Osolsobě, Klára

Automatická morfologická analýza a strojový slovník češtiny

In: Osolsobě, Klára. *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny.* Vyd. 1. Brno: Masarykova univerzita, 2011, pp. 13-14

ISBN 9788021055650

II.

Automatická morfologická analýza a strojový slovník češtiny

Tvarotvorná analýza češtiny je v současnosti již poměrně dobře formálně popsána (Hajič 1994; Osolsobě 1996). Popis české formální morfologie aplikovaný a testovaný na rozsáhlém strojovém slovníku češtiny (Osolsobě 1996) otevřel cestu k dalšímu bádání i dalším aplikacím. Jednou z nich je i výzkum mezí a možností automatického zpracování oblasti, která je tradičně nejblíže formální morfologii (tvarosloví), totiž slovotvorbě. Strojový slovník (Osolsobě 1996) je morfologický slovník obsahující 170 000 kmenů, z nichž každý kmen má přiřazeno pravidlo (morfologický vzor), pomocí kterého se generují uspořádané trojice: základní tvar (lemma) – generovaný tvar (slovní tvar) – slovní druh a další slovnědruhově závislé interpretace (morfologická značka/tag).

Tento slovník se stal lingvistickou bází automatického morfologického analyzátoru *ajka*[2] (Sedláček 2004) a prostřednictvím tohoto softwarového nástroje je možné s ním dále pracovat[3].

Morfologický slovník zahrnoval v rámci definic pravidel tvoření tvarů slov podle jednotlivých tvarotvorných vzorů i definice některých pravidelných derivací. Tak například součástí definice tvarotvorných vzorů substantiv pojmenovávajících osoby byla pravidla odvozování adjektiv na *-ův* (maskulina životná) a na *-in* (feminina označující převážně živé osoby). Součástí definic tvarotvorných vzorů adjektiv bylo propojení s derivačními vzory definujícími derivace a) tvarů komparativu a superlativu, b) adverbií paradigmaticky tvořených od adjektiv, c) tvoření tvarů komparativu a superlativu příslušných adverbií, d) komplexní vzory pro tvoření tvarů číslovek určitých i derivaci jednotlivých druhů číslovek[4]. Dobře patrná byla komplexnost tvarotvorných a slovotvorných pravidel (vzorů) na definicích vzorů sloves. Na základě pravidel přiřazených jednotlivým kmenům se generovaly jak jednoduché tvary určité (tvary indikativu prézentu/futura aktiva a imperativu), tak neurčité (tvary participia l-ového, participia pasivního, přechodníků přítomného i minulého a infinitivu). Obdobný přístup byl aplikován i ve slovníku používaném pro značkování korpusů Českého národního korpusu (Hajič 2004)[5].

2 Analyzátor *ajka* je přístupný přes DebDict – webový prohlížeč slovníků.

3 Použitý systém značek (tagset) viz příloha A.

4 Srv. více Osolsobě 1995.

5 Srv. popis morfologických značek – poziční systém Jana Hajiče na http://ucnk.ff.cuni.cz/bonito/znacky.php. Informace zachycené na druhé pozici značky (detailní určení slovního druhu) jsou v řadě případů informace týkající se tvoření slov odvozováním tradičně v gramatikách řazených do popisu slovotvorby (adjektiva posesivní, adjektiva tvořená od přechodníků, některé druhy zájmen a číslovek,

Cílem naší práce je prozkoumat meze a možnosti automatické analýzy některých pravidelných typů derivací v češtině.

Popisy tvoření slov v češtině obsažené v moderních českých gramatikách (zejména v Mluvnici češtiny 2) se vesměs opírají o teoretická východiska shrnutá v Dokulilově koncepci. Jsou zaměřeny na klasifikaci slov motivovaných z hlediska významových změn realizovaných v procesu tvoření slov odvozených od slov základových (mutace, modifikace, transpozice) a dále třídění utvořených slov na základě jejich obecného významu do slovotvorných tříd a na základě formálních prostředků do slovotvorných typů.

Jádrem popisů je slovní charakteristika typu (obecný význam a formant) opřená o ilustrativní příklady centrálních jevů a výjimek. Na rozsahu práce pak závisí úplnost popisu. Mnohdy zůstávají opomenuty jevy okrajové, jindy je jejich zachycení v rámci jednoho popisu nejednotné (frekventované okrajové jevy zachyceny jsou, nefrekventované nikoli).

Utvářenost slovní zásoby češtiny ve slovníkových pracích (Slavíčková 1974, Šiška 1998) zachycuje teoreticky zdůvodněnou morfematickou segmentaci slova, nikoli interpretaci segmentů i celku. Navíc korpusy, z nichž obě uvedená díla vycházejí, jsou nesrovnatelně menší než ty, které jsou v současné době k dispozici.

Naším cílem je formální popis (otevřený), s jehož pomocí lze testovat pokrytí slovotvorných vztahů (formálních i významových) na masových datech.

Hlavním cílem práce je návrh jisté metodologie zpracování slovní zásoby z hlediska utvářenosti slovních jednotek. Formální popis vybraného úseku slovotvorby testovaný na masových datech slouží k ověření teoretických předpokladů týkajících se vztahů formy a významu slov základových a odvozených v měřítku překračujícím možnosti starších popisů. Metoda formálního popisu slovotvorných vztahů je obecná, lze ji tudíž aplikovat na další (v práci nezahrnuté) slovotvorné třídy a typy.

Formální popis představený v naší práci se tak může stát východiskem pro různé formy automatického zpracování přirozeného jazyka (NLP).

klasifikace adverbií dle +/- stupňovatelnosti atd.). Zařazení stupňovaní (i stupňovatelnosti) adjektiv i adverbií mezi informace zprostředkované morfologickou značkou (pozice 10 – stupeň) svědčí o tomtéž (srv. k tomuto tématu Osolsobě 2008[1]).