ČASOPIS PRO MODERNÍ FILOLOGII 106, 2024, Č. 1, S. 44–59

OPEN ACCESS

Korpus OnomOs: principy a příklady aplikací[1]

Michal Místecký – Jaroslav David –
Jana Davidová Glogarová – Tereza Klemensová (Ostrava)

THE ONOMOS CORPUS: PRINCIPLES AND EXAMPLES OF APPLICATIONS
The study introduces OnomOs, a new corpus of Czech texts with annotation of proper names. The corpus was compiled by onomasticians from the Department of Czech Language, Faculty of Arts, University of Ostrava, and made available by the Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. The paper briefly discusses the content and structure of the corpus, the selection of texts for inclusion, and the onomastic-geographical classification of the identified names. The text consists chiefly of three preparatory analyses, which focus on the most frequent surnames, collocations found in Western and Eastern countries in the pre-1989 period, and the declension patterns of three types of onyms. In the summary, further possibilities of onomastic corpus research are presented.

KEYWORDS
quantitative linguistics; corpus; onomastics; proper names; collocation; declension

KLÍČOVÁ SLOVA
kvantitativní lingvistika; korpus; onomastika; vlastní jména; kolokace; deklinace

DOI
https://doi.org/10.14712/23366591.2024.1.3

1. ÚVOD A MATERIÁL

V české i slovanské onomastice dlouhodobě chybí uživatelsky přístupný a profesionálně zpracovaný korpus textů, v němž by byla anotována propria (vlastní jména).[2] Tato skutečnost brzdí kvantitativně zaměřené výzkumy, jejichž potřeba začíná být v současné vědě o propriích čím dál zřejmější (srov. Motschenbacher, 2020; David —

1 Sestavení korpusu bylo podpořeno Filozofickou fakultou Ostravské univerzity (projekt SGS02/FF/2023 *OnomOs — ostravský korpus vlastních jmen*). Analytická část článku (části 3.–5.) byla zpracována v rámci řešení projektu GAČR 22-09310S Kvantitativní onomastika: východiska, koncepty, aplikace.

2 Na Filozofické fakultě Masarykovy univerzity vznikl v roce 2012 korpus GEOGRAF, který obsahuje webové stránky měst, obcí, historických a přírodních památek ad. Motivací k výběru textů byl předpokládaný zvýšený výskyt toponym (v širším slova smyslu, včetně urbanonym). Vlastní jména však v korpusu značkována nejsou, vyhledávání je tak možné jen pomocí sekvence znaků v jazyce CQL (Geržová, 2016; Pličková, 2017). Rovněž existuje korpus CNEC (Czech Named Entity Corpus), který obsahuje přes 35 000 ručně značkovaných vlastních jmen, ale není přístupný v rámci databází ČNK (Ševčíková a kol., 2014).

Klemensová — Místecký a kol., 2022; viz přehled dosavadních výzkumů David — Místecký, 2023). Tuto mezeru se snaží zaplnit nově vzniklý korpus OnomOs (= Onomastika Ostrava), který sestavili a koncepčně připravili badatelky a badatelé z Katedry českého jazyka Filozofické fakulty Ostravské univerzity pod vedením Jaroslava Davida, Jany Davidové Glogarové, Terezy Klemensové a Michala Místeckého[3] a za technické podpory a spolupráce pracovníků Ústavu Českého národního korpusu FF UK Tomáše Jeziorského a Michala Křena.

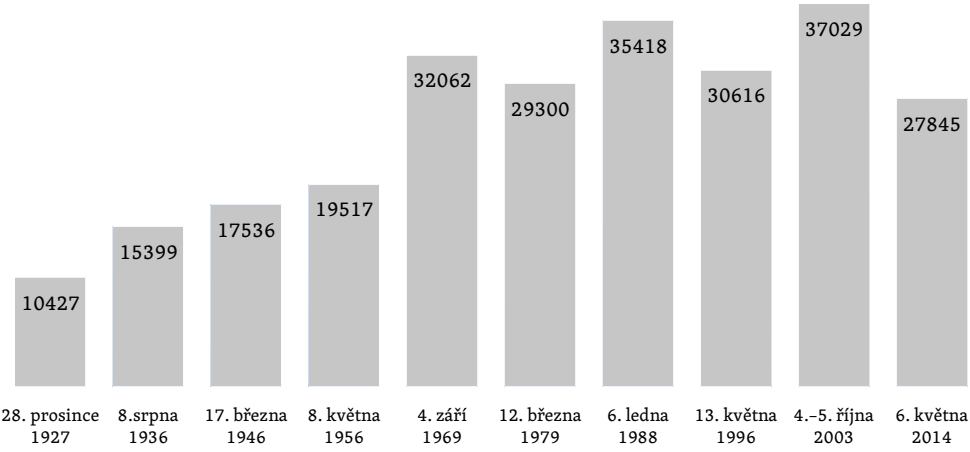
Nový korpus, který je jednou z veřejně dostupných databází Českého národního korpusu, je založen na textech z deníku Rudé právo (založeno 1920, od roku 1995 Právo) jakožto periodika, které pokrývá velkou část 20. století, vycházelo po většinu doby legálně a je vydáváno do současnosti. V korpusu je tak možné provádět onomastické analýzy nejen na jazyce současném, jímž rozumíme — podobně jako česká korpusová lingvistika — texty od roku 1989, ale též v diachronní perspektivě minulého století. Každé desetiletí z období 1920–2019 je reprezentováno jedním číslem (viz obrázek 1), které bylo vybráno náhodně, nicméně do budoucna se počítá s rozšířením databáze a větší saturací jednotlivých časových údobí. Nahodilost výběru zaručuje neutrálnost korpusu vůči zlomovým historickým událostem (struktura proprií může být např. v Rudém právu z 25. či 26. února 1948 jiná než v běžném čísle z téhož roku). Vzhledem k růstu významu Rudého práva v období po druhé světové válce — a zejména po roce 1948 — jsou dekády předcházející komunistickému převratu z hlediska velikosti korpusu podreprezentovány, a těžiště databáze se tak posouvá směrem k současnějšímu jazyku.

Aktuálně (k 31. 1. 2024) obsahuje korpus OnomOs 255 149 tokenů; v průběhu roku 2024 plánujeme rozšíření korpusu na dvojnásobnou velikost[4] a v budoucnu zapojení dalších periodik, která vycházela po větší část 20. a 21. století a jež budou s to vyvážit jeho nutnou politickou i stylistickou jednostrannost. V současném stavu považujeme korpus za provizorní, pročež je třeba i analýzy, které budeme dále v článku prezentovat, brát spíše jako inspiraci k budoucím velkoplošným výzkumům a jako představení možností, jichž lze v korpusově založené onomastice využít.

Korpus primárně zahrnuje komunikáty s převahou souvislého žurnalistického textu. Vypuštěny jsou inzerce, programy kin a divadel, komiksy, křížovky, romány na pokračování a další útvary (např. sportovní články s převahou enumerací výsledků soutěží, soupisky hráčů); vylučujeme rovněž reklamy a citáty umístěné do textu, které jej informačně dublují. Respektujeme autorské členění textu na odstavce; kratší texty slučujeme do jednoho celku — rubriky (atribut text.section). Z žánrového hlediska vyčleňujeme pouze zprávy (news) a popisky obrázků (photo), protože v určitých obdobích je např. rozlišení mezi zprávami a komentáři problematické. V případě popisků obrázků zahrnujeme pouze ty, které jsou tvořeny výpovědí s určitým slovesným tvarem.

3 V této souvislosti by autoři studie rádi poděkovali studentkám Haně Halatové, Janě Pavlištíkové, Karolíně Poláškové a Magdaleně Strnadlové a doktorandkám Kristýně Březinové, Jarmile Mádrové a Agatě Reclik, které se podílely na přípravě textové složky korpusu.

4 V průběhu roku 2024 bude korpus rozšířen o jedno číslo (Rudého) Práva v každém desetiletí, přičemž toto číslo je náhodně vybráno vždy z jiného pětiletí, než tomu bylo v první verzi korpusu.



OBRÁZEK 1. Struktura korpusu OnomOs (v tokenech)

Při klasifikaci proprií jsme vycházeli z třídění autorů programu NameTag 2 (Straková — Straka — Hajič, 2019; Ševčíková — Žabokrtský — Krůza, 2007), který jsme v korpusu využili k identifikaci vlastních jmen. Výchozí třídění jsme však přizpůsobili onomastickým zvyklostem i současným trendům, směřujícím k redukci oborové terminologie (viz David — Klemensová — Místecký, 2021). Z tohoto důvodu mezi vlastní jména nezahrnujeme např. názvy dní v týdnu, měsíců či bibliografické údaje. Pro podrobnosti k transformaci původního třídění v NameTagu 2 do kategorizace, kterou využíváme v korpusu OnomOs, viz wikistránku o daném korpusu (Kocek, 2023); zde je též dostupný návod pro vyhledávání. Tato kategorizace zahrnuje skupiny dvou řádů — vyšší reprezentují antroponyma (A), chrématonyma (C) a toponyma (T), nižší pak podrobnější třídění těchto tří ústředních jednotek (např. AF: rodná jména). Klasifikaci společně s příklady z korpusu shrnuje tabulka 1. Detailněji se jednotlivým kategoriím věnujeme v druhé části článku.

V současné verzi nepracuje korpus OnomOs s tzv. containers, takže jednotlivé členy víceslovných onymických jednotek jsou považovány za samostatná propria. Například onymum „Pražské jaro 1996“ tvoří tři vlastní jména klasifikována jako CC (conferences, contests and events), onymum „Česká republika“ je zase považováno za dvě toponyma (konkrétně za dva zástupce kategorie TT — territories). Tento přístup není ideální, protože znemožňuje identifikaci jména s příslušným pojmenovaným objektem (tzv. named entity grounding). V budoucnu je proto ke zvážení implementace anotace víceslovných jednotek v korpusu.

Aplikace programu Name Tag 2 na naše jazyková data byla automatická a výsledky neprocházely manuální korekcí — nutně se mezi nimi tudíž objevují chybně klasifikované pojmenované entity (viz příklady 1, 2 a 3), případně falešně negativní výstupy (tedy pojmenované entity, které anotátor jako vlastní jména nevyhodnotil; viz příklad 4). První zmíněné chyby jsou dány i skutečností, že mezi některými kategoriemi druhého řádu existují různé, nejen hierarchické vztahy (např. kategorii CN — periodicals lze zahrnout pod kategorii CF — companies, AM — religious and

Skupiny vyššího řádu	**Skupiny nižšího řádu**	**Příklady z korpusu OnomOs**
anthroponyms (A)	AF: first names	*Václav*, *Roger*, *Anna*, *M.* [Miroslav Ransdorf], *Luciano*
	AI: inhabitants	*Němec*, *Moravan*, *Rom*, *Trnavští*
	AM: religious and mythological names	*Mojžíš*, *sv. Václav*, *Justinián*
	AS: surnames	*Zeman*, *Klaus*, *Zieleniec*, *Ortman*, *Dienstbier*
	AX: underspecified anthroponyms	*Dzurinda*, *Kennedy*, *Toufar*, *Jidášové* [sic!]
chrematonyms (C)	CA: art products	*Banánové rybičky*, *Čekání na Godota*, *Harry Potter a Fénixův řád*
	CC: conferences, contests and events	*Pražské jaro 1996*, *MS* [= mistrovství světa], *Liga mistrů*
	CD: directives and norms	*Postupimská dohoda*, *SALT 2*, [dopis] *VZN — Šk — Ká — 1267*
	CF: companies	*Česká rafinérská*, *Chemopetrol*, *Siemens*, *VW*
	CH: feasts	*dny Páně*, *Nový rok*
	CI: cultural and educational institutions	*Rudolfinum*, *Centrum volného času*, *OA* [= Obchodní akademie]
	CM: currencies	*Kčs*, *USD*, *Kč*
	CN: periodicals	*Financial Times*, *Vlastivědný věstník*, *Právo*, *ABC*
	CP: politics	*ODS*, *ČSSD*, *Levý blok*, *Československá obchodní banka*
	CR: products	*Guttalax*, *Zetor*, *Cabernet Sauvignon*, [rakety] *D5*
	CT: radios and TVs	*Nova*, *CBS*, *Český rozhlas*, *Rossija*
	CX: underspecified institutions	*RVHP*, *IAAF*
	CY: underspecified artifacts	*Queen Mary* [loď], *Wellington* [steak]
toponyms (T)	TN: nature names	*Ural*, *Divoký západ*, *Kalahari*, *Barentsovo moře*
	TS: settlements	*Atlanta*, *Bonn*, *Rokycany*, *Dakar*
	TT: territories	*Česká republika*, *Florida*, *Čečensko*, *Evropa*, *Kroměřížsko*
	TU: urbanonyms	*Palackého ulice*, *Bratislavská*, *Nuselský most*, *Můstek*
	TX: underspecified toponyms	*Diamantový důl* [pozemek], *Fukušima* [jaderná elektrárna], *Želivka* [řeka]

TABULKA 1. Klasifikace proprií v korpusu OnomOs (skupiny seřazeny podle abecedy; názvy uvádíme anglicky, aby byla srozumitelná motivace zkratky)

mythological names zase osciluje mezi AF — first names a AS — surnames), případně též výskytem homonym („s/Svoboda"; viz příklad 4).

[1] Konečně, přečtěme si, co napsal 18. srpna list Guardian [CN]: „V minulých dnech používal Husák [AS] sdělovací prostředky k tomu, aby v ČSSR [TT] vytvořil ovzduší hrůzy. Lid ČSSR [TT] byl vystaven hysterickým hrozbám." Je docela pochopitelné, že klid a pořádek v naší zemi hatí plány těchto pánů. Poslední dny potvrdily, že výzvy ke klidu měly své oprávnění. Jenže pánové z Guardianu [CF] chtěli na ulicích Prahy [TS] vidět více krve, než jí, bohužel, proteklo. (*Rudé právo*, 4. září 1969)

[2] Teď ještě přijde Jan [AF] Nepomucký [AS] s uříznutým jazykem. (*Rudé právo*, 28. prosince 1927)

[3] Jsem dnes už tak otrlý, že chodím na sv. Jana Nepomuckého [AM] do průvodu a tlačím se pokaždé mezi zánovní panny. (*Rudé právo*, 28. prosince 1927)

[4] Svoboda ztrácí nervy, soudí Grebeníček [AS]. Podle předsedy KSČM [CP] Miroslava [AF] Grebeníčka [AS] šéf lidovců Cyril [AF] Svoboda [AS] ztrácí nervy ze strachu, že na nadcházejícím sjezdu KDU-ČSL [CP] neobhájí svůj post. *(Právo*, 4.–5. října 2003)

2. STRUKTURA KORPUSU

V této části studie představíme základní frekvenční charakteristiky proprií v korpusu. Jejich struktura, kterou zachycuje tabulka 2, odpovídá intuitivnímu předpokladu, že (Rudé) Právo se zaměřuje především na politické události. Důraz je kladen na osobnosti (AS), chrématonyma pojmenovávající státní úřady a politické strany (CP) a geopolitické souvislosti (TT). Určité pozornosti se dostává i kultuře (CI; viz příklad č. 5). Na konci žebříčku se naopak objevují jména, která jsou očekávatelná spíše v jiném typu textů (např. kategorie CD se bude objevovat v administrativě), případně taková pojmenování, která tagger nedokáže úžeji zařadit (typ CY). Náboženská a mytologická jména (AM) zase v novinářských textech fungují jen omezeně, např. v metaforách nebo přirovnáních. Překvapivé je však nízké zastoupení svátků a slavností (CH), které může souviset s levicovou orientací periodika a s dominancí textů z období po roce 1948, pro které byl charakteristický silný vliv komunistické ideologie.

[5] Cyklem symfonických básní Má vlast [CA] Bedřicha [AF] Smetany [AS] bylo v neděli večer v Rudolfinu [CI] tradičně zahájeno Pražské jaro 1996 [CC]. Avšak poprvé v jedenapadesátileté historii hlavního českého hudebního festivalu byl tento úkol svěřen zahraničnímu tělesu, britskému orchestru London Classical Players [CI] s dirigentem Rogerem [AF] Norringtonem [AS]. (*Právo*, 13. května 1996)

Užší klasifikace proprií	Absolutní frekvence
AS: surnames	4 152
CP: politics	2 848
TT: territories	2 661
AF: first names	2 483
TS: settlements	2 188
CI: cultural and educational institutions	1 639
CF: companies	1 095
CN: periodicals	643
CA: art products	595
CC: conferences, contests and events	490
AI: inhabitants	323
CR: products	279
TN: nature names	228
AX: underspecified anthroponyms	139
TU: urbanonyms	137
CM: currencies	105
CD: directives and norms	79
AM: religious and mythological names	49
CT: radios and TVs	38
CY: underspecified artifacts	33
TX: underspecified toponyms	24
CX: underspecified institutions	8
CH: feasts	7

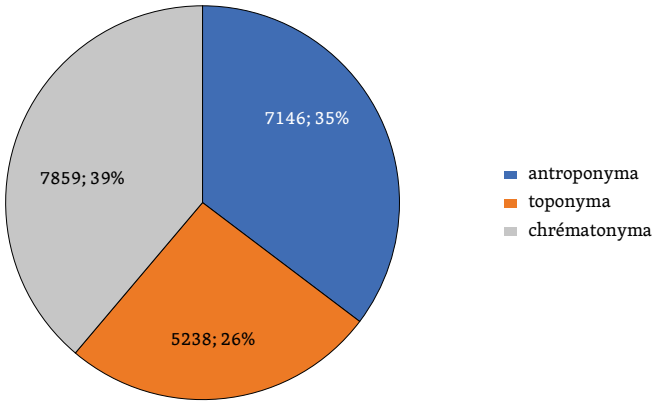
TABULKA 2. Složení proprií v korpusu OnomOs

Sledujeme-li distribuci proprií v korpusu OnomOs z perspektivy kategorií vyššího řádu (viz obrázek 2), je patrná tendence k vyrovnanosti jejich podílu. V korpusu sice mírně převažují chrématonyma (39 %), avšak ani jejich dominance není výrazná. Výsledek lze nicméně interpretovat především strukturou klasické zprávy, která odpovídá na základní quintiliánovské otázky „kdo/co" (antroponymum, případně chrématonymum), „kde" (toponymum) apod. Výsledky naší sondy tak rovněž otevírají další výzkumnou otázku — zdali je daný trend stylistickým rysem žurnalistiky, nebo je charakteristický pouze pro zkoumané periodikum.

V následující části studie se prostřednictvím prvních analýz a heuristických sond pokusíme ukázat, jaká je využitelnost korpusu OnomOs při rozboru fungování vlastních jmen v textu.

3. VÝZKUM 1 — FREKVENČNÍ ANALÝZA PŘÍJMENÍ

První analýza bude zkoumat rozdíly v distribuci příjmení (kategorie AS) v obdobích prvorepublikového, komunistického a polistopadového (Rudého) Práva; budeme tedy



OBRÁZEK 2. Rozdělení proprií v korpusu OnomOs podle tradičních onomastických kategorií

pracovat se třemi subkorpusy (RP 1927 a 1936; RP 1956, 1969, 1979 a 1988; a P 1996, 2003 a 2014). Údaje o AS zachycujeme rovněž pomocí relativizované frekvence i.p.m. (Kováříková, 2021), která vyjadřuje teoretickou frekvenci daného jevu v hypotetickém korpusu o rozsahu jeden milion tokenů, a pomocí relativní frekvence lemmat, kterou počítáme jako podíl daného počtu lemmatizovaných AS a celkového počtu lemmat v daném subkorpusu (např. RP 1927, 1936). Výsledky prezentuje tabulka č. 3.

	celkem AS	i.p.m.	celkem AS (lemmata)	Relativní frekvence lemmat
RP 1927, 1936	317	12 274,45	201	0,0387
RP 1956, 1969, 1979, 1988	1 590	13 671,89	1 066	0,0816
P 1996, 2003, 2014	1 956	20 483,82	1 007	0,0792

TABULKA 3. Frekvenční struktura příjmení ve třech studovaných subkorpusech (AS = příjmení)

Na základě výsledků lze konstatovat, že počet příjmení v čase roste, ale jejich pestrost, kterou lze determinovat na základě lemmat, zůstává v polistopadovém období ve srovnání s komunistickou periodou spíše konstantní. Po pádu komunistického režimu tak Právo pracuje s mírně menším množstvím osobností, ale zmiňuje je častěji; to může souviset s dlouhodobými politickými kariérami prominentních politiků (*Zeman*, *Klaus*, *Havel*, *Putin*) či s proměnami žurnalistického stylu od faktografické roztříštěnosti k propojenějším tematickým linkám, které můžeme klást do souvislosti s určitou tendencí k „příběhovosti" současné novinařiny.

Konkrétní proměny skladby jmen orientačně představuje tabulka č. 4 (vypuštěna byla interpunkce a posesivní adjektiva — typ *Zemanův*). Velké množství frekventovaných jmen, které je typické pro polistopadové Právo, odpovídá předpokladu, že v těchto číslech se častěji používá menší počet příjmení. Obsahové metamorfózy odpovídají historickým transformacím 20. století — za první republiky se temati-

zují nepřátelé komunismu (nacistický diktátor Adolf *Hitler*, italský diktátor Benito *Mussolini*, řecký autokratický politik Ioannis *Metaxas* — viz příklad 6) a spisovatel Maxim *Gorkij*; v komunistické epoše se prosazují především domácí politické osobnosti (prezident Ludvík *Svoboda*, generální tajemník KSČ Gustav *Husák*) a sportovci (cyklokrosaři František *Klouček*, Miroslav *Kvasnička* a Peter *Hric*, francouzská atletka Colette *Bessonová* — viz příklad 7); v polistopadové době pak dominují znovu politické osobnosti, tentokrát však s větším podílem zahraničních lídrů (ruští prezidenti Boris *Jelcin* a Vladimir *Putin*, americký prezident George W. *Bush*, německý prezident Joachim *Gauck*). Pozoruhodný je návrat Adolfa *Hitlera*, který je připomínán v souvislosti s interview s Arnoldem Schwarzeneggerem (viz příklad 8).

[6] Generální stávka, vyhlášená na 4. srpna, byla projevem odporu země proti fašistickým metodám vlády. Metaxas [AS] odpověděl na nespokojenost lidu násilím. Násilím vnutil zemi fašistický, diktátorský režim. Odpor lidu má být zlomen krvavým terorem. (*Rudé právo*, 8. srpna 1936)

[7] Sotva byl ohlášen závod na 400 m, vrhli se fotoreportéři jako hejno kobylek na slavnou tmavovlásku s čokoládovýma očima Collete [sic!; AF] Bessonovou [AS]. Také hlediště se sympatiemi a zvědavostí sledovalo každý pohyb téhle čiperné dívky z Bordeaux [TS], jejíž stylový běh — kdesi jsem to četl — umí právě tak sladce omámit jako víno z kraje, v němž se narodila. (*Rudé právo*, 4. září 1969)

[8] Pár dní před úterními volbami kalifornského guvernéra se postavení vedoucího republikánského kandidáta Arnolda [AF] Schwarzeneggera [AS] zkomplikovalo další aférou. Vyšlo totiž najevo, že v interview před 18 lety měl vyjádřit obdiv Adolfu [AF] Hitlerovi [AS], což ale herec okamžitě ostře popřel. (*Právo*, 4.–5. října 2003)

Závěrečný výzkum příjmení se zaměří na tzv. pluralizaci, v jejímž rámci se jméno užívá v množném čísle, aby označilo nositele podobných vlastností (Švejk > švejkové) — funguje tedy jako apelativum, resp. apelativizované proprium. Po manuální selekci se ukázalo, že tento způsob práce s příjmeními je přes konfrontační styl především předválečného Rudého práva velmi minoritní, jelikož se v korpusu objevily jen dva případy, které naší definici odpovídají (viz příklady 9 a 10; je k diskusi, zdali za příjmení považovat jméno markýze *Géra*). Je otázkou, zda není takové využívání antroponym typické spíše pro komentářovou publicistiku, nebo jestli je v současné žurnalistice vůbec produktivní (srov. David — Klemensová — Místecký, 2023).

