ČASOPIS PRO MODERNÍ FILOLOGII 97, 2015, Č. 2, S. 136–145

Korpusy jako zdroje dat pro úpravy nástrojů automatické morfologické analýzy (Slovotvorné varianty adjektiv na [(o*u*)|*í*]c*í* z hlediska morfologického značkování)[1]

Klára Osolsobě (Brno)

CORPORA AS DATA SOURCES FOR THE UP-GRADING OF MORPHOLOGICAL TAGGING
Adjectives ending with *-oucí/-ící* are regularly derived from verbs and hence are not usually listed in any of the Czech monolingual dictionaries. On the level of automatic morphological analysis (the dictionary) of Czech they should be generated from verbal roots and tagged as verbal adjectives (pos tag: AG.*). The data from Czech corpora prove a) inconsistencies in tagging and b) gaps in the dictionary. The main cause of both kinds of insufficiency is the existence of variants on the level of verbal forms from which the verbal adjectives are potentially derived. Consequently, text corpora are a significant source of knowledge about the formation and use of adjectives with endings *-oucí/-ící* that can be important for both a) automatic morphological analysis of Czech and b) theoretical description of Czech grammar (derivational morphology). Our goal is to present a corpus-based study of the Czech gerund, i.e. verbal adjectives with *-oucí/-ící*. The link between the inflected and the word-formation variants will be demonstrated using material from the SYN corpus (2,6 billion tokens of written Czech) and the large web corpus czTenTen12 (5,2 billion tokens of Czech text from the Internet — cleaned and deduplicated).

KEYWORDS
gerund/deverbal adjective, pos tagging, automatic morphological analysis, variant, derivational morphology

KLÍČOVÁ SLOVA
verbální adjektivum, morfologické značkování, automatická morfologická analýza, varianta, slovotvorba

1. ÚVOD

Jednou z charakteristik morfologického systému češtiny je velké množství variant a dublet. Na rovině popisu formální morfologie (tvarosloví) je toto bohatství poměrně velmi dobře zachyceno nástroji automatické morfologické analýzy (téměř vyčerpávajícím způsobem jsou zachyceny tvaroslovné varianty kodifikované a částečně jsou zpracovány i varianty nekodifikované). Přesto některé varianty z nejrůznějších důvodů zůstaly opomenuty. Ukazuje se, že nedostatky na rovině popisu tvarosloví souvisí s nedostatky popisu pravidelných slovnědruhových transpozic. Některé

1 Tento text vznikl za podpory grantu Slovník afixů užívaných v češtině (GAČR, reg. č. 13-07138S).

slovotvorné varianty nejsou konzistentně označkovány nebo nejsou rozpoznány automatickou morfologickou analýzou.

Naším cílem bude ukázat, jak lze na základě pozorování korpusových dat (konkrétně tagování adjektiv jednoho derivačního typu) shromáždit relevantní data a připravit podklady pro systematické opravy slovníku automatického morfologického analyzátoru[2] užívaného pro anotaci korpusů Českého národního korpusu a Pražského závislostního korpusu.

Konkrétně se zaměříme na tzv. procesuální adjektiva na *-oucí/-ící*, jejichž tvary se v češtině tvoří pravidelně (paradigmaticky) od slovesných základů. Budeme sledovat jejich značkování v korpusu SYN[3].

2. PROCESUÁLNÍ ADJEKTIVA NA *-OUCÍ/-ÍCÍ* Z HLEDISKA MORFOLOGICKÉHO ZNAČKOVÁNÍ

Popis tvarů adjektiv odvoditelných od jedné z variant slovesného kmene přítomného (nikoli jen od tvaru přechodníku přítomného, viz Dokulil a kol., 1986, s. 321) je na úrovni slovníků automatické morfologické analýzy užívaných v českém prostředí řešen tak, že adjektivní tvary jsou pomocí formálních pravidel „rozgenerovány" od slovesného kmene a je jim v důsledku takové operace přidělena značka. V „pražském systému" morfologického značkování má značka konkrétní podobu AG.* (přídavné jméno odvozené od slovesného tvaru přítomného přechodníku), odlišnou od dalších značek adjektiv[4].

