Klára Osolsobě (Brno) – Hana Žižková (Brno)

HOMONYMIE MEZI APELATIVY A PROPRII JAKO PROBLÉM AUTOMATICKÉ MORFOLOGICKÉ ANALÝZY ČEŠTINY[1]

HOMONYMY AMONG CZECH COMMON AND PROPER NOUNS AS THE PROBLEM OF AUTOMATIC MORPHOLOGICAL ANALYSIS

The aim of this paper is to provide a corpus-based analysis of one type of Czech proper nouns (type *Zubří*). We will argue that the adequate annotation (lemmatisation and morphological tagging) of proper nouns type *Zubří* depends on several circumstances:
1) the coverage of the dictionary of the automatic analyser;
2) the accurate description of the variability of inflexion forms;
3) the non-trivial disambiguation of numerous homonymous word forms.
We believe that while meeting the first two conditions is possible, the adequate disambiguation goes beyond the possibilities of automatic morphological analysis.

Key words
toponyms; tokenisation; lemmatisation; disambiguation; corpus linguistics

Klíčová slova
toponyma; tokenizace; lemmatizace; desambiguace; korpusová lingvistika

Úvod

Jazykové korpusy slouží již více než čtvrt století v české lingvistice jako zdroje poznání fungování jazyka (češtiny). Jedním ze základních požadavků na jazykový korpus v moderním slova smyslu je vedle elektronického uložení a přístupnosti, reprezentativního rozsahu a obsahu požadavek na kvalitní anotace. V tomto příspěvku chceme upozornit na problémy, s nimiž se potýká korpusová lingvistika při budování nástrojů automatické morfologické analýzy. Představíme problém adekvátního lingvistického popisu jednoho typu proprií, a sice deadjektivních substantiv tvořených z druhově posesivních adjektiv, a to a) konverzí a b) substantivizací provázenou změnou flektivního typu *jarní* na typ *stavení*.

Tři kroky automatické analýzy

Automatická morfologická analýza je jedním z typů zpracování textů v rámci počítačového zpracování přirozeného jazyka (Natural Language Processing – NLP),

[1] Text vznikl za podpory projektu Čeština v jednotě synchronie a diachronie – 2020

potažmo korpusové lingvistiky. Zahrnuje tři kroky. Prvním je **tokenizace**, tedy vyčlenění jednotek – tokenů, které odpovídají přibližně grafickým slovům na straně jedné a interpunkčním znakům na straně druhé. Druhým je **přiřazení lingvistických interpretací** (většinou na základě porovnání se slovníkem automatického morfologického analyzátoru, dále jen slovníkem) ve formě **lemmatu** (reprezentativní tvar, většinou nominativ singuláru jmen, infinitiv sloves) a **tagu** (značka zahrnující informace o slovním druhu a dalších slovnědruhově závislých gramatických významech na základě předem navrženého systému značek). Pokud je kombinací lemma + tag vícero (analyzovaný slovní tvar je slovnědruhově a/nebo morfologicky víceznačný/homonymní, a interpretací přiřazených na základě slovníku je tudíž více), následuje jako třetí krok tzv. **desambiguace**. Při desambiguaci je vybrána jediná, pokud možno kontextově správná, interpretace. Automatická morfologická analýza pracuje pouze s formálním kritériem určení slovního druhu, kritérium syntaktické a sémantické bývá někdy použito při desambiguaci, pravidla jsou ovšem mnohdy obtížně formalizovatelná. V případě, že pravidla nelze aplikovat, je desambiguace opřena o statistické metody a nejnověji i o využití neuronových sítí.

Všechny tři uvedené kroky jsou prováděny strojově a mají vliv na výslednou podobu morfologické interpretace v podobě lemmatu a tagu, a tudíž i na veškerý lingvistický výzkum korpusu, který s výsledky automatické morfologické analýzy jakkoli pracuje.

Tokenizace a propria

Tokenizace vlastních jmen úzce souvisí s velmi rozsáhlou oblastí automatického zpracování tzv. MWE (Multiwords Expressions – víceslovných výrazů) jako např. *Karlovy Vary, Nové Město nad Metují, Kostelec nad Černými lesy,* …, kterou v rámci této studie okrajově zmíníme, a to v souvislosti s názvy typu *Medvědí hora/kámen/román/pohádka*. Tokenizéry většinou pracují pouze s jednoslovnými jednotkami. Analýza víceslovných jednotek bývá buď zcela vyloučena, nebo je prováděna samostatným analytickým modulem.

