovesa odvozená od sloves: vymezení slovotvorných čeledí

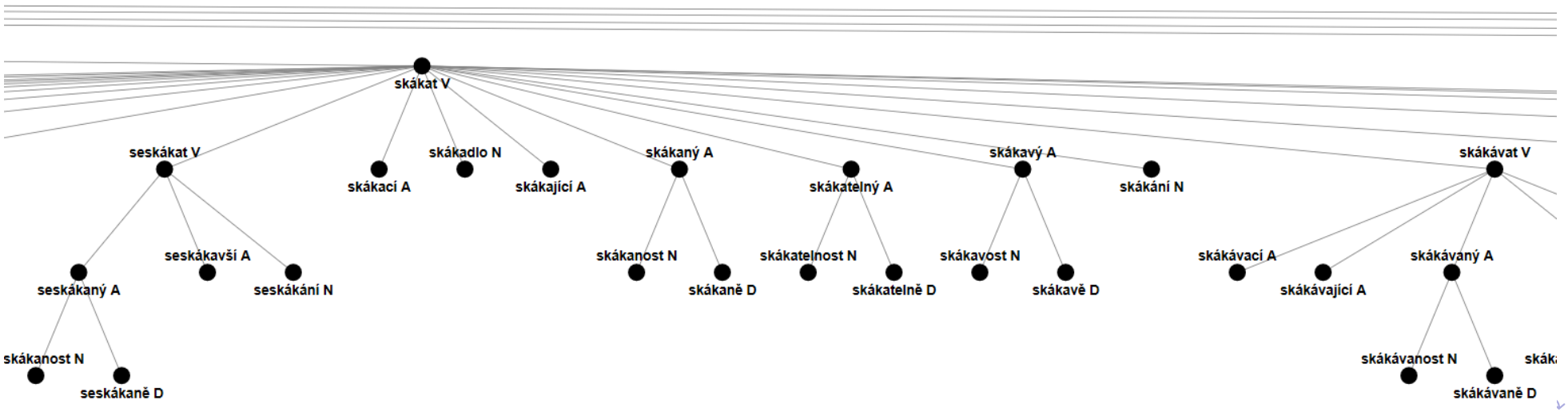
- extrakce vidových dvojic z Vallexu (Lopatková a kol. 2017), identifikace párů koncových řetězců
  - *skočit, skákat*
- seznam prefixů
- dohledání odvozených sloves ve Vallexu
  - *skočit, skákat, naskočit, naskakovat, odskočit, odskakovat, poskočit, poskakovat, přeskočit, přeskakovat, přeskákat, vyskočit, vyskakovat, zaskočit, zaskakovat*
- pak v celých datech DeriNetu
  - *skočit, skákat, skákávat, doskočit, doskákat, doskakovat, doskakovávat, naskočit, naskakovat, naskakovávat, naskákat, nadskočit, nadskakovat, nadskakovávat, obskočit, obskakovat, obskakovávat, obskákat, odskočit, odskakovat, odskakovávat, odskákat, poodskočit, podskočit, podskakovat, podskakovávat, poposkočit, poposkakovat, poposkakovávat, poskočit, poskakovat, poskakovávat, poskákat, povyskočit, povyskakovat, povyskakovávat, proskočit, proskakovat, proskakovávat, předskočit, předskakovat, předskakovávat, předskákat, přeskočit, přeskakovat, přeskakovávat, přeskákat, přiskočit, přiskakovat, přiskakovávat, přiskákat, rozskočit, rozskočovat, rozskakovat, rozskakovávat, rozskákat, rozeskákat, seskočit, seskakovat, sekakovávat, seskákat, uskákat, uskočit, uskakovat, uskakovávat, veskákat, vskákat, vskákávat, vskočit, vskakovat, vskakovávat, vyskočit, vyskakovat, vyskákat, zaposkakovat, zaposkakovávat, zaskočit, zaskakovat, zaskákat, zaskakovávat*

Organizace příbuzných sloves do derivačního stromu

- kořenem stromu bezpříznakový člen vidové dvojice
  - ve vidové dvojici tvořené sufixem: dokonavé nebo nedokonavé sloveso
    $skočit_{pf} \rightarrow skákat_{impf}$
    $štěkat_{impf} \rightarrow štěknout_{pf}$
  - ve vidové dvojici tvořené prefixem: nedokonavé sloveso
    $psát_{impf} \rightarrow napsat_{pf}$
- prefigovaná perfektiva deriváty neprefigovaného protějšku (pf nebo impf):
  $skočit \rightarrow naskočit/vyskočit/poskočit/\ldots$
  $skákat \rightarrow naskákat/vyskákat/poskákat/\ldots$
- sekundární imperfektiva deriváty prefigovaných perfektiv:
  $naskočit \rightarrow naskakovat$
- iterativa deriváty imperfektiv:
  $skákat \rightarrow skákávat$
  $zaskakovat \rightarrow zaskakovávat$