[9] Waltrováci toto protidělnické jednání ostře odsuzují a chystají se k dalším krokům, od nichž je neodvrátí ani tucet Kadleců [AS]. Případ Kadlecova [AS] zásahu ukazuje všem kovákům, že je svrchovaný čas dělat v organisaci pořádek a vyrovnat se s těmi, kteří sabotují usnesení své kovodělnické i Amsterdamské internacionály [CP] o jednotě a snaží se jednotné akce znemožnit. (*Rudé právo*, 8. srpna 1936)

Prvorepublikové RP		Komunistické RP		Polistopadové P			
Lemma	**i.p.m.**	**Lemma**	**i.p.m.**	**Lemma**	**i.p.m.**	**Lemma**	**i.p.m.**
Hitler	813,13	Svoboda	266,56	Zeman	617,87	Hausmann	62,83
Metaxas	271,04	Husák	111,78	Klaus	586,45	Roušal	62,83
Kahánek	271,04	Klouček	85,99	Havel	230,39	Kennedy	62,83
Křivánek	232,32	Tito	85,99	Jelcin	230,39	Beneš	62,83
Gorkij	232,32	Novák	85,99	Kott	209,45	Bartoš	62,83
Mussolini	193,60	Majakovský	85,99	Kožený	198,97	Hojdar	62,83
		Bessonová	77,39	Sobotka	167,56	Dyba	62,83
		Šimůnek	60,19	Putin	157,08	Vojtková	52,36
		Kosygin	60,19	Bush	146,61	Janukovyč	52,36
		Hric	60,19	Gross	146,61	Vaclík	52,36
		Adamec	60,19	Saudek	136,14	Tichý	52,36
		Havlíčková	51,59	Špidla	125,67	Kiska	52,36
		Brandl	51,59	Zjuganov	94,25	Mussolini	52,36
		Veselý	51,59	Hitler	94,25	Prodi	52,36
		Dvořák	51,59	Škromach	83,78	Pöpperle	52,36
		Indra	51,59	Koller	83,78	Pouzar	52,36
		Kramer	51,59	Kolář	83,78	Dienstbier	52,36
		Gromyko	51,59	Gauck	73,31	Thatcherová	52,36
		Honecker	42,99	Brožová	73,31	Šosták	52,36
		Kvasnička	42,99	Hudeček	73,31	Pilař	52,36
		Jakeš	42,99	Janeček	73,31	Kalvoda	52,36
		Gorbačov	42,99	Babiš	73,31	Šplíchal	52,36
		Spychalský	42,99	Železný	73,31	Wilson	52,36
		Rogers	42,99	Svoboda	73,31	Zelenka	52,36
		Schneiderová	42,99	Pokorný	62,83	Šmerdová	52,36
		Šváb	42,99	Čech	62,83	Schwarzenegger	52,36
		Orejou	42,99	Kemel	62,83	Nečas	52,36
		Hašek	42,99	Budík	62,83	Staňa	52,36
				Ransdorf	62,83		

TABULKA 4. Nejčastější příjmení v jednotlivých subkorpusech (jen příjmení s absolutní frekvencí > 5 v daném subkorpusu)

[10] Jsme dlužníky Slezska [TT], jež nám bylo věřitelem trpělivým a důvěřivým. Bylo národní výspou, která jak na Opavsku [TT], tak na Těšínsku [TT] čelila steré smrti, rozhlodávána národnostně i sociálně, štvána všemi zastaveními křížové cesty, od legendárních markýzů Gerů [AS] a Larischů [AS] tavírnou jejich hutí a dolů, nesnesitelným útlakem jejich úředníků a dozorců, až po pochopy a hajné v hlubinách jejich lesů, kde všude zotročovali, vyssávali, poněmčovali a popolšťovali slezský lid. (*Rudé právo*, 17. března 1946)

4. VÝZKUM 2 — VÝCHOD A ZÁPAD VE SVĚTLE TERITORIÁLNÍCH NÁZVŮ

Druhá sonda bude lingvisticky analyzovat, jestli se do korpusu promítl napjatý vztah mezi západní Evropou a USA a sovětskými satelity během studené války (srov. David, 2013). Korpus bude znovu rozdělen na dva menší celky zahrnující Rudá práva vyšedší před sametovou revolucí (ročníky 1956, 1969, 1979 a 1988) a po ní (1996, 2003 a 2014). Za země východního bloku považujeme Albánii, Bulharsko, Československo, Maďarsko, Polsko a Rumunsko, přičemž pracujeme rovněž se zavedenými zkratkami plných jmen států (ALR, BLR, ČSSR, MLR, PLR, RLR, RSR, NDR a SSSR). Neuvažujeme Německo, protože se jedná o významově nejednoznačné lemma (může označovat jak NDR, tak NSR), a víceslovné názvy, např. Sovětský svaz nebo Velkou Británii. Jako země euroatlantického Západu jsme vyčlenili Andorru, Belgii, Dánsko, Francii, Irsko, Island, Itálii, Lichtenštejnsko, Lucembursko, Nizozemí/Nizozemsko, Norsko, Portugalsko, Rakousko, Řecko, Španělsko, Švédsko, Turecko a USA. Neutrální státy (Švýcarsko, Finsko) a země s problematickým zařazením (Jugoslávie) jsme z výzkumu vyloučili.

V první analýze se zaměříme na frekvenční obraz států v korpusu. Tabulka č. 5 zachycuje změny četností názvů zkoumaných zemí v předlistopadové a polistopadové době. Zatímco u východního bloku registrujeme významný frekvenční pokles, který potvrzuje i statisticky signifikantní výsledek chí-kvadrátového testu (χ^2 = 258,79, p <<< 0,05), v případě evropského Západu k očekávanému nárůstu nedochází (mírné snížení nebylo vyhodnoceno jako statisticky významné: χ^2 = 2,22, p = 0,14; Cvrček, 2021). Výsledek, který však stojí pouze na omezených datech, může naznačovat, že Právo se po sametové revoluci zaměřuje primárně na domácí politickou scénu, nereflektuje tolik širší geopolitická témata a klade důraz na témata zcela nepolitická; nižší frekvence teritoriálních názvů si lze rovněž vysvětlit důrazem na osobnosti v čele daných zemí, který se propisuje i do četnosti příjmení, již jsme analyzovali ve třetí části (místo „Francie jednala s Německem" se pak může objevit např. „Macron jednal s Merkelovou"). Za poklesem prominence východních zemí může stát také zánik některých státních celků (např. ČSSR, SSSR), nebo opět již zmiňovaná nahodilost výběru čísel periodika do korpusu.

	Východ	Východ (i.p.m.)	Západ	Západ (i.p.m.)
RP (1956, 1969, 1979, 1988)	396	3405,07	177	1521,97
P (1996, 2003, 2014)	26	272,28	122	1277,62

TABULKA 5. Frekvenční rozdíly mezi východními a západními zeměmi ve zkoumaných subkorpusech

V druhé fázi výzkumu jsme provedli kolokační analýzu na datech z čísel Rudého práva komunistického období (1956, 1969, 1979, 1988). Hledali jsme kolokáty v rozmezí –3/+3 od ústředního lemmatu (názvu státu, případně zkratky), přičemž jako minimální frekvenci kolokátu i kolokace v korpusu jsme stanovili hodnotu 3. Kolokace byly vygenerovány na základě míry logDice, která preferuje exkluzivní sousloví a je nezávislá na frekvencích hlavy a kolokátu (Brezina, 2018). Minimální hodnotou

míry (cut-off skóre) bylo 8. Analýza byla provedena pro východoevropské a západoevropské státy zvlášť; do vygenerovaných seznamů kolokátů jsme nezasahovali, protože i slova synsémantická a interpunkci pokládáme za interpretačně nosné. Výsledky jsme roztřídili do tří skupin — kolokáty společné, kolokáty charakteristické pro západní země a ty, které definují státy východního bloku (viz tabulku č. 6).[5]

Kolokáty společné	*ČSSR, mezi, ministr,), návštěva, SSSR, (, a, státní, návrh, zahraniční, vláda, USA, —, Rakousko*
Kolokáty spjaté s Východem	*prezident, Polsko, předseda, velvyslanec, NDR, rada, v, s, delegace, 20, M, 1968, lidový, z, PLR, pozvání, rok, navštívit, ., přátelský, ministerstvo, komunistický, A, událost, nad, území, Československo, dnes, Gustáv, Spychalský, reprezentantka, spolupráce, k, W, Husák, organizace, do, reprezentant, strana, Maďarsko, spolkový, 12, družstvo, hospodářství, zvítězit, věc, přestavba, Francie, 8, představitel, tým, vládní, národ, o, my*
Kolokáty spjaté se Západem	*Britannia, Británie, Itálie, KS, NSR, velký, 6, Kanada, Japonsko, severní, Švédsko, Belgie, 23, opět, list, oficiální, tajemník, svaz, republika, sovětský, politika, proti*

TABULKA 6. Klasifikace kolokátů spjatých se jmény zemí rozdělených železnou oponou

Ve skupině sdílených výrazů se objevují *ČSSR* a *SSSR* jakožto politické svorníky tehdejší československé diplomacie, formální a očekávatelné *státní návštěvy* a společné politické funkce (*ministr*, *vláda*). Důležitý je kolokát *mezi*, s nímž se u západních a východních států pracuje odlišně — zatímco u jmen zemí východního bloku je akcent položen na vzájemnou spolupráci (viz příklad 11), v případě kapitalistických zemí se jím naznačuje postupné normalizování vztahů mezi dvěma stranami železné opony (viz příklad 12). S tímto fenoménem souvisí i kolokát *USA*, který odkazuje k vzájemné kontrole vojenských arzenálů mezi tehdejšími světovými supervelmocemi (viz příklad 13), a částečně také *Rakousko*, jehož prezident zavítal v roce 1979 na státní návštěvu Československa. Závorky zase odkazují ke stylu sportovních komentářů, které potvrzují neutrální roli sportu v kontextu světové politiky (viz příklad 14).

[11] Ředitel zdejšího stánku Centrotexu [CF] ing. Z. [AF] Černý [AS] to dokazuje na několika číslech: tak letos dosáhne vzájemná výměna textilu mezi ČSSR [TT] a NDR [TT] výše téměř půl miliardy korun. (*Rudé právo*, 4. září 1969)

[12] Z Bukurešti [TS] odcestoval do vlasti po ukončení třídenní oficiální návštěvy Rumunska [TT] francouzský prezident Giscard [AF] d'Estaing [AS]. Ve společném komuniké je věnována hlavní pozornost hospodářské spolupráci mezi Rumunskem [TT] Francií [TT]. (*Rudé právo*, 12. března 1979)

5 Kompletní výsledky kolokační analýzy jsou dostupné na https://github.com/KCJFFOU/David_et_al_OnomOs.

[13] Tento návrh stanoví, aby byl přibližně ve tříletém období podstatně snížen stav ozbrojených sil a množství výzbroje obvyklého typu především ozbrojené síly a výzbroj pěti velmocí [sic!]. Stanoví, že nejvyšší úroveň ozbrojených sil pro SSSR [TT], USA [TT] a Čínu [TT] má být stanovena na 1 — 1 a půl milionu mužů a pro Velkou Britannii [TT] a Francii [TT] po 650.000 mužů. (*Rudé právo*, 8. května 1956)

[14] Dále se ze zahraničních účastnic líbily Voljaniková [AS] (SSSR [TT]), reprezentantky NDR [TT] Plötzová [AS] a Heffnerová [AS] a Rumunka [AF][6] Ignatovová [AS]. Po stránce choreografie a kompozice zaujalo také cvičení mladých Američanek [AI]. (*Rudé právo*, 12. března 1979)

V oblasti kolokátů specifických se obě skupiny pojí s dalšími státy svého bloku, případně se zeměmi sdílejícími hodnotovou orientaci (*Polsko*, *NDR*, *Maďarsko*; *Itálie*, *NSR*, *Švédsko*, *Japonsko*). U východního bloku se objevují výrazy zdůrazňující vřelost a důležitost kontaktů a charakter států (*přátelský*, *spolupráce* — viz příklad 15; *komunistický*, *lidový*), ale také diskuse o nových konceptech (*přestavba*, *20* — viz příklad 16), kdežto u západních zemí se setkáváme s neutrální či kritickou perspektivou (*oficiální*, *politika*, *proti* — viz příklad 17). V normalizačním Rudém právu se pak probírá především sportovní zápolení obou bloků (*reprezentant*, *reprezentantka*; *opět*, *proti* v jiném kontextu — viz příklad 18).

[15] Na pozvání předsedy státní rady Polské lidové republiky [TT] maršála Polska [TT] M. [AF] Spychalského [AS] vykonal ve dnech 1. — 3. září 1969 přátelskou návštěvu v Polsku [TT] president Československé socialistické republiky [TT] armádní generál L. [AF] Svoboda [AS] s chotí. (*Rudé právo*, 4. září 1969)

[16] A je snad možné hovořit o podobnosti toho, co se dostalo do politického slovníku jako československé události v roce 1968, s přestavbou v Sovětském svazu [TT]? Srovnejme například úlohu strany v životě společnosti nyní v SSSR [TT] a před 20 lety v Československu [TT]. (*Rudé právo*, 6. ledna 1988)

[17] Korejský představitel zdůraznil, že po hrdinné vlastenecké a osvobozenecké válce proti USA [TT] v letech 1950 až 1953 bylo vybudováno silné samostatné hospodářství. Průmyslová výroba, jejíž jádro tvoří strojírenství, vzrostla v roce 1967 ve srovnání se stavem při vzniku republiky 22krát. (*Rudé právo*, 4. září 1969)

[18] Mužstvo ČSSR [TT] nastoupilo ke svému druhému zápasu proti Francii [TT] a opět neuspělo a překvapivě podlehlo. Ani potřetí čs. hráči nebodovali a prohráli s celkem SSSR [TT]. (*Rudé právo*, 6. ledna 1988)

6 Nametag 2 zde nesprávně vyhodnotil vlastní jméno „Rumunka" jako rodné jméno (správnou kategorií by bylo jméno obyvatelské — AI).

Kolokační sonda do názvů východních a západních zemí ukázala na potenciál, který má výzkum vlastních jmen pro historii, politologii či politickou geografii; demonstruje rovněž stylistické a lexikální proměny, jimiž Rudé právo v předlistopadové epoše procházelo.

5. VÝZKUM 3 — SROVNÁNÍ PÁDOVÝCH DISTRIBUCÍ U VYBRANÝCH SKUPIN ANTROPONYM, CHRÉMATONYM A TOPONYM

Třetí sonda je zaměřena teoretičtěji — věnuje se pádovým distribucím u tří nejčetnějších kategorií onym (AS, CP a TT), které zároveň patří do různých skupin onomastických (antroponyma, chrématonyma a toponyma). Pádové distribuce vlastních jmen jsou v současné době rozvíjející se oblastí nejen onomastického výzkumu (viz např. David — Místecký, 2023; Janda — Fidler — Cvrček — Obukhova, 2022), příkladně chrématonyma však byla korpusově zkoumána zatím jen minimálně; jejich systémové charakteristiky — např. kolokabilita, gramatické preference — jsou proto spíše předmětem kvalitativně založených odhadů (např. Kuba — Šrámek, 1989; Knappová, 1995; Kolářová, 1999; Mitter, 2003).

Ze srovnání našich vzorků (vyloučeny byly tvary, které nebyly korpusovým taggerem rozpoznány) vyplývá pro sledované kategorie několik rysů (viz tabulku č. 7), které by měly být v budoucnu ověřeny na robustnějších datech. Jde především o:

a) rozdíl v dominantním pádu (u antroponym jde o nominativ, u chrématonym a toponym o genitiv, i když u toponym je téměř vyrovnán podíl druhého pádu a lokálu);
b) nízkou až mizivou frekvenci dativu a vokativu u všech tří onymických typů;
c) výraznější postavení lokálu u chrématonym oproti antroponymům.

Kvantitativní převahu genitivu nad nominativem u chrématonym lze vysvětlit jednak vazbami některých frekventovaných předložek (*od*, *z*, *bez*; typ „politik XY ze [strany] ABC“), jednak jejich postavením v pozici adnominálního přívlastku (*předseda ODS*, *rozšíření NATO*, *Úřad práce*). Na základě abstraktních prepozic či podnikových sídel lze interpretovat i vyšší frekvenci lokálu (*o*, *na*, *v*; pro genitiv i lokál u chrématonym viz příklad 19). U toponym jsou vysoké frekvence genitivu a lokálu očekávatelné, protože souvisejí s místními předložkami, které se s těmito pády pojí (*z*, *do*, *v*).

Kvaziabsence vokativu pravděpodobně souvisí s obecnějšími rysy publicistického stylu — pátý pád je charakteristický pro konverzace a literární dialogy (Cvrček a kol., 2020, s. 40–41), které se v novinách, s možnou výjimkou interview, neobjevují. S dativem se pak pojí jen omezený repertoár předložek (typicky *k*, *proti*; viz příklad 20) a v jiných významech se jedná především o příjemce děje (sémantickou roli beneficienta), což omezuje jeho použití na životná podstatná jména, v našem výzkumu reprezentovaná pouze kategorií AS (Dvořák, 2017; viz příklad 21).

[19] Dvě protichůdné linie, které dnes v Atlantickém paktu [CP] proti sobě stojí, se výrazně projevily právě při debatě o této otázce. Podle zpráv západních agen-

tur předložil na příklad britský ministr zahraničí Selwyn [AF] Lloyd [AS] návrh na „rozšíření činnosti" Severoatlantického paktu [CP]. (*Rudé právo*, 8. května 1956)

[20] Jestliže tato zpráva potvrdí, že Izrael [TT] nadále porušuje [sic!] v kolonizaci okupovaného území, musí rada vyhlásit proti Izraeli [TT] sankce v souladu s článkem 7 Charty OSN [CD]. (*Rudé právo*, 12. března 1979)

[21] Mnohem bezmocnější jsou české úřady. Těm se kvůli neexistenci mezinárodní smlouvy o právní pomoci mezi ČR [TT] a Bahamami [TT] za dva roky nepodařilo doručit Koženému [AS] ani usnesení o zahájení jeho trestního stíhání kvůli vytunelování Harvardského průmyslového holdingu [CF]. (*Právo*, 4.–5. října 2003)

	AS		CP		TT	
Pád	**Frekvence**	**Podíl**	**Frekvence**	**Podíl**	**Frekvence**	**Podíl**
nominativ	2330	65,05 %	360	21,74 %	409	19,37 %
genitiv	592	16,53 %	910	54,95 %	680	32,21 %
dativ	117	3,27 %	26	1,57 %	92	4,36 %
akuzativ	255	7,12 %	159	9,60 %	121	5,73 %
vokativ	4	0,11 %	0	0,00 %	0	0,00 %
lokál	43	1,20 %	133	8,03 %	660	31,26 %
instrumentál	241	6,73 %	68	4,11 %	149	7,06 %

TABULKA 7. Deklinační distribuce v užších třídách AS (příjmení) a CP (názvy vládních a politických institucí) a TT (názvy teritorií)

6. SHRNUTÍ

Cílem článku bylo formou tří sond představit možnosti, které nabízí nový český onomasticky značkovaný korpus OnomOs, jehož postupné rozšiřování je plánováno nejen pro rok 2024, ale také v následujících letech. Nabízí se rozšíření o texty např. Lidových novin, Národních listů, Mladé fronty DNES či Blesku. Periodika, která vycházela po omezenou dobu, plánujeme spárovat s obdobně zaměřenou tiskovinou z jiné doby (např. obdobou Blesku v dnešní skladbě deníků mohly za první republiky být bulvární deníky Expres nebo Polední list).