Lemmatizace, morfologické značkování a propria

Druhý krok automatické morfologické analýzy je lemmatizace a interpretace na rovině morfologické značky. Tvar izolovaný při tokenizaci je porovnán se slovníkem. Už v tomto kroku automatické morfologické analýzy může dojít k chybě. Důvodem bývá nedostatečné pokrytí slovníku, takže interpretace tvaru neodpovídá potenci jazyka, kterou prezentuje korpus. Například tvar, který může být interpretován jak jako tvar apelativního adjektiva (<***baraní***> *střeva*), tak jako substantivizované proprium (*osada* <***Baraní***>), je ve slovníku interpretován výhradně

jako proprium. V takových případech automatická analýza selže z důvodů nedostatečného pokrytí slovníku (tvar s malým písmenem není rozpoznán automatickou morfologickou analýzou). Nedostatky v pokrytí slovníku mohou být ovšem ještě komplikovanější. I tvarům psaným s velkým počátečním písmenem, které jsou morfologickými adjektivy a jsou součástí víceslovného propria (<***Baraní/N***[2]> *sedlo/kopa*), je chybně přiřazena substantivní interpretace. O nedostatečném pokrytí slovníku svědčí nerozpoznané tvary adjektivní flexe (<***Baraního/X***[3]> *sedla*).

Pokud je ve slovníku interpretovaný tvar doložen/interpretován pouze jako proprium, pak lemma začíná velkým písmenem, takže např. tvaru *Kožlí* přiřadí automatická morfologická analýza lemma *Kožlí*. Proprium je následně adekvátně interpretováno i na rovině morfologické značky jako substantivum. V případě propria *Kožlí* jde o jedinou a správnou interpretaci. Výsledky analýzy jsou v pořádku.

Pokud je interpretovaný tvar doložen jak jako apelativum (adjektivum), tak jako proprium (substantivum), následuje další krok automatické analýzy – desambiguace.

Desambiguace a propria

Některá lemmata jednoslovných substantivních proprií, která se podle platné kodifikace píší s velkým počátečním písmenem, jsou uložena ve slovníku s velkým počátečním písmenem. Výsledky desambiguace tj. přiřazení jednoznačné interpretace v případě, že existuje více řešení (např.: <***Zaječí***> *úmysly mě opustily.* × <***Zaječí***> *osolíme a opepříme.* × <***Zaječí***> *je obec s dlouhou vinařskou tradicí.* × <***Zaječí***> *pokaždé, když šlápne do vody.*), nemusí být zdařilé. V uvedených případech je analyzovaný tvar na začátku věty, píše se tudíž s velkým počátečním písmenem, a nelze tedy na základě pravopisných vlastností budovat rozumné desambiguační pravidlo. Jednoduché řešení nemusí nastat ani v případě, že se tvar s velkým písmenem objeví uvnitř věty – formálně tehdy, když nenásleduje po znaku „.“ nebo „!“ nebo „?“ nebo „:“ (*Projekt lipové aleje pod* <***Zaječím***> *vrchem jsme začali realizovat už zhruba před dvěma lety.* × *Dole pod* <***Zaječím***> *u nádrže Nové Mlýny pak první koupání.*). Při lemmatizaci a slovnědruhovém značkování proprií je tedy poměrně nesnadně řešitelný problém lemmatizace proprií, která v jazyce figurují (mohou figurovat) jako jednoslovná substantivní propria i jako adjektivní součást víceslovného propria (adjektivum + substantivum: *Zaječí vrch*).

[2] N je značka slovního druhu (POS – Part of Speech) pro substantivum.
[3] X je značka slovního druhu (POS) přiřazená slovu, které není ve slovníku uloženo.

Propria z druhově posesivních/široce vztahových adjektiv od názvů zvířat

V naší studii se ovšem budeme věnovat pouze jedné z otázek, která souvisí s homonymií apelativum/proprium. Na základě analýzy vybraných proprií doložených v korpusech ČNK (pracovali jsme s korpusem SYN v6, Křen a kol, 2017), konkrétně proprií z druhově posesivních/široce vztahových adjektiv od názvů zvířat, se budeme snažit doplnit seznamy výjimek, které mohou pomoci zabránit chybným interpretacím na úrovni automatické morfologické analýzy. Příspěvek lze chápat jako doplněk obrazu morfologických zvláštností proprií (viz též Osolsobě, 2008).

Materiálová základna

Pro potřeby naší analýzy jsme provedli excerpci knihy *Zeměpisná jména v Čechách, na Moravě a ve Slezsku: slovník vybraných zeměpisných jmen s výkladem jejich původu a historického vývoje* (Lutterer, Šrámek, 2004), dále jsme pracovali se seznamem kostelů[4] a se seznamem obcí ČR,[5] k excerpci jsme využili také portál www.mapy.cz a na závěr jsme provedli korpusovou sondu. Zvolili jsme korpus SYN v6 ČNK, což byl v době konání šetření největší korpus řady SYN, a vybrali jsme celkem 28 proprií, která jsou kandidáty na sledovaný typ (*Baraní, Hovězí, Jazevčí, Jelení, Jestřabí, Jestřebí, Kančí, Kobylí, Kozí, Krahulčí, Kunčí, Medvědí, Ovčí, Pstruží, Račí, Rybí, Sokolí, Srní, Sýkoří, Ščučí, Štičí, Telecí, Veveří, Vlčí, Vůsí, Vydří, Zaječí, Zubří*). Nezařazena: *Medvězí, Nedvězí, Kožlí, Podkozí, Povydří, Ptení, Rabí, Ryboví, Rybitví, Vrábí*.

Morfologické vlastnosti proprií tvořených z druhově posesivních/široce vztahových adjektiv od názvů zvířat

Předpokladem pro fungování automatické morfologické analýzy vybraného typu proprií jsou zásadní následující faktory: a) doplnění lemmat proprií, která chybí ve slovníku automatického morfologického analyzátoru a rozgenerování tvarů podle typu flexe v úzu;[6] b) doplnění lemmat apelativ, která chybí ve slovníku a rozgenerování tvarů podle typu flexe v úzu;[7] c) implementace desambiguačních pravidel.

[4] Dostupný z: www.kostelycz.cz/seznam.htm.

[5] Dostupný z: http://wwwinfo.mfcr.cz/ares/obce/ares_obce.html.cz.

[6] Neplatí jednoznačné pravidlo adjektivní užití = adjektivní flexe (*jarní*), substantivní užití = substantivní flexe (*stavení*). Viz Internetová jazyková příručka poznámka k heslu *Kobylí* a dalším (http://prirucka.ujc.cas.cz/?slovo=Kobyl%C3%AD).

[7] Korpusové analýzy ukazují, že k některým propriím evidovaným ve slovníku existují homonymní apelativa, která slovník nezaznamenává.

Korpusová sonda

Cílem korpusové sondy je zmapovat: a) zastoupení sledovaných slovních tvarů v korpusech řady SYN; b) jejich interpretace na rovině automatické morfologické analýzy a c) jejich morfologické vlastnosti, jak se zrcadlí v reálných užitích v českých psaných textech.

Na základě pozorování korpusových dat si klademe za cíl navrhnout doplnění slovníku,[8] popřípadě formulovat doporučení, která by se mohla stát podkladem pro pravidlovou desambiguaci.

Jako nejméně zatížený postup analýzy korpusových dat se nám technicky jevil dotaz na tvary vybraných proprií tvořených z druhově posesivních/široce vztahových adjektiv od názvů zvířat, který zahrnuje jak tvary podle adjektivní flexe vzoru *jarní*, tak tvary podle substantivní flexe typu *stavení* a odhlíží přitom od interpretací na rovině lemmatu a tagu, které do korpusu vkládá automatická morfologická analýza.

Dotaz formulovaný v jazyce CQL (Corpus Query Language) zní:

[lc="(baran|hověz|jazevč|jelen|jestř[ae]b|kanč|kobyl|koz|krahulč|kunč|medvěd|ovč|pstruž|rač|ryb|sokol|srn|sýkoř|telec|ščuč|štič|veveř|vlč|vůs|vydř|zaječ|zubř)(í|ího|ímu|ím|ích)"]

Tabulka 1 ukazuje frekvenční analýzu lemmat *hovězí/Hovězí* v závislosti na počátečním velkém/malém písmenu lemmatu, slovního tvaru a slovnědruhové interpretaci v korpusu SYN v6.

Tabulka 1

lemma	počáteční písmeno	POS	Frekvence
hovězí	h	A	40507
		N	17415
	H	A	5238
		N	1790
Hovězí	H	N	5966

Tvary nezaznamenané na rovině slovníku lze získat jednoduchým dotazem za použití pozitivního filtru a značky pro nerozpoznané tvary (**[pos="X"]**). Jejich přehled uvádí Tabulka 2.

