



MORFOLOGIE ČESKÉHO SLOVESA A TVOŘENÍ DEVERBATIV JAKO PROBLÉM STROJOVÉ ANALÝZY ČEŠTINY

Klára Osolobě

t\$/k5e 1 mlčet (611), mlčelivý (4), mlčelivost (1)

t\$/k5e 2 snášet (573), snášelivý (0), snášelivost (1)

t\$/k5e 3 trpět (657), trpělivý (144), trpělivost (1147)

t\$/k5e 4 zdržet (250), zdrželivý (4), zdrželivost (2)

*t\$/k5eA.*mF>livý/k2eAgMnSc1d1>livost/k1gFnSc1*

*t\$/k5eA.*mF>livý/k2eAgMnSc1d1>livost/k1gFnSc1*

*t\$/k5eA.*mF>livý/k2eAgMnSc1d1>livost/k1gFnSc1*

OPERA UNIVERSITATIS MASARYKIANAE BRUNENSIS
FACULTAS PHILOSOPHICA

SPISY MASARYKOVY UNIVERZITY V BRNĚ
FILOZOFICKÁ FAKULTA

Číslo 401

muni
PRESS

MORFOLOGIE ČESKÉHO SLOVESA
A TVOŘENÍ DEVERBATIV
JAKO PROBLÉM STROJOVÉ ANALÝZY
ČEŠTINY

Klára Osolsobě



Masarykova univerzita
Brno 2011

© 2011 Klára Osolsobě

© 2011 Masarykova univerzita

ISBN 978-80-210-8213-7 (online : pdf)

ISBN 978-80-210-5565-0 (brožovaná vazba)

ISSN 1211-3034

Poděkování

Na tomto místě bych ráda poděkovala kolegům z Fakulty informatiky Masarykovy univerzity, zejména Karlu Palovi, který mě nikdy nepřestal provázet na cestě mezi lingvistikou a počítačovým zpracováním přirozeného jazyka, a Pavlu Šmerkovi, autorovi derivačního rozhraní Deriv.

Můj dík patří též kolegům z Ústavu českého jazyka filozofické fakulty Masarykovy univerzity, především Zdeňce Hladké, Petru Karlíkovi a Janě Pleskalové, kteří mi dali svou důvěru a nepřestávali mě trpělivě povzbuzovat, jakož i všem ostatním, díky nimž jsem mohla pokračovat ve své práci v prostředí přátelském a tvůrčím.

Chtěla bych též poděkovat Marku Nekulovi, nejen za to, že mi umožnil prezentovat některé dílčí výsledky mé práce na půdě mimo moji mateřskou univerzitu, ale především za stálý kolegiální zájem.

Děkuji také svým dvěma doktorandkám, Pavlíně Vališové a Kateřině Najbrtové. Kateřině Najbrtové za korektorské přečtení textu, oběma pak za to, že jsem je mohla učit a diskutovat s nimi o některých problémech na úrovni hodné univerzity.

Na závěr chci ovšem poděkovat svým rodičům za pomoc a morální podporu ve chvílích, kdy se výsledky mé práce zdály být až příliš vzdálené, a svému tchánovi za tichou radost, s níž mou práci z dálky sledoval.

Můj největší dík patří mému muži Petrovi a dětem Lukášovi a Bernadetě za to, že se mnou „nesli tíži dne a horka“ a nereptali, že se „denár“, který jim jako manželka a maminka dlužím, tenčí.

