JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA

SLOVENSKEJ AKADÉMIE VIED

4

ROČNÍK 72, 2021

sciendo

SAP
SLOVAK ACADEMIC PRESS

JAZYKOVEDNÝ ČASOPIS
VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS
SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s. r. o.
- elektronicky vo vydavateľstve Sciendo – De Gruyter
  https://content.sciendo.com/view/journals/jazcas/jazcas-overview.xml

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
Kontakt: gabriela.mucskova@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: http://www.juls.savba.sk/ediela/jc/

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2021/72/4) – jún 2022

CiteScore 2020: 0,4
SCImago Journal Rank (SJR) 2020: 0,186
Source Normalized Impact per Paper (SNIP) 2020: 0,876

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is covered by the following services: Baidu Scholar; Cabell's Directory; CEJSH (The Central European Journal of Social Sciences and Humanities); CEEOL (Central and Eastern European Online Library); CNKI Scholar (China National Knowledge Infrastructure); CNPIEC – cnpLINKer; Dimensions; DOAJ (Directory of Open Access Journals); EBSCO (relevant databases); EBSCO Discovery Service; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); Genamics JournalSeek; Google Scholar; IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); International Medieval Bibliography; J-Gate; JournalGuide; JournalTOCs; KESLI-NDSL (Korean National Discovery for Science Leaders); Linguistic Bibliography; Linguistics Abstracts Online; Microsoft Academic; MLA International Bibliography; MyScienceWork; Naver Academic; Naviga (Softweco); Primo Central (ExLibris); ProQuest (relevant databases); Publons; QOAM (Quality Open Access Market); ReadCube; SCImago (SJR); SCOPUS; Semantic Scholar; Sherpa/RoMEO; Summon (ProQuest); TDNet; Ulrich's Periodicals Directory/ulrichsweb; WanFang Data; WorldCat (OCLC).

ISSN 0021-5597 (tlačená verzia/print) **ISSN 1338-4287 (verzia online)** **MIČ 49263**

JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA
SLOVENSKEJ AKADÉMIE VIED

4

ROČNÍK 72, 2021

Mimoriadne číslo Jazykovedného časopisu je venované problematike budovania webových korpusov ako zdrojov lingvistických výskumov a ich aplikácií.

Prizvaní editori:
Radovan Garabík
Zuzana Puchovská

sciendo

SAP
SLOVAK ACADEMIC PRESS

OBSAH

CONTENT

Štúdie
Studies

sciendo
DOI 10.2478/jazcas-2022-0010

KONZISTENCE MORFOLOGICKÉHO SLOVNÍKU MORFFLEX[1]

JAROSLAVA HLAVÁČOVÁ – MARIE MIKULOVÁ – BARBORA ŠTĚPÁNKOVÁ
Matematicko-fyzikální fakulta, Univerzita Karlova, Praha, Česká republika

HLAVÁČOVÁ, Jaroslava – MIKULOVÁ, Marie – ŠTĚPÁNKOVÁ, Barbora: Consistency of morphological dictionary MorfFlex. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 855 – 861.

Abstract: Language corpora usually contain, in addition to their own texts, various types of annotations. The most common one is a morphological annotation, which consists in assigning a lemma and a morphological tag to each wordform. For morphological tagging, morphological dictionaries are traditionally used. Our paper presents a new version of the so-called "Prague" morphological dictionary MorfFlex used for tagging many Czech corpora (particularly Prague Dependency Treebanks, corpora published by the Institute of the Czech National Corpus in Prague or large Czech web corpora of the Aranea series). Three basic principles were used to update the dictionary: the Golden Rule of Morphology, the Principle of Paradigm Unity, and the Principle of Paradigm Uniqueness.

Key words: morphological dictionary, morphological analysis, language corpus, the Czech language

1. MOTIVACE

MorfFlex je morfologický slovník, který se používá v celé řadě nástrojů automatického zpracování češtiny. Jde zejména o morfologickou analýzu mnoha českých korpusů, která se provádí pomocí nástroje MorphoDiTa (Straka et al., 2014; Straková et al., 2014). Nejznámější jsou korpusy řady SYN (Hnátková et al., 2014) vydávané Ústavem Českého národního korpusu FF UK v Praze nebo velké české webové korpusy řady Aranea (Benko, 2014). Je také základem pro ruční anotaci Pražských závislostních korpusů (Hajič et al., 2017) vytvářených v Ústavu formální a aplikované lingvistiky MFF UK.

Jeho základy a koncepce byly položeny v posledních dvou desetiletích 20. století (Hajič, 2004). Slovník vznikal postupně a na jeho budování se podílela řada přispěvatelů. Jeho současná velikost je více než 125 mil. slovních tvarů (125 348 901 – prosinec 2020). Za zhruba 30 let jeho používání jak pro tagování, tak i pro generování českých textů se mnohé změnilo. Ukazuje se navíc, že slovník samotný již není pro nástroje automatického zpracování přirozených jazyků (např. pro rozpoznávání jednotlivých slovních forem) tak důležitý, neboť dnešní automatické nástroje se umí samy učit

[1] Tento příspěvek byl podpořen projektem Ministerstva školství, mládeže a tělovýchovy ČR: LINDAT/CLARIAH-CZ (LM2018101).

z textů. Existence a dobrý stav slovníku je však zásadní pro stanovení norem morfologického značkování.

Během mnohaletého používání morfologického slovníku v nejrůznějších úlohách automatického zpracování českých textů se ukázalo, že některé principy, podle kterých byl vytvořen, by bylo vhodné změnit, některé přidat, některé dokonce vypustit. Bezprostřední motivací se stal projekt ruční morfologické anotace velkého objemu dat pro korpus PDT-C 1.0 (Hajič et al., 2020a). Za dlouhá léta došlo při údržbě slovníku k tomu, že některé změny v něm provedené přestaly být konzistentní s daty v korpusech. Původní pokyny pro anotaci se rozcházely s tím, co bylo skutečně ve slovníku. Proto jsme přistoupili k zásadní revizi slovníku. Ta probíhala paralelně s anotací korpusových dat.

Cílem bylo doplnit slovník o chybějící slovní tvary, a především učinit slovník konzistentním, aby stejné jazykové jevy byly popisovány vždy stejným způsobem. Revize tedy měla tři hlavní části:

1. přidání „neznámých" slov, tedy slov, která slovník neobsahoval,
2. upravení morfologické značky přesnějšími definicemi jednotlivých morfologických kategorií a jejich hodnot,
3. zajištění konzistence slovníku jako celku.

1.1 Nová slova

První požadavek není třeba příliš rozvádět. Snahou bylo obsáhnout co nejvíce slov. Zdrojem pro nová slova byl především vznikající korpus PDT-C 1.0.

1.2 Zpřesnění definic

Bylo třeba přesně vymezit definice všech morfologických kategorií a jejich hodnot. Zavedli jsme čtyři nové slovní druhy (cizí slovo, segment, zkratka, samostatné písmeno), přidali kategorii slovesného vidu ke všem slovesům, zpřesnili jsme značkování variant a zkratek, nově jsme zavedli značkování tzv. agregátů (např. *naň, byls*), které nebyly dosud uspokojivě popsány v žádném českém morfologickém slovníku. Všechny zásadní změny týkající se morfologického popisu jednotlivých slovních tvarů byly prezentovány na konferenci Slovko 2019 (Hlaváčová et al., 2019). Podrobná dokumentace je v práce Mikulová et al. (2020).

1.3 Konzistence slovníku

Některá pravidla, podle kterých byl slovník dlouhá léta budován, nebyla dodržována, některá dokonce podmínku konzistence tak, jak ji chápeme dnes, nesplňovala, proto bylo třeba je změnit nebo i vypustit, jiná naopak přidat.

Principy, kterými jsme se řídili při úpravách morfologického slovníku, jsou:

- Zlaté pravidlo morfologie
- Jednota paradigmatu
- Jedinečnost paradigmatu

2. ZLATÉ PRAVIDLO MORFOLOGIE

Základním pravidlem je tzv. Zlaté pravidlo morfologie, které říká, že jedna dvojice <lemma, morfologická značka> nemůže popisovat více než jeden slovní tvar.

K porušení zlatého pravidla ve „starém" slovníku docházelo poměrně často. Důvodem bylo nedůsledné rozlišování variant slovních tvarů. Např. lemma *Sedlec* mělo pro stejnou značku NNIS6-----A---- dva různé tvary, totiž *Sedleci* a *Sedlci*. Slovní tvary *nervosní*, *nervózní* měly ve slovníku stejné lemma (*nervózní*) a stejný vzor, což mělo za důsledek, že dokonce každá značka spolu s tímto lemmatem popisovala dva rozdílné slovní tvary. Oba právě uvedené příklady je možno najít v referenčních korpusech řady SYN, např. SYN2015 (dostupný z www.korpus.cz).

Aby nedocházelo k porušování Zlatého pravidla morfologie, je třeba především jednoznačně popsat varianty. Zejména je třeba důsledně rozlišovat mezi variantami flektivními (příklad *Sedlci* x *Sedleci* výše) a globálními (příklad *nervosní* x *nervózní* výše).[2]

2.1 Globální varianty (full paradigm variants)

Globální varianty se liší na úrovni lemmatu a projevují se stejně v celém paradigmatu (např. *citron*, *citrón*). Popisujeme je jako samostatné jednotky, ovšem s odkazem na příslušnou „základní" variantu (viz dále). Odkaz se pak stává součástí lemmatu.

Kódování odkazu využívá již zavedeného formalismu užívaného pro derivační odkazy (detaily viz Mikulová et al., 2020, s. 31). Typ odkazu na variantu může být buď DD (víceméně rovnocenná varianta), GC (varianta nespisovná nebo jinak příznaková), nebo DS (častý překlep, nebo jinak „poškozené" lemma). Všechny typy doprovází upřesnění stylu. V příkladu uvedeném v tabulce 2 má styl se značkou _,s význam „další standardní varianta".

Jedna ze dvou (případně více) variant byla vždy vybrána jako „základní". Pravidlo pro výběr základní varianty nebylo určeno striktně, protože jednoduché a jednoznačné určení není snadné. Pokud to bylo rozlišitelné, jako základní byla zvolena varianta spisovná (oproti nespisovné), běžnější (oproti méně frekventované), modernější (oproti zastaralé). V některých případech, především u cizích vlastních jmen, která mají často více globálních (zejména ortografických) variant, byl výběr té základní víceméně náhodný. V tabulce 1 uvádíme nejběžnější typy globálních variant s příklady.

2.2 Flektivní varianty (wordform variants)

Jestliže jsou dvě různé formy patřící do téhož paradigmatu (tzn. mající stejné lemma) popsány stejnými hodnotami všech relevantních morfologických kategorií, říkáme jim flektivní varianty. Jejich morfologické značky se musí lišit – k popisu flektivních variant je určena 15. pozice morfologické značky. Příkladem mohou být

[2] Používáme zde terminologii J. Hlaváčové (2009). V dalším textu v závorce uvádíme anglický ekvivalent, který jsme zavedli pro anglickou verzi popisu (porov. Mikulová et al., 2020).

tvary *obchodu*, *obchodě*, které mají stejné lemma (*obchod*), stejný rod (mužský neživotný), číslo (jednotné) i pád (6).

Typ	Příklady		Typ	Příklady	
dlouhá-krátká	*Abrahám*	*Abraham*	tvrdá-měkká	*student*	*študent*
	acetylén	*acetylen*		*vlaštovka*	*vlašťovka*
	salón	*salon*		*dolík*	*ďolík*
	apetýt	*apetyt*		*Bardejov*	*Bardějov*
	alexandrín	*alexandrin*		*čtyřhranný*	*čtyrhranný*
	přezůvky	*přezuvky*		*zbrzďování*	*zbržďování*
	Plútarchos	*Plutarchos*	t-th	*tema*	*thema*
	regulérní	*regulerní*	á-e	*originální*	*originelní*
é-í/ý	*kolébka*	*kolíbka*	ý-ej	*mýdlo*	*mejdlo*
	okénko	*okýnko*	protetické v	*olej*	*volej*
z-s	*klauzule*	*klausule*	jiné	*Afganistan*	*Afghanistan*

Tab. 1. Příklady nejběžnějších typů globálních variant v češtině

Je zřejmé, že globální i flektivní varianty se mohou libovolně kombinovat.

Příklad je uveden v tabulce 2. Druhý a třetí sloupec říká, o jaký typ varianty pro uvedený slovní tvar jde. 0 znamená základní varianta, + varianta vyznačená v lemmatu (pro globální variantu) nebo v morfologické značce na 15. pozici (pro flektivní variantu).

Slovní tvar	Flekt. var.	Glob. var.	Lemma	Morfologická značka
intensivními	0	+	intensivní_,s_^(^DD**intenzívní)	AAFP7----1A----
intensivníma	+	+	intensivní_,s_^(^DD**intenzívní)	AAFP7----1A---6
intenzívními	0	0	intenzívní	AAFP7----1A----
intenzívníma	+	0	intenzívní	AAFP7----1A---6

Tab. 2. Příklad kombinací globálních a flektivních variant

3. JEDNOTA PARADIGMATU

Pravidlo jednoty paradigmatu říká, že některé morfologické kategorie musí mít v celém paradigmatu stejnou hodnotu. Konkrétně je to slovní druh, u sloves vid a u podstatných jmen rod.

3.1 Jednota slovního druhu

Jestliže se dvě homonymní lemmata[3] liší slovním druhem, je třeba je odlišit pomocí číselného indexu. Příkladem jsou lemmata *hnát-1* (sloveso ve významu „utíkat“) a *hnát-2* (podstatné jméno ve významu „pařát“).