[8] Analyzátor používaný pro anotace korpusů řady SYN pracuje se slovníkem *MorfFlex* (Hajič, Hlaváčová, 2013). V rámci projektu NOVAMORF (Osolsobě et al., 2017) se počítá mj. s revizí uvedeného slovníku.

Tabulka 2

Lemma	Slovní tvar	POS	Frekvence
baraní	baraní	X	12
Baraního	Baraního	X	2
baraního	baraního	X	1
baraních	baraních	X	2
baraním	baraním	X	1
Baranímu	Baranímu	X	1
jestřabí	jestřabí	X	2
jestřebí	jestřebí	X	7
jestřebích	jestřebích	X	2
krahulčí	krahulčí	X	1
Krahulčího	Krahulčího	X	1
Kunčí	Kunčí	X	179
KUNČÍ	KUNČÍ	X	1
Kunčího	Kunčího	X	4
Kunčím	Kunčím	X	4
Ščučí	Ščučí	X	38
Ščučím	Ščučím	X	2
Vůsí	Vůsí	X	54
VŮSÍ	VŮSÍ	X	3

Tvary zaznamenané na rovině slovníku interpretované jako substantiva lze získat jednoduchým dotazem za použití pozitivního filtru a značky pro substantiva (**[pos="N"]**). Jejich přehled uvádí Tabulka 3.

Tabulka 3

Lemma	Slovní tvar	POS	Frekvence
Baraní	Baraní	N	66
Baraní	Baraních	N	3
Baraní	Baraním	N	5
hovězí	hovězí	N	7798
hovězí	Hovězí	N	1575
Hovězí	Hovězí	N	5428
hovězí	HOVĚZÍ	N	63
Hovězí	HOVĚZÍ	N	389
hovězí	hovězího	N	7698
hovězí	Hovězího	N	74
hovězí	HOVĚZÍHO	N	24
hovězí	hovězím	N	1596

hovězí	Hovězím	N	43
Hovězí	Hovězím	N	141
Hovězí	HOVĚZÍM	N	7
hovězí	HOVĚZÍM	N	2
hovězí	hovězímu	N	323
hovězí	Hovězímu	N	4
hovězí	HOVĚZÍMU	N	5
Jazevčí	Jazevčí	N	6
Jeleň	Jeleních	N	2
Jestřebí	Jestřebí	N	12283
Jestřebí	JESTŘeBÍ	N	1
Jestřebí	JESTŘEBÍ	N	344
Jestřebí	Jestřebích	N	1
Jestřebí	Jestřebím	N	270
Jestřebí	JESTŘEBÍM	N	1
Kančí	Kančí	N	17
Kančí	KANČÍ	N	2
Kančí	Kančího	N	2
Kančí	Kančímu	N	1
Kobylí	Kobylí	N	4
Krahulčí	Krahulčí	N	1342
Krahulčí	KRAHULČÍ	N	92
Krahulčí	Krahulčím	N	37
Rybí	Rybí	N	1
Srní	Srní	N	5647
Srní	SRní	N	2
Srní	SRNÍ	N	117
Srní	Srním	N	174
Srní	SRNÍM	N	1
telecí	telecí	N	1381
telecí	Telecí	N	771
telecí	TELECÍ	N	27
telecí	telecího	N	767
telecí	Telecího	N	166
telecí	TELECÍHO	N	5
Telek	TELECÍCH	N	1
telecí	telecím	N	211
telecí	Telecím	N	290
telecí	TELECÍM	N	5
telecí	telecímu	N	124
telecí	Telecímu	N	1

telecí	TELECÍMU	N	1
Zaječí	Zaječí	N	482
Zaječí	ZAJEČÍ	N	3
Zaječí	Zaječích	N	1
Zaječí	Zaječím	N	55
Zubří	Zubří	N	44207
Zubří	ZUbří	N	1
Zubří	ZUBŘÍ	N	1420
Zubří	Zubřím	N	2521
Zubří	ZUBŘÍM	N	4

Tvary zaznamenané na rovině slovníku interpretované jako adjektiva lze získat jednoduchým dotazem za použití pozitivního filtru a značky pro adjektiva (**[pos="A"]**). Přehled tří nejfrekventovanějších lemmat uvádí Tabulka 4.