OBSAH

I. Úvod	11
Automatická analýza přirozeného jazyka na ÚJČ FF MU	11
II.	
Automatická morfologická analýza a strojový slovník češtiny	13
III.	
Pravidelnost derivace a strojové zpracování přirozeného jazyka	15
IV.	
Vzájemné vztahy formální morfologie a derivace z hlediska možností automatické slovtvorné analýzy	17
Morfematická segmentace slovesných tvarů a deverbativ	17
V.	
Alomorfie (variantnost lexikálních kořenů a tvarotvorných kmenů) jako tvarotvorné a slovtvorné spoluformanty	20
Alternace kmenotvorné přípony a samohlásky – vokálu v základu – alomorfie slovesného základu a slovesného kmene	21
Poznámka k alternaci <i>e/o</i>	22
Alternace finální souhlásky kořene (kf)	22
Popis pravidel alternací při tvoření tvarů sloves (dle slovesných subparadigmat) a derivátů tvořených paradigmaticky (pravidelně přímo od slovesných tvarů)	23
Samohláskové alternace kořenové samohlásky (KoV) a kmenotvorné přípony (KmV)	24
Alternace finální souhlásky kořene (kf), iniciální souhlásky kořene (ki) a alternace souhlásky, která je součástí kmenotvorné přípony (kt)	32
Grafické alternace	37
Kvantita ve slovesných prefixech	38
Střídání <i>0/e</i> ve slovesných prefixech	40
VI.	
Deriv – softwarový nástroj pro testování mezí a možností automatické slovtvorné analýzy	45
Extrakce dat z morfologického slovníku	45

Vyhledávací funkce	45
Další zpracování automaticky extrahovaných dat	48
Práce s obsahem souborů, se soubory, s adresáři	48
Ruční analýza automaticky generovaných slovotvorných vztahů	50
Systematické značkování ručně analyzovaného materiálu	52
Automatické nástroje pro udržení konzistentních řešení při ručním zpracování automaticky získaných výsledků	53
Korpusy jako zdroje dat pro ověření platnosti navržených pravidel	53

VII.

Formální popisy deverbativních jmen tvořených od slovesného tvaru, od slovesného kmene a od slovesného základu (kořene)	55
Substantiva od slovesného tvaru a od slovesného kmene a od slovesného základu (kořene)	57
Slovesné substantivum na <i>-ní/-tí</i>	58
Činitelská jména na <i>-l</i> a další deverbativa od <i>l</i> -ových příčestí	61
Životná substantiva na <i>-[aeě]n-ec, -[nm]ut-ec, -[aeěiyu]t-ec</i> s významem osoby zasažené dějem (patiens)/nositele vlastnosti plynoucí z děje	64
Neživotná deverbativní substantiva na <i>-[aeě]n-ec, -[nm]ut-ec,</i> <i>-[aeěiyu]t-ec</i> s významy objektu děje	70
Činitelská jména na <i>-tel</i>	73
Činitelská jména na <i>-č</i>	76
Jména prostředků na <i>-tel</i>	79
Jména prostředků na <i>-č</i>	81
Jména prostředků a dějů na <i>-čka</i>	84
Deverbativa: jména prostředků, činitelská, dějů, míst na <i>-dlo</i>	87
Jména prostředků na <i>-tko</i>	92
Činitelská jména na <i>-ce</i>	95
Činitelská jména na <i>-ec</i>	102
Činitelská jména na <i>-čí</i>	105
Neživotná deverbativní substantiva na <i>-ec</i> s významy prostředku nebo výsledku děje	108
Adjektiva od slovesného tvaru a od slovesného kmene a od slovesného základu (kořene)	111
Adjektivizované přechodníky přítomné (adjektiva procesuální) na <i>-[(ou) í]cí</i>	111
Adjektivizované přechodníky minulé na <i>-(v)ší</i>	114
Adjektivizovaná příčestí trpná (<i>n-/t-ová</i>) <i>-[aeě]ný, -[áiyu]tý</i>	116
Adjektivizovaná příčestí činná (<i>l-ová</i>)	119
Dezaktualizovaná adjektiva na <i>-cí/-cný (-ou-cný, -í-cný, -ují-cný, -ají-cný,</i> <i>-elějí-cný)</i>	121

