


DeriNet: Lexikální databáze českých derivátů

Magda Ševčíková, Zdeněk Žabokrtský
{sevcikova,zabokrtsky}@ufal.mff.cuni.cz

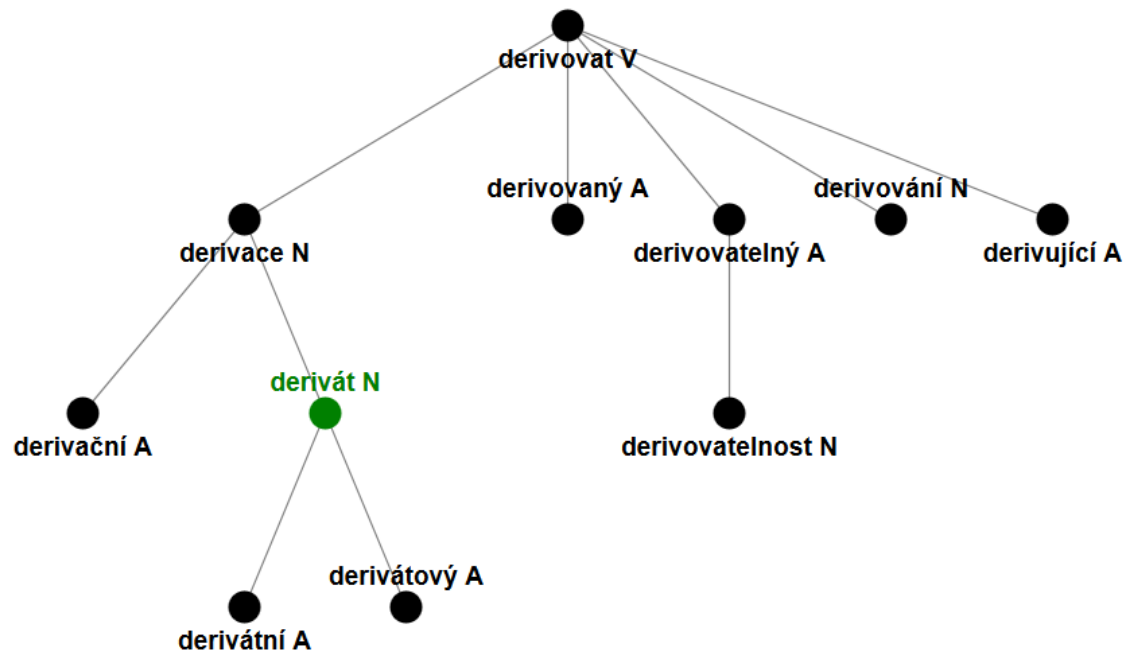
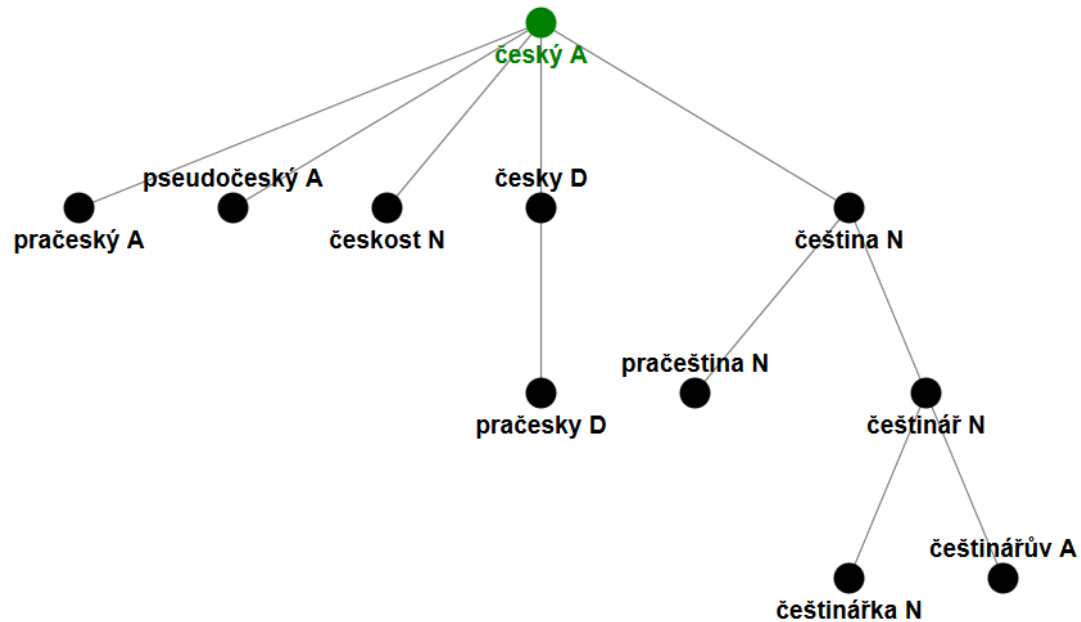
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Seminář formální lingvistiky
15. prosince 2014