Tabulka 4

Lemma	Slovní tvar	POS	Frekvence
hovězí	hovězí	A	19149
hovězí	hověZí	A	1
hovězí	Hovězí	A	4692
hovězí	HoVěZí	A	1
hovězí	HOVĚZÍ	A	388
hovězí	hovězího	A	17105
hovězí	Hovězího	A	59
hovězí	HOVĚZÍHO	A	21
hovězí	hovězích	A	1256
hovězí	Hovězích	A	14
hovězí	HOVĚZÍCH	A	3
hovězí	hovězím	A	2345
hovězí	Hovězím	A	11
hovězí	HOVĚZÍM	A	27
hovězí	hovězímu	A	521
hovězí	Hovězímu	A	14
hovězí	HOVĚZÍMU	A	1
jelení	jelení	A	6309
jelení	jElEní	A	1
jelení	Jelení	A	12263
jelení	JeLenÍ	A	1
jelení	JeLeNÍ	A	1
jelení	JELENÍ	A	303

jelení	jeleního	A	613
jelení	Jeleního	A	882
jelení	JELENÍHO	A	4
jelení	jeleních	A	561
jelení	Jeleních	A	655
jelení	JELENÍCH	A	5
jelení	jelením	A	366
jelení	Jelením	A	1058
jelení	JELENÍM	A	6
jelení	jelenímu	A	38
jelení	Jelenímu	A	51
Jestřabí	Jestřabí	A	3606
Jestřabí	JESTŘABÍ	A	78
Jestřabí	Jestřabího	A	3
Jestřabí	Jestřabích	A	1
Jestřabí	Jestřabím	A	67
Jestřebí	Jestřebí	A	1731
Jestřebí	JestŘeBÍ	A	1
Jestřebí	JESTŘEBÍ	A	147
Jestřebí	Jestřebího	A	8
Jestřebí	Jestřebích	A	1150
Jestřebí	JESTŘEBÍCH	A	5
Jestřebí	Jestřebím	A	24
Jestřebí	JESTŘEBÍM	A	1
Jestřebí	Jestřebímu	A	4

Tvary zaznamenané na rovině slovníku interpretované jako jiný slovní druh než substantiva a adjektiva lze získat jednoduchým dotazem za použití negativního filtru (**[pos="[XNA]"]**). Jejich přehled uvádí Tabulka 5.

Tabulka 5

Lemma	Slovní tvar	POS	Frekvence
zaječet	zaječím	V	54
zaječet	Zaječím	V	26
zaječet	zaječí	V	20
zaječet	Zaječí	V	1

Při zapojení lingvistické fantazie lze ovšem předpokládat, že uvedená data nemusí odpovídat skutečnosti, neboť jsou výsledkem automatické morfologické analýzy se všemi problémy uvedenými výše. Příkladem budiž jednoznačná adjektivní

interpretace tvaru *vlčí*, a to navzdory výskytům typu (*Při požáru domu ve <**Vlčí/A**[9]> na Plzeňsku se popálil muž a jeho paní. Ve městě je hůř „Děti <**vlčí/A**>," lamentuje moje babička, když slyší, co se děje (nejen) ve školách.*). Přesto lze uvedené statistiky použít k formulaci rámcových závěrů korpusové analýzy.

Tvary nezaznamenané ve slovníku

Ukazuje se, že několik proprií zkoumaného typu není ve slovníku zastoupeno. Nepřekvapí patrně, že pouze propriální výskyt v korpusech mají substantiva *Kunčí, Ščučí* a *Vůsí*, jejich doplnění do slovníku není problematické.

Důležitější informaci lze ovšem získat z korpusů o slovech *Baraní/baraní, jestřebí* a *jestřabí*. Slovník interpretuje pouze tvary *Baraní, Jestřebí* a *Jestřabí*, a to jako substantivní propria s flexí podle typu *stavení*. Analyzovaná korpusová data ovšem dokládají jak tvary adjektiva *baraní* (*U odborné poroty nakonec zvítězil javořinský kotlíkový guláš, druhé místo získal <**baraní**> guláš Lukašovič a bronz si odvezl tým Fešáků za hovězí klasiku.*), tak výskyt substantivního propria *Baraní* s flexí podle typu *jarní* (*Už za války byla na <**Baraním**> vytýčena hranice mezi Protektorátem Čechy a Morava a Slovenským státem.*). Korpusová sonda dále ukazuje řadu dokladů adjektiva ve víceslovných pojmenováních propriálního charakteru (*Tento projekt obsahuje hráz na řece Šišemce, poldr a úpravu terénu u <**Baraního** kopce>, což částečně zabrání přívalu vody.*). Adjektivum ve dvouslovném propriu se píše s velkým písmenem, morfologicky je ovšem správné je interpretovat jako adjektivum, čemuž odpovídá i adjektivní typ flexe (*jarní*). V případě doplnění lemmat *Baraní/N* a *baraní/A* do slovníku je třeba počítat se všemi doloženými tvary. Automatická desambiguace (s ohledem na vysoký počet homonymních tvarů typu *jarní* i typu *stavení*) bude patrně velmi obtížná, ne-li nemožná.