Adjektiva účelová na <i>-[aiěuy]cí</i> od minulého kmene	124
Adjektiva vyjadřující vlastnost plynoucí z možnosti zasažení dějem na <i>-telný</i> (<i>-ova-telný, -a-telný, -i-tel, -nu-telný</i>)	126
Adjektiva na <i>-vý</i> (<i>-ova-vý, -a-vý, -i-vý, -li-vý, -ja-vý, -ji-vý</i>) (od kmene)	130
Adjektiva na <i>-čný</i> (<i>-ova-čný, -a-čný, -i-čný, -[eě]-čný</i>)	134
Adjektiva na <i>-livý, -lavý</i> (od kořene)	137
Adjektiva na <i>-čí</i> od slovesného základu (kořene)	141
Adjektiva na <i>-ný</i> tvořená od slovesného základu (kořene)	144

VIII.

Kvantitativní analýza automaticky generovaných dat	169
Míra přegenerování jednotlivých sufixů připojovaných ke slovesnému tvaru (základ+(KmV)+tvarová koncovka) – derivace od tvaru, ke slovesnému kmene (základ+KmV) – stem derivation a ke slovesnému základu (kořeni) – root derivation	170
Substantiva a adjektiva tvořená paradigmaticky od slovesného tvaru	170
Substantiva životná odvozená od kmene/slovesného tvaru a od slovesného základu (kořene)	170
Substantiva neživotná odvozená od kmene/slovesného tvaru a od kořene	171
Adjektiva odvozená od kmene/slovesného tvaru a od kořene	172
Míra přegenerování jednotlivých sufixů	173
Porovnání míry přegenerování strukturovaných a nestrukturovaných derivačních formantů	173
Porovnání míry přegenerování homonymních (polyfunkčních) a monofunkčních sufixů	175
Porovnání míry přegenerování pravidel bez alternací a s alternacemi	181

IX.

Derivační slovník deverbativ analyzovaných typů	186
---	-----

X.

Závěr	190
The Morphology of the Czech Verb and Verb Derived Nouns and Adjectives as a Problem of the Formal Description and Automatic Analysis of the Czech Language	193

Bibliografie	195
Elektronické zdroje:	199

PŘÍLOHY

Příloha A: Systém použitých morfologických značek	201
Příloha B: <i>Deriv</i> – webové rozhraní	208
Příloha C: Ukázky automaticky generovaných dat	211

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 slovtvorby uvedené ve druhé části naší práce (Kap. 7). Tyto mohou posloužit k testování pravidelných slovtvorných 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 nástrojů¹.

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 práci přinášejí publikace (Osolsobě – 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 „standards“ přirozeného jazyka (spisovný jazyk tištěných textů, přepis mluveného jazyka, psaný jazyk neformálních ne-korigovaných projevů) se soustřeďují studie (Osolsobě 2001, Osolsobě 2005, Osolsobě 2006, Hlaváčková – Osolsobě 2008).

K tématu teorie morfologického značkování se vrací studie porovnávací systémy morfologických značek (tagsety) používané ke značkování v českém/slovenském prostředí (Osolsobě 2007¹, 2008¹).

Průběžné sledování mezi a možností značkování jazykových korpusů z hlediska zachycení morfologických vlastností nezachycených explicitně v systému značek glosuje řada studií (Osolsobě 1999, 2002, 2007³, 2008¹, 2009^{1, 2, 4}).

Automatické slovtvorné 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 slovtvorby (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).

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*² (Sedláček 2004) a prostřednictvím tohoto softwarového nástroje je možné s ním dále pracovat³.

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 číslovek⁴. 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)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 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 nejjednodušší (frekventované okrajové jevy zachyceny jsou, nefrekventované nikoli).

Utváření slovní zásoby češtiny ve slovníkových pracích (Slavičková 1974, Šiška 1998) zachycuje teoreticky zdůvodněnou morfematically 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 2008¹).

III.

Pravidelnost derivace a strojové zpracování přirozeného jazyka

Cílem naší práce je formálně popsat realizované⁶ 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 přegenerování.