[3] V této větě chápeme lemma jako základní slovní tvar nějakého paradigmatu, tedy nikoli jako abstraktní jednotku, což by bylo možná přesnější, avšak vedlo by to k méně srozumitelnému vyjádření.

3.2 Jednota vidu

Dodržení slovesného vidu napříč celým paradigmatem je také přirozeným požadavkem. Homonymních sloves s odlišným videm není mnoho (např. dokonavé *napovídat-1* s významem „hodně mluvit“, nedokonavé *napovídat-2* „ve škole“).

3.3 Jednota rodu

Požadavek jednoty rodu napříč celým paradigmatem substantiva vypadá také na první pohled přirozeně, ale vede k ne zcela přirozenému rozdělení paradigmat podstatných jmen, jejichž rod kolísá. Kvůli zachování jednoty rodu podstatných jmen v rámci stejného paradigmatu jsme museli taková lemmata rozdělit do dvou. Jde např. o kolísání mezi ženským a mužským rodem u lemmatu *kredenc*, ale i kolísání mezi mužským životným a neživotným rodem u lemmatu *tenor* (zpěvák vs. hlas). Výsledkem použití pravidla jednoty rodu jsou tedy dvojice lemmat *kredenc-1* s paradigmatem rodu mužského neživotného a *kredenc-2* v rodě ženském. Podobně *tenor-1_(zpěvák)* v rodě mužském životném a *tenor-2_(hlas)* v rodě mužském neživotném.

Toto pravidlo si vyžádalo též poněkud komplikovaný popis slovních tvarů, jejichž rod v jednotném čísle je jiný než v čísle množném. Takových slov má čeština naštěstí málo. Jsou to: *oko*, *ucho*, *dítě* (včetně odvozenin, např. *biodítě*).

Oko, ucho

Obě lemmata mají v jednotném čísle vždy střední rod, v množném čísle mají ale dvojí skloňování. Pravidelné skloňování si zachovává i v množném čísle rod střední (*oka*, *ucha*), ve významu části těla je však skloňování odlišné (*oči*, *uši*), které odpovídá rodu ženskému. Zjevně zde dochází k porušení pravidla jednoty paradigmatu, proto bylo třeba paradigmata rozdělit a přiřadit jim rozdílná lemmata. Máme tedy lemma *oko-1*, které se skloňuje v jednotném i množném čísle pravidelně, a má tedy v celém paradigmatu rod střední. Kromě toho existuje lemma *oko-2*, jehož paradigma obsahuje jen tvary množného čísla, všechny rodu ženského. Stejným způsobem jsme rozdělili i *ucho*.

Dítě

Lemma *dítě* je odlišné, protože nemá „celé“ paradigma (jednotné i množné číslo ve stejném rodě) vůbec. Výsledkem rozdělení je tedy lemma *dítě-1*, jehož paradigma obsahuje jen tvary jednotného čísla (a středního rodu), a lemma *dítě-2*, jehož paradigma má jen tvary množného čísla (a ženského rodu).

Všechny tři uvedené případy změny rodu uvnitř paradigmatu mohly být pojaty ještě jinak, a to zvolením jiného lemmatu. Místo lemmatu s číslem -2 jsme uvažovali o lemmatu v množném čísle, tedy *oči, uši, děti*. Zvolené řešení je odůvodněno tím, že jsme chtěli zachovat příslušnost množného čísla k „základnímu“ tvaru čísla jednotného.

4. JEDINEČNOST PARADIGMATU

Pokud má nějaké slovo v rámci jednoho slovního druhu více významů, ale všechny mají stejnou sadu slovních tvarů popsaných stejnou sadou morfologických značek, ve slovníku mu náleží pouze jediné lemma.

Příkladem může být *kolej*, která má dva rozdílné významy, vzniklé rozdílným původem obou slov. Vzhledem k tomu, že MorfFlex je primárně morfologický slovník, významy rozlišujeme jen tehdy, když mají příslušná lemmata rozdílná paradigmata. Homonymní *kolej* tedy není třeba dělit do více paradigmat s odlišným lemmatem (pomocí číslování), protože v obou významech je množina tvarů stejná. Toto pravidlo nám umožnilo zredukovat počet slov, která se dostala jako samostatná lemmata do jedné z prvních verzí slovníku ze starších českých slovníků (např. Slovník spisovného jazyka českého, 1971).

Uplatňování pravidla jedinečnosti paradigmatu se vztahuje pouze k flexi, nikoli k derivačnímu chování. V našem příkladu s lemmatem *kolej* je sice množina tvarů pro oba významy stejná, ovšem každý význam má jiné odvozeniny. *Kolej* jakožto ubytovací zařízení tvoří přídavné jméno *kolejní*, které není možno vztáhnout k druhému významu *kolej* jako kolejnice, případně stopa vyhloubená kolem vozu. Druhý význam také tvoří přídavné jméno, ale *kolejový*, které zase nelze vztáhnout k prvnímu významu. Podobný „rozpor" bychom našli u lemmatu *matka*, jehož jeden význam (máma) umožňuje vytvořit přídavné jméno přivlastňovací (*matčin*), zatímco druhý (matice ke šroubu) takovou možnost neskýtá.

Rozhodli jsme se však odhlédnout i od derivačních vzorců a pojmout slovník jako čistě morfologický, takže rozdílné derivační chování nebereme při rozdělování paradigmat do úvahy.

5. ZÁVĚR

Aktualizace morfologického slovníku MorfFlex[4] (Hajič et al., 2020b) probíhaly paralelně s ruční morfologickou anotací konsolidovaného vydání Pražského závislostního korpusu PDT-C 1.0 (Hajič et al., 2020a).[5] Hlavním cílem bylo uvést do souladu obsah slovníku s anotacemi v korpusu. Morfologický slovník MorfFlex verze 2.0 i korpus PDT-C 1.0 jsou uloženy v datovém úložišti LINDAT.

Obsah slovníku MorfFlex 2.0 nyní vyhovuje všem třem principům představeným na začátku tohoto příspěvku.

Bibliografie

BENKO, Vladimír: Aranea: Yet Another Family of (Comparable) Web Corpora. In: Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September

[4] http://hdl.handle.net/11234/1-3186

[5] http://hdl.handle.net/11234/1-3185

8–12, 2014. Proceedings. Eds. P. Sojka – A. Horák – I. Kopeček – K. Pala. LNCS, Springer International Publishing Switzerland 2014. s. 247–256.

HAJIČ, Jan: Disambiguation of Rich Inflection. (Computational Morphology of Czech). Praha: Karolinum 2004.

HAJIČ, Jan – HAJIČOVÁ, Eva – MIKULOVÁ, Marie – MÍROVSKÝ, Jiří: Prague Dependency Treebank. In: Handbook on Linguistic Annotation. Eds. N. Ide – J. Pustejovsky. Berlin: Springer Verlag 2017, s. 555–594.

HAJIČ, Jan – BEJČEK, Eduard – BÉMOVÁ, Alevtina – BURÁŇOVÁ, Eva – FUČÍKOVÁ, Eva – HAJIČOVÁ, Eva – HAVELKA, Jiří – HLAVÁČOVÁ, Jaroslava – HOMOLA, Petr – IRCING, Pavel – KÁRNÍK, Jiří – KETTNEROVÁ, Václava – KLYUEVA, Natalia – KOLÁŘOVÁ, Veronika – KUČOVÁ, Lucie – LOPATKOVÁ, Markéta – MAREČEK, David – MIKULOVÁ, Marie – MÍROVSKÝ, Jiří – NEDOLUZHKO, Anna – NOVÁK, Michal – PAJAS, Petr – PANEVOVÁ, Jarmila – PETEREK, Nino – POLÁKOVÁ, Lucie – POPEL, Martin – POPELKA, Jan – ROMPORTL, Jan – RYSOVÁ, Magdaléna – SEMECKÝ, Jiří – SGALL, Petr – SPOUSTOVÁ, Johanka – STRAKA, Milan – STRAŇÁK, Pavel – SYNKOVÁ, Pavlína – ŠEVČÍKOVÁ, Magda – ŠINDLEROVÁ, Jana – ŠTĚPÁNEK, Jan – ŠTĚPÁNKOVÁ, Barbora – TOMAN, Josef – UREŠOVÁ, Zdeňka – VIDOVÁ, Hladká Barbora – ZEMAN, Daniel – ZIKÁNOVÁ, Šárka – ŽABOKRTSKÝ Zdeněk: Prague Dependency Treebank – Consolidated 1.0. Data/software, LINDAT/CLARIAH-CZ digital library. Prague: Charles University 2020a. Dostupné na: http://hdl.handle.net/11234/1-3185.

HAJIČ, Jan – HLAVÁČOVÁ, Jaroslava – MIKULOVÁ, Marie – STRAKA, Milan – ŠTĚPÁNKOVÁ Barbora: MorfFlex CZ 2.0. Data/software, LINDAT/CLARIAH-CZ digital library. Prague: Charles University 2020b. Dostupné na: http://hdl.handle.net/11234/1-3186.

HLAVÁČOVÁ, Jaroslava – MIKULOVÁ, Marie – ŠTĚPÁNKOVÁ, Barbora – HAJIČ Jan: Modifications of the Czech morphological dictionary for consistent corpus annotation. In: Jazykovedný časopis, 2019, roč. 70, č. 2, s. 380–389.

HLAVÁČOVÁ, Jaroslava: Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Ph.D. thesis, Praha: FF UK 2009.

HNÁTKOVÁ, Milena – KŘEN, Michal – PROCHÁZKA, Pavel – SKOUMALOVÁ, Hana: The SYN-series corpora of written Czech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC‘14). Reykjavík: ELRA 2014, s. 160–164.

MIKULOVÁ, Marie – HAJIČ, Jan – HANA, Jiří – HANOVÁ, Hana – HLAVÁČOVÁ, Jaroslava – JEŘÁBEK, Emil – ŠTĚPÁNKOVÁ, Barbora – VIDOVÁ, Hladká Barbora – ZEMAN, Daniel: Manual for Morphological Annotation, Revision for the Prague Dependency Treebank – Consolidated 2020 release. Technical report no. 2020/TR-2020-64, Praha: Institute of Formal and Applied Linguistics, Charles University 2020.

Slovník spisovného jazyka českého. Hl. red. B. Havránek. Praha: Nakl. Československé akademie věd 1960–1971.

STRAKA, Milan – STRAKOVÁ, Jana: MorphoDiTa: Morphological Dictionary and Tagger. Data/software, LINDAT/CLARIAH-CZ digital library. Prague: Charles University 2014. Dostupné na: http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0.

STRAKOVÁ, Jana – STRAKA, Milan – HAJIČ, Jan: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics 2014, s. 13–18.

SYN. ÚČNK FF UK, Praha. Dostupný z http://www.korpus.cz [25. 06. 2018]

sciendo
DOI 10.2478/jazcas-2022-0011

TYP *KLADENŠTÍ* JAKO PROBLÉM AUTOMATICKÉ MORFOLOGICKÉ ANALÝZY[1]

KLÁRA OSOLSOBĚ – HANA ŽIŽKOVÁ
Filozofická fakulta, Masarykova univerzita, Brno, Česká republika

OSOLSOBĚ, Klára – ŽIŽKOVÁ, Hana: *Kladenští* type as a problem of automatic morphological analysis. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 862 – 872.

Abstract: The aim of our paper is to demonstrate the procedures by which the data needed to refine tools for automatic morphological analysis of Czech can be obtained using a corpus, namely the Araneum Bohemicum IV Maximum (Czech, 20.03) 7.10 G web corpus of the ARANEA series and Araneum Bohemicum Maximum (Czech, 15.04) 3,20 G (hereinafter Araneum). Particularly, we will focus on propria of the Kladenští type, i.e., substantivized adjectives of denoting groups of persons according to affiliation. The goal of the probe into the Aranea web corpus is: 1) a corpus-based description of frequented properties of the *Kladenští* type, which can be used as a starting point for rule disambiguation; 2) creating a list of the most frequent lemmas belonging to the *Kladenští* type, which can then be included into dictionaries of automatic morphological analyzers (e.g. the MorfFlex dictionary by Hajič and Hlaváčová). We believe that the probe can help improve the results of tools for automatic morphological analysis of Czech.

Key words: automatic morphological analysis, derivational type *Kladenští*, part of speech transition

1. AUTOMATICKÁ MORFOLOGICKÁ ANALÝZA V PŘÍPADĚ SLOVNĚDRUHOVÝCH PŘECHODŮ

K příčinám nedostatků automatické morfologické anotace patří: a) nedostatečné pokrytí slovníku a b) chybná desambiguace. V několika studiích (Osolsobě – Žižková, 2019; Osolsobě – Žižková, 2020) jsme se zabývali problémem obého, a sice v případě slovnědruhového přechodu adjektivum → substantivum. Jedním z typů, které zůstaly stranou našeho zájmu, byla vlastní jména typu *Kladenští*.

1.1 Substantiva typu *Kladenští*

O substantivním typu *Kladenští* se v oddílu věnovaném tvoření jmen osob zmiňuje *Velká akademická gramatika spisovné češtiny I* (Štícha a kol., 2018, s. 280n.). Píše se zde: „Tento typ názvů osob tvoří vlastně samostatnou kategorii. O jména obyvatelská ve vlastním slova smyslu jde u nich jen z malé části. Obvykle se totiž tímto typem názvu označují nikoli obecně (všichni či typičtí, vybraní apod.) obyvatelé dané obce,

[1] Tento text vznikl za podpory grantu MUNI/A/1181/2020 *Gramatika a lexikon češtiny*.

nýbrž bývají to nejčastěji organizované sociální skupiny lidí nějak příslušících k dané obci: nejčastěji jsou to sportovní kluby či oddíly, dále umělecké soubory (divadelní soubory, orchestry), pracovní kolektivy (např. zaměstnanci podniku), zastupitelské sbory apod. Z morfologického hlediska jde o jména pomnožná, nemají v daném významu singulár; obvykle se užívají v nominativu.“ Dodáváme, že z hlediska pravopisu se píší s velkým počátečním písmenem (viz Velká písmena – jména živých bytostí a přídavná jména od nich odvozená v Internetové jazykové příručce, 2020).