Přestože je korpus rozsahem prozatím omezený, výsledky pilotních výzkumů otevírají nové badatelské otázky v oblasti proprií a naznačují směr dalšího využití kvantitativního přístupu k propriálním analýzám. V korpusu OnomOs tak nabízíme nástroj, jenž má v průběžně doplňované podobě potenciál obohatit tradiční vhled do problematiky používání proprií o kvantitativní aspekty; může také vést k formulování obecných principů jejich chování v textech a identifikovat jejich gramatické charakteristiky. Výstavba takovéto gramatiky onym může být v budoucnu stěžejní pro

všechny obory, v nichž se dá s rolí vlastních jmen počítat, od stylistiky a pragmatiky po medicinsko-lingvistický výzkum specifických jazykových projevů (např. pacientů s neurodegenerativními poruchami či schizofrenií; viz Collins a kol., 2023). Vedle ryze onomastických výzkumů lze nový korpus využít i při výzkumu jazyka totality a jeho proměn, metamorfóz publicistického stylu v průběhu 20. století a v historicky zaměřených analýzách.

LITERATURA

Brezina, V. (2018): *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Collins, L. — Brezina, V. — Demjen, Z. — Semino, E. — Woods, A. (2023): Corpus linguistics and clinical psychology: Investigating personification in first-person accounts of voice-hearing. *International Journal of Corpus Linguistics*, 28, s. 28–59.

Cvrček, V. — Laubeová, Z. — Lukeš, D. — Poukarová, P. — Řehořková, A. — Zasina, A. J. (2020): *Registry v češtině*. Praha: NLN.

Cvrček, V. (2021): *Calc 1.04*. URL: https://www.korpus.cz/calc/ (poslední přístup: 8. 11. 2023).

David, J. (2013): Historická sémantika proprií. In: J. David a kol., *Slovo a text v historickém kontextu. Perspektivy historickosémantické analýzy jazyka*. Ostrava — Brno: Ostravská univerzita — Host, s. 100–125.

David, J. — Klemensová, T. — Místecký, M. (2021): Věda mnoha jmen: onomastické termíny v publicistice Českého národního korpusu. *Jazykovedný časopis*, 72, s. 114–123.

David, J. — Klemensová, T. — Místecký, M. (2023): Appellativization of Proper Names — In the Perspective of Corpus Analysis. *Jazykovedný časopis*, 74, s. 32–42.

David, J. — Klemensová, T. — Místecký, M. a kol. (2022): *Od etymologie ke krajině. Onomastika pro 21. století*. Brno: Host.

David, J. — Místecký, M. (2023): Prolegomena ke kvantitativní onomastice. *Acta onomastica*, 64, s. 301–320.

Dvořák, V. (2017): Dativ. In: P. Karlík — M. Nekula — J. Pleskalová (eds.), *CzechEncy — Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/DATIV (poslední přístup: 8. 11. 2023).

Geržová, H. (2016): *Analýza jazykového materiálu v korpusu GEOGRAF*. Bakalářská práce, Filozofická fakulta, Masarykova univerzita.

Janda, L. A. — Fidler, M. — Cvrček, V. — Obukhova, A. (2022): The case for case in Putin's speeches. *Russian Linguistics*, 47, s. 15–40.

Knappová, M. (1995): Obchodní jméno jako fenomén onomaziologický a sociologický. *Slovo a slovesnost*, 56, s. 276–284.

Kocek, J. (2023): *Korpus OnomOs*. URL: https://wiki.korpus.cz/doku.php/cnk:onomos (poslední přístup: 12. 2. 2024).

Kolářová, I. (1999): Jméno obce jako součást názvů územních celků a institucí (konkurence nominativu jmenovacího a lokálu). In: K. Klímová — H. Kneselová (eds.), *Propria v systému mluvnickém a slovotvorném*. Brno: Pedagogická fakulta Masarykovy univerzity, s. 121–126.

Kováříková, D. (2021): *i.p.m.* URL: https://wiki.korpus.cz/doku.php/pojmy:ipm (poslední přístup: 8. 11. 2023).

Kuba, L. — Šrámek, R. (eds.) (1989): *Chrématonyma z hlediska teorie a praxe*. Brno: Onomastická komise ČSAV.

Mitter, P. (2003): Možnost aplikace vztahových modelů u názvů restaurací. *Acta onomastica*, 44, s. 48–52.

Motschenbacher, H. (2020): Corpus Linguistic Onomastics. A Plea for a Corpus-Based Investigation of Names. *Names*, 68, s. 88–103.

Plíčková, K. (2017): *Vyvážení a rozšíření korpusu GEOGRAF*. Bakalářská práce, Filozofická fakulta, Masarykova univerzita.

Straková, J. — Straka, M. — Hajič, J. (2019): Neural Architectures for Nested NER through Linearization. In: A. Korhonen — D. Traum — L. Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florencie: Association for Computational Linguistics, s. 5326–5331.

Ševčíková, M. — Žabokrtský, Z. — Krůza, O. (2007): Named Entities in Czech: Annotating Data and Developing NE Tagger. In: V. Matoušek — P. Mautner (eds), *Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science*. Berlin — Heidelberg: Springer, s. 188–195.

Ševčíková, M. — Žabokrtský, Z. — Straková, J. — Straka, M. (2014): *Czech Named Entity Corpus 2.0*. URL: http://hdl.handle.net/11858/00-097C-0000-0023-1B22-8 (poslední přístup: 27. 4. 2024).

Michal Místecký | Katedra českého jazyka Filozofické fakulty Ostravské univerzity |
Reální 5, 701 03 Ostrava
ORCID ID: 0000-0002-9183-4435
michal.mistecky@osu.cz

Jaroslav David | Katedra českého jazyka Filozofické fakulty Ostravské univerzity |
Reální 5, 701 03 Ostrava
ORCID ID: 0000-0003-2443-5351
jaroslav.david@osu.cz

Jana Davidová Glogarová | Katedra českého jazyka Filozofické fakulty Ostravské univerzity |
Reální 5, 701 03 Ostrava
ORCID ID: 0000-0002-3946-1379
jana.davidova@osu.cz

Tereza Klemensová | Katedra českého jazyka Filozofické fakulty Ostravské univerzity |
Reální 5, 701 03 Ostrava
ORCID ID: 0000-0002-9635-0851
tereza.klemensova@osu.cz