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 mluvnících, jsou pro potřeby formálního popisu v mnoha aspektech neúplné. Z tohoto důvodu bylo nejdříve třeba vypracovat metodu postupu pro formální 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 mluvčích 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 případech 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 zachyceny⁷ a b) splňují danou hypotézu.

7 Morfologický slovník analyzátoru *ajka* zahrnuje přibližně 400 000 lemmat, z nichž lze na základě morfologických vzorů 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 Pała 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).

IV.

Vzájemné vztahy formální morfologie a derivace z hlediska možností automatické slovtvorné analýzy

Automatická slovtvorná analýza se opírá o pravidelnosti formálních změn, k nimž dochází při derivaci. Derivaci lze formálně popsat jako záměnu/záměny části/částí slova základového, které mají za následek vznik slova odvozeného. Slova základová, od nichž potenciálně mohou být derivována slova odvozená, lze definovat na rovině gramatické abstrakce zakódované v morfologických značkách. Tyto vztahy jsou do různé hloubky popisovány v klasických popisech české slovtvorby. Předložený text si klade za cíl na základě pozorování dat strojového slovníku češtiny a jazykových korpusů vytvořit formální popis a otestovat a) možnosti a meze formalizace; b) doplnit stávající popisy o případy, které dosud nebyly zaznamenány.

Morfematická segmentace slovesných tvarů a deverbativ

Řešení otázky segmentace slovesných tvarů a deverbativ je pro popis tvarosloví sloves a tvoření deverbativ klíčové. Týká se i problematiky morfologických alternací (alomorfie), konkrétně nejrůznějších alternací, které doprovázejí tvoření slovesných tvarů (konjugace) i některých pravidelných derivací od slovesného kmene (adjektivum slovesné na *-ný/-tý* – VA, substantivum slovesné na *-ní/-tí* – VSB). Těm se budeme věnovat v následujících kapitolách. V této kapitole se chceme zmínit o problémech spjatých se segmentací (morfémovou analýzou) slova v oblasti strojové analýzy přirozeného jazyka.

Otázce pravidel segmentace slovního tvaru (morfémové analýzy) jsou věnovány úvodní studie řady morfematických a retrográdních slovníků (Worth – Kozak – Johnson 1970), (Slavičková 1974), (Tichonov 1985), (Šiška 1998), (Sokolová a kol. 1999).

Hloubka morfémové analýzy je mnohdy závislá na účelu příslušné analýzy. Při segmentaci slovesných tvarů a interpretaci segmentů se přidržíme systému použitého v Mluvnici češtiny 2 (Komárek a kol. 1986). Jak jsme uvedli v úvodu, jádrem naší práce bude analýza mezí a možností automatického zpracování derivace některých typů deverbativ. Za tímto účelem se v rámci této kapitoly budeme věnovat zásadám segmentace slovesného tvaru a deverbativ.

U sledovaných typů deverbativ se přidržíme zásad segmentace E. Slavičkové (Slavičková 1974), protože tyto zásady (aplikované na korpusy nesrovnatel-

ně menšího rozsahu, než jsou ty, které máme dnes k dispozici) se nám nejeví v žádném směru jako překonané.⁸

Termíny, které zavádíme v následujícím přehledu, odpovídají (až na výjimky – konekt) termínům použitým v Mluvnici češtiny 2 (Komárek a kol. 1986). Zavedení segmentu, který nazýváme konekt, vychází ze zkušenosti se segmentací některých okrajových typů derivací v partiích věnovaných popisu tvoření slov v českých mluvnicích i v některých výše uvedených (českých) pracích slovníkového charakteru.