1.2 Poznámka k výskytu nenominativních tvarů substantiv typu *Kladenští*

Výše jsme uvedli, že typ *Kladenští* zahrnuje z morfologického hlediska jména pomnožná, která nemají v daném významu singulár; obvykle se užívají v nominativu. Přesto lze v korpusech najít i řídké doklady na nenominativní užití.

Vyskytují se také **v genitivu**:

*... To se jim vymstilo v 52. minutě, kdy se z ojedinělé šance **Brněnských** trefil Petr Lainka ...;*

*... smlouva **Brněnských** s táborským hejtmanem na Moravském Krumlově ...;*

*... Velmi důrazně bránili a znemožňovali tak úspěšné zakončování **Brněnských** ...;*

*... Třeba se nám a domácí Náměšti podaří náskok **Brněnských** vymazat! ...;*

*... a u **Brněnských** se zase nejvíce prosadili hoši ...;*

v **dativu**:

*... **Brněnským** nepomohla ani účast slovenských juniorských posil ...;*

*... ale na hřištích soupeřů se **Brněnským** zatím nedaří tolik jako v domácím prostředí ...;*

*... Boleslavští ze třetí příčky Gambrinus ligy byli favority proti **Brněnským**, kteří se krčí téměř na dně II. ligy ...;*

v **akuzativu**:

*... takže už jen závěrečná připomínka pro všechny **Brněnské** ...;*

*... Za nepříznivého stavu pro **Brněnské** se navíc na začátku sedmnácté minuty po srážce s Polanským skácel k zemi Jiří Trvaj ...;*

*... Ve druhém kole čeká na **Brněnské** další tým z Ostravy ...;*

zřídka i v **lokále**[2]:

*... a to neplatí jen o **Brněnských** ...;*

v **instrumentálu**:

*... Před **Brněnskými** tentokrát smekám klobouk ...;*

*... mezi **Brněnskými** se nekoná žádné snadné rozhodování ....*

1.3 Korpus Araneum jako zdroj dat pro výzkum substantiv typu *Kladenští*

Korpus Araneum skýtá, jak se domníváme, vhodná data pro náš výzkum. Zahrnuje totiž množství textů, v nichž je uvedený typ hojně zastoupen. Výraznou výhodou je také to, že se jedná o texty stažené z webu, které na rozdíl od textů v korpu-

[2] Upozorňujeme na homonymii typu *... Kostel na **Brněnské** zachránili věřící ...*

sech řady SYN méně vyhovují kodifikačním předpisům. Z tohoto aspektu je na korpusu Araneum možné testovat, nakolik je u zkoumaného typu desubstantivních adjektiv dodržována pravopisná norma (psaní s velkým počátečním písmenem).

1.4 Typy homonymie

Substantiva typu *Kladenští* (viz níže v příkladech 2 a 5) se vyznačují homonymními tvary s částí (plurálové tvary maskulin životných) paradigmatu adjektiv zakončených na *[sc]ký* tvořených z názvů obcí / městských částí / zemí (*brněnští*, *bratislavští*, *žďárští* aj.), která se z různých důvodů realizují s velkým počátečním písmenem (viz níže příklady 1 a 3). Kromě toho existuje ještě další typ homonymie, a sice substantivní příjmení končící na *[sc]ký* (příklad 4):

(1) *... je možno zhlédnout výstavu **Brněnští** starostové ...*
(2) *... **Brněnští** vtěsnali všechny své góly do rozmezí od 39. do 47. minuty ...*
(3) *... **Žďárští** mladíci odehráli o víkendu další várku ...*
(4) *... páni Florián, Hynek a Jan Jiří **Žďárští**, obvinění, že dne 1. října 1619 složili přísahu ...*
(5) *... **Žďárští** si tak po třech prohrách opět připisují tři body ...*

V tomto příspěvku se zaměříme na desambiguaci homonym i na vylepšení pokrytí slovníku *MorfFlex* (Hajič – Hlaváčová, 2016) užívaného v řadě nástrojů automatické morfologické analýzy češtiny.

2. RELEVANTNÍ LEMMATA ZASTOUPENÁ V KORPUSU ARANEUM

V prvním kroku jsme zkoumali možnosti, jak zjistit, která substantiva se v analyzovaném korpusu vyskytují. Vytvořili jsme CQL dotazy s cílem získat seznam tvarů, které patří k analyzovanému typu substantiv:
Dotaz: **[word="[[:upper:]].*[šč]tí" & lemma=".*[sc]ký"]**

Další dotazy jsme specifikovali na tvary zakončené na *[šč]tí* tak, že jsme se snažili definovat vlastnosti kontextu, který vylučuje adjektivní a podporuje substantivní interpretaci tvaru.

Ověřili jsme platnost následujících tvrzení:

2.1 Tvrzení a):

Stojí-li v bezprostředním pravém kontextu za tvarem končícím na *[šč]tí* s velkým počátečním písmenem tvar, který lze interpretovat jako jméno, maskulinum životné v nominativu plurálu,[3] pak je na místě interpretovat tvar zakončený na *[šč]tí* jako adjektivum s lemmatem končícím na *[sc]ký*.

[3] Pravidlo je použitelné pouze tam, kde je tvar v pravém kontextu za tvarem končícím na *[šč]tí* s velkým počátečním písmenem jednoznačný. Jinak je pravidlo závislé na desambiguaci víceznačného tvaru (v případě jmen s flexí typu *jarní* jako např. *rozhodčí*, *cestující*, *první*, ...).

Dotaz ověřující uvedené tvrzení:
[word="[[:upper:]].*[šč]tí" & lemma=".*[sc]ký"] [tag="..MP1.*"]

Výjimku z uvedeného pravidla představují případy, kdy po sobě následují dva tvary v nominativu, které netvoří syntaktický celek.[4] Platnost pravidla dále narušují např. chyby v interpunkci a aktualizovaný slovosled (postpozice zájmenného shodného přívlastku[5]). Ambiguitní čtení mohou mít i doklady, kdy za tvarem zakončeným na *[šč]tí* následuje valenční adjektivum, které může,[6] ale nemusí[7] mít i substantivní platnost.

2.2 Tvrzení b):

Stojí-li v bezprostředním pravém kontextu za tvarem zakončeným na *[šč]tí* tvar, který lze interpretovat[8] jako sloveso, adverbium, zájmeno (nikoli tvar maskulinum životné v nominativu plurálu), číslovku (nikoli tvar maskulinum životné v nominativu plurálu), nebo interpunkci, pak je na místě interpretovat tvar končící na *[šč]tí* buď jako substantivum s lemmatem zakončeným na *[šč]tí*, nebo jako substantivum s lemmatem končícím na *[sc]ký*.
Dotaz ověřující uvedené tvrzení:
[word="[[:upper:]].*[šč]tí" & lemma=".*[sc]ký"] [tag="[VDPCZ].*" & tag!="..MP1.*"]

Platnost pravidla narušují např. chyby v psaní velkého počátečního písmene u tvarů s adjektivní platností v postponovaném přívlastku.[9] Tvary, jimž předchází substantivum buď maskulinum v nominativu plurálu,[10] nebo koordinovaná skupina definovaná jako tvary s velkým počátečním písmenem proložené spojovacími tvary (čárka, spojka *a* nebo *i*),[11] mají platnost substantiva propria příjmení. Jejich lemmatem by měl být tvar zakončený na *[sc]ký* se substantivní plat-

[4] Např.: *... A od počátku zápasu byli **Vimperští** ti, co diktovali tempo hry ...; ... skončili **Hradečtí** druzí za Libercem ...*

[5] Např.: *... A dej, jakož **Valdenští** tito od počátku tobě věrní ...*

[6] Např.: *... Čeští **účinkující** v připravených stáncích na náměstíčku Piazza Duomo po dva dny zde nabízeli své produkty ...*

[7] Např.: *... Lisabonští **zuřící** nad zkázou svého města našli obětního beránka ...*

[8] V ambiguitních případech je pravidlo závislé na desambiguaci.

[9] Např.: *... Varani **Komodští** jsou však jediní veleještěři světa ...; ... synové **Izraelští** neposlechli mne ...*

[10] Např.: *... Bratři **Kaczynští** považují Německo za hrozbu ...; ... manželé **Škvorečtí** vydali jeho první českou knížku ...* U dokladu *... aby pánové **Velvarští** je i faráře jejich omluvena míti ráčili ...* narážíme na problematickou interpretaci šlechtických přídomků ve funkci příjmení, u nichž hraje roli znalost encyklopedického rázu, kterou lze těžko zachytit na úrovni automatické desambiguace (viz např. pravopisná chyba v dokladu *... Měšťané **Velvarští** užívajíce branného práva ...*). Další problém představují propria příjmení v textech s nejasnými hranicemi syntaktických celků: *... Vítejte na stránkách Včelařství **Vozdečtí** Náš obchod nabízí ...*

[11] Např.: *... sourozenci Ema a Mirek **Toužimští** byli od sebe odtrženi ...*

ností.[12] Substantivně interpretované tvary končící na *[šč]tí* se mohou vyskytnout i po tvarech dalších jmen v nominativu plurálu.[13] U nich je obtížné definovat pravidla pro desambiguaci mezi typem *Kladenští* a příjmením.

Pokud se v desambiguaci budeme opírat o morfologické značky tvarů, které se nacházejí v kontextu tvaru zakončeného na *[šč]tí*, narazíme na chyby ve značkování,[14] jejichž důvody jsou rozmanité.

2.3 Tvrzení c):

Stojí-li v bezprostředním pravém kontextu za tvarem zakončeným na *[šč]tí* tvar, který lze interpretovat[15] jako předložku a zároveň nejde o předložku *z(e)* následovanou tvarem s počátečním velkým písmenem, pak je na místě interpretovat tvar končící na *[šč]tí* jako substantivum s lemmatem zakončeným na *[šč]tí*, nebo jako substantivum s lemmatem zakončeným na *[sc]ký*.

Dotaz ověřující uvedené tvrzení:

[word="[[:upper:]].*[šč]tí" & lemma=".*[sc]ký"][tag="R.*"], n-filtr: **[lemma ="z"] [word="[[:upper:]].*"]**

Platnost pravidla narušují např. chyby v psaní velkého počátečního písmene u tvarů s adjektivní platností v postponovaném přívlastku.[16] Pro upřesnění pravidla lze opět využít filtrování substantiv, popřípadě dalších konstrukcí v levém kontextu, viz výše. V některých případech je ovšem desambiguace příjmení opřená o pravidla obtížná[17] a využití slovníku, který by zahrnoval seznam příjmení a seznam proprií typu *Kladenští*, se jeví jako dobrá volba.

---

[12] Kandidáty na interpretace proprium – příjmení je možné filtrovat např. tak, že budeme: a) filtrovat doklady, u nichž se v levém kontextu <-3, -1> nachází fráze
[word="[[:upper:]].*"][word="a"][word="[[:upper:]].*"];
b) filtrovat doklady, u nichž se v levém kontextu <-1, -1> nachází slovní tvar
[lc="bratř[íí]|.*manželé|sourozenci|páni|pánové"].
Kandidáty na interpretace adjektivum je možné filtrovat tak, že budeme filtrovat doklady, u nichž se v levém kontextu <-1, -1> nachází slovní tvary
[lc="synové|muži|obyvatelé|občané|měšťané|chlapci|mudrci|děkanové|koně|sedláci|mlynáři|vojvodové|bohové|preláti|hrdinové"].
Do seznamu lze přidat i další substantiva maskulina v nominativu plurálu.

[13] Např.: *... Původnější **Smiřičtí** se ze svého rodiště sice vystěhovali ...; ... První **Hodičtí** se objevují ve 14. století ...; ... U nich byli druzí **Sloupští** a třetí Papuč Team ...; ... když favorizovaní **Lipovečtí** také prohráli ...*

[14] Například v dokladu *... LP (DIY) **Brněnští** My Dead Cat fungují s menšími přestávkami od roku 2001 ...* je tvar anglického zájmena *my* interpretován automatickou morfologickou analýzou jako nominativ plurálu českého zájmena *my*, přičemž celé spojení *My Dead Cat* je víceslovný název, k němuž je tvar *brněnští* adjektivním přívlastkem chybně napsaným s velkým písmenem.

[15] V ambiguitních případech je pravidlo závislé na desambiguaci.

[16] Např.: *... Nadto zajali též synové **Izraelští** z bratří svých dvakrát sto tisíc žen ...*

[17] Např.: *... **Kopřivničtí** od té chvíle již tahali za kratší konec ...; ... O pozdější vrchnosti se dočítáme, že se psala jako **Sedlečtí** od Dubu ...; ... jenž si **Chomutovští** pro hokejisty připravili ...; ... Dům postavili bratři František, otec režiséra Zdeňka, a Jan **Podskalští** pro svou matku Marii ...*

2.4 Tvrzení d):

Stojí-li v bezprostředním pravém kontextu za tvarem končícím na *[šč]tí* tvar, který lze interpretovat[18] jako spojku a zároveň nejde o spojku *[ai]* následovanou tvary s počátečním velkým písmenem nebo tvary, které lze interpretovat jako maskulinum životné v nominativu plurálu, pak je na místě interpretovat tvar zakončený na *[šč]tí* jako substantivum s lemmatem zakončeným na *[šč]tí*, nebo jako substantivum s lemmatem zakončeným na *[sc]ký*.

Dotaz ověřující uvedené tvrzení:

[word="[[:upper:]].*[šč]tí" & lemma=".*[sc]ký"][tag="J.*" & word!="[ai]"] [word!="[[:upper:]].*"|tag!="..MP1.*"][word!="[[:upper:]].*"|tag!="..MP1.*"][19]

Platnost pravidla narušují např. chyby v psaní velkého počátečního písmene u tvarů s adjektivní platností v postponovaném přívlastku.[20] Problematické jsou víceznačné[21] případy.

Pro upřesnění pravidla (odlišení typu *Kladenští* od příjmení) lze opět využít filtrování substantiv, popřípadě dalších konstrukcí v levém kontextu,[22] viz výše.

3. NÁVRH NA DOPLNĚNÍ SLOVNÍKU O FREKVENTOVANÁ SUBSTANTIVA TYPU *KLADENŠTÍ*

Pokrytí proprií je obecně Achillovou patou slovníků automatických morfologických analyzátorů, *MorfFlex* nevyjímaje. Doplňování proprií není vždy systematické. Sonda, kterou jsme realizovali na datech korpusu Araneum, ukazuje dotazy, díky nimž lze najít kandidáty na substantiva typu *Kladenští* a sestavit jejich seznam pro doplnění slovníku. Kromě adekvátní lemmatizace a morfologického značení by bylo možné použít i sémantické značení (typ *;E*: příslušník národů, měst aj.). Právě korpusy Aranea totiž umožňují vyhledávání i podle sémantického značení.

Pro potřeby doplnění slovníku jsme vytvořili seznam substantivizovaných adjektiv typu *Kladenští* s výskytem 10 a více, která jsou doložena v korpusu Araneum.

Vyšli jsme z předpokladu, že substantiva typu *Kladenští* se podle platné pravopisné konvence píší na rozdíl od homonymních adjektivních tvarů s velkým počáteč-

---

[18] V ambiguitních případech je pravidlo závislé na desambiguaci.

[19] Viz https://kontext.korpus.cz/view?ctxattrs=word&attr_vmode=visible&pagesize=40&q=~aT4DZdBj7ZZx&viewmode=kwic&attrs=word%2Clemma%2Ctag&corpname=aranea%2Faranbohe_cs_ar13__a_a&attr_allpos=kw

[20] Např.: *... rozhodčí **Lounští** nebo dodaní Čba tak už jdite do P.!!! ...; ... Orli Moršti ale i domácí mazlíčci ...*

[21] Např.: *... Starší **Mariánskolázeňští** však hrají tuto kategorii mimo soutěž ...*

[22] Např. *... Šárka a Petr **Záhrobští** nebo ...; ... manželé Milan a Marie **Petrovští** jako ...; ... jako jsou bratři **Cidlinští** nebo ...; ... bratři **Lipští** coby režisér a herec ...*

ním písmenem.[23] Následně jsme vyloučili doklady, kdy za vyhledanými tvary stojí tvar, který lze interpretovat jako maskulinum životné v nominativu plurálu.[24] Poté jsme odstranili doklady, kdy se vlevo od vyhledaného slova zakončeného na *[šč]tí* vyskytuje trojice slovních tvarů, kdy první slovní tvar začíná velkým počátečním písmenem, za ním následuje tvar *a* a za ním stojí slovní tvar začínající velkým počátečním písmenem.[25] Dále jsme eliminovali doklady, kdy se vlevo od vyhledaného slova končícího na *[šč]tí* vyskytují slovní tvary jako *manželé*, *sourozenci*, *synové*, *muži*, ...[26] Poté jsme odstranili doklady, kdy za vyhledanými tvary stojí lemma *z* (předložka) následované tvarem s velkým počátečním písmenem.[27] Pokračovali jsme eliminací dokladů, kdy se vpravo od vyhledaného slova zakončeného na *[šč]tí* vyskytuje trojice, která začíná slovními tvary *a* nebo *i*, za nimiž následuje opakovaně buď tvar s velkým počátečním písmenem, nebo tvar, který lze interpretovat jako maskulinum životné v nominativu plurálu.[28] V posledním kroku jsme odstranili doklady frekventovaných příjmení jako *Wacho[vw]ští*, *Kinští*, *Škvorečtí*, ...[29]

Sérií dotazů uvedených v poznámkách výše jsme získali seznam kandidátů,[30] z nichž lze vybrat např. podle ARF výrazy, které je možné následně doplnit do slovníku *MorfFlex* s lemmatem zakončeným na *[šč]tí* a substantivní interpretací. Se-

---

[23] Dotaz: **[word="[[:upper:]].*[šč]tí" & lemma=".*[sc|ký"]**. O tom, že tato norma bývá porušována, viz níže.

[24] Dotaz: n-filtr <1,1> **[tag="..MP1.*"]**. Jsme si vědomi, že takový postup se opírá o výsledky automatické morfologické analýzy.

[25] Dotaz: n-filtr <−3, −1> **[word="[[:upper:]].*"][word="a"][word="[[:upper:]].*"]**. Tento dotaz má za cíl odfiltrovat doklady typu ... *Antonín a Jan **Klatovští*** ..., ale i doklady jako ... *A. a B. Strugačtí* .... Dotazem ovšem odstraníme i řídké doklady typu ... *Obec Sudějov a SHŠ Vyšehradští zvou* ...

[26] Dotaz: n-filtr <−1, −1> **[lc="bratř[ií]|.*manželé|sourozenci|synové|muži|obyvatelé|občané|měšťané|chlapci|mudrci|děkanové|koně|sedláci|mlynáři|vojvodové|bohové|preláti|hrdinové|páni|pánové"]**. Dotaz má za cíl odfiltrovat jednak příjmení, jednak případy, kdy se tvar zakončený na *[šč]tí* nachází v pozici postponovaného přívlastku za substantivem a má adjektivní platnost. Jsme si vědomi toho, že uvedené omezení lze dále rozšiřovat o další substantiva.

[27] Dotaz: n-filtr <1,2> **[lemma="z"] [word="[[:upper:]].*"]**. Tento dotaz má za cíl odfiltrovat doklady typu ... *kterými se stali **Holičtí** ze Šternberka* ..., tedy doklady šlechtických přídomků.

[28] Dotaz: n-filtr <1,3> **[word="[ai]"][word="[[:upper:]].*"|tag="..MP1.*"][word="[[:upper:]].*"|tag="..MP1.*"]**. Dotaz má za cíl odfiltrovat doklady typu ... ***Plzeňští** a pardubičtí úředníci případ vidí jinak ...*,; ... ***Pražští** a brněnští diváci mohou ...; ... **Wachowští** i Tom Tykwer si příběhy rozdělili ...* Odfiltrují se tak ovšem i doklady jako ***Boskovičtí** i dobří lidé z okolí letos opět nezklamali – vánoční sbírka dárků pro boskovickou psí záchytnou stanici se zase vydařila ...; ... před kostelem se shromažďují **Olomoučtí** i hosté třímající v rukou pozvánky ...; ... Zatímco **Chomutovští** a hráči Komety ...; ... posílili **Chomutovští** i Slovákem Matúšem Kostúrem* ..., které jsou v pořádku. Takových není mnoho.

[29] Dotaz: n-filtr <0,0> **[word="Wacho[vw]ští|Kinští|Strugačtí|Škvorečtí|Kopečtí|Farští|Stránští|Kaczyńští|Smiřičtí|Hostinští|Rychetští"]**. Dotaz má za cíl odfiltrovat frekventovaná příjmení.

[30] Viz https://www.korpus.cz/kontext/freqs?maincorp=aranea%2Faranbohe_cs_ar13__ a_a&viewmode=kwic&pagesize=40&attrs=word%2Clemma%2Ctag&attr_vmode=visible-kwic&base_viewattr=word&q=~wEeIUyEwKi28&fcrit=lemma%2Fe%200~0%3E0&flimit=1&fpage=1&ftt_include_empty=1.

znam prvních 20 výrazů podle frekvence spolu s ARF uvádí tabulka 1 (data jsou z korpusu Araneum Bohemicum Maximum (Czech, 15.04) 3,20 G).

Tabulka 1

substantivum na *[šč]tí*	**frekvence**	**ARF**	**substantivum na *[šč]tí***	**frekvence**	**ARF**
Vsetínští	1087	427,05	Budějovičtí	657	332,44
Chomutovští	1012	458,84	Kladenští	648	321,98
Boleslavští	1011	438,19	Ostravští	615	343,75
Plzeňští	1005	519,52	Letenští	614	330,05
Třinečtí	948	435,03	Karvinští	534	250,85
Zlínští	905	438,47	Čáslavští	512	212,14
Liberečtí	872	466,15	Vítkovičtí	512	256,39
Hradečtí	799	370,48	Jihlavští	489	251,16
Pardubičtí	758	390,25	Brněnští	478	270,38
Českobudějovičtí	725	319,35	Jablonečtí	461	241,03

4. CHYBY V PRAVOPISU SUBSTANTIV TYPU *KLADENŠTÍ* – KORPUSOVÁ SONDA

Výše popsaná sonda do korpusu Araneum byla založena na datech, u nichž jsme předpokládali regulérní užití počátečního velkého písmena. Sonda odhalila i chyby v pravopisu (viz např. psaní velkého počátečního písmene v adjektivním užití). Zajímalo nás tudíž, nakolik je dodržení pravopisné normy v datech webového korpusu běžné. U prvních pěti nejfrekventovanějších lemmat (viz tabulka 1) jsme sledovali doklady, kdy se tvar zakončený na *[šč]tí* ve stejně vymezeném kontextu realizuje s malým počátečním písmenem. K pěti nejfrekventovanějším tvarům jsme vyhledali kontextově stejně definované tvary s malým počátečním písmenem.[31] Zjistili jsme, že v naprosté většině případů jde o pravopisné chyby. Výjimkou jsou ty případy, kdy za tvarem zakončeným na *[šč]tí* následuje substantivum v nominativu plurálu, které není rozpoznáno automatickou morfologickou analýzou,[32] nebo koordinovaná skupina životných maskulin.[33] Snažili jsme se tyto případy odstranit tak, že jsme filtrovali v pravém kontextu tvaru zakončeného na *[šč]tí* tvary napsané s velkým počátečním písmenem. Až na tři doklady[34] šlo o správně napsané tvary. Odstranili jsme tedy z bezpro-

[31] Viz zde https://kontext.korpus.cz/view?ctxattrs=word&attr_vmode=visible&pagesize=40&q=~EgEvNqVTlP08&viewmode=kwic&attrs=word%2Clemma%2Ctag&corpname=aranea%2Faranbohe_cs_ar13__a_a&attr_allpos=kw.

[32] Např. nebyly odstraněny správně napsané tvary v dokladech jako: *... Po třech extraligových kolech měli* ***boleslavští*** *Billy Boy na svém kontě nulu ...; ... Na té druhé akci se představí i* ***boleslavští*** *náladotvůrci ...; ... Mít* ***vsetínští*** *fans víc důvtipu, mohli na oplátku pokřikovat ...*

[33] Např. nebyly odstraněny správně napsané tvary v dokladech jako: *... Ale ligu obohacovali svou kvalitou stejně jako rok předtím i liberecký Štajner,* ***plzeňští*** *Bakoš, Horváth ...; ... Do té doby propásli brankové příležitosti* ***boleslavští*** *Kulič, Chramosta a Kysela ...*

[34] *Např.: ... dokázali* ***boleslavští*** *Chomutov porazit v obou utkáních ...; ... V 53. minutě kopali* ***boleslavští*** *Procházkou trestný kop ze třiceti metrů ...; ... ani ti* ***vsetínští*** *Zlín zrovna nešetří ...*

středního pravého kontextu tvarů zakončených na *[šč]tí* tvary napsané s velkým počátečním písmenem a tvar *fans*.[35] Několik málo dalších dokladů jsme neodstranili.[36] Na číselné údaje uvedené v tabulce 2 je třeba pohlížet s touto rezervou.

Tabulka 2 ukazuje počty tvarů končících na *[šč]tí* pěti vybraných lemmat napsaných s velkým a malým písmenem získaných stejným postupem z korpusu Araneum. Má ilustrovat přibližný procentuální poměr správně a chybně napsaných tvarů substantiv typu *Kladenští* ve sledovaném korpusu. Data v tabulce 2 jsou z korpusu Araneum Bohemicum Maximum (Czech, 15.04) 3,20 G.

Tabulka 2

filtrované tvary s velkým počátečním písmenem	**počet výskytů / %**	**ARF**	**filtrované tvary s malým počátečním písmenem**	**počet výskytů / %**	**ARF**	**celkem výskytů**
Vsetínští	1087 79 %	427,05 80 %	vsetínští	235 21 %	108,58 20 %	1122
Chomutovští	1012 74 %	458,84 74 %	chomutovští	346 26 %	162,81 26 %	1358
Boleslavští	1011 86 %	438,19 83 %	boleslavští	164 14 %	87,78 17 %	1175
Plzeňští	1005 69 %	519,52 69 %	plzeňští	448 31%	235,77 31 %	1453
Třinečtí	948 84 %	435,03 81 %	třinečtí	183 16 %	103,53 19 %	1131

Na základě uvedených šetření je patrné, že v psaní velkého písmene se u tohoto typu proprií v textech stažených z webu vyskytují poměrně často chyby (procentuální rozpětí je 20 až 30 %). Pravidla pro vyhledávání chybně napsaných dokladů nabídneme, aby mohla být využita pro připravovaný webový korektor (viz více Hlaváčková a kol. 2019).

5. ZÁVĚR

Naším cílem bylo zmapovat možnosti a meze automatické morfologické analýzy v případě typu *Kladenští*, tedy substantiv označujících skupiny osob podle pří-

[35] N-filtr <1,1> **[word="[[:upper:]].*"| word="fans"]**.

[36] Např. správně napsané doklady jako: *... Na ní představí své nové EP (12 tracků) „Woodland Journey" **plzeňští** blackouši Panychida ...; ... V dnešní době chodí **plzeňští** skins povzbuzovat klub Senco Doubravka ...; ... Mezi kandidáty chybí **plzeňští** rozhodčí Rejthar a Šindler ...; ... teenagery milovaní **plzeňští** mandrage ...; ... Na té druhé akci se představí i **boleslavští** náladotvůrci ...; ... české zástupce skywalker z liberce narychlo doplnili **třinečtí** scound a stay true ...*

slušnosti. Tato substantiva vznikla slovnědruhovým přechodem z adjektiv končících na *[sc]ký*. Na úvod jsme prověřili výskyt substantiv typu *Kladenští* v nenominativních tvarech. Ověřili jsme platnost tvrzení, že ve srovnání s tvary v nominativu se vyskytují velmi zřídka. Vyšli jsme z dat webového korpusu Araneum. Popsali jsme typy homonymie, s nimiž se setkáváme u tvarů s velkým počátečním písmenem zakončených na *[šč]tí*. Otestovali jsme série dotazů, kterými lze v korpusech hledat relevantní doklady analyzovaného typu substantiv. Ve formulaci dotazů jsme se maximálně snažili vyjít z formálních vlastností klíčových slov.[37] Při filtrování dat jsme použili informace vložené do korpusů v procesu automatické morfologické analýzy.[38] Jsme si vědomi omezení, která takový postup přináší.[39]

Sestavili jsme seznam adjektivizovaných substantiv, která se vyskytují v korpusu Araneum. Tento seznam lze využít k vylepšení (doplnění) slovníků užívaných v nástrojích automatické morfologické analýzy češtiny (například slovník *MorfFlex;* Hajič – Hlaváčová, 2016).

Ze seznamu jsme vzali pět nejfrekventovanějších substantiv a vyhledali jsme jejich tvary s malým počátečním písmenem ve stejných kontextech. Výskyty s malým počátečním písmenem jsme dále filtrovali a ručně procházeli. Zjistili jsme, že v naprosté většině dokladů získaných výše popsaným způsobem došlo k porušení kodifikace (psaní substantiv typu *Kladenští* s malým počátečním písmenem místo písmene velkého). Stanovili jsme, že na základě rešerše lze předpokládat, že k porušení kodifikace psaní velkých písmen u substantiv typu *Kladenští* dochází v průměru ve více než 25 % případů.

Domníváme se, že popsané postupy, které jsme použili v našich sondách, se mohou stát podkladem pravidlové desambiguace a mohou přispět ke zlepšení fungování automatických nástrojů pro zpracování přirozeného jazyka (tagery, korektory).

Bibliografie

BENKO, Vladimír: Araneum Bohemicum IV Maximum (Czech, 20.03) 7.10 G. Bratislava: Comenius University 2020. Dostupný na: http://unesco.uniba.sk/aranea_about [cit. 14. 12. 2020].

BENKO, Vladimír: Aranea: Yet Another Family of (Comparable) Web Corpora. In: TSD 2014, LNAI 8655. Eds. P. Sojka – A. Horák – I. Kopeček – K. Pala. Springer International Publishing 2014, s. 257–264.

[37] Například typický výskyt tvarů nominativu plurálu nás vedl k tomu, abychom kladené dotazy omezili na slovní tvary zakončené na *[šč]tí*. Podobně platná kodifikační norma (psaní s velkým počátečním písmenem) nás vedla k tomu, abychom se zaměřili na testování výskytu tvarů napsaných s velkým počátečním písmenem.

[38] Například jsme odstranili z pravého kontextu klíčových slovních tvarů s velkým počátečním písmenem končících na *[šč]tí* slovní tvary tagované jako nominativ plurálu životných maskulin.

[39] Například u tvarů tagovaných jako nominativ plurálu životných maskulin je interpretace u lemmat, jejichž tvary se tvoří podle vzoru *jarní* (např. substantiva typu *mluvčí*, *cestující*, ...), výsledkem desambiguace, která nemusí být bezchybná. Podobně tvary, které nejsou zaznamenány ve slovníku použitého morfologického analyzátoru, nejsou označkovány správně a navržená pravidla opřená o výsledky automatické morfologické analýzy je nezachytí.

HAJIČ, Jan – HLAVÁČOVÁ, Jaroslava: MorfFlex CZ 161115. Praha: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics(ÚFAL), Faculty of Mathematics and Physics, Charles University 2016. Dostupné na: http://hdl.handle.net/11234/1-1834 [cit. 14. 12. 2020].

HLAVÁČKOVÁ, Dana – HRABALOVÁ, Barbora – MACHURA, Jakub – MASOPUSTOVÁ, Markéta – MRKÝVKA, Vojtěch – VALÍČKOVÁ, Marie – ŽIŽKOVÁ, Hana: New Online Proofreader for Czech. In: Slavonic Natural Language Processing in the 21st Century. Eds. A. Horák – P. Rychlý – A. Rambousek. Brno: Tribun EU 2019, s. 79–92.

HNÁTKOVÁ, Milena – JELÍNEK, Tomáš – KOPŘIVOVÁ, Marie – PETKEVIČ, Vladimír – ROSEN, Alexandr – SKOUMALOVÁ, Hana – VONDŘIČKA, Pavel: Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. In: Korpus – gramatika – axiologie, 2018, roč. 17, s. 3–22.

Internetová jazyková příručka. Praha: Ústav pro jazyk český AV ČR, v. v. i. (2008–2020). Dostupný na: https://prirucka.ujc.cas.cz/?id=181 [cit. 14. 12. 2020].

JELÍNEK, Tomáš – KOPŘIVOVÁ, Marie – PETKEVIČ, Vladimír – SKOUMALOVÁ, Hana: Variabilita českých frazémů v úzu. In: Časopis pro moderní filologii, 2018, roč. 100, č. 2, s. 151–175.

PETKEVIČ, Vladimír – HLAVÁČOVÁ, Jaroslava – OSOLSOBĚ, Klára – SVÁŠEK, Martin – ŠIMANDL, Josef: Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. In: Jazykovedný časopis, 2019, roč. 70, č. 2, s. 358–369.

OSOLSOBĚ, Klára – ŽIŽKOVÁ, Hana: Improving Nominalized Adjectives Tagging. In: Jazykovedný časopis, 2019, roč. 70, č. 2, s. 370–379.

OSOLSOBĚ, Klára – ŽIŽKOVÁ, Hana: Homonymie mezi apelativy a proprii jako problém automatické morfologické analýzy češtiny. In: Acta onomastica, 2020, roč. LXI, č. 1, s. 161–174.

ŠTÍCHA, František a kol.: Velká akademická gramatika spisovné češtiny I. Morfologie: Druhy slov / Tvoření slov. Část I (VAGSČ I). Praha: Academia 2018.

sciendo
DOI 10.2478/jazcas-2022-0012

JAZYKOVÁ INTERPRETÁCIA NEMECKÉHO MIGRAČNÉHO DISKURZU (V KOMPARAČNOM POHĽADE ROKOV 2019 A 2015/16)[1]

EVA MOLNÁROVÁ – JANA LAUKOVÁ
Filozofická fakulta Univerzity Mateja Bela, Banská Bystrica, Slovensko

MOLNÁROVÁ, Eva – LAUKOVÁ, Jana: Language interpretation of German migration discourse (in comparison view of the years 2019 and 2015/16). Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 873 – 881.

Abstract: The presented paper is a research dive into the topic of web corpora as well as an analysis of linguistic grasp of the issue of migration from the perspective of social, cultural and cognitive linguistics. The presented research reflects the problem of the construction of the language grasp of this issue in Europe in a selected German mass media discourse. We compare the phenomenon of migration in 2015/2016, when record migration flows to the EU were recorded, and in 2019, when migration kept increasing. The analysis of language grasp of the issue of migration is a part of our scientific research within the project VEGA Xenisms in German and Slovak communications.

Key words: linguistic interpretation, web corpora, migration, German political discourse

ÚVOD

Parciálny výskum prezentovaný v predloženom príspevku je vedeckovýskumným ponorom do tematiky webových korpusov v digitálnej podobe ako efektívnych nástrojov na získavanie aktuálnych relevantných dát. Práca s webovými korpusmi sa postupne stáva neoddeliteľnou súčasťou metodologického aparátu diskurznej analýzy a poskytuje okrem kvalitatívnej analýzy aj možnosti rozsiahlej analýzy kvantitatívnej, čo podľa G. Mautnerovej (Mautner, 2015, s. 180 – 204) znamená prínos najmä v oblasti reprezentatívnosti dát.

Z hľadiska metodiky sa odvolávame na monografiu *Jazykový obraz migrácie v nemeckom masmediálnom diskurze* (Lauková – Molnárová, 2018), kde vychádzame zo základnej koncepcie komparatívnej kvantitatívno-kvalitatívnej lingvistickej analýzy. Kvalitatívna analýza nám umožnila odhaliť špecifické charakteristické znaky jazykového uchopovania migrácie, t. j. aké jazykové prostriedky používajú komunikanti nemeckého diskurzu, keď píšu o migrácii. Kvantitatívna analýza je potrebná pri zisťovaní početnosti, frekvencie výskytu zistených po-

[1] Príspevok je publikovaný v rámci projektu VEGA *Xenizmy v nemeckých a slovenských komunikátoch* (ITMS: 1/0472/20).

znatkov alebo ich intenzity a poslúžila nám na rozbor získaných a numericky vyjadrených údajov, na ktorých sme si overovali a aj overujeme naše teoretické hypotézy.

Fenomén migrácie bol v centre našej pozornosti v roku 2019, v ktorom migrácia opäť patrila medzi aktuálne témy. Skúmame ju v komparačnom kontexte s rokmi 2015 a 2016, kedy boli zaznamenané rekordné migračné toky do Európskej únie. Spracovanú tému považujeme za aktuálnu a v spoločnosti veľmi diskutovanú, reflektuje výskumný problém konštrukcie jazykového uchopovania problematiky migrácie v Európe vo vybranom nemeckom masmediálnom diskurze.

1. XENOLINGVISTICKÝ EXKURZ DO PROBLEMATIKY

Jednou z motivácií rozpracovania výskumu tohto druhu, t. j. s fokusom na atribút cudzosti (inakosti) v rámci xenolingvistiky v súvislosti s témou migrácie je aj skutočnosť, že fenomén migrácie je veľmi citlivou témou, bezprostredne sa týka samotnej Európskej únie a v ostatných rokoch aj Slovenska (Dobrík, 2018), aj keď aktuálne v roku 2020 je nepochybne v úzadí z dôvodu prepuknutia pandémie COVID-19.

Migráciu v tomto kontexte vnímame ako pohyb osôb alebo skupín osôb v geografickom a sociálnom priestore spojený s prechodnou alebo trvalou zmenou miesta pobytu.[2] Riešenie migrácie je pre EÚ jednou z jej dlhodobých priorít. Európa vo svojej histórii zaznamenala niekoľko migračných vĺn. V novodobej histórii sa s migračnými vlnami vyrovnávala najmä v rokoch 2015 a 2016, kedy pohraničná a pobrežná stráž zaznamenala viac ako 2,3 milióna nezákonných prekračovaní vonkajších hraníc EÚ a do Európy sa dostalo viac ako milión osôb. V tom čase však EÚ ešte nemala spoločnú politiku týkajúcu sa riadenia migrácie a bezpečnosti hraníc.[3] V roku 2019 klesol počet nelegálnych prekročení hraníc na svoju najnižšiu úroveň od roku 2013 – 141 846. Uvedené fakty uvádzame zámerne, pretože z týchto údajov vyplýva, že v roku 2019 zaznamenala Európa jednoznačný pokles prílevu migrantov na svoje územie a tieto skutočnosti podľa nášho názoru nepochybne ovplyvnili aj samotný diskurz.

Výraznou kognitívnou oporou pri našich teoretických úvahách je xenolingvistická koncepcia, poznatková báza týkajúca sa jazykových xenizmov obsiahnutá v publikáciách Juraja Dolníka (2010, 2012, 2015). Práve Dolníkov rozsiahly aplikačný potenciál súčasného stavu poznania v slovenskej xenolingvistike, konkrétne tzv. teórie cudzosti a inakosti, využívame pri našom výskume za účelom analýzy vybraných komunikátov, ktoré sa konštruujú v nemeckom sociokultúrnom priestore v období roku 2019 v porovnaní s rokmi 2015 a 2016.

Pojem xenizmus je produkt xenologického nazerania na jazyk, ktoré je špecifi-

[2] Porov. https://www.iom.sk/sk/pre-media/zakladne-pojmy-o-migracii.html#migracia

[3] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM: 2019:0481:FIN:SK:PDF

kované tým, že pozornosť sa sústreďuje na interpretovanú aktualizovanú cudzosť. Extenzia tohto termínu teda koreluje s rozsahom pojmu cudzosť. Cudzosť sa spravidla spája s interkultúrnym vzťahom (cudzia kultúra, cudzí jazyk, cudzie slová a pod.). Cudzosť navodzuje stav neistoty, ale je aj príťažlivou silou. Bežne možno konštatovať, že cudzie niekoho priťahuje, fascinuje, ale spravidla sa pritom výraz „cudzie" chápe ako niečo neznáme a nepoznané. Toto neznáme a nepoznané je tzv. potenciálnou cudzosťou. Reálnou cudzosťou sa stáva, keď je objektom interpretácie a kladie odpor proti asimilácii, proti tomu, ako mu interpretujúci rozumel (porov. Dolník, 2010, s. 23). V určení cudzosti je aj potenciálne hodnotenie, ktoré sa často aktualizuje. Prisudzovaný príznak cudzosť funguje aj ako hodnotiaci parameter. Pragmatizácia cudzosti sa ohraničuje zámernou interpretáciou inakosti ako cudzosti v mene istého partikulárneho záujmu. Kategória cudzosť sa tu zreteľne ukazuje vo svetle kategórie moc.

Podľa Dolníkovej koncepcie (2010) môžeme rozlišovať intrakultúrne a interkultúrne xenizmy, t. j. jazykové znaky v rámci jednej kultúry alebo v rámci iných kultúr. Intrakultúrne xenizmy existujú pravdepodobne vo všetkých krajinách, pretože každá predstavuje špecifický sociokultúrny priestor (Dobrík, 2018, s. 20).

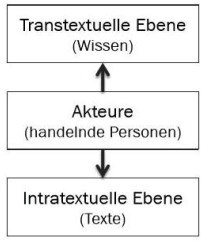
V súvislosti s príznakom cudzosti analyzujeme v rámci nášho mikrovýskumu, v akých súvislostiach a kontextoch sa o migrácii píše, aká je typická tematická štruktúra takýchto správ a aké výrazové jazykové prostriedky aktéri komunikácie (autori článkov) využívajú. Je migrácia v politickom diskurze vnímaná len ako „cudzí element", alebo skôr aj v istých tzv. stereotypných a stereotypizujúcich kontextoch? Vyskytuje sa v nemecky hovoriacom diskurze v rokoch 2019 a 2015/2016 skôr pozitívne konotovaný obraz migranta ako cudzinca, ktorý má záujem sa integrovať do spoločnosti alebo skôr negatívne konotovaný obraz cudzinca, ktorý sa snaží „parazitovať" na sociálnom systéme?

2. JAZYKOVÉ UCHOPENIE MIGRÁCIE VO VYBRANOM NEMECKOM WEBOVOM KORPUSE

Diskurz vnímame v kontexte nášho výskumu ako zložitý kognitívno-komunikačný fenomén a jeho súčasťou nie je len samotný text, ale aj rôzne extralingvistické faktory (poznanie sveta, názory, hodnotové orientácie), ktoré sú dôležité na pochopenie a percepciu informácií. V rámci nášho výskumu nazeráme na diskurz ako na komunikačný a sociálny rámec, ktorý je vymedzený jednou (makro)témou – (súčasná) migrácia. Pri analýze takéhoto diskurzu pracujeme s korpusom, ktorého texty obsahujú kľúčové slovo *migrácia* (nem. *Migration*) a/alebo kľúčové slovo *utečenec* (nem. *Flüchtling*).

Pri našich analýzach sme ako metodologický model, ktorý nám umožní orientáciu v diskurze, použili model diskurznej viacúrovňovej analýzy DIMEAN (Spitzmüller – Warnke, 2011, s. 136). Táto metodológia spája rovnako analýzu vzťahujúcu sa

na jazyk, ako aj analýzu vzťahujúcu sa na kogníciu a zároveň vymedzuje aktérov ako hlavnú dimenziu diskurzu (Lauková – Molnárová, 2018, s. 36 – 40). Model rozlišuje transtextovú rovinu (poznanie, kogníciu), intratextovú rovinu (texty) a aktérov (konajúce, činné osoby).



Obr. 1. Tri roviny diskurznej analýzy (Spitzmüller – Warnke, 2011, s. 136)

Prístup k dátam sme získali cez *Nemecký referenčný korpus DeReKo*, ktorý zostavil v prevažnej miere Inštitút nemeckého jazyka Leibniz (nem. skratka IDS). Využili sme prístup cez Windows-verziu Cosmas II. Naše korpusy spĺňajú kritérium podľa P. Bakera (2008, s. 277), ktorý sa pri výstavbe korpusov opiera o kritérium mediálneho typu (druh novín = bulvárne noviny, noviny veľkého formátu a pod.).

Všetky analyzované texty sú súčasťou mediálneho (makro)žánru. Vytvorili sme ich vo forme virtuálnych korpusov z nasledovných troch periodík:

a) týždenníka Focus (nemecký seriózny mienkotvorný spravodajský týždenník so sídlom v Berlíne, ktorý vydáva Hubert Burda Media, názorovo je občianskoliberálny);

b) denníka Süddeutsche Zeitung (skratka SZ, tzv. mienkotvorný nemecký denník, ktorý vydáva holdingová spoločnosť Süddeutscher Verlag so sídlom v Mníchove, názorovo sa hlási k liberalizmu a ľavému stredu);

c) denníka Dresdner Neueste Nachrichten (skratka DNN, regionálny denník vydávaný vo východnej časti Nemecka, v meste Drážďany, bývalej NDR, vydavateľstvo Verlag Dresdner Nachrichten GmbH & Co. KG).

Korpusy sú tvorené periodikami s rôznou územnou pôsobnosťou, rôznou frekvenciou vydávania, všetky však patria k tzv. serióznej tlači. Ich súčasťou sú texty rovnako z printovej ako aj online verzie novín rôznych publicistických žánrov (spravodajská, analytická, beletristická publicistika, napr. komentár, glosa, reportáž, úvodník ale aj listy čitateľov a pod.). Predpokladali sme určitú rozdielnosť v záujme o utečeneckú problematiku vo vybraných periodikách, ktorá by mohla súvisieť napr. s lokalizáciou sídiel spoločností, ktoré ich vydávajú. Süddeutsche Zeitung majú síce celoštátnu platnosť, ale ich sídlom je Mníchov, hlavné mesto spolkovej krajiny Bavorsko, ktorej sa utečenecká problematika výraznejšie dotýka aj v dôsledku jej „záchytnej, resp. tranzitnej“ polohy na území Nemecka. Dresdner Neueste Nachrichten majú sídlo v meste Drážďany, ktoré sú spájané so vznikom protiislamistického hnu-

tia Pegida a zároveň v minulosti patrili k územiu bývalej NDR.

2.1. Korpusová analýza

Prvý krok našej analýzy predstavoval sémantické vymedzenie korpusov, pretože sú pre nás relevantné iba texty s tematikou migrácie. Preto sme v korpusoch Süddeutsche Zeitung 2019, 2015 a 2016 zadali do vyhľadávania lemy *migrácia* (nem. *Migration*) a *utečenec* (nem. *Flüchtling*). Získali sme nasledovné kvantitatívne údaje, pričom skúmaný materiál predstavoval 23 968 661 slov (2019)**,** 23 879 964 slov (2015) a 24 423 013 slov (2016)**.**

Süddeutsche Zeitung	**2019** (počet výskytov)	**2015** (počet výskytov)	**2016** (počet výskytov)
Migration	1574 (65, 67/ipm)	1112 (46,57/ipm)	1469 (60,15/ipm)
Flüchtling	3730 (155, 6/ipm)	16 750 (701/ipm)	15 984 (654,5/ipm)

Tab. 1. Lemy *Migration* a *Flüchtling* v denníku Süddeutsche Zeitung

Prvý číselný údaj v tabuľke predstavuje počet výskytov lemy *migrácia* (nem. *Migration*) a *utečenec* (nem. *Flüchtling*) v skúmaných korpusoch. V komparácii s rokmi 2015/2016 je počet výskytov lemy *Migration* v roku 2019 mierne vyšší, čím môžeme preukázať aktuálnosť migračnej problematiky aj v tomto roku. Jednoznačný pokles výskytov lemy *Flüchtling* poukazuje na zmenu jazykových prostriedkov – nosnej lemy –, ktorými je modelovaný migračný diskurz v korpuse Süddeutsche Zeitung 2019.

Rovnako sme postupovali v prípade korpusov Focus a Dresdner Neueste Nachrichten 2019, 2015 a 2016. Do vyhľadávania sme zadali lemu *migrácia* (nem. *Migration*) a *utečenec* (nem. *Flüchtling*) a v prípade korpusu Focus sme získali nasledovné kvantitatívne údaje, pričom skúmaný materiál predstavoval 1 915 527 slov (2019), 2 122 769 (2015) a 1 961 118 (2016).

Focus	**2019** (počet výskytov)	**2015** (počet výskytov)	**2016** (počet výskytov)
Migration	137 (71,52/ ipm)	140 (65,95/ ipm)	86 (43,85/ ipm)
Flüchtling	272 (142,5/ ipm)	1740 (819/ ipm)	1323 (674,6/ ipm)

Tab. 2. Lemy *Migration* a *Flüchtling* v týždenníku Focus

V prípade korpusu Dresdner Neueste Nachrichten predstavoval skúmaný materiál 10 188 104 slov (2019), 7 817 945 (2015) a 12 584 591 (2016).

Dresdner Neueste Nachrichten	**2019** (počet výskytov)	**2015** (počet výskytov)	**2016** (počet výskytov)
Migration	584 (57,32/ ipm)	309 (39,52/ ipm)	596 (47,36/ ipm)
Flüchtling	1556 (152,7/ ipm)	5255 (672,2/ ipm)	7884 (626,5/ ipm)

Tab. 3. Lemy *Migration* a *Flüchtling* v denníku Dresdner Neueste Nachrichten

Ak počet výskytov lemy *Flüchtling* v roku 2019 porovnáme s výskytmi v roku 2015 a 2016 vidíme jednoznačný nižší výskyt tejto lemy vo všetkých korpusoch. Naopak, v prípade výskytu lemy *Migration* zaznamenávame vo všetkých korpusoch v roku 2019 mierny nárast v porovnaní s rokmi 2016 a 2015.

V prípade slovných tvarov lemy *Flüchtling*, ktorou sme obsahovo vymedzili skúmané korpusy, sa najčastejšie vyskytuje tvar množného čísla, čiže tvar *Flüchtlinge* (slov. *utečenci*), čo potvrdzuje aj našu analýzu z hľadiska aktérstva (porov. model DIMEAN), prípadne vytvárania obrazu utečencov, v ktorej sme dospeli k záveru, že utečenci sú v textoch prezentovaní väčšinou ako skupina. Ďalšími tvarmi lemy *Flüchtling,* ktoré prevládajú v korpuse, sú jednoznačne kompozitá.

Kompozitá, ako je všeobecne známe, sa v nemeckom jazyku používajú veľmi často. Sú vytvorené buď kolokáciami (napr. *Sonntag – sonniger Tag*, *Schulbesuch – Schule besuchen*) alebo majú vlastnosti charakteristické pre kolokácie, napr. asociatívnosť (*Flüssiggas*) alebo čiastočná idiomaticita (*hundemüde*). Hoci ide o kombináciu dvoch významov, od kolokácií sa líšia tým, že nejde o syntaktické spojenia slov, ale o morfematické spojenia (Vajíčková, 2019, s. 32).

S najvyšším počtom výskytov vo všetkých korpusoch z rokov 2015 a 2016 dominuje kompozitum *Flüchtlingskrise* (slov. *utečenecká kríza*), kým v roku 2019 je to *Flüchtlingsfrage* (slov. *utečenecká otázka*). Samotné základné slovo kompozita *Flüchtlingskrise* je negatívne konotované, v kompozite *Flüchtlingsfrage* táto negatívna konotácia nie je prítomná. V textoch z roku 2019 (korpusy Focus 2019 a Süddeutsche Zeitung 2019) sa kompozitum *Flüchtlingskrise* vyskytuje výhradne na označenie obdobia v rokoch 2015 a 2016. K ďalším kompozitám s určujúcim slovom *Flüchtling* v textoch z roku 2019 patria *Flüchtlingsdebatte* (slov. *diskusia o utečencoch*), *Flüchtlingsheim* (slov. *domov*, resp. *ubytovňa pre utečencov*) a *Flüchtlingscamp* (slov. *utečenecký tábor*). Kompozitum *Flüchtlingsdrama* (slov. *utečenecká dráma*) sa vyskytuje v textoch tematizujúcich situáciu na území Turecka a na gréckych ostrovoch, čiže mimo územia Nemecka.

Výskum kolokácií v oblasti korpusovej lingvistiky je koncentrovaný na problém automatickej detekcie relevantných slovných spojení v textových korpusoch (porov. Jarošová, 2007, s. 81).

Východiskom pri našom výskume kolokácií je skutočnosť, že pri komunikácii sa len ojedinele používajú samostatné slová (porov. Vajičková – Luta, 2019, s. 17).

Slová sa vyskytujú v spoločnosti iných slov, ktoré sú do istej miery vopred určené. Spájateľnosť slov je teda daná nielen ich gramatickými vlastnosťami (morfologickými a syntaktickými kategóriami), ale predovšetkým ich sémantikou, ktorá limituje možnosť spájateľnosti slov.

Podľa M. Ivanovej (2018, s. 137) „Štúdium kolokácií predstavuje iný prístup ku skúmaniu viacslovných pomenovaní. Vychádza sa pri ňom zo súvyskytu slov – teda z faktu, že isté slová sa navzájom predpokladajú, to znamená, že isté slovo sa objavuje v kolokácii s iným slovom častejšie, ako to je štatisticky pravdepodobné, napr. konečné rozhodnutie, výsledný efekt a pod.“ P. Ďurčo a M. Vajičková (2017, s. 26) definujú kolokácie ako ustálené spojenia tvoriace lexikálno-syntagmatické kategórie, ktoré sa v lexikálnom systéme jazyka nachádzajú medzi voľnými slovnými spojeniami a idiómami.

Pri vyhľadávaní kolokácií v našom korpuse sme sa opierali o kľúčové slovo a v korpuse sme zisťovali jeho najfrekventovanejšie pravostranné a ľavostranné kolokáty, susediace slová, a to na základe stanoveného rozsahu (-5 vľavo od KWIC a + 5 vpravo od KWIC) s celkovou minimálnou frekvenciou v korpuse 5 výskytov. Zo štatistických (asociačných) mier sme využili LLR-Wert (porov. Tomášková, 2019, s. 183), v ktorej hodnote je prepočítaný pomer frekvencie spoluvýskytu slov v nastavenom kontexte s frekvenciami obidvoch slov. Za kolokáciu teda považujeme štatisticky preukázateľný spoluvýskyt dvoch slov v nami určenom rozpätí, pričom kolokácia nemusí byť kontaktná, čiže dané slová nemusia stáť v bezprostrednej blízkosti v rámci konkordancie.

Ako sme už v našom príspevku uviedli, v migračnom diskurze v roku 2019 sme v porovnaní s rokmi 2015 a 2016 zaznamenali rozdiely v používaných lexikálnych prostriedkoch. Prejavilo sa to teda aj vo výskyte kolokácií. Za ustálené kolokácie vo všetkých skúmaných rokoch môžeme považovať kolokácie *Migranten und Flüchtlinge* (slov. *migranti a utečenci*). Ku kolokáciám s pomerne vysokým počtom výskytov patrí aj *Merkels Flüchtlinge* (slov. *Merkelovej utečenci*), ktorá sa vyskytuje najmä v textoch posudzujúcich politické rozhodnutia a vystúpenia nemeckej kancelárky. Ak by sme pre sprehľadnenie a zostručnenie analýzy v príspevku mali kolokátory zoskupiť do tematických skupín, potom najväčší rozdiel sme zaznamenali vo verbalizácii spôsobu príchodu utečencov do Nemecka a do Európy. V roku 2019 sú už celkovo v korpusoch v diskurze menej verbalizované spôsoby a miesta, odkiaľ a ako utečenci do Európy prichádzajú. V korpuse je vyšší výskyt iba kolokácie *syrische Flüchtlinge* (slov. *sýrski utečenci*). Z hľadiska slovných druhov použitých lexikálnych jednotiek je v korpusoch Focus, Süddeutsche Zeitung a Dresdner Neueste Nachrichten 2015 a 2016 príchod utečencov vyjadrený väčšinou podstatným menom napr. *Flut* (slov. *príliv*), *Zuzug* (slov. *prílev*), *Einreise* (slov. *vstup*), *Andrang* (slov. *nával*), *Zustrom* (slov. *príval*, *prítok*), *Schlepper* (slov. *remorkér*), v korpuse Dresdner Neueste Nachrichten dominujú propriá *Türkei* (slov. *Turecko*) a *Griechenland* (slov. *Grécko*). Len v niektorých

prípadoch je príchod utečencov vyjadrený slovesným tvarom, napr. *kommen* (slov. *prísť*), *strömen* (slov. *prúdiť*), *drängen* (slov. *tlačiť sa*). V korpuse Focus 2019 a Süddeutsche Zeitung 2019 naopak prevládajú štylisticky bezprízнakové slovesné tvary ako *kommen* (slov. *prísť*), *sich treffen* (slov. *stretnúť sa*), *suchen* (slov. *hľadať*), v korpuse Dresdner Neueste Nachrichten prevláda slovesný tvar *aufnehmen* (slov. *prijať*). Kolokátor *Aufnahme* (slov. *prijatie*) sa aj naďalej vyskytuje v korpuse Dresdner Neueste Nachrichten vo vyššej miere ako kolokátor *Integration (slov. integrácia)*, ktorý v korpusoch Focus 2019 a Süddeutsche Zeitung 2019 nahradil kolokátor *Aufnahme* (slov. *prijatie*), čo signalizuje posun a zmenu prioritných tém v riešení utečeneckej problematiky.

ZÁVER

Cieľom nášho príspevku bolo stručne načrtnúť východiská, metodológiu a výsledky čiastkového výskumu problematiky jazykového uchopovania témy migrácie v nemeckom masmediálnom diskurze v komparácii rokov 2019 a 2015/2016. Ako už bolo uvedené, jednou z motivácií rozpracovania výskumu s fokusom na atribút cudzosti (inakosti) v rámci xenolingvistiky v súvislosti s témou migrácie je aj fakt, že fenomén migrácie bol predovšetkým v období rokov 2015 a 2016 pertraktovanou témou. Každá inakosť nesie v sebe symbolický potenciál konštruovať cudzosť, a tak aj potenciálne nepriateľstvo a konflikty.

V skúmaných webových korpusoch sme sa zamerali na výskyt lemy *migrácia* (nem. *Migration*) a *utečenec* (nem. *Flüchtling*). Získané kvantitatívne údaje potvrdili, že diskurz o migrácii bol považovaný za jeden z najzávažnejších problémov predovšetkým v roku 2016. Zhrňujúco možno konštatovať, že vo všetkých analyzovaných webových korpusoch v komparácii rokov 2015/2016 a 2019 pozorujeme podobné diskurzné stratégie podporované rôznymi špecifickými jazykovými prostriedkami. V analyzovaných korpusoch dochádza k syntéze subjektívnych a objektívnych parciálnych obrazov o migrácii, na základe ktorej vznikajú určité formy percepcie cudzosti a tieto sa ďalej v textoch realizujú napr. aj vo forme zovšeobecnení, stereotypizácie a pod., pričom je celkovo prezentovaných niekoľko zovšeobecňujúcich obrazov utečencov. Nesporným faktom zostáva, že jazykový obraz utečencov je obrazom skupinovým. V prvom rade je to obraz homogénnej skupiny, ktorá je problémová – nevie alebo nechce sa adaptovať na nové prostredie a prispôsobiť sa pravidlám majoritnej skupiny.

Bibliografia

Azyl a migrácia v EÚ – fakty a čísla. Európsky parlament. Spravodajstvo. Dostupné na: https://www.europarl.europa.eu/ news/sk/headlines/society/20170629STO78630/azyl-a-migracia--v-eu-fakty-a-cisla [cit. 02.11.2020].

BAKER, Paul et al.: A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. In: Discourse & Society, 2008, roč. 19, č. 3, s. 273 – 306.

DOBRÍK, Zdenko: Cudzosť a inakosť v jazykovej komunikácii. Banská Bystrica: Belianum: vydavateľstvo UMB 2018. 126 s.

DOLNÍK, Juraj: Jazyk – človek – kultúra. Bratislava: Kalligram 2010. 224 s.

DOLNÍK, Juraj: Sila jazyka. Bratislava: Kalligram 2012. 368 s.

DOLNÍK, Juraj a kol.: Cudzosť – jazyk – spoločnosť. Bratislava: IRIS 2015. 316 s.

ĎURČO, Peter – VAJIČKOVÁ, Mária et al.: Kollokationen im Unterricht. Ein Lehr- und Übungsbuch. 2. Aufl. Nümbrecht: Kirsch-Verlag 2017. 274 s.

Európska únia. EURACTIV Slovensko. Dostupné na: https://www.europskaunia.sk/migracia [cit. 02. 11. 2020].

IVANOVÁ, Martina: Kolokácie v korpuse, viacslovné pomenovania v slovníku (Úvodné poznámky k príprave slovníka viacslovných pomenovaní). In: Jazyk je zázračný organizmus... Metamorfózy jazyka a jazykovedy. Eds. M. Imrichová – J. Kesselová. Prešov: Filozofická fakulta PU 2013, s. 132 – 147. Dostupné na: https://www.pulib.sk/web/kniznica/elpub/dokument/Imrichova1/subor/Ivanova.pdf [cit. 02. 11. 2020].

JAROŠOVÁ, Alexandra: Problém vymedzenia kolokácií. In: Jazykovedný časopis, 2007, roč. 58, č. 2, s. 81 – 102.

LAUKOVÁ, Jana – MOLNÁROVÁ, Eva: Jazykový obraz migrácie v nemeckom masmediálnom diskurze. Banská Bystrica: UMB Belianum 2018. 162 s.

MAUTNER, Gerlinde: Checks and balances: how corpus linguistics can contribute to CDA. In: Methods of critical discourse studies. Los Angeles – London: Sage Publications 2015, s. 154 – 179.

Medzinárodná organizácia pre migráciu. Dostupné na: https://www.iom.sk/sk/pre-media/zakladne-pojmy-o-migracii.html#migracia [cit. 02. 11. 2020].

Nemecký referenčný korpus DeReKo. COSMAS II. 2020. Version 2.3.5 Mannheim: Institut für Deutsche Sprache (IDS). Dostupné na: https://cosmas2.ids-mannheim.de/cosmas2-web/ [cit. 02. 11. 2020].

Oznámenie Komisie Európskemu Parlamentu, Európskej rade a Rade. Správa o pokroku pri vykonávaní európskej migračnej agendy. Brusel: Európska komisia 2019. Dostupné na: https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2019:0481:FIN:SK:PDF [cit. 02.11.2020].

SPITZMÜLLER, Jürgen – WARNKE, Ingo H.: Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin: Walter de Gruyter 2011. 230 s.

TOMÁŠKOVÁ, Simona: Pragmatische Aspekte der Kollokationen in mündlicher Kommunikation. In: Kollokationen im Sprachsystem und Sprachgebrauch. Eds. P. Ďurčo – M. Vajičková – S. Tomášková. Nümbrecht: Kirsch-Verlag 2019, s. 163 – 190.

VAJIČKOVÁ, Mária: Theoretische Aspekte der Kollokationen. In: Kollokationen im Sprachsystem und Sprachgebrauch. Eds. P. Ďurčo – M. Vajičková – S. Tomášková. Nümbrecht: Kirsch-Verlag 2019, s. 11 – 51.

VAJIČKOVÁ, Mária – LUȚA, Marilena Felicia: Ako ovplyvňuje typologická príslušnosť tvorbu kolokácií v cudzom jazyku. Na pozadí slovenčiny, nemčiny a rumunčiny. In: PHILOLOGIA XXIX, 2019, č. 1 – 2, s. 17 – 34.

Turecko odstupuje od migračnej dohody s Úniou. EURACTIV. Dostupné na: https://euractiv.sk/section/vonkajsie-vztahy/news/turecko-odstupuje-od-migracnej-dohody-s-uniou/ [cit. 02. 11. 2020].

sciendo
DOI 10.2478/jazcas-2022-0013

TEMATICKÉ SLOVÁ V PREDVOLEBNEJ KAMPANI NA FACEBOOKU[1]

NATÁLIA KOLENČÍKOVÁ
Jazykovedný ústav Ľudovíta Štúra SAV, Bratislava, Slovensko

KOLENČÍKOVÁ, Natália: Thematic words in the Slovak pre-election campaign on Facebook. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 882 – 893.

Abstract: The aim of this paper is 1. to describe/specify and compare thematic orientations of 735 pre-election microblogs published on the virtual profiles of major six Slovak political parties and 2. based on this description and comparison to identify and sketch features of Slovak political discourse. The conceptual and methodological frame consists of thematic words, that is, autosemantics above the so-called h-point, and the qualitative analysis of these thematic words. The identified features of the general Slovak pre-election communication include: populist communication, ego presentation of the party, leader or the candidate, conflict between government parties and opposition parties, image of Slovakia as a country facing troubles but also hiding potential to solve them.

Key words: h-point, Facebook, microblog, political communication, political discourse, pre-election campaign, thematic words

Aktuálny príspevok zameraný na tematické slová v predvolebnej kampani na Facebooku vzniká ako súčasť širšie koncipovaného výskumu mapujúceho jazykové a komunikačné špecifiká predvolebnej kampane v prostredí sociálnych médií. Hoci výskum politicko-mediálneho diskurzu si nachádza priestor aj v slovenskej lingvistike (pozri napr. Odaloš, 2003; Patráš, 2003; Ogoňová – Dolník, 2010; Macho, 2012; Rašová, 2013; Štefančík – Dulebová, 2017; Molnárová, 2019), systematickejšie zameranie na politickú komunikáciu sprostredkúvanú sociálnymi médiami u nás doteraz nepozorujeme. Situácia je prekvapujúca, keďže sociálnomediálne komunikačné prostredie je dnes imanentnou súčasťou politickej komunikácie (Jaworowicz, 2016) a keďže vedecky ide o pomerne jednoducho dostupný zdroj výskumného materiálu, na ktorý možno nazerať z najrozmanitejších perspektív, a spájaním parciálnych výskumov tak o ňom postupne vytvárať komplexný obraz.

Na dôležitú úlohu sociálnych médií v predvolebnej kampani sa často upozorňuje v spojitosti s víťazstvami amerických prezidentov B. Obamu (YouTube) a D. Trumpa (Twitter). V slovenskom geopolitickom kontexte sa tento aspekt prvýkrát spomína pri prezidentskej predvolebnej kampani I. Radičovej v roku 2009 (Facebook) či nečakanom volebnom úspechu politickej strany Sloboda

[1] Príspevok vzniká v rámci projektu podporeného Fondom Štefana Schwarza a v rámci projektu VEGA č. 2/0016/21 *Slovník súčasného slovenského jazyka – 7. etapa (Koncipovanie a redigovanie slovníkových hesiel a s tým spojený lexikologicko-lexikografický výskum).*

a Solidarita v parlamentných voľbách v roku 2010 (Facebook). Dôvody prirodzeného včleňovania sociálnych médií do politickej komunikácie sú podmienené špecifikami, ktorými sa tento typ médií odlišuje od pôvodných elektronických médií (napr. rozhlas, televízia).[2] Vo vzťahu k téme príspevku ide predovšetkým o to, že autorom ľubovoľného, či už spoločensky relevantného alebo triviálneho a značne individualizovaného obsahu môže byť, pri dodržaní základných technických a momentálne stále relatívne uvoľnených legislatívnych pravidiel, ktokoľvek, politické strany nevynímajúc. Komunikačný model šírenia správy prostredníctvom siete participantov pritom umocňuje vyrovnávanie komunikačnej pozície autora a adresáta, vďaka čomu môže volič pri virtuálnom kontakte s politikom alebo politickou stranou nadobúdať dojem aktívnej participácie na politickom dianí. Témy, ktoré politické strany v predvolebnom období vo svojich mikroblogoch či statusoch spracúvajú, tak môžu byť podnetným predmetom skúmania, a to v nadväznosti na volebné úspechy jednotlivých politických strán aj na kľúčové témy predvolebného diskurzu komplexne. S oporou o tieto úvahy sa v príspevku usilujeme špecifikovať a porovnať tematickú orientáciu predvolebnej sociálnomediálnej komunikácie úspešných politických strán a na tomto základe identifikovať základné znaky slovenského predvolebného diskurzu.[3]

Pri dosahovaní stanovených zámerov nachádzame funkčnú metodologickú oporu v kvantitatívnej lingvistike, ktorá operuje s tzv. tematickými slovami. Tie sa nachádzajú v skupine najpočetnejšie sa objavujúcich lem, preto je najprv potrebné zostaviť rangovú frekvenčnú distribúciu textu, ktorú sme východiskovo vytvorili nástrojom *LancsBox* (Brezina – McEnery – Wattam, 2015; Brezina – Weill-Tessier – McEnery, 2020) a následne manuálne upravili zistené nepresnosti. Najvyššie priečky štandardne obsadzujú neplnovýznamové slová, ktoré zabezpečujú bezproblémové gramatické fungovanie textu. Plnovýznamové slová naopak, pribúdajú s klesajúcou frekvenciou. Hranicu medzi výskytom synsémantík a autosémantík predstavuje tzv. h-bod – miesto, v ktorom sa rang lemy rovná jej frekvencii (Popescu, 2007). Ak takýto bod nie je možné určiť zo samotného frekvenčného zoznamu, vypočítame ho pomocou vzorca

[2] V odbornej literatúre sa môžeme stretnúť s viacerými prístupmi k osobitostiam sociálnych médií. D. McQuail upozorňuje na vzájomné väzby participantov, dostupnosť pre individuálnych používateľov v pozícii wridrov, interaktívnosť, rozmanitosť spôsobov používania, otvorenosť, všadeprítomnosť, priestorovú neohraničenosť a delokalizáciu (McQuail, 2007). K podobným vlastnostiam dospieva aj A. Mayfield, keď hovorí, že sociálne médiá sú mediálne platformy, ktoré sa vyznačujú participáciou používateľov, otvorenosťou, obojsmernosťou komunikácie, formovaním komunít a prepojenosťou na iné stránky, zdroje a ľudí (Mayfield, 2008). L. Manovich sa zase zameriava na päť technologicky podmienených princípov – numerická reprezentácia, modulárnosť, automatizácia, variantnosť a transkódovanie (Manovich, 2001). Mediálne faktory (so zameraním na synchrónnosť, prenos informácie, trvácnosť komunikátu, rozsah komunikátu, kanály komunikácie, súkromné nastavenia, anonymitu a štruktúru komunikácie) a situačné faktory (so zameraním na štruktúru participácie, charakteristiky participácie, zámer, tému, tón, normy a kód) sociálnomediálnej komunikácie potom podrobne konkretizujú R. Page, D. Burton, J. W. Unger a M. Zappavigna (2014).

[3] Diskurz chápeme ako „integrovaný celok textu a kontextu, ako spojenie jazykovej realizácie interakcie a jej kontextového presahu“ (Hoffmannová, 1997, s. 8).

$$h = \frac{f(i) \times r_j - f(j) \times r_i}{r_i - r_j + f(i) - f(j)},$$

v ktorom je r_i najväčší rang, pre ktorý platí $r_i < f(i)$, a r_j je najmenší rang, pre ktorý platí $r_j > f(j)$. V našom prípade k takejto situácii dochádza iba pri politickej strane Sme rodina. 29. priečku vo frekvenčnom zozname zastáva lema „do“, ktorú sme vo zvolenom výskumnom materiáli identifikovali 30-krát. Po nej, teda na 30. mieste, sa nachádza lema „pre“, ktorej frekvencia je 29. Po vyriešení rovnice

$$h = \frac{30 \times 30 - 29 \times 29}{30 - 29 + 30 - 29}$$

získavame hodnotu h-bodu, ktorá je 29,5. Rangovú frekvenčnú distribúciu štyridsiatich najpočetnejších lem pre komunikáciu jednotlivých politických strán sprístupňujeme prostredníctvom elektronického repozitára GitHub na nasledujúcom linku: https://github.com/NataliaKolencikova/Tematicke-slova-v-predvolebnej-kampani-na-Facebooku/blob/main/Tematicke_slova_rang_h-bod.pdf.

Napriek uvedenému tvrdeniu o autosémanticko-synsémantickej rozdeľovacej funkcii h-bodu sa stáva, že nadeň preniknú aj plnovýznamové slová, pričom práve tie považujeme za tematické slová.[4] Vzhľadom na predpokladaný výskyt autosémantík a synsémantík v rangovej frekvenčnej distribúcii sú teda tematické slová istou anomáliou, no odrážajú osobitosti textu. V kvantitatívnej lingvistike slúžia ako podklad meraní tematickej koncentrácie textov, ktoré už boli realizované aj na politicky motivovanom materiáli (napr. Čech, 2014; Kubát – Čech, 2016; Dai – Liu, 2019).

Materiálovú základňu výskumu tvorí 735 mikroblogov[5] publikovaných na facebookových profiloch šiestich politických strán úspešných v parlamentných voľbách v roku 2020. Objem materiálu je ovplyvnený metodologickými parametrami, podľa ktorých je určovanie tematických slov perspektívne v textoch s rozsahom od 200 do 6 500 tokenov (Čech – Kubát, 2016); na základe toho sme pre každú politickú stranu zostavili textový korpus požadovanej veľkosti, pričom zber začínal „odzadu“ – mikroblogmi uverejnenými najbližšie k dňu parlamentných volieb (29. 2. 2020), a pokračoval smerom do minulosti, až kým sme nenarazili na uvedenú metodologickú hranicu.[6] Skúma sa tak relatívne obmedzený, avšak metodologicky prípustný materiál, ktorý by zrejme bolo možné obsiahnuť aj výlučne kvalitatívnym spôsobom, no

[4] Do skupiny tematických slov zvyčajne nebývajú zaradené ani zámená a príslovky, hoci, ako je známe z pragmaticky orientovaných výskumov, ide o diskurzne dôležité vyjadrovacie prostriedky. V aktuálnom príspevku medzi tematickými slovami zámená a príslovky chýbajú takisto, no skôr z dôvodu úsilia o obmedzenie jeho rozsahu.

[5] Mikroblogom alebo statusom rozumieme hybridný textový útvar z komunikačného prostredia sociálnych médií, ktorý sa konštituuje na priesečníku viacerých funkčných štýlov a žánrov (k jeho textovotypologickým charakteristikám bližšie pozri Kolenčíková, 2018).

[6] Korpusy použité v aktuálnom výskume boli väčšinou vytvorené z o niečo rozsiahlejších korpusov zahŕňajúcich mikroblogy publikované od 1. 1. 2020 do 29. 2. 2020. Rozsah korpusov niektorých politických strán sa preto pohybuje na spodnej hranici metodologicky odporúčaného rozsahu.

práve kvôli možnosti exaktného vyjadrenia napríklad v tlači intuitívne vyvodzovaných tematických rámcov jednotlivých politických strán si volíme kvantitatívne podmienený metodologický základ. Relevantné informácie už spolu so zoznamom zistených tematických slov pre konkrétne politické strany súhrnne uvádzame v tabuľke č. 1. Na jej základe sa až následne pokúšame o podrobnejšiu kvalitatívnu analýzu tematických slov jednotlivých politických strán. Všetky citované doklady uvádzame v autentickej podobe bez akýchkoľvek korektorských zásahov.

Tab. 1. Súhrnná tabuľka relevantných dát

politická strana	volebný výsledok	dátum uverejnenia prvého analyzovaného mikroblogu	počet mikroblogov	počet tokenov	h-bod	tematické slová
OĽaNO	25,02 %	30. 1. 2020	202	6 440	32	človek, oľano, igor, kandidát, bratislava, zmena, slovensko, kresťanský, voľba
SMER – SD	18,29 %	1. 1. 2020	136	4 315	24	slovensko, človek, strana, peter, pellegrini, sociálny, smer – sd
Sme rodina	8,24 %	31. 1. 2020	142	6 428	29,5	človek, rodina, boris, slovensko
ĽSNS	7,97 %	1. 1. 2020	10	209	4	kotleba
SaS	6,22 %	31. 1. 2020	153	6 412	30	sas, rok, slovensko, bratislava
Za ľudí	5,77 %	15. 2. 2020	92	6 463	27	človek, slovensko, smer, kiska, andrej, program, rok

a) OĽaNO

Tematické slová politickej strany OĽaNO sú značne ovplyvnené tým, že strana frekventovane publikuje tzv. platený obsah, o čom má legislatívnu povinnosť informovať. Robí tak prostredníctvom schematizovaného doplnku k vecne a formálne inak rôznorodým príspevkom, ktorý je možné vnímať ako prejav administratívneho štýlu. Vďaka jeho predpísanej podobe (*Objednávateľ: OBYČAJNÍ ĽUDIA a nezávislé osobnosti (OĽANO), NOVA, Kresťanská únia (KÚ), ZMENA ZDOLA, Zámocká 14, Bratislava IČO:42287511 Dodávateľ: Facebook Ireland Limited, 4 Grand Canal Square, Írsko, IČO: 462962*) potom v skupine tematických slov identifikujeme aj lemy, ktoré sa v komunikácii takmer nepoužívajú v iných kontextoch; v tomto prípade to je „bratislava“, „zmena“ a „kresťanský“. Hoci teda nemôžeme hovoriť o priamej tematizácii týchto slov, nepochybne tiež naznačujú špecifiká komunikácie tejto strany.

S administratívnymi charakteristikami mikroblogov súvisí aj to, že najfrekventovanejším tematickým slovom politickej strany OĽaNO je slovo „človek“, ktoré je v príslušnom gramatickom tvare konštitučnou zložkou plného názvu strany. Toto slovo sa však používa tiež v iných kontextoch a časté odvolávanie sa na „ľudí“ môže byť indikátorom populistických tendencií (Jagers – Walgrave, 2007). Dôležité je pritom určiť, na koho sa týmto slovom referuje. V komunikácii OĽaNO sa ľuďmi väčšinovo myslia voliči, občania, obyvatelia Slovenska, ktorí sú hybnou silou politických rozhodnutí (*V NOVEJ VLÁDE PRESADÍME TO, O ČOM ROZHODNÚ ĽUDIA!*) a motiváciou, dôvodom, pre ktorý strana vyvíja svoje aktivity (*OĽANO robí politiku pre ľudí.*). Použitie v slovenskom politickom diskurze bežnej kolokácie „náš človek“, nesúcej príznak kritiky korupčného správania predvolebnej vlády, zároveň umožňuje pozorovať pre populistickú komunikáciu príznačné budovanie konfliktu medzi ľudom a elitou. Lema „človek“ je ale vo vzťahu ku kontextu často uplatňovaná aj bezpríznakovo – jednoducho označuje mysliacu živú bytosť (*Vždy som vyzývala ľudí, aby bojovali za spravodlivosť.*) alebo formálne slúži na konkretizáciu istej skupiny ľudí (*Útoky a ponižovania nás za tie roky zocelili, už sme imúnni. A rovnako ľudia, ktorí nám fandia.*).

Signálom egoprezentačného charakteru komunikácie je prítomnosť skrátenej podoby názvu politickej strany medzi jej tematickými slovami. Strana OĽaNO samú seba tematizuje dvoma spôsobmi. V prvom, a podporené je to aj formálne subjektovou alebo objektovou pozíciou vo výpovedi, sa prezentuje ako kolektív, celok, tím, ktorý reflektuje požiadavky svojich voličov (*ĽUDIA ROZHODNÚ: OĽANO spúšťa hlasovanie o programových prioritách*) a ktorý pre seba predpokladá priaznivý povolebný stav (*Majú oprávnenú hrôzu z nástupu OĽANO, lebo u nás sa na stranícky kabát pozerať nebude (...)*). V druhom spôsobe politická strana čerpá z prezentácie svojho kandidáta, a to buď explicitným vyjadrením jeho straníckej príslušnosti (*V roku 2016 bol Marek Krajčí vďaka vašim hlasom na kandidátke hnutia OĽANO zvolený za poslanca Národnej rady SR*.) alebo využitím technologických možností sociálnomediálneho prostredia, teda tzv. otagovaním virtuálneho profilu kandidáta, ktorý sa v jeho oficiálnom názve hlási k svojej politickej strane (*@Jaro Naď – OĽANO*). S tým sa spája tiež to, že „kandidát“ sa takisto nachádza v skupine tematických slov, vďaka čomu pozorujeme aj repetitívnosť predvolebnej sociálnomediálnej komunikácie (*DISKUSIA S KANDIDÁTOM OĽANO* ako titulok viacerých obsahovo príbuzných, no s odstupom času zverejňovaných mikroblogov).

S egoprezentáciou možno spájať aj antroponymum „igor“, ktorého najčastejším kolokátom je síce „matovič“, no táto lema sa v zozname tematických slov nