Osolsobě, Klára

**Závěr**

In: Osolsobě, Klára. *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny.* Vyd. 1. Brno: Masarykova univerzita, 2011, pp. 190-192

ISBN 9788021055650

# X. ZÁVĚR

V naší práci, která nese název *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny*, jsme se snažili ukázat meze a možnosti automatického zpracování české morfologie, konkrétně tvoření slov odvozováním – sufixací vč. morfologických alternací, které derivaci doprovázejí a podílejí se na ní jako spoluformanty.

Úvodní kapitoly shrnují výsledky práce na poli automatické morfologické analýzy češtiny, a to jak na rovině teoretické, tak praktické. Algoritmický popis české morfologie a strojový slovník češtiny (Osolsobě 1996) se stal lingvistickou bází nejrůznějších nástrojů v oblasti strojového zpracování přirozeného jazyka (NLP). Zkušenosti s využitím aplikací založených na uvedeném popisu v oblasti korpusové lingvistiky, zejména automatické lemmatizace a automatického morfologického značkování korpusů, ukázaly na některé nedořešené otázky a problémy a otevřely tak cestu dalšímu bádání na poli strojového zpracování přirozeného jazyka. Přes rostoucí technické možnosti se stále ukazuje, že se ani zdaleka nepodařilo odhalit všechny taje a zákoutí přirozeného jazyka a popsat je natolik exaktně, aby se, metaforicky řečeno, stroje nedopouštěly směšných chyb. Mnohdy se ale ukazuje, že naopak v jazyce existují zákonitosti a pravidla, která lze formalizovat tak, že přinášejí nové pohledy na jazyk jako systém i na jeho fungování.

Za zásadní přínos práce pokládáme to, že jsme navrhli a otestovali formální popis tvoření slov derivací spoluutvářenou řadou morfologických alternací na různých úrovních. Vyšli jsme z vlastního systematického popisu pravidel alomorfie ve slovesném tvarosloví založeného na analýze dat rozsáhlého strojového slovníku češtiny a korpusů. Na základě uvedených pravidel jsme dále sledovali uplatnění alomorfů slovesných tvarů při derivaci, popřípadě další případy alomorfie v deverbativech. Tato pravidla jsme zahrnuli do formálního popisu tvoření tvarů sloves a tvoření vybraných typů deverbativ.

Ve spolupráci lingvistů a informatiků vznikl nástroj *Deriv – webové rozhraní*, který umožňuje komunikovat s a) strojovým morfologickým slovníkem češtiny, b) elektronickými verzemi tištěných slovníků a c) jazykovými korpusy. Tento nástroj slouží k testovaní formálních substitučních pravidel popisujících derivační vztahy v češtině.

Jádro práce tvoří komentované přehledy formálních substitučních pravidel derivace vybraných typů deverbativ (substantiv a adjektiv tvořených od sloves). Formální popis vychází z klasických tradičních popisů slovotvorby v té míře, že sleduje tradiční slovotvorné typy a třídy, přičemž si všímá a komentuje některá sporná místa v dřívějších popisech. Jedná se především o zpřesnění popisu de-

rivace od sloves patřících do uzavřených slovesných tříd/vzorů. Především jde o přesnější určení mezí a možností derivace některých typů od sloves 1. třídy vzorů *nést, péct,* 2. třídy a 3. třídy vzoru *krýt*. Důležité je tedy zahrnutí okrajových případů. V korpusech je doložena řada případů okazionálně utvořených slov, jejichž podrobnější analýza pomáhá doplnit/opravit obraz tradičních popisů slovotvorby. Za novum pokládáme především podrobné systematické popisy nejrůznějších morfologických alternací, které jsou spoluformanty zkoumaných derivačních typů a jejich zahrnutí do substitučních pravidel. Navržená substituční pravidla jsou testována na rozsáhlých datech strojového morfologického slovníku češtiny a výsledky jsou porovnávány se slovníky a korpusy.

Přehledy substitučních pravidel popisují jistý výsek tvoření slov odvozováním. Volba jednotlivých derivačních typů byla promyšlená a směřovala k tomu, aby byly zastoupeny typy s větší i menší produktivitou i frekvencí, ale především s formálně odlišnými základy. Sloveso jako jediný slovní druh má v češtině velmi dobře zachovanou strukturu tradičně popisovanou jako kmen (základ/kořen + kmenotvorná přípona), k němuž se pak připojuje buď osobní, nebo tvarová koncovka. Slovesné tvarosloví zahrnuje několik tvarových subsystémů (subparadigmat). Toto bohatství na úrovni tvarosloví pokračuje i při odvozování nových slov od slovesa (deverbativ). Základem pro odvozování se může stát a) slovesný tvar (jeho část), b) slovesný kmen (stem derivation) a c) slovesný základ/kořen (root derivation). Při formálním popisu je možné se opřít o řadu vlastností, které mají materiální realizaci a jsou tudíž dobře zachytitelné na úrovni formálního popisu. Formální popis lze automaticky testovat pomocí nástrojů NLP (webové rozhraní *Deriv*).

Formální popis i jeho následné testování ovšem naráží na obecný problém přirozeného jazyka, kterým je mnohoznačnost na všech úrovních (homonymie). Při testování substitučních pravidel jsme se setkávali zejména s homonymií způsobenou překryvy vycházejícími z různých možností formální segmentace analyzovaných jednotek. Teoreticky i prakticky jsou otázky morfémové segmentace češtiny vyřešeny v dosud nepřekonaném díle E. Slavíčkové (Slavíčková 1974). Automatizace procesu segmentace má ovšem meze dané právě výše zmíněnou homonymií. Druhým typem homonymie, která s prvním typem souvisí, je homonymie na úrovni jednotlivých typů morfů (lexikálních i derivačních). Konkrétní příklady jsou bohatě dokumentovány v poznámkách k výsledkům získaným testováním jednotlivých substitučních pravidel. Tyto poznámky nechtějí být ilustrací známého problému NLP, jímž je přegenerovávání, to by bylo příliš málo. Shromážděný materiál totiž ukazuje na možné cesty řešení. V řadě případů je přegenerovávání závislé na velké míře homonymie (nejrůzněji pojaté) jednotlivých lexémů (srv. počet přegenerovaných dvojic, do nichž vstupují např. slovesa *jíst, jít, pít, dít, …* vč. jednotlivých prefigovaných variant), jindy je naopak podmíněno typově (srv. např.

přegenerovávání v případě dvojic příbuzných sloves, a to zejména sloves uvnitř 4. třídy slovesné). Příklady chyb tak přispívají k poznání slabých míst, která se mnohdy opakují. Analýza chyb vede a) k optimalizaci formálního popisu a b) stává se bází pro vypracování optimální strategie tam, kde formální popis naráží na své meze a kdy je třeba zajistit konzistentní řešení zpracování jiného (např. ručního).

Naše práce se snažila formou několika jednoduchých statistik hodnotit možnosti a meze formálního popisu tvoření slov. S ohledem na výše jmenované typy derivací (derivace od tvaru, kmene a základu) jsme zkoumali závislost přegenerovávání na bohatství struktury odvozovacích prostředků, na jejich homonymii a na zahrnutí morfologických alternací jakožto spoluformantů derivace.

Formální pravidla publikovaná a otestovaná v naší práci jsou formulována jako substituční pravidla. Zvolený formát se příliš neliší od jazyka, jímž lze zadávat dotazy v korpusových manažerech. Pro čtenáře, kteří jsou zvyklí na práci s jazykovými korpusy, by neměl být problém využít nabídnutý popis slovotvorného systému v oblasti korpusové lingvistiky orientované k analýze slovotvorby. Korpusy jsou často značkovány na úrovni lemmat a gramatických kategorií včetně některých kategorií slovotvorných. Značkování korpusů nemá a nemůže být poslední instancí, je pouze praktickou pomůckou při práci s rozsáhlými jazykovými daty. Řada zákonitostí jazyka i výjimek z nejrůznějších pravidel na úrovni tvarosloví i slovotvorby, které jsou formulovány v kapitolách 5 a 7, může být podnětná pro práci s jazykovými korpusy tehdy, když z jakýchkoliv důvodů není žádoucí či možné anotace na úrovni morfologického značkování, či lemmatizace používat.

Nedílnou součástí naší práce je derivační slovník deverbativ uvedených typů. Ten je popsán v 9. kapitole. Fyzicky jsou data uložena na serveru Fakulty informatiky Masarykovy univerzity http://deb.fi.muni.cz/deriv/ v příslušných podadresářích (srv. výše Kapitola 9). Tento slovník zahrnuje automaticky generovaná a ručně zpracovaná data ve formě dvojic fundující sloveso/deverbativum (více než 96 000 dvojic). Ke každému členu dvojice lze zobrazit frekvenci uvedené jednotky v korpusu a zároveň je možné interaktivně přepnout na definici v různých slovnících přístupných přes webový prohlížeč slovníků DebDict. Součástí prohlížeče je i přístup k morfologickému analyzátoru *ajka.* Slovník představuje jádro derivačního slovníku češtiny.