I. ÚVOD

Automatická analýza přirozeného jazyka na ÚJČ FF MU

Cílem tohoto textu je podat přehled o výsledcích, k nimž jsme dospěli v oboru strojového zpracování přirozeného jazyka, zejména v oblasti propojení formální morfologie a tvoření slov. Jádrem práce jsou formální popisy vybrané oblasti české slovotvorby uvedené ve druhé části naší práce (Kap. 7). Tyto mohou posloužit k testování pravidelných slovotvorných procesů, a to jednak na úrovni elektronických morfologických databází, jednak na úrovni elektronicky přístupných korpusů.

Práce navazuje na dlouholetý výzkum na poli automatického zpracování přirozeného jazyka, zvláště pak morfologie. Ten započal na půdě Kabinetu počítačové lingvistiky Ústavu českého jazyka Masarykovy univerzity (dříve Univerzity Jana Evangelisty Purkyně – UJEP) koncem 80. let minulého století. Výsledky jedné jeho etapy jsou shrnuty v disertační práci ***Algoritmický popis české formální morfologie a strojový slovník češtiny*** (Osolsobě 1996). Od poloviny 90. let se dále rozvíjel především v rámci širší spolupráce akademických pracovišť účastnících se řešení grantových projektů směřujících k budování jazykových korpusů a korpusových nástrojů[1].

Strojový slovník češtiny (Osolsobě 1996) se stal lingvistickou bází některých aplikací v oblasti strojového zpracování přirozeného jazyka (NLP) realizovaných v rámci Laboratoře zpracování přirozeného jazyka Fakulty informatiky Masarykovy univerzity (LZPJ FI MU) a v současné době Centra zpracování přirozeného jazyka tamtéž. Nejvýznamnější z nich je automatický morfologický analyzátor *ajka* (Sedláček 2004), používaný mimo jiné k anotacím korpusů budovaných na Fakultě informatiky a na Filozofické fakultě Masarykovy univerzity. Tento analyzátor je součástí dalších aplikací, v nichž slouží jako modul zajišťující automatickou morfologickou analýzu. Návrh na nový formát dat podal v disertační práci Pavel Šmerk (Šmerk 2010).

Výsledky výzkumu na poli kvantitativních charakteristik češtiny založené na frekvenční analýze morfologických typů a podtypů definovaných pro potřebu automatické morfologické analýzy v citované v disertační práce přinášejí publikace (Osolsosobě – Pala – Rychlý 1998[1, 2]).

1 Jednalo se o tyto grantové projekty: 1. GAČR č. 405/93/0218 „Počítačový korpus českých psaných textů“ (úspěšně ukončen v r. 1995); 2. GAČR č. 405/96/K214 „Textové korpusy a lexikální i gramatická základna pro rozvoj češtiny v 21. století“ (úspěšně ukončen v r. 2001) 3. GAČR č. 405/03/0248 „Současná soukromá korespondence. Vytvoření databáze a zpracování vybraných jevů z pohledu lexikologicko-lexikografického a dialektologického“ (úspěšně ukončen v r. 2005).

V souvislosti s budováním speciálních korpusů na Filozofické fakultě MU – (Brněnský mluvený korpus (bmk) a Korpus soukromé korespondence (ksk) v rámci projektu Českého národního korpusu (srv. více Hladká 2005) byl algoritmický popis morfologie i strojový slovník rozšířen a modifikován o některé substandardní jevy (Hlaváčková 2001). Na problematiku využití automatických nástrojů pro různé „standardy" přirozeného jazyka (spisovný jazyk tištěných textů, přepis mluveného jazyka, psaný jazyk neformálních nekorigovaných projevů) se soustřeďují studie (Osolsobě 2001, Osolsobě 2005, Osolsobě 2006, Hlaváčková – Osolsobě 2008).

K tématu teorie morfologického značkování se vrací studie porovnávající systémy morfologických značek (tagsety) používané ke značkování v českém/slovenském prostředí (Osolsobě 2007[1], 2008[1]).

Průběžné sledování mezí a možností značkování jazykových korpusů z hlediska zachycení morfologických vlastností nezachycených explicitně v systému značek glosuje řada studií (Osolsobě 1999, 2002, 2007[3], 2008[1], 2009[1, 2, 4]).

Automatické slovotvorné analýze češtiny je věnováno několik studií. Ty lze rozdělit do dvou skupin. První zahrnuje studie referující o aplikacích v oblasti strojového zpracování přirozeného jazyka (češtiny) vzniklých ve spolupráci lingvistů a informatiků (Osolsobě – Pala – Sedláček – Veber 2002, Hlaváčková – Osolsobě – Pala – Šmerk 2009[1, 2]). Druhá sleduje lingvistické problémy počítačového zpracování přirozeného jazyka, konkrétně slovotvorby (Osolsobě 2008[2, 3, 4], 2009[1, 2, 3, 4]).

Do širšího kontextu matematické lingvistiky v bohemistice jsou výzkumy na poli automatického zpracování češtiny, na nichž jsme se podíleli, zařazeny v kapitole *Matematická lingvistika* uveřejněné v monografii *Kapitoly z dějin české jazykovědné bohemistiky* (Pleskalová – Krčmová – Večerka – Karlík 2007 : 447n).