VI.

Deriv – softwarový nástroj pro testování mezí a možností automatické slovotvorné analýzy

V této kapitole popíšeme možnosti, které skýtá automatický nástroj *Deriv – webové rozhraní* (Šmerk 2009) pro testování mezí a možností automatické slovotvorné analýzy. Nástroj byl vyvíjen ve spolupráci lingvistů a programátorů (Hlaváčková – Osolsobě – Pala – Šmerk 2009[1,2]).

Deriv – webové rozhraní je víceúčelový nástroj pro automatické zpracování přirozeného jazyka. Interaktivní webové rozhraní má dvě základní sady funkcí. První sada jsou funkce vyhledávací. Druhá sada umožňuje automaticky extrahovaná data prohlížet, třídit, editovat, ukládat.

V naší práci publikujeme řadu výsledků získaných pomocí nástroje *Deriv*. Jsou to především testy pokrytí formálních pravidel popisujících vybrané případy derivací v češtině. Testování sloužilo zároveň k hledání cest vedoucích k optimalizaci navržených funkcí i případnému doplnění dosud nerealizovaných funkcí, které se během práce ukázaly být užitečné pro automatické zpracování přirozeného jazyka.

Extrakce dat z morfologického slovníku

Vyhledávací funkce slouží k prohledávání morfologického slovníku kmenů (Osolsobě 1996), který je součástí automatického morfologického analyzátoru *ajka* (Sedláček 2004). Data jsou vyhledávána a extrahována na základě zadavatelných parametrů. Tyto parametry jsou a) formálně definovatelné vlastnosti hledaných jednotek a b) interpretace definovaných jednotek uložené v podobě tzv. morfologických značek.

Vyhledávací funkce

Vyhledávací funkce nástroje *Deriv* umožňují vyhledávání jednotek na různé úrovni abstrakce. Požadované jednotky lze definovat jednak na základě jejich formálních vlastností (jako řetězce znaků s použitím regulárních výrazů), jednak na základě jejich gramatických vlastností.

Pro formální definice hledaných jednotek se využívá regulárních výrazů (programovací jazyk Perl 5.106). Přehledně jsou zachyceny v následující tabulce.[105]

znak	**interpretace**
. (tečka)	libovolný znak
[abc]	jeden z vyjmenovaných znaků
[[:upper:]]	velké písmeno
[[:lower:]]	malé písmeno
[^abc]	libovolný znak mimo vyjmenované
^	začátek řetězce
$	konec řetězce
(a | b)	první nebo druhý znak
?	předchozí znak může/nemusí být přítomen
*	libovolné opakování předcházejícího znaku
+	libovolné opakování, alespoň jednou
$C	libovolný konsonant včetně ch
$V	libovolný vokál včetně ou
$L	libovolný dlouhý vokál včetně ou
$S	libovolný krátký vokál

Gramatické vlastnosti jsou zachyceny v morfologických značkách. Ve slovníku automatického morfologického analyzátoru jsou uloženy jednotky definované jako **slovní tvar/lemma/tag**.

Morfologická značka **(tag)** má striktně stanovenou formu. Je to posloupnost příslušných atributů a jejich hodnot. Pokud chceme zadat pouze některý z atributů (například všechny tvary označené jako substantiva – [tag="k1.*"], všechny tvary mající značku signalizující množné číslo [tag=".*nP.*"], všechny tvary jmen v dativu [tag=".*c3.*"]), pak je třeba použít patřičným způsobem regulární výrazy. Pro potřeby vyhledávání podle značek vystačíme se sekvencí „.*", kde tečka „." představuje jeden libovolný znak a hvězdička „*" představuje libovolný počet (0 a více) opakování předchozího znaku nebo výrazu.

Systematický popis použitých morfologických značek i s příklady uvádíme v Příloze A, viz níže.

Jak vypadá v praxi extrakce dat ukážeme na následujícím příkladu.

Chceme vyhledat všechna substantiva skloňovaná podle vzoru *stavení*, protože se chceme dozvědět něco o deverbativních jménech. Po formální stránce se nespokojíme s tím, že víme, že lemmata substantiv, která definujeme, lze popsat pomocí morfologické značky jakožto substantiva, neutra s tvarem nominativu singuláru **(k1gNnSc1)** zakončená řetězcem **[nt]í$**, ale zahrneme i empiricky doloženou a gramaticky podmíněnou podmínku. Ta zní, že před

105 Pro ty čtenáře, kteří jsou zvyklí pracovat s jazykovými korpusy, je řada výrazů (nikoliv všechny) shodná s výrazy, které se běžně používají při práci s korpusovými manažery.

koncovým řetězcem **[nt]í$** může předcházet pouze vybraná samohláska, popř. vybraná kombinace souhláska + samohláska **([áeě]n|[aáeěiuy]t|nut|mut)$**.

Při ruční analýze automaticky generovaných dat zjistíme, že kromě homogenní skupiny verbálních substantiv požadované formální vlastnosti splňují i další slova jako např. *století, úpatí,* vlastní jména jako *Rokytí* atd. Jde o případy spadající pod pojem přegenerovávání.[106]

Negativní příklady přitom slouží jak k mapování případů, kdy jedna a táž forma má více interpretací (homonymie), tak především k optimalizaci dotazů pro automatické vyhledávání (k nalezení rovnováhy mezi pře- a podgenerováváním[107]). Pokud zjistíme, že přegenerovávání spadá na vrub nedostatku na úrovni formulace vyhledávacího dotazu, lze na základě pozorování chyb zadání dotazu upravit a operaci zopakovat. Za optimální pokládáme dotazy, které vyloučí co největší počet nesprávně generovaných jednotek, aniž by došlo k tomu, že existující (doložené) jednotky zůstanou nezachyceny.

V rámci vyhledávacích funkcí lze definovat i nejrůznější vztahy mezi formálními a gramatickými vlastnostmi jednotek. K tomuto účelu slouží funkce pro vyhledávání dvojic (popř. n-tic), jež lze definovat jako nahrazovací pravidla. Substitučním pravidlem lze postihnout vlastnosti dvojic **základové slovo/ odvozené slovo.** Klasické popisy v mluvnicích intuitivně naznačují vzájemné vztahy významu i formy **základového** a **odvozeného slova.** Na této intuici zakládáme předpoklad, že vznik **odvozeného slova** lze popsat formálním substitučním pravidlem.

Substituční pravidla vycházejí tedy z předpokladu, že existují dvojice definovatelné pomocí řetězců znaků takové, že: nahradíme-li část/části řetězce tvořenou/tvořené přesně definovatelnou/definovatelnými posloupností/posloupnostmi znaků řetězcem/řetězci tvořeným/tvořenými přesně definovatelnou/ definovatelnými posloupností/posloupnostmi znaků, výsledná dvojice bude pravděpodobně **základové slovo/odvozené slovo**.

Uvedeme příklad:

Naším cílem je formulovat vztah sloves a od nich odvozených dějových jmen neuter na **[nt]í$**. K tomuto účelu použijeme funkci pro vyhledávání dvojic. Substitučním pravidlem se budeme snažit zachytit formální a gramatické vztahy dvojice základové slovo (slovo definované pomocí regulárních výrazů

106 Jedním z podstatných rysů aplikací automatické analýzy přirozeného jazyka je tzv. přegerovávání. Formální definici odpovídají jednotky, které tvoří homogenní skupinu (tu, kterou se prostřednictvím formálního zadání snažíme definovat), ale i jednotky, které jsou vůči této skupině heterogenní. Tento jev spadá na vrub obecné vlastnosti přirozeného jazyka, jíž je nejednoznačnost (homonymie) na všech úrovních.

107 Podgenerovávání je případ, kdy formální zadání je vymezeno příliš úzce, takže nejsou zachyceny jednotky, které se jeho prostřednictvím snažíme definovat.

– RE a gramatických vlastností) – odvozené slovo (slovo definované pomocí regulárních výrazů – RE a gramatických vlastností). Základové slovo/slovní tvar můžeme definovat pomocí morfologické značky jakožto pasivní participium[108], maskulinum životné, singulár **(k5eA.*mNgMnS)** končící na **[áeě]n$** nebo **[aáeěiuy]t$** nebo **[n|m]ut$**. Odvozené slovo definujeme pomocí morfologické značky jakožto substantivum, neutrum, jehož tvar nominativu singuláru **(k1gNnSc1)** končí řetězcem **[áeě]ní$** nebo **[aáeěiuy]tí$** nebo **[n|m]utí$**. Pomocí substitučních pravidel můžeme zachytit kombinace substitucí, k nimž při derivaci dochází. Podrobně se jimi budeme zabývat v kapitole 7 věnované formálnímu popisu derivace některých typů deverbativ, který má podobu substitučních pravidel popisujících derivační vztahy.