Derivace sloves: změny v datech

- 23K+ nových derivačních vztahů
- významné změny ve velikosti stromů
  - 80 stromů po více než 500 uzlech (*dát* 1288, *hodit*, *vést*, *jet*, ...)
  - nepoužitelná vizualizace v nástrojích DeriNet Search a Viewer
- evidentní ústřední funkce sloves v české slovní zásobě



Na závěr: Uzly bez rodičů v DeriNet 1.4

		z toho bez rodiče
lexémy celkem	1.011.965	238.602
- s počátečním velkým písmenem	210.354	124.566
- s počátečním malým písmenem	801.611	114.036

- malopísmenné lexémy bez rodičů
  - slova s více alternacemi:
    vejc-e→vaječ-ný
    sníh→sněž-ný
  - slova s jinými změnami:
    Prah-a→praž-ský (→mimo-praž-ský)
  - dějová substantiva odvozená od sloves:
    vysloužit/vysluhovat→výslužba
  - kompozita
  - ...

Na závěr: Uzly bez rodičů v DeriNet 1.4

		z toho bez rodiče
lexémy celkem	1.011.965	238.602
- s počátečním velkým písmenem	210.354	124.566
- s počátečním malým písmenem	801.611	114.036

- malopísmenné lexémy bez rodičů
  - slova s více alternacemi:
    vejc-e→vaječ-ný
    sníh→sněž-ný
  - slova s jinými změnami:
    Prah-a→praž-ský (→mimo-praž-ský)
  - dějová substantiva odvozená od sloves:
    vysloužit/vysluhovat→výslužba
  - kompozita
  - ...

Na závěr: další kroky

- úprava vizualizace
- nástroj pro ruční editaci
  - zpracování nespojených derivátů
- změna datové reprezentace
  - reprezentace kompozice
  - další informace k uzlům
  - ohodnocení hran (vč. vidu)
- kompozice (Adéla Kalužová), morfematická segmentace (Jonáš Vidra), sémantické značkování
- sítě derivačních vztahů pro další jazyky

`http://ufal.mff.cuni.cz/derinet`; repozitář Lindat/Clarin

Literatura

- Baayen, R. H. et al. (1995): *The CELEX lexical database* (release 2). Data/software. Philadelphia, PA: LDC.
- Baranes, M. – Sagot, B.: A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. *LREC* 2014:2793–2799.
- Bednaříková, B. (2009): *Slovo a jeho konverze*. Olomouc: UPOL.
- Cvrček, V. – Vondřička. P. (2013): Nástroj pro slovotvornou analýzu jazykového korpusu. In: *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.
- Hajič, J. – Hlaváčová, J. (2013): *MorfFlex CZ*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9
- Hathout, N. – Namer, F.: Démonette, a French Derivational Morpho-Semantic network. *LiLT* 2014:11, 125–168.
- Komárek, M. a kol. (1986): *Mluvnice češtiny 2. Tvarosloví.* Praha, Academia.
- Litta, E. et al. (2016): Formatio formosa est. Building a Word Formation Based Lexicon for Latin. In: *CLiC-it 2016*: 185–189.
- Lopatková, M. a kol. (2017): *Valenční slovník českých sloves Vallex*. Praha: Karolinum.
- Osolsobě, K. (2002): Alternace hlásková. In: P. Karlík et al. (eds.): *Encyklopedický slovník češtiny*. Praha: NLN, pp. 35–36.
- Osolsobě, K. et al. (2009): Exploring Derivational Relations in Czech with the Deriv Tool. In *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Bratislava: Tribun, pp. 152–161.
- Pala, K. – Hlaváčková, D. (2007): Derivational Relations in Czech WordNet. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*. Prague: ACL, pp. 75–81.
- Pala, K. – Šmerk, P. (2015): Derivancze – Derivational Analyzer of Czech. In: *TSD 2015*: 515–523.
- Šnajder, J. et al.: DerivBase.Hr: A High-Coverage Derivational Morphology Resource for Croatian. *LREC* 2014: 3371–3377.
- Zeller, B. et al.: DErivBase: Inducing and evaluating a derivational morphology resource for German. *ACL* 2013: 1201–1211.

Prezentovaný výzkum je prováděn za podpory projektu GAČR 16-18177S.