2.1 OVĚŘENÍ KONZISTENCE POPISU A POKRYTÍ SLOVNÍKU NA ZÁKLADĚ ANALÝZY KORPUSU

Naším cílem bude a) vyhledat všechna adjektiva, která končí na *-cí*, b) pozorovat, jak jsou značkována, a c) na základě pozorování ověřit konzistenci a pokrytí slovníku použitého pro morfologické značkování.

Všechna adjektiva tvořená od sloves s oporou ve tvaru 3. os. pl. / přechodníku přítomného mají být v korpusech řady SYN označkována jako AG.*. Pokud tomu tak

2 Slovník automatického morfologického analyzátoru, o kterém bude v celém dalším textu řeč, je dílem kolektivu vedeného Janem Hajičem (více Hajič, 2004). Odborná veřejnost užívá označení „pražský systém" (Hlaváčová, 2009). Na tento slovník navazují nejrůznější další automatické nástroje. Slovník sám ovšem v zásadě neprošel výraznějšími úpravami. Kromě tohoto slovníku se v českém prostředí užívá slovník, jehož autorkou je Klára Osolsobě (Osolsobě, 1996), který je součástí automatických morfologických analyzátorů užívaných v řadě aplikací na FI MU, jako jsou např. morfologický analyzátor *ajka* (Sedláček, 2004), morfologický analyzátor *majka* (Šmerk, 2010). Odborná veřejnost (Hlaváčová, 2009) o něm hovoří jako o „brněnském systému".

3 Korpus SYN je *nereferenční* spojení textů všech *referenčních* synchronních psaných korpusů. V tomto článku je použita verze 3, zpřístupněná v lednu 2014 (více http://ucnk.ff.cuni.cz/syn.php).

4 K morfologickému značkování srov. http://wiki.korpus.cz/doku.php/seznamy:tagy.

není, příčina tkví a) v nekonzistenci a b) v neúplnosti slovníku používaného pro automatickou morfologickou analýzu (tagování).

V následující analýze budeme pracovat s vyhledávacím programem pro práci s korpusy KonText. Zvolíme-li „Typ dotazu cql", do dotazovacího řádku napíšeme dotaz ve formě [lemma="((.*oucí)|(.*ící))"] a podíváme se na frekvenční distribuci lemmat a morfologických značek, zjistíme, že existují lemmata zakončená na *-oucí/-ící*, která v korpusech řady SYN nejsou přiřazena ke tvarům označkovaným jako AG.*. Pomocí negativního filtru tedy odstraníme tvary označkované jako AG.*.

2.2 POZOROVÁNÍ RELEVANTNÍCH DAT A JEJICH ANALÝZA

Projdeme-li seznam relevantních lemmat[5] a budeme-li si všímat morfologického značkování, zjistíme, že se objevují jak lemmata relevantní pro výše stanovený cíl (sledování konzistence a pokrytí slovníku automatické morfologické analýzy), tak lemmata, která s tímto cílem nikterak nesouvisejí. Do první skupiny patří např. pravidelně utvořená adjektiva jako *plovoucí, mizející, visící,* do druhé skupiny bychom nepochybně zařadili adjektivní číslovku *tisící* nebo adjektivum *letoucí* v etymologické figuře tzv. hebrejského superlativu *leta letoucí*.