U slovesného tvaru budeme tedy rozlišovat následující typy segmentů:

prefix (_{1-n})	tvarotvorný základ (kořen)	kmenotvorná přípona (vč. 0)	konekt	tvarotvorný formant	tvarová koncovka
<i>ne-u-po-</i>	<i>-třeb-</i>	<i>-i-</i>		<i>-l-</i>	<i>-a</i>
<i>za-</i>	<i>-kry-</i>	<i>-0-</i>		<i>-t-</i>	<i>-a</i>
<i>ne-u-po-</i>	<i>-třeb-</i>	<i>-í-</i>			<i>-me</i>
<i>ne-na-</i>	<i>-sáz-</i>	<i>-ej-</i>			<i>-te</i>
<i>u-</i>	<i>-kry-</i>		<i>-v-</i>	<i>-š-</i>	<i>-e</i>

U deverbativ odvozených od slovesného tvaru konverzí budeme rozlišovat následující typy segmentů:

prefix (_{1-n})	tvarotvorný základ (kořen)	kmenotvorná přípona (vč. 0)	konekt	tvarotvorný formant	tvarotvorný formant	tvarová koncovka
<i>u-za-</i>	<i>-vř-</i>	<i>-e-</i>	<i>-v-</i>	<i>-š-</i>	<i>-í-</i>	<i>-ho</i>
<i>o-</i>	<i>-pi-</i>	<i>-0-</i>		<i>-l-</i>	<i>-ý-</i>	<i>-m</i>
	<i>nes-</i>	<i>-0-</i>		<i>-ouc-</i>	<i>-í-</i>	<i>-mi</i>
<i>u-</i>	<i>-tř-</i>	<i>-e-</i>		<i>-n-</i>	<i>-ý-</i>	<i>-ch</i>
	<i>kut-</i>	<i>-i-</i>		<i>-l-</i>	<i>-0-</i>	<i>-em</i>
<i>vy-</i>	<i>klouz-</i>	<i>-0-</i>		<i>-0-</i>	<i>-0-</i>	<i>-ovi</i>

U deverbativ odvozených od slovesného tvaru sufixací budeme rozlišovat následující typy segmentů:

prefix (_{1-n})	tvarotvorný základ (kořen)	kmenotvorná přípona (vč. 0)	konekt	tvarotvorný formant	slovotvorný formant	tvarová koncovka
<i>roz-</i>	<i>-mázl-</i>	<i>-0-</i>		<i>-en-</i>	<i>-0c-</i>	<i>-i</i>
<i>o-</i>	<i>-žr-</i>	<i>-a-</i>		<i>-l-</i>	<i>-0c-</i>	<i>-em</i>
<i>o-</i>	<i>pi-</i>	<i>-0-</i>		<i>-l-</i>	<i>-ec-</i>	<i>-0</i>
<i>za-</i>	<i>-ry-</i>	<i>-0-</i>		<i>-t-</i>	<i>-ec⁹</i>	

8 Zásady segmentace slovního tvaru jsou bez jasného zdůvodnění masivně měněny vzhledem k dosavadní praxi české mluvnické i lexikografické tradice v kapitole věnované tvoření slov v Mluvnici současné češtiny (Cvrček a kol. 2010, s. 81–124).

9 Upozorňujeme na segmentaci substantiva *zarytec*, které chápeme jako odvozené od tvaru participia/adjektiva slovesného : *zarýt/zaryt/zarytj/zarytec*, na rozdíl od níže uvedeného substantiva *rytec*.

U deverbativ odvozených od slovesného kmene budeme rozlišovat následující typy segmentů:

prefix (_{1-n})	tvartvorný základ (kořen)	kmenotvorná přípona (vč. 0)	konekt	tvartoslovný formant (_{1-n})	tvartvorný formant	tvárová koncovka
<i>s-</i>	<i>-běr-</i>	<i>-a-</i>	<i>0</i>	<i>-tel-</i>	<i>-0-</i>	<i>-em</i>
	<i>hnět</i>	<i>0</i>	<i>-a-</i>	<i>-č-</i>	<i>-0-</i>	<i>-ů</i>
<i>ne-</i>	<i>-trp-</i>	<i>-ě-</i>	<i>-l-</i>	<i>-iv-</i>	<i>-ý-</i>	<i>-mi</i>
<i>nej-ne-</i>	<i>-pře-</i>	<i>-j-</i>		<i>-íc-n-ějš-</i>	<i>-í-</i>	<i>-mi</i>

U deverbativ odvozených od slovesného kořene budeme rozlišovat následující typy segmentů:

prefix (_{1-n})	kořen	konekt	tvartoslovný/tvartoslovné formant/formanty	tvartvorný formant	tvárová koncovka
	<i>soud-</i>		<i>-c-</i>		<i>-e</i>
	<i>ry-</i>	<i>-t-</i>	<i>-0c-</i>		<i>-em</i> ¹⁰
<i>vý-</i>	<i>-běr-</i>	<i>0</i>	<i>-č-</i>	<i>-l-</i>	<i>-m</i>
	<i>smír-</i>		<i>-č-</i>	<i>-l-</i>	<i>-ho</i>
<i>pří-</i>	<i>-sa-</i>	<i>-v-</i>	<i>-n-</i>	<i>-ý-</i>	<i>-ch</i>
<i>ú-</i>	<i>-spě-</i>	<i>-š-</i>	<i>-n-</i>	<i>-ě-</i>	<i>-m</i>

10 Upozorňujeme na segmentaci substantiva *rytec*, které chápeme jako odvozené od základu (kořene) *rytí/rytec*, na rozdíl od výše uvedeného substantiva *zarytec*.

Alomorfie (variantnost lexikálních kořenů a tvarotvorných kmenů) jako tvarotvorné a slovotvorné spoluformanty

Při tvoření slov může docházet a často dochází k hláskovým změnám (hláskovým/morfologickým alternacím). Morfologické alternace lze definovat jako pravidelné střídání určitých fonémů, popř. jejich skupin, přičemž alternující hlásky (skupiny) utvářejí jeden morfoném, na těchto alternacích je založena alomorfie. Alternace hlásek sama o sobě není téměř nikdy samostatným slovotvorným formantem, je vázána na existenci jiného slovotvorného postupu, tedy např. na rozšíření odvozovacího kmene o slovotvornou příponu, záměnu slovotvorné přípony, či pouhou změnu tvaroslovné charakteristiky slova (konverze).

Alternace lze rozčlenit na kombinační (syntagmatické), v nichž jsou hláskové proměny vázány na hláskové okolí základu, a volné (paradigmatické), které na hláskovém okolí základu nezávisí a představují mnohonásobně početnější skupinu.¹¹

V této kapitole nastíníme pravidla morfologických alternací, k nimž dochází při tvoření slovesných tvarů a některých derivací od slovesných základů. Na tato pravidla navážeme v jednotlivých kapitolách věnovaných formálnímu popisu tvoření vybraných typů deverbativ, a sice v tabulkových přehledech alternací u příslušného derivačního typu dle jednotlivých typů morfů.

Při tvoření tvarů sloves a při derivaci deverbativ dochází k hláskovým alternacím, které pracovně rozdělíme podle typu morfu, jehož se týkají. Budeme rozlišovat alternace v kořeni, kmenotvorné příponě a prefixu, popř. na švu prefixu a kořene, dále alternace finály kořenového morfu, přičemž si budeme všimnout i jejich grafických realizací¹².

- a) alternace kmenotvorné přípony (**KmV**) a její grafické realizace vč. alternací souhlásky, která je součástí kmenotvorné přípony (**kt**);
- b) alternace samohlásky/diftongu v základu (**KoV**);
- c) alternace finální souhlásky kořene (**kf**), popř. iniciální souhlásky kořene (**ki**);
- d) alternace samohlásek v prefixu (**PV**), alternace (včetně grafických a *e/0*) na švu prefix/základ (**P0e**).

11 Srv. Dokulil (1962 : 159–162), Dokulil a kol. (1986 : 319–320).

12 Formální pravidla se testují na psané formě jazyka, proto je třeba brát tuto podobu v úvahu.