Problematické interpretace ve slovníku

Zatímco tvary *jestřebí* a *jestřabí* nejsou rozpoznány automatickou morfologickou analýzou, tvary *Jestřebí/Jestřabí* jsou interpretovány jak jako substantiva, tak jako adjektiva.

Podobně tvar *Krahulčího* není rozpoznán automatickou morfologickou analýzou, přestože lemma *Krahulčí* se substantivní flexí podle *stavení* je automatickou morfologickou analýzou rozpoznán a interpretován jako substantivum.[10]

Výsledky desambiguace nejsou vždy správné, jak dokazuje adjektivní interpretace (např. *Projíždíte-li <**Jestřebím/A**> směrem na Bořitov, odbočíte ještě před*

[9] A je značka slovního druhu (POS) adjektivum.
[10] Tvary *v Krahulčím* jsou v důsledku toho nesprávně tagovány jako tvary instrumentálu.

hospodou doprava a dále se poptáte, jak najít sypanou cestu ke Kameňáku.) a substantivní interpretace (např. *Nezbývá než uzavřít sedmý ročník Mistrovství Úpicka konstatováním, že v roce 2006 bude šipkařům z mikroregionu <**Jestřebí/N** hory> vládnout zaslouženě letos bezkonkurenční Jiří Holík.*).

Do slovníku je třeba doplnit apelativní lemmata *jestřebí, jestřabí* (viz v korpusech doložené doklady *<**jestřebí**> hnízdo, <**jestřabí**> oko*) s příslušnými rozgenerovanými tvary podle adjektivní flexe a adjektivní interpretací na rovině tagu.

U lemmat *Jestřabí, Jestřebí* a *Krahulčí* je třeba zvážit tvary flexe podle *jarní* v substantivní platnosti, a to jak na rovině slovníku, tak na rovině desambiguačních pravidel, která nelze jednoduše formulovat.

Ve slovníku mají lemmata *Hovězí/hovězí, Kobylí/kobylí, Rybí/rybí, Srní/srní, Zaječí/zaječí, Zubří/zubří* dvojí interpretaci (lemma s velkým počátečním písmenem = substantivní proprium s flexí typu *stavení* a lemma s malým písmenem = adjektivní apelativum s flexí typu *jarní*). Tvary substantivních proprií s adjektivní flexí (*ze <**Zubřího**>, v <**Zubřím**>*) jsou mylně desambiguovány jako tvary adjektiv, přestože jde o substantivní propria s adjektivní flexí. Opět platí to, co bylo uvedeno v závěru předešlého odstavce.

Ve slovníku je lemma *telecí*, které má substantivní i adjektivní interpretaci. Rozlišení na proprium (lemma s velkým počátečním písmenem) a apelativum (lemma s malým počátečním písmenem) chybí. Podle výsledků desambiguace se navíc zdá, že u substantivní interpretace se počítá výhradně s flexí podle *jarní,* a to přesto, že v korpusech je doložena flexe propria *Telecí* jak podle typu *jarní* (*V <**Telecím**> mají malovanou kroniku.*), tak podle typu *stavení* (*V <**Telecí**> dnes mají černošský ples.*)

Ve slovníku jsou lemmata *jelení, pstruží, račí, rybí, veveří, vlčí, vydří* pouze s adjektivní interpretací a lemmatem s malým počátečním písmenem (apelativum). V korpusech jsou ovšem doložena i substantivní užití proprií *Jelení, Pstruží, Račí, Rybí, Veveří, Vlčí, Vydří*.

U proprií *Hovězí*, *Jelení, Jestřabí, Jestřebí, Pstruží, Vlčí, Zubří* jsou doloženy jak tvary podle *jarní*, tak tvary podle *stavení*:

*Na horké půdě v <**Jelení**> se těžko vyhrává a přesvědčí se o tom asi i fotbalisté Kameniček.*

*Vladimír Kubík byl v <**Jelením**> nejen jako zástupce jednoho z parohatých měst, ale také jako prezident paroháčů.*

*K další dopravní nehodě došlo v <**Pstruží**>, kde havarovalo osobní auto.*

*„Jsem ze <**Pstružího**>,“ usmívala se slečna v šatičkách a třpytivým lakem ve vlasech.*

*Na výstavě ve <**Vlčí**> mohli lidé obdivovat například severočeský, jihočeský, východočeský a západočeský betlém vystříhaný z papíru.*

*Na stavební parcelu ve <**Vlčím**> na Lounsku vnikl zloděj.*

Doložené tvary pouze podle *stavení* mají propria *Krahulčí, Sokolí, Srní, Ščučí, Veveří, Vůsí a Vydří.*

V případě doplnění lemmat proprií *Jelení, Pstruží, Veveří, Vlčí, Vydří* je třeba brát v úvahu komplikovaný tvaroslovný systém. Adekvátní popis tvarosloví těchto substantiv na rovině slovníku zvýší míru tvarové homonymie a znesnadní tak následnou desambiguaci.

Závěr

Korpusová sonda odhalila, že pouhé mechanické doplňování proprií do slovníku nemůže být jediným krokem ke zlepšení automatické morfologické analýzy. Ukázala, že homonymie mezi apelativními a propriálními tvary má několik aspektů. Obecně komplikovaná situace morfologického značkování substantivizovaných adjektiv (Žižková, 2017; Richterová, 2017; Osolsobě a Žižková, 2019) je v případě proprií **tvořených z druhově posesivních/široce vztahových adjektiv od názvů zvířat** znesnadněna dvěma faktory: Prvním faktorem je stav flexe jednoslovných deadjektivních substantivizovaných proprií v úzu. Ta mohou v naprosté většině případů morfologicky fungovat buď jako substantivizovaná adjektiva typu *jarní* (*Mluvčí sice zmínila, že loni v úseku od <**Jestřebího**> k Mladé Boleslavi došlo k instalování nových pachových ohradníků, které zvěř odpuzují.*), nebo jako substantiva typu *stavení* (*Řidič motorky, jedoucí ve směru od <**Jestřebí**> na Českou Lípu, nezvládl u Zahrádek průjezd prudkou pravotočivou zatáčkou*.). Rozkolísanost morfologického úzu je prvním problémem, s nímž je třeba při snaze o zlepšení výsledků automatické morfologické analýzy počítat. Vzhledem k nedostatečné kodifikaci propriálního tvaroslovného systému je to problém řešitelný jen dosti obtížně. Doplnit do morfologického slovníku všechny doložené interpretace jak apelativních, tak propriálních užití je zajisté možné. V případě dvojí flexe substantiv by měl být lemmatem tvar na *-í* s velkým počátečním písmenem (tedy např. *Jestřebí*) a rozgenerované tvary (variantní tvary podle typu *stavení* i *jarní*) by měly mít substantivní interpretace (POS=N). Doplnění mnoha víceznačných tvarů ovšem samo o sobě povede patrně spíše ke zhoršení výsledků automatické morfologické analýzy. Desambiguaci propriálních užití nelze (jak plyne z korpusově podloženého výzkumu) mechanicky opírat o jednoduchá pravidla zohledňující pravopisné zvyklosti psaní proprií s velkým počátečním písmenem. Zatímco jednoslovná substantivní propria mají rozkolísanou flexi (vyskytují se jak tvary podle typu *jarní*, tak tvary podle typu *stavení*), ve víceslovných názvech (z hlediska automatické morfologické analýzy MWE) se adjektivní komponenty víceslovných proprií píší sice s počátečním velkým písmenem, nicméně morfologicky se vždy chovají pouze jako adjektiva (skloňují se podle typu *jarní*, a to ve všech třech rodech a obou číslech – v gramatické shodě s rozvíjeným jménem). Je tudíž třeba, aby byla desambiguována jako adjektiva (POS=A) a aby jejich lemma bylo s počátečním

malým písmenem (tedy například ve spojení *Jelení příkop* by tvar *Jelení* měl mít lemma *jelení* a měl by být označkován jako adjektivum). Identifikace víceslovných proprií (například pomocí databáze LEMUR[11]) je jednou z cest, jimiž by se měly snahy o zlepšení stavu desambiguace ubírat. Databázi víceslovných proprií je možné budovat na základě průzkumu korpusů, popřípadě některých databází proprií (viz výše). Užitečné mohou být i analýzy kolokací klíčových proprií.

Přesto se domníváme, že výsledky desambiguace nebudou v oblasti sledovaného typu proprií natolik uspokojivé, aby o ně bylo možné v dohledné budoucnosti opřít adekvátní analýzu. Onomastikům, kteří ve svém výzkumu vycházejí z anotovaných korpusových dat, tudíž radíme, aby se v práci s nimi řídili známým heslem „důvěřuj, ale prověřuj".

LITERATURA

ARES přehled obcí v ČR [online] <http://wwwinfo.mfcr.cz/ares/obce/ ares_obce.html.cz> cit. 2019-01-12.

HAJIČ, J. – HLAVÁČOVÁ, J. (2013): MorfFlex CZ, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, [online] <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>.

HNÁTKOVÁ, M. – JELÍNEK, T. – KOPŘIVOVÁ, M. – PETKEVIČ, V. – ROSEN, A. – SKOUMALOVÁ, H. – VONDŘIČKA, P. (2018): Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. *Korpus – gramatika – axiologie*, 17, s. 3–22.

JELÍNEK, T. – KOPŘIVOVÁ, M. – PETKEVIČ, V. – SKOUMALOVÁ, H. (2018): Variabilita českých frazémů v úzu. *Časopis pro moderní filologii*, 100, 2, s. 151–175.

Kobylí, Internetová jazyková příručka [online] <http://prirucka.ujc.cas.cz id=Kobylí&ref=kobylí> cit. 2019-01-12.

Kostelycz.cz [online] <www.kostelycz.cz/seznam.htm> cit. 2019-01-12.

KŘEN, M. a kol. (2017a): *Korpus SYN, verze 6 z 18. 12. 2017*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z: http://www.korpus.cz.

LUTTERER, I. – ŠRÁMEK, R. (2004): *Zeměpisná jména v Čechách, na Moravě a ve Slezsku: slovník vybraných zeměpisných jmen s výkladem jejich původu a historického vývoje*. Havlíčkův Brod: Tobiáš.

Mapy.cz [online] <www.mapy.cz> cit. 2019-01-12.

OSOLSOBĚ, K. (2008): Propria (příjmení na -č) – problém automatické morfologické analýzy. In: M. ČORNEJOVÁ – P. KOSEK (eds.), *Jazyk a jeho proměny*. Brno: Host, s. 205–216.

[11] Péčí Ústavu teoretické a komputační lingvistiky FF UK a Ústavu Českého národního korpusu FF UK vznikla databáze víceslovných lexikálních jednotek: Multiword Expressions Lexical Database – LEMUR, srov. např. Jelínek et al., 2018; Hnátková et al., 2018.

OSOLSOBĚ, K. – HLAVÁČOVÁ, J. – PETKEVIČ, V. – ŠIMANDL, J. – SVÁŠEK, M. (2017): Nová automatická morfologická analýza češtiny. *Naše řeč*, 4, s. 225–234.

OSOLSOBĚ, K. – ŽIŽKOVÁ, H. (2019): Improving Nominalized Adjectives Tagging. In *SLOVKO 2019. NLP, Corpus Linguistics, Language Dynamics and Change.* 2019. doi:10.2478/jazcas-2019-0066.

PETKEVIČ, V. (2017): Morfologická analýza. In: P. KARLÍK – M. NEKULA –J. PLESKALOVÁ (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita [cit. 2019-01-14]. Dostupné z: https://www.czechency.org/slovnik/MORFOLOGICKÁ_ANALÝZA.

PETKEVIČ, V. (2017a): Lemmatizace. In: P. KARLÍK – M. NEKULA –J. PLESKALOVÁ (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita [cit. 2019-01-14]. Dostupné z: https://www.czechency.org/slovnik/LEMMATIZACE.

PETKEVIČ, V. (2017b): Tagset. In: P. KARLÍK – M. NEKULA –J. PLESKALOVÁ (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita [cit. 2019-01-14]. Dostupné z: https://www.czechency.org/slovnik/TAGSET.

RICHTEROVÁ, O. (2017): *Od slovesa ke jménu a předložkám Deparcipiální formy v češtině: forma, funkce, konkurence*. FF UK. Disertační práce.

ŽIŽKOVÁ, H. (2017): Substantivizace adjektiv jako problém strojové analýzy češtiny. *Bohemica Olomucensia 2 – Philologica Juvenilia*, 9, 2, s. 172–179.

Klára Osolsobě
osolsobe@phil.muni.cz
Ústav českého jazyka
Filozofická fakulta
Masarykova univerzita
Arna Nováka 1
602 00 Brno

Hana Žižková
zizkova@phil.muni.cz
Ústav českého jazyka
Filozofická fakulta
Masarykova univerzita
Arna Nováka 1
602 00 Brno