Automaticky vyhledaná data lze ukládat do souborů a ty do strukturovaných adresářů a lze je následně automaticky i ručně zpracovávat. Podrobněji o významu těchto funkcích pojednáme v kapitolách věnovaných ručnímu zpracování automaticky generovaných dat a automatickým nástrojům pro udržení konzistence dat.

Další zpracování automaticky extrahovaných dat

Data automaticky extrahovaná ze slovníku automatického morfologického analyzátoru lze, jak jsme uvedli výše, dále zpracovávat. Automaticky generovaná data (slova, slovní tvary, dvojice základové slovo/odvozené slovo) lze uložit do souborů, které můžeme uchovávat, prohlížet a třídit v systému adresářů. Adresáře a podadresáře lze vytvořit pomocí funkce **práce s adresáři**. Dále lze zvolit buď funkci **práce s obsahem souborů**, nebo funkci **práce se soubory**.

Práce s obsahem souborů, se soubory, s adresáři

Automatický nástroj *Deriv* umožňuje pomocí funkce **práce s adresáři** navrhnout dostatečně strukturovaný systém podadresářů, což usnadní orientaci a systematické ukládání při práci s masovými daty.

108 Vycházíme z předpokladu, že základovým slovem derivace dějových jmen je forma odpovídající potencionálnímu tvaru pasivního příčestí. Jsme si vědomi toho, že zatímco tvoření pasivních příčestí je v češtině omezeno gramatickými vlastnostmi příslušného slovesa, u tvoření dějových jmen na *-ní/-tí* takováto omezení neplatí (paradigmatické tvoření). Srv. příslušnou pasáž věnovanou tvoření dějových jmen na *–ní/-tí* v Příruční mluvnici češtiny (Karlík – Nekula – Rusínová 1995 : 148), která uvádí na pravou míru tvrzení, že dějová jména se tvoří od trpných příčestí (srv. např. Šmilauer 1972 : 208, Čechová a kol. 1995 : 93).

Funkce **práce se soubory** umožňuje v rámci systému adresářů jednotlivé soubory přesunovat, slučovat stejnorodé soubory do jednoho, popřípadě rozdělovat data z jednoho souboru do více souborů a konečně soubory mazat.

Funkce **práce s obsahem souborů** umožňuje prohlížení uložených souborů. Při prohlížení lze zvolit třídění dat uložených v prohlíženém souboru a) abecední/retrográdní b) s/bez frekvence/frekvencí jednotek v korpusu. Prohlížená data lze ručně zpracovat. Každý řádek je opatřen okénkem, do něhož lze vepsat text[109]. Tímto textem může být značka, pomocí které lze pak data automaticky roztřídit. Kromě toho lze zvolit funkci **upravit obsah souboru** a data uložená v souboru editovat ručně.

Při prohlížení automaticky generovaných dat lze a) označovat případy, kdy dochází k nejrůznějším typům přegenerovávání, tj. k případům, kdy automaticky extrahovaná data sice formálně splňují příslušné požadavky, ale vzhledem k celku jsou heterogenní (viz příklady výše) a b) opravovat případné chyby. Ručně označkovaná data lze opětovně uložit a rozdělit do jednoho souboru označkovaná data (např. přegenerované případy), která jsou náhodnými případy formální shody, a do druhého souboru neoznačkovaná data, která jsou správně generována.

Automaticky získaná data lze navíc interaktivně sledovat. Automatický nástroj *Deriv* je propojen jednak s internetovým prohlížečem *DebDict* (chrome://debdict/content/debdict.xul), který umožní zobrazit hesla odpovídající automaticky generovaným jednotkám v řadě výkladových slovníků (SSJČ, SSČ, PSJČ) i ve slovníku automatického morfologického analyzátoru *ajka*, jednak s korpusem SYN2000 (popřípadě dle potřeby i s dalšími korpusy), takže lze jednoduše přepnout do korpusu na seznam konkordančních řádků, přičemž klíčovým slovem (KWIC) je libovolný požadovaný člen automaticky generované dvojice[110].

Výsledky automatické analýzy lze dále testovat pomocí automatického nástroje *Deriv* z hlediska možných chyb v pokrytí pravidel. K tomuto účelu slouží, jak už bylo řečeno výše, porovnání počtu potencionálních derivátů a počtu správných derivací pokrytých navrženými pravidly.

Uvedeme příklad:

Při formulaci pravidel popisujících derivační vztahy a jejich testování postupujeme následovně. Nejdříve ověříme počet potencionálních odvozených jednotek. Poté navrhneme formální (substituční) pravidla a uložené výsledky automatického generování ručně procházíme. Pokrytí pravidel zjišťujeme tak, že porovnáváme počet potencionálních odvozených slov *(X)* s počtem automaticky generovaných dvojic (n-tic) *(Y)* zmenšeným o počet případů, v nichž

109 Viz Kap. 9.
110 Viz Příloha B.

došlo k přegenerovávání *(Z)*. Navíc sledujeme, které jednotky ze seznamu potencionálních odvozených jednotek nejsou podchyceny pravidly a operativně navrhujeme další pravidla tak, abychom zachytili maximum jednotek. Takto získáme přehled o případných výjimkách a chybách v datech slovníku automatického morfologického analyzátoru. Velikost rozdílu *(X-(Y-Z))* je nepřímo úměrná pokrytí pravidel. V jednotlivých případech provádíme stručnou analýzu dat „nepokrytých" pravidly (viz níže).

Ruční analýza automaticky generovaných slovotvorných vztahů

Ruční analýza materiálu se opírá o interpretace v a) mluvnických popisech, b) ve slovnících a c) v korpusech, přičemž fakta z „objektivních zdrojů" dotváří jazykové povědomí kontrolujícího subjektu.

Typologie chyb, k nimž při automatickém generování slovotvorných vztahů dochází (analýza nejrůznějších případů přegenerovávání), je nesmírně důležitým zdrojem zkušeností pro zkvalitnění ručního zpracování automaticky generovaných dat[111].

První okruh chyb lze charakterizovat jako chyby související s široce pojatou homonymií přirozeného jazyka na jeho nejrůznějších úrovních. Ty lze dále klasifikovat.

Od základového slova se aplikací formálního pravidla vytvoří slovo odvozené nikoli od tohoto slova, ale od jeho slova základového. Může nastat několik variant:

A. Od dvou příbuzných základových slov téhož slovního druhu se aplikací formálního pravidla vytvoří slovo odvozené, které lze vztáhnout k oběma členům dvojice. Například od sloves 4. třídy vzorů *prosit/trpět/sázet,* jsou-li dvojice sloves patřící k různým vzorům synonymní.
B. Od dvou příbuzných základových slov téhož slovního druhu se aplikací formálního pravidla vytvoří slovo odvozené, které lze vztáhnout pouze k jedinému členu dvojice. Například od sloves 4. třídy vzorů *prosit/trpět/sázet*, nejsou-li dvojice sloves patřící k různým vzorům synonymní.

111 Automatickému zpracování lingvistických dat slouží řada různě koncipovaných automatických nástrojů. Zpracování automatickými nástroji je doplňováno v různých fázích tzv. ručním zpracováním, kdy do procesu vstupuje „lidský faktor". Znalost typologie chyb, k nimž na úrovni automatického zpracování dat dochází, vede a) k optimalizaci automatického zpracování, b) k usměrňování pozornosti lidského korektora. Nejen automat, ale i člověk se dopouští chyb. Zdrojem lidských chyb při zpracování hromadných dat je a) neznalost, b) nepozornost, jejíž příčinou kromě únavy bývá i „nepřehlednost" opravovaných chyb. Analýza chyb umožňuje lepší orientaci v opravovaných datech.

Od základového slova se aplikací formálního pravidla vytvoří slovo odvozené nikoli od tohoto slova, ale od jiného slova základového. Může nastat několik variant:

C. Od dvou nepříbuzných základových slov téhož slovního druhu se aplikací formálního pravidla vytvoří slovo odvozené, které lze vztáhnout pouze k jedinému základovému slovu. Tento typ chyb souvisí s homonymií alternujících variant různých kořenů a s omezeními danými podobou výchozích dat strojového slovníku (data ve strojovém slovníku nejsou segmentována aniž by segmenty byly interpretovány, nelze tudíž např. rozlišit, zda vokál před sufixem je KoV nebo KmV). Například od slovesa *řadit* se tvoří deverbativum *řadič*. U tohoto typu derivace fakultativně dochází k alternacím KoV i KmV (*pálit/palič, mlít/mleč*), takže prostřednictvím pravidla pro derivaci se samohláskovou alternací KoV se vygeneruje i nesprávná dvojice *řádit/řadič*. V jiném případě jde navíc o to, že morfologický slovník, z nějž extrahujeme data, nemá žádnou morfematickou segmentaci (na úrovni pravidel nelze rozlišit KoV u sloves bez kmenotvorné přípony ve kmeni minulém a sloves s KmV). Od slovesa *dojit* se pravidelně tvoří sloveso *dojič*. Zároveň ale pravidlo pro generování deverbativ na *-č* od sloves bez kmenotvorné přípony ve kmeni minulém, u nichž se derivační sufix *-č* připojuje přímo k otevřenému kořeni a u nichž zároveň může docházet k alternacím KoV (např. *bít/bič*), vygeneruje i chybnou dvojici *dojít/dojič*.

D. Existují dvě slova různých slovních druhů (mohou i nemusí být příbuzná). Aplikací formálního pravidla na slovo patřící k jednomu slovnímu druhu vznikne slovo derivované od příbuzného slova jiného slovního druhu. Tento typ chyb souvisí jak s homonymií základů, tak s homonymií derivačních prostředků (přičemž může jít o různé kombinace). Například od adjektiv (*zelený*) se tvoří sloveso (*zelenat se*), nebo existuje substantivum *vole,* které má náhodně shodný základ se slovesem *vol-* *-a-t*. Od sloves se tvoří názvy činitelské a prostředků sufixem *-č* s řadou variant, přičemž před *-č* musí předcházet vokál (pouze některé kombinace jsou přípustné), který je interpretovaný buď jako kmenotvorná přípona (derivace od kmene např. *maz-á-č, top-i-č*), nebo jako kořenový vokál nebo konekt (v případě sloves bez kmenotvorné přípony pro derivaci od kmene minulého *bi-0-č, rý-0-č, hrá-0-č, pek-á-č, klad-e-č*). Různé varianty jsou homonymní např. se sufixy *-áč, -ič,* které slouží mj. k derivaci desubstantivních jmen podle význačného rysu (*vole/voláč* nikoli *volat/voláč*), desubstantivních názvů prostředků vzniklých zkracováním či univerbizací (*snowboard/snoubič* nikoli *snoubit/snoubič*) a deadjektivních názvů nositelů vlastností (*zelený/zelenáč* nikoli *zelenat/zelenáč*).

Systematické značkování ručně analyzovaného materiálu

Automatický nástroj *Deriv* umožňuje editaci automaticky generovaných dat (**práce s obsahem souborů**), a to jednak vkládání informací k automaticky vygenerovaným seznamům, jednak opravy dat v automaticky generovaných seznamech. K systematickému odstranění chybně generovaných dokladů derivačních vztahů (přegenerovávání navržených a testovaných pravidel) slouží editace formou vkládání značek (např. označení chybných řádků, označení různých typů chyb).

Data lze třídit, přičemž klíčem třídění je a) abecední/retrográdní uspořádání b) značka.

Uvedeme příklad: Testujeme substituční pravidlo pro automatické generovaní dvojic infinitiv (A)/substantivum maskulinum životné v nominativu (B), pro které platí, že řetězec (B) lze generovat náhradou koncového řetězce x řetězce (A) řetězcem y, takže (A-x)+y=B a zároveň dochází k alternacím uvnitř obou řetězců, kdy přesně definovaný podřetězec řetězce A je nahrazen přesně definovaným podřetězcem.

Pravidlo má podobu: **ou(.)it/k5.*mF>u$1ič/k1gMnSc1.**

Můžeme jej číst tak, že hledáme dvojice infinitivů končících řetězcem *it*, které mají uvnitř řetězce podřetězec *ou* a skupinu substantiv maskulin životných takových, že oddělíme-li koncový řetězec infinitivu *it* a nahradíme-li jej řetězcem *ič* a zároveň nahradíme podřetězec *ou* uvnitř infinitivu podřetězcem *u*, pak bychom měli dostat kandidáty na dvojice infinitiv slovesa/činitelské jméno na *č* od téhož slovesa takové, že derivace je provázena alternací *ou/u* KoV.

Výsledkem jsou data v následující tabulce. Chybně generované dvojice označíme (X). V případě, že chceme označit chyby různého charakteru (viz výše), je možné zvolit více různých značek (např. X, ?) a následně je třídit podle různých kombinací zvolených značek.

X	boudit	budič
	bouřit	buřič
	brousit	brusič
	hloubit	hlubič
?[112]	koulit	kulič
	loupit	lupič
	snoubit	snubič
	soudit	sudič

112 Vztah motivace u dvojice *koulit/kulič* je složitější než u dvojic *loupit/lupič* nebo *brousit/brusič*, které jsou generovány tímtéž pravidlem. Definice obou členů dvojice ve slovníku SSJČ znějí takto:

Označený seznam uložíme, vytřídíme označené řádky a opět uložíme.

	bouřit	buřič
	brousit	brusič
	hloubit	hlubič
	loupit	lupič
	snoubit	snubič
	soudit	sudič

Výsledný seznam obsahuje pouze správně generované dvojice.

Automatické nástroje pro udržení konzistentních řešení při ručním zpracování automaticky získaných výsledků

K automatické kontrole konzistence řešení při ručním zpracování automaticky získaných dat slouží funkce „porovnat“. Při ručním zpracování hromadných dat dochází, jak bylo řečeno výše, k chybám, které spadají na vrub člověku, který automaticky zpracovaná data kontroluje (korektor). Výsledky více korektorů, popřípadě opakované výsledky jednoho a téhož korektora lze automaticky porovnat, a poté vyhodnotit rozdíly. Tímto postupem mnohdy vyplynou na povrch nejen chyby „z nepozornosti“, ale i případy, u nichž je řešení sporné. Analýza druhého typu chyb je krokem k optimalizaci ručního zpracování.

Korpusy jako zdroje dat pro ověření platnosti navržených pravidel

Jak bylo řečeno výše, nástroj *Deriv* je propojen s korpusy (konkrétně s korpusem SYN2000). V korpusech lze interaktivně pozorovat data získaná automatickou analýzou ze slovníku morfologického analyzátoru (*ajka*). Korpusy mohou ovšem sloužit k optimalizaci navržených pravidel i mimo tuto interaktivní funkci nástroje *Deriv.*

Jazykové korpusy byly a jsou zdrojem pro studium jazyka, přičemž velkou roli hraje studium lexika (korpusová lexikografie). Jednou z oblastí obohacování slovní zásoby je i tvoření slov, tedy tvoření nových slov na základě slov

kouliti (se) v. kuliti (se)
kuliti ned. (3. mn. -í) **1.** (zř. též kouliti) (co) *koulet, kutálet:* k. míč; k. klubko; k. klády **2.** expr. k. oči *poulit, valit, vyvalovat* **3.** slang. expr. (koho) *klamat, balamutit:* vy mě zřejmě kulíte; **--- kuliti se** ned. expr. *zvolna, batolivě jít; batolit se:* dítě se kulilo pomalu se schodu na schod (o) předp. **do-, do- se, na-, na- se, od-, od- se, pře-, pře- se, při-, při- se, s-, s- se, vy-, za**
kulič-e m. sklář. dělník, který provádí zdobení skla vybrušováním na brusných kotoučích
Definice ve slovníku ukazují sice na společnou motivaci slovesa a substantiva, zároveň je z nich patrné, že vztah je poněkud uvolněný. V korpusu je doloženo pouze příjmení *Kulič*.

již existujících. Charles J. Fillmore (1992 : 35) napsal: „I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore … [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way."[113] V tomto smyslu jsou jazykové korpusy (díky rozsahu i pestrosti dat, která zahrnují) zdrojem, o nějž lze opřít naši představivost o možném i nemožném v oblasti slovotvorby. Je tomu tak z mnoha důvodů. Pravidla tvoření nových pojmenování i jejich užití jsou namnoze natolik evidentní pro rodilé mluvčí, že utvořené jednotky nejsou (pravidelně) zaznamenány ve výkladových slovnících. Totéž platí i o morfematických slovnících založených na korpusech o několik řádů menších než ty, které máme dnes k dispozici (Slavíčková 1974). Pravidla odvozování slov jsou popsána v oddílech věnovaných slovotvorbě, které jsou pravidelnou součástí českých moderních gramatik. Pravidla obsažená v gramatikách ovšem narážejí na některé nedostatky způsobené omezeným počtem ilustrativních příkladů, který je dán jednak rozsahem příslušné gramatiky, jednak rozsahem jazykového materiálu, na němž je gramatika založena. S ohledem na otevřenost procesu tvoření slov je každý nový korpus zdrojem nových dokladů tam, kde je jich z jakéhokoli důvodu nedostatek. O tom, jak lze v korpusech hledat doklady tvoření nových slov podle existujících modelů, jak lze na základě analýzy korpusových dokladů doplňovat popisy fungování slovotvorného systému srv. více (Osolsobě 2008[3], 2009[3], 2011).

113 Nemyslím si, že by mohly existovat sebevětší korpusy, v nichž bych našel všechny informace o slovníku a gramatice angličtiny, které chci zkoumat … (ale) i ten nejmenší korpus, který jsem měl možnost zkoumat, mě přivedl k věcem, na něž bych jinak nebyl býval přišel.