

## II.

### Automatická morfoloická 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 mezi a možnostmi automatického zpracování oblasti, která je tradičně nejbliže formální morfologii (tvarosloví), totiž slovtvorbě. Strojový slovník (Osolsobě 1996) je morfoloický slovník obsahující 170 000 kmenů, z nichž každý kmen má přiřazeno pravidlo (morfoloický vzor), pomocí kterého se generují uspořádané trojice: základní tvar (lemma) – generovaný tvar (slovní tvar) – slovní druh a další slovnědruhově závislé interpretace (morfoloická značka/tag).

Tento slovník se stal lingvistickou bází automatického morfoloického analyzátoru *ajka*<sup>2</sup> (Sedláček 2004) a prostřednictvím tohoto softwarového nástroje je možné s ním dále pracovat<sup>3</sup>.

Morfoloický 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) adverbii paradigmaticky tvořených od adjektiv, c) tvoření tvarů komparativu a superlativu příslušných adverbii, d) komplexní vzory pro tvoření tvarů číslovek určitých i derivaci jednotlivých druhů číslovek<sup>4</sup>. Dobře patrná byla komplexnost tvarotvorných a slovtvorných pravidel (vzorů) na definicích vzorů sloves. Na základě pravidel přiřazených jednotlivým kmenům se generovaly jak jednoduché tvary určité (tvary indikativu přítomného/futura aktiva a imperativu), tak neurčité (tvary participia I-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)<sup>5</sup>.

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 morfoloických 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 slovtvorby (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 vytvořených slov na základě jejich obecného významu do slootovorných tříd a na základě formálních prostředků do slootovorných typů.

Jádrem popisu 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).

Utváření slovní zásoby češtiny ve slovníkových pracích (Slavičková 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 slootovorných 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 slootovorby 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 slootovorných vztahů je obecná, lze ji tudíž aplikovat na další (v práci nezahrnuté) slootovorné třídy a typy.

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

klasifikace adverbíí dle +/- stupňovatelnosti atd.). Zařazení stupňování (i stupňovatelnosti) adjektiv i adverbíí mezi informace zprostředkované morfologickou značkou (pozice 10 – stupeň) svědčí o tom-též (srv. k tomuto tématu Osolsobě 2008<sup>1</sup>).