Lemmata zakončená na *-oucí/-ící*, která v korpusech řady SYN nejsou přiřazena tvarům označkovaným jako AG.*, lze roztřídit do následujících skupin:

a) substantivizovaná adjektiva (včetně kompozit) tvořená od sloves s oporou ve tvaru 3. os. pl. /přechodníku přítomného (např. substantiva *vedoucí, kolemjdoucí*);
b) kompozita s druhým členem adjektivním tvořeným od sloves s oporou ve tvaru 3. os. pl. / přechodníku přítomného (např. adjektiva *srdcervoucí, dechberoucí*);
c) adjektiva tvořená od substantiv v etymologické figuře hebrejského superlativu (např. *leta letoucí, pravda pravdoucí*[6]);
d) adjektiva tvořená od sloves, u nichž je opora ve tvaru 3. os. pl. / přechodníku přítomného synchronně zastřená nebo mají jiné nepravidelnosti (např. *žádoucí*, neboť synchronně pravidelně je utvořeno *žádající, horoucí*, neboť synchronně pravidelně je utvořeno *hořící*);
e) adjektiva náhodně zakončená řetězcem *-ící* (např. *tisící,* kompozitum *růžolící*);
f) adjektiva tvořená od sloves s oporou ve tvaru 3. os. pl. / přechodníku přítomného, která se používají převážně v dezaktualizovaných významech (např. *vroucí*)[7];
g) adjektiva tvořená od sloves s oporou ve tvaru 3. os. pl. / přechodníku přítomného (např. *visící*).

5 Dotaz [lemma="((.*oucí)|(.*ící))" & tag!="AG.*"] dává seznam 6312 různých lemmat a 6320 různých lemmat + pos.

6 Více Osolsobě (2013).

7 Na tomto místě se podrobněji dezaktualizovanými adjektivy nebudeme zabývat (více k této problematice viz Osolsobě, 2009). Užití procesuálních adjektiv v dezaktualizovaných významech spadá převážně do oblasti disambiguace, zatímco nás zajímá otázka konzistentního zpracování slovníku automatické morfologické analýzy.

V následující analýze se soustředíme na skupinu g), tedy na procesuální adjektiva pravidelně tvořená ze sloves, která zůstala opominuta při budování slovníku automatického morfologického analyzátoru.

Jejich seznam můžeme dále rozšířit cílenými sondami do dalších korpusů (webový korpus czTenTen12 a internet). Nebudeme se (v omezené míře) vyhýbat ani intuici rodilého mluvčího, z níž budeme vycházet v hodnocení potenciality jazyka na rovině slovotvorby.

2.3 KORPUSOVĚ ZALOŽENÁ ANALÝZA DAT

Z pozorování dat vyplývá, že adjektiva zakončená na *-oucí/-ící*, která v korpusech řady SYN nejsou označkována jako AG. *****, jsou poměrně často adjektiva tvořená od sloves s oporou ve tvaru 3. os. pl. / přechodníku přítomného, v případě, že opěrný tvar má variantu/dubletu/tripletu.

Příčinou tohoto stavu je, že příslušný opěrný tvar není rozgenerován na úrovni morfologického slovníku[8] a adjektivum je pak a) uloženo jako „obyčejné adjektivum" (má značku AA.*) nebo b) ve slovníku uvedeno není a na úrovni automatické morfologické analýzy je tvaru přidělena značka X.*****.[9]

Morfologické varianty jsou pro českou flexi typické. Varianty tvarů 3. osoby indikativu prézentu aktiva představují formálně popsatelný systém založený na principu třídění sloves podle kmene přítomného do slovesných tříd a podle kmene minulého ke slovesným vzorům (Komárek a kol., 1986). Varianty jsou podmíněny především přechody sloves mezi třídami a vzory, kolísáním mezi vzory/třídami a působením nejrůznějších analogií. Jistou roli sehrávají i posuny v hodnocení variant (kodifikace)[10].

Podrobněji zdokumentujeme nejlépe doložený případ, a sice adjektiva od sloves podle vzoru *sázet* a adjektiva kolísající mezi vzory *prosit/trpět/sázet*. Dále naznačíme, že typologicky stejný problém představuje existence variant v důsledku kolísání mezi vzory *brát/dělat* a *mazat/dělat* a variantnost koncovek i prézentních kmenů dalších vzorů sloves I. třídy podle kmene prézentního (*péci* a *umřít*).

8 Toto tvrzení lze opřít o fakt, že v řadě případů existuje korelace mezi značkováním variantního slovesného tvaru a značkováním adjektiva derivovaného s oporou v tomtéž tvaru. Uvedeme příklad značkování variantních tvarů 3. os. pl. slovesa *plešatět* v korpusu SYN: *… protože se nadýchali otravných plynů, z čehož hubnou, <plešatějí/plešatějí/X.*> …* a *Proč někteří muži <plešatí/plešatět/VB-P---3P-AA---I>*. Více též Osolsobě (2011, 2014).

9 Praxe automatické morfologické analýzy „pražský systém" řeší nedostatky v pokrytí slovníku automatického morfologického analyzátoru tak, že u tvarů nenalezených ve slovníku je jako lemma uveden tvar sám a ve značce je na první pozici **X** = neznámý, neurčený, neurčitelný slovní druh.

10 Významný kodifikační posun představuje změna hodnocení tvarů 3. os. pl. ind. préz. akt. vzoru *sázet* v Pravidlech českého pravopisu (Hlavsa a kol., 1993), viz též http://prirucka.ujc.cas.cz/.

KLANÍCÍ SE / KLANĚJÍCÍ SE

Tabulky 1, 2, 3 zachycují adjektiva derivovaná od sloves, která patří ke vzoru *sázet*, případně kolísají mezi vzory *prosit/trpět* a *sázet*. V důsledku tohoto kolísání, ať už je stav současné kodifikace jakýkoliv, existují slovotvorné varianty procesuálních adjektiv na *-cí* odpovídající variantám opěrného tvaru (tvar 3. osoby pl. / přechodníku přítomného). Jedná se o 33 dvojic (slovotvorných synonym / dublet), v jejichž značkování je rozpor. U 31 dvojic (tabulka 1) je jeden člen dvojice označkován AG.* a druhý buď AA.*, nebo X.*. U dvou dvojic je značkování sice ve shodě, obě adjektiva mají značku AA.* (tabulka 2), nicméně konzistentní řešení je, aby měla obě značku AG.*. Pouze u 13 dvojic (tabulka 3) je značkování konzistentní. Variantní lemmata jsou převážně z korpusů, a to zejména z korpusu SYN, dále z korpusu czTenTen12. Některé varianty jsou doloženy na internetu. Uvádíme i varianty, které pokládáme za možné z hlediska rodilého mluvčího, a označujeme je „?“.

Uvedené dvojice adjektiv (***ící/[ě]jící*** tabulka 1, 2, 3) jsme vesměs získali automaticky prostřednictvím nástroje Morfio (Cvrček — Vondřička, 2012)[11]. Mapují tudíž situaci na datech z korpusů SYN2010 a SYN2005. Značkování některých tvarů níže uvedených adjektiv v korpusu SYN2005 se odlišuje od značkování týchž tvarů v korpusu SYN, a to tak, že v korpusu SYN2005 je značka AA.*, nebo AG.* a v korpusu SYN je značka X.*. Značkování v korpusu SYN2005 se nezakládá na úpravě slovníku, ale na práci guesserů (více Jelínek, 2008), a tudíž na ně neupozorňujeme (nejsou uvedeny v tabulce 3). Některé dvojice jsme doplnili ruční analýzou dat z celého korpusu SYN.

bdít	*bdící*	AG.*	*bdějící*	X.*
bytnět	*bytnící*	AG.*	*bytnějící*	X.*
civět	*civící*	AG.*	*civějící*	AA.*
čnít	*čnící*	AG.*	*čnějící*	AA.*
čumět	*čumící*	AG.*	*čumějící*	AA.*
dřevnatět	*dřevnatící*	AG.*	*dřevnatějící*	AA.*
fičet	*fičící*	AG.*	*fičející*	X.*
hanět	*hanící*	AG.*	*hanějící*	X.*
hladovět	*hladovící*	X.*	*hladovějící*	AG.*
hovět (si)	*hovící*	AG.*	*hovějící*	X.*
hřmít/hřmět	*hřmící/hřmoucí*[12]	X.*/-	*hřmějící*	AG.*

11 http://morfio.korpus.cz/ba2PBDTG a http://morfio.korpus.cz/pISbe460

12 Tvar *hřmoucí* jsme našli doložen v korpusu CzTenTen12 (*... Po kilometru slyším v dálce <hřmoucí> lavinu ...*). Může jít o analogii k tvarům jako *skvoucí, horoucí, žádoucí ...*. Diachronně souvisí některé slovotvorné dublety (*horoucí/hořící, žádoucí/žádající*) s historickým vývojem slovesných tříd a vzorů (z původních 4 se vyvinulo současných 5, více Bauer — Lamprecht — Šlosar, 1986, s. 202). Kromě historického vývoje a stavu v dialektech může spolupůsobit i „tlak systému“ (*Systemzwang*), projevující se různými analogiemi.

chtít	*chtící/ ?chcoucí*	AG.*/ czTenTen12[13]	*chtějící*	X.*
chvít/chvět	*chvící*	X.*	*chvějící*	AG.*
churavět	*churavící*	X.*	*churavějící*	AG.*
chybět	*chybící*[14]	AA.*	*chybějící*	AG.*
kálet	*kálící*	X.*	*kálející*	AG.*
kamenět	*kamenící*	X.*	*kamenějící*	AG.*
klanět (se)	*klanící*	AG.*	*klanějící*	AA.*
koulet (se)	*koulící*	X.*	*koulející*	AG.*
kráčet	*kráčící*	X.*	*kráčející*	AG.*
kvílet	*kvílící*	X.*	*kvílející*	AG.*
kypět	*kypící*	AG.*	*kypějící*	X.*
lpět	*lpící*	AG.*	*lpějící*	X.*
míjet	*míjící*	AA.*	*míjející*	AG.*
mizet	*mizící*	AG.*	*mizející*	AA.*
mohutnět	*mohutnící*	AG.*	*mohutnějící*	AA.*
pelášit	*pelášící*	AG.*	*pelášející*	X.*
plešatět	*plešatící*	AG.*	*plešatějící*	X.*
pouštět	*pouštící*	X.*	*pouštějící*	AG.*
rezavět[15]	*rezavící*	X.*	*rezavějící*	AG.*
růžovět	*růžovící*	X.*	*růžovějící*	AG.*
stavět	*stavící*	AG.*	*stavějící*	X.*
strmět/strmit	*strmící*	AA.*	*strmějící*	X.*
supět	*supící*	AG.*	*supějící*	AA.*
svádět	*svádící*	X.*	*svádějící*	AG.*
šedivět	*šedivící*	X.*	*šedivějící*	AG.*
šílet	*šílící*	X.*	*šílející*[16]	AA.*/AG.*
šumět	*šumící*	AG.*	*šumějící*	X.*
temnět	*temnící*	AG.*	*temnějící*	X.*

13 Okrajově se objevují i doklady jako např. *… se jenom tak od oslavy na chvilku ztratila mezi dav <chcoucí> podpisy a fotky …*, které sice porušují současnou kodifikaci, ale lze je zdůvodnit jak historicky (srov. též dnešní slovenské *chcúci*), tak analogií (dublety na *-oucí/-ící* u dalších nepravidelných sloves jako *vidět, vědět, dát*).

14 Např. *… Křížová cesta by vyžadovala postavení pomníčku a doplnění <chybících> obrazů u několika zastavení …* Adjektivum *chybící* od slovesa *chybět* je ovšem homonymní s adjektivem tvořeným pravidelně od slovesa *chybit*:*… deset hlavní ručnic, nikdy <nechybících> , hledí na ně.* Takových případů není mnoho. Jsme si ovšem vědomi toho, že jejich konzistentní zaznamenání přinese další problémy na rovině disambiguace.

15 V korpusu SYN je doloženo adjektivum *rezivějící/AG.**, máme však za to, že správně lze utvořit též *rezivící* (srov. slovesný tvar z korpusu SYN *… Značky totiž <reziví> a blednou …*) podobně jako k *děravějící/AG.** je pravidelně tvořeno *děravící* (*… Nikde vrchcábe mé víry <děravící> …*).

16 Chyba v lemmatizaci: tvary *šílející.** jsou částečně lemmatizovány správně jako *šílející* a mají značku AG.*, částečně jsou ale lemmatizovány chybně lemmatem *„šílelící“* a mají značku AA.*.

válet	*válící*	X.*	*válející*	AG.*
večeřet	*večeřící*	AG.*	*večeřející*	X.*
viset	*visící*	AA.*	*visející*	AG.*
vjíždět	*vjíždící*	X.*	*vjíždějící*	AG.*
vonět	*vonící*	AG.*	*vonějící*	X.*
závidět	*závidící*	AG.*	*závidějící*	X.*
želet	*želící*	AG.*	*želející*	X.*

TABULKA 1

nenávidět	*nenávidící*	AA.*	*nenávidějící*	AA.*
souznít/souznět	*souznící*	AA.*	*souznějící*	AA.*

TABULKA 2

náležet	*náležící*	AG.*	*náležející*	AG.*
přináležet	*přináležící*	AG.*	*přináležející*	AG.*
příslušet	*příslušící*	AG.*	*příslušející*	AG.*
řeřavět	*řeřavící*	AG.*	*řeřavějící*	AG.*
slzet	*slzící*	AG.*	*slzející*	AG.*
souviset	*souvisící*	AG.*	*související*	AG.*
svářet (se)	*svářící*	AG.*	*svářející*	AG.*
svrbět	*svrbící*	AG.*	*svrbějící*	AG.*
úpět	*úpící*	AG.*	*úpějící*	AG.*
vláčet	*vláčící*	AG.*	*vláčející*	AG.*
záležet	*záležící*	AG.*	*záležející*	AG.*
záviset	*závisící*	AG.*	*závisející*	AG.*
znít/znět	*znící*	AG.*	*znějící*	AG.*

TABULKA 3

Tvary registrované v tabulkách 1, 2 a 3 představují neúplný seznam sloves, u kterých jsou doloženy (korpusy, internet) slovotvorné varianty zkoumaného typu. Tento seznam lze dále rozšířit, např. na základě údajů v *Internetové jazykové příručce* (http://prirucka.ujc.cas.cz/), popřípadě na základě seznamu sloves extrahovaných ze slovníku automatických morfologických nástrojů užívaných na FI MU (viz výše).

Podobným způsobem lze postupovat i v dalších případech. V tabulce 4 uvádíme dva příklady slovotvorných variant adjektiv tvořených od sloves, která kolísají mezi vzory *brát* a *dělat*.

	-oucí		-ající	
klepat	*klepoucí*	X.*	*klepající*	AG.*
plavat	*plavoucí*	AA.*	*plavající*	AG.*

TABULKA 4

V tabulce 5 jsou příklady slovotvorných variant adjektiv tvořených od sloves, která kolísají mezi vzory *mazat* (tato slovesa mívají dubletu opěrného tvaru *oni pláčou / oni pláčí*) a *dělat*.

	-oucí		-ící		-ající	
plakat	*pláčoucí*	int[17]/?	*pl[aá]čící*	AA.*/X.*	*plakající*	AG.*
mazat	*mažoucí*	int[18]	*mažící*	czTenTen12[19]	*mazající*	AG.*

TABULKA 5

V tabulce 6 uvádíme příklady slovotvorných variant adjektiv tvořených od sloves patřících ke vzorům *péci* a *umřít*, která mívají dubletu opěrných tvarů (oni *mrou/mřou, pekou/pečou*).