- DeriNet
- Omezení na derivaci
 - dokulilovské pojetí derivace
- Automatická identifikace základových slov s manuální (lingvistickou) kontrolou

- Technické prostředky pro vývoj sítě DeriNet
- Pracovní postup
- Kvantitativní vlastnosti verze 0.9



Milan Straka: *DeriNet Viewer*

<http://ufallab.ms.mff.cuni.cz/~straka/derinet-viewer/>

DeriNet

- word-formation network for Czech (LREC 2014), database of Czech derived words (SLE 2014), lexical network of Czech derived words; databáze českých derivátů, lexikální síť českých derivátů
- databáze českých lexémů
 - lexémy extrahovány z korpusových dat
 - **mezi dvěma lexémy (uzly) vytvořen vztah (hrana), pokud se jedná o slovo základové a slovo od něj odvozené**
 - stromy odpovídají slovním čeledím / slovotvorným hnízdům

DeriNet: „historie“, verze

- vznik v souvislosti s projektem GAČR P406/12/P175
 - „*Vybrané derivační vztahy pro automatické zpracování češtiny*“
 - postdoktorský projekt, PI Magda Ševčíková, 2012–2014
- technická realizace: **Zdeněk Žabokrtský**
- DeriNet vyvíjen od podzimu 2012
 - DeriNet 0.5 (květen 2014)
 - DeriNet 0.9 (prosinec 2014)
 - DeriNet 1.0 (leden 2015)
- DeriNet 0.9
 - celkem 305 781 uzlů, 117 327 hran
 - z toho 98 683 adjektivních uzlů, 26 019 adjektivních derivátů

Pouze derivace

- derivace v češtině nejčastějším a nejvíce produktivním slovotvorným procesem
 - několik set slovotvorných formantů (převážně přípony)
 - např. názvy vlastností tvořeny až 14 příponami (-ost, -ství, -oba, -ota, -da, -ka, -ina, -í; -ita, -ismus, -ika, -ura, -ance/-ence, -ie)
- teoretický popis založený na Dokulilově onomasiologickém přístupu k slovotvorbě
 - M. Dokulil, *Tvoření slov v češtině 1: Teorie odvozování slov*, Academia 1962
 - Daneš et al. 1967, Šmilauer 1971, Hauser 1986, Dokulil et al. 1986, Grepl et al. 2000, Cvrček et al. 2010, Čermák 2012, Štícha et al. 2013
- zachycení derivačních vztahů v DeriNetu
 - v souladu s dokulilovskou teorií: cílem data podávající teoreticky správný obraz o české derivaci
 - se základovým slovem nejsou spojovány lexémy, které jsou produktem kombinovaných slovotvorných procesů (kompozice s derivací), ani kompozita
 - *mořeplavec, rychlovarný, velkovýroba*
 - hrany tvořeny poloautomaticky

Derivace v češtině

- velké množství formantů
- komplikovaný vztah formy a významu
 - mnoho formantů má několik významů
 - *-ka*: *učitel* > *učitelka*, *skříň* > *skříňka*
 - mnoho významů vyjádřeno hned několika formanty
 - „ženský protějšek“: *učitel* > *učitelka*, *soudce* > *soudkyně*
- komplikovaná kombinatorika
 - mnoho formantů se spojuje se základovými slovy z několika slovních druhů
 - *-ař*: *houba* > *houbař*, *tesat* > *tesář*
 - *-ina*: *ryba* > *rybina*, *pevný* > *pevnina*, *třetí* > *třetina*, *vidět* > *vidina*
- derivace doprovázena
 - hláskovými alternacemi, vkládáním/vypouštěním hlásek a/nebo změnou velkého počátečního písmena na malé
 - *stuha* > *stužka*, *list* > *lístek*, *žít* > *žití*, *léčba* > *léčebna*, *Vimperk* > *vimperský*
 - možnost „zakázat“ pouze na prvním písmeni (*hena* > *ženský*)

Slovotvorná fundace/motivace

- fundace/motivace jako formální/významové zakládání se jednoho slova na druhém
 - Dokulil (1962:11–14)
 - „Jedna forma se zakládá na druhé tehdy, (1) jestliže kmen této druhé formy je plně zahrnut v rozšířeném kmeni první formy – to je synchronní průmět genetického procesu odvozování (předponami nebo příponami); (2) jestliže kmen této formy obměňuje kmen druhé formy hláskově, a to ve smyslu směru takových zvukových alternací (např. *c*, *č* se může zakládat na *k*, *č*, kromě toho rovněž na *c*, nikoli však naopak) – to je synchronní průmět genetického procesu fonetického tvoření slov. Jsou-li kmeny (základy) jedné i druhé formy totožné, nelze zpravidla z formy samé rozhodnout, je-li jedna forma fundována druhou, popřípadě která forma je fundovaná (odvozená v nejširším smyslu) a která fundující (základová). Směr fundace (odvození) je pak dán pouze v rovině významu.“
 - „Jeden význam se zakládá na druhém, když je-li dán tento druhý význam, je tím určen i význam první, jinými slovy, je-li první význam odvoditelný.“
 - *Encyklopedický slovník češtiny* (2002:144)
 - „slovo, které je obvykle složitější významově i formálně, je tak fundováno slovem jednodušším“

Automatická identifikace základových slov s manuální (lingvistickou) kontrolou

- ve slovní zásobě spousta pravidelností
 - automaticky rozpoznatelné
 - pouze malá část z nich slovotvorně (fundačně/motivačně) relevantních
- derivační „pravidla“ vygenerovaná na základě porovnání dvojic lexémů v databázi
 - shoda v dostatečně dlouhé sekvenci písmen u dostatečně velkého množství dvojic
 - povoleny hláskové alternace
 - př.:
 - N- N-ová *Novák > Nováková*
 - A-ý N-ost *hladký > hladkost*
 - V- pře-V- *dělat > předělat*

Základová slova substantiv na *-ství/-tví*

- substantiva s příponou *-ství/-tví* popisována
 - jako deriváty adjektiv (*Mluvnice češtiny 1 1986, Příruční mluvnice češtiny 2000*)
 - jako deriváty substantiv (Štícha et al. 2013)
- DeriNet 0.5
 - 1262 substantiv s příponou *-ství/-tví*
 - jen pro 152 (12 %) z nich automaticky identifikován jediný kandidát na fundující slovo
 - adjektivum na *-ský/-ký*
 - v databázi pro 1139 z 1262 derivátů (90,3 %; pravidlo A-ký > N-tví)
 - substantivum (nejčastěji masc.anim.) označující osobu
 - 721 (57,1 %) N- > N-ství: př. *amatérství, barbarství*
 - 245 (19,4 %) N-k > N-ctví : př. *dobráctví, estébáctví*
 - 187 (14,8 %) N-ík > N-ictví: př. *básnictví, dobrovolnictví*
 - 87 (6,9 %) N- N-tví: př. *sobectví, bezdomovectví*
 - popř. další: 104 (8,2 %) N- N-nictví: *čaroděj > čarodějnictví* (vs. N-ík > N-ictví: *čarodějník > čarodějnictví*, N-e > N-tví: *čarodějnice > čarodějnictví*)

Základová slova substantiv na *-ství/-tví* (ii)

- celkem 40 pravidel vedoucích k substantivu na *-ství/-tví*
 - adjektivum na *-ský/-ký* a název osoby
 - 3. a další fundující slovo pro jednotlivé deriváty různé
 - deminutivum
 - femininum
 - sloveso
 - ...

– *sobectví* {
– *sobecký*
– *sobec*
– *sobeček*

– *státnictví* {
– *státnický*
– *státník*
– *státní*
– *statný*
– *stát*

– *ředitelství* {
– *ředitelský*
– *ředitel*
– *ředitelka*
– *ředitelovat*
– *ředitelný*

– *služebnictví* {
– *služebnický*
– *služebník*
– *služebný*
– *služební*
– *služebníček*
– *služebna*
– *služebka*

Základová slova lexémů s cizím formantem, s prefixy ad.









- formanty cizího původu:

- impresionismus  impresionista
 impresionistický

- fotbalista  fotbalistický

- negace, další prefixace:

- mačkový  mačkovost
nemačkový   nemačkovost 

- biologie   biologický   biologicky
mikrobiologie   mikrobiologický   mikrobiologicky 