# III.

## Pravidelnost derivace a strojové zpracování přirozeného jazyka

Cílem naší práce je formálně popsat realizované[6] případy derivace vybraných typů českých deverbativ a otestovat tak meze a možnosti automatizace analýzy přirozeného jazyka na úrovni tvoření slov.

Odvozování slov (derivace) hraje v obohacování slovní zásoby češtiny významnou roli. Rodilý mluvčí je od útlého věku schopen využívat existující modely tvoření slov tak, že umí podle těchto modelů jednak tvořit nová slova, jednak dovozovat významy slov, se kterými se setkává poprvé. Ti, pro které čeština není mateřským jazykem, tuto schopnost získávají postupně s prohlubováním jazykových znalostí. Problémem je, že malé děti, kreativní jedinci, nebo lidé bez dostatečné znalosti češtiny užívají jednotky utvořené podle modelů pro pravidelné derivace i tam, kde se v jazyce běžně užívají jednotky jiné. Zkrátka řečeno, „slovotvorný stroj" má své meze.

Na tyto meze narážejí rovněž pokusy automatické analýzy/syntézy v oblasti strojového zpracování přirozeného jazyka (NLP), kde bývají takové případy označovány termínem přegenerovávání.

Vztahy mezi slovem základovým a slovem odvozeným, u nichž klasická lingvistika rozlišuje dva aspekty, vztah na úrovni formy a významu (fundaci – základové slovo je součástí slova odvozeného a motivaci – význam slova odvozeného lze odvodit na základě významu slova základového), jsou popsány v českých klasických gramatikách v oddílech věnovaných tvoření slov. Tyto popisy byly vodítkem pro formulování specifického popisu předloženého v této práci. Ukázalo se, že popisy uváděné v klasických mluvnicích, jsou pro potřeby formálního popisu v mnoha aspektech neúplné. Z tohoto důvodu bylo nejdříve třeba vypracovat metodu postupu pro formalní popis pravidel tvoření slov derivací v češtině.

Východiskem formálního popisu jsou vzájemné vztahy (formy a významu) slova základového a slova odvozeného. Na rovině formy vycházíme z grafické

6 Adjektivum realizovaný chápeme tak, že testy formálních pravidel jsou prováděny na materiálu slov uložených ve strojovém slovníku češtiny. Není pochyb o tom, že tento slovník nezahrnuje všechny přípustné derivace. Sondy do jazykových korpusů, ale i znalost jazyka (češtiny) rodilých mluvčí je toho důkazem. Naopak strojový slovník, s nímž pracujeme, má mnoho nevýhod. Za hlavní pokládáme tu, že za jeho základ byl použit heslář Slovníku spisovného jazyka českého (SSJČ), který v mnoha ohledech neodpovídá synchronnímu stavu jazyka (viz též následující poznámka). Přes tato omezení je metoda formálního popisu otevřená, takže je možné ji v případě potřeby modifikovat.

podoby slova/slovního tvaru. Jak slovo základové, tak slovo odvozené lze chápat jako řetězec grafémů (písmen). Na rovině významu vycházíme z obecného významu slovního druhu a dalších obecných významů, které jsou zachyceny v interpretaci (morfologické značce) každého tvaru v morfologickém slovníku.

Změny, k nimž dochází při derivaci na úrovni formy, lze popsat jako systém záměn částí řetězce (základového slova) takových, aby jejich výsledkem byl nový řetězec (slovo odvozené). Změny, k nimž dochází na úrovni významu, lze popsat jako podmínky doprovázející změny na úrovni formy.

Slovotvorné vztahy jsou zachyceny formálně v podobě nahrazovacích (substitučních) pravidel zahrnujících popis substitucí na úrovni formy za určitých podmínek. Pravidla zachycují slovotvorné procesy jakožto operace nad řetězci grafických znaků/písmen (slovních tvarů uložených ve strojovém slovníku morfologického analyzátoru *ajka*), jejichž podmínkou jsou definovatelné vlastnosti zadaných řetězců (gramatické informace obsažené v gramatických značkách). Na základě lingvisticky stanovené hypotézy, tedy souboru pravidel záměn (substitucí), k nimž dochází za definovaných podmínek, lze z morfologického strojového slovníku automaticky extrahovat n-tice jednotek, které a) jsou ve slovníku zachyceny[7] a b) splňují danou hypotézu.

7 Morfologický slovník analyzátoru *ajka* zahrnuje přibližně 400 000 lemmat, z nichž lze na základě morfologických vzoru generovat 60 000 000 slovních tvarů. Morfologický slovník je jednou z aplikací strojového slovníku kmenů (Osolsobě 1996). Tento slovník byl budován od konce 80. let 20. století na Katedře českého jazyka FF UJEP, později FF MU (více Pala 1992). Jádrem slovníku byl heslář Slovníku spisovného jazyka českého, k němuž byla připojena některá další lemmata získaná během první poloviny 90. let z korpusů budovaných v rámci grantových projektů podporujících vznik Českého národního korpusu (ČNK). Systém pravidel definujících morfologické vzory je podrobně popsán v disertační práci (Osolsobě 1996). Dnešní podoba morfologického slovníku prošla řadou úprav a kontrol (Bartůšková – Hlaváčková – Ungermannová 2004), stále ovšem nese původní rysy hesláře SSJČ (mnohá lemmata jsou z dnešního pohledu zastaralá).