	.*roucí		.*řoucí		.*řící	
dřít	*droucí*	AG.*	*dřoucí*	X.*	*dřící*	AG.*
mřít	*mroucí*	AG.*	*mřoucí*	czTenTen12[20]	*mřící*	czTenTen12[21]
	.*[kh]oucí		.*[čž]oucí		.*čící	
péci	*pekoucí*	AG.*	*pečoucí*	int.[22]	*pečící*[23]	AG.*
síci/ séci	*sekoucí*	czTenTen12[24]	*sečoucí*	int.[25]	*sečící*	AG.*

TABULKA 6

3. ZÁVĚR

Na základě pozorování dat získaných z označkovaného korpusu jsme ukázali typické případy nekonzistencí a nedostatečného pokrytí morfologického značkování procesuálních adjektiv.

Diagnostikovali jsme hlavní příčinu nedostatků automatické morfologické analýzy, která tkví v nedostatečné reflexi tvarových dublet opěrného tvaru pro tvoření

17 *… a TV Hamás posílá AL Jazire záběry <pláčoucích> žen a dětí.*

18 *… nebo naopak všude lepící, ale nedržící tekoucí a vše <mažoucí> hmota …*

19 *… troufl bych si odhadnout, že uživatelé <mažící> po sobě jednotlivé příspěvky, jsou naprostou menšinou …*

20 *… Představa ovšem hladem <mřoucího> děcka, …*

21 *… a pod nohama by mi tančila rudá barva <mřícího> sluníčka.*

22 *… Táborák, pivko, pohodička, vůně <pekoucího> se masa.*

23 Často použito chybně na místě účelového adjektiva *pečicí*. Najdou se ovšem i korektní užití: *… vůně <pečících> se špízů ….*

24 *… ještě skřípějí fůry s úrodou, ještě zvoní kosy <sekoucí> otavu;*

25 *… osamocení zaměstnanci <sečoucí> trávu křovinořezem nebo motorovou sekačkou …*

procesuálních adjektiv v češtině na rovině slovníku automatického morfologického analyzátoru. Adjektiva na *-oucí/-ící,* tvořená pravidelně od sloves s oporou ve tvaru 3. osoby plurálu prézentu / přechodníku přítomného, jsou na rovině slovníku automatického analyzátoru popsána v rámci tzv. rozgenerování tvarů od příslušného slovesného základu. Pravidla pro rozgenerování tvarů nezachycují všechny potence jazyka (konkrétně možnost tvořit příslušná adjektiva od variantních opěrných tvarů). Jenom v menším množství případů jsou zachyceny a adekvátně označkovány slovotvorné varianty (*související/AG.** a *souvisící/AG.**), v mnoha případech frekventovaných adjektiv tvořených variantně je jedna z variant označkována jako AG.***** (*mizící/AG.**), zatímco u druhé je značka AA.***** (*mizející/AA.**). V řadě případů je jedna z variant správně označkována (*lpící/AG.**) a druhá není rozpoznána automatickou morfologickou analýzou (*lpějící/X.**).

Na základě analýzy a) nekonzistencí a b) nedostatečného pokrytí slovotvorných dublet na úrovni slovníku lze navrhnout cílená vylepšení slovníku automatického morfologického analyzátoru (pravidla pro „rozgenerování" tvarů příslušných adjektiv i jejich interpretací na úrovni morfologického slovníku). Po implementaci navržených úprav do slovníku automatického morfologického analyzátoru lze předpokládat, že se a) zvýší konzistence morfologického značkování, b) opraví některé drobné chyby ve značkování i lemmatizaci adjektiv na *-oucí/-ící*, c) rozšíří pokrytí slovníku a d) neměl by narůst (až na několik drobností[26]) počet homonymních tvarů nabízených k disambiguaci.

Lze očekávat, že výsledky tagování, které má nyní k dispozici běžný uživatel korpusu, budou v důsledku námi navržených úprav transparentnější, než tomu bylo dosud. Zobecnění pozorování korpusových dat může pomoci k formulování přesnějších pravidel popisu variant adjektiv na *-oucí/-ící* nejen pro potřeby popisu těchto variant na poli automatické morfologické analýzy češtiny, ale může být východiskem i pro kodifikační doporučení opřená o zjištění stavu úzu reprezentovaného rozsáhlými korpusovými daty[27].

POUŽITÉ KORPUSY

Křen, M. — Čermák, F. — Hlaváčová, J. — Hnátková, M. — Jelínek, T. — Kocek, J. — Kopřivová, M. — Novotná, R. — Petkevič, V. — Procházka, P. — Schmiedtová, V. — Skoumalová, H. — Šulc, M. (2014): *Korpus SYN*, 27. 1. 2014. Praha: Ústav Českého národního korpusu FF UK. Dostupný z: http://www.korpus.cz

FI MU — *czTenTen12*. Centrum zpracování přirozeného jazyka FI MU, Brno. Dostupný z: http://ske.fi.muni.cz/bonito

26 Homonymie se týká adjektiv *dající < dát/dát se, chybící < chybit/chybět, jedoucí < jet/jíst, pářící < pářit (se)/párat (se), smějící < smět/smát se, stávající < stávat/stávat se, vážící < vážit (se)/vázat (se)*.

27 Žádná kodifikační doporučení analyzovaného slovotvorného typu v kodifikačních příručkách zahrnuta nejsou.

LITERATURA

BAUER, J. — LAMPRECHT, A. — ŠLOSAR, D. (1986): *Historická mluvnice češtiny*. Praha: SPN.

CVRČEK, V., VONDŘIČKA, P. (2012): Morfio. Dostupný z: http://morfio.korpus.cz

DOKULIL, M. a kol. (1986): *Mluvnice češtiny 1.* Praha: Academia.

HAJIČ J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum Charles University Press.

HLAVÁČOVÁ, J. (2009): *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. Disertační práce. Praha: UK.

HLAVSA, Z. a kol. (1993): *Pravidla českého pravopisu (PČP)*. Praha: Academia.

Internetová jazyková příručka. Dostupná z: http://prirucka.ujc.cas.cz.

JELÍNEK, T. (2008): Nové značkování v Českém národním korpusu. *Naše řeč*, 91, 1, s. 13–20.

KOMÁREK, M. a kol. (1986): *Mluvnice češtiny 2.* Praha: Academia.

OSOLSOBĚ, K. (1996): *Algoritmický popis české formální morfologie a strojový slovník češtiny*. Disertační práce. Brno: MU.

OSOLSOBĚ, K. (2009): Kajícný a nevěřícný — adjektiva na *-cí/-cný*: slovníky, gramatiky, korpusy. In: D. HLAVÁČKOVÁ — A. HORÁK — K. OSOLSOBĚ — P. RYCHLÝ (eds.), *After Half a Century of Slavonic Natural Language Processing*, Brno: Masarykova univerzita, s. 173–183.

OSOLSOBĚ, K. (2011): *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny*. Brno: MU.

OSOLSOBĚ, K. (2013): Korpusy a internet jako zdroje dat pro výzkum produktivity periferního slovotvorného typu: adjektiva typu *hrůzoucí (hrůza)* v korpusech a na internetu. *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.

OSOLSOBĚ, K. (2014): *Česká morfologie a korpusy*. Praha: Karolinum.

PETKEVIČ, V. (2006): Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In: M. ŠIMKOVÁ (ed.), *Insight into the Slovak and Czech Corpus Linguistics*. Bratislava: Veda, s. 26–44.

SEDLÁČEK, R. (2004): *Morphematic analyser for Czech.* Disertační práce. Brno : FI MU.

SPOUSTOVÁ, D. — HAJIČ, J. — VOTRUBEC, J. — KRBEC, P. — KVĚTOŇ, P. (2007): The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. Praha: ACL, s. 67–74.

ŠMERK, P. (2010): *K počítačové morfologické analýze češtiny*. Disertační práce. Brno: FI MU.

Klára Osolsobě | Ústav českého jazyka, FF MU | Arna Nováka 1, 602 00 Brno
osolsobe@phil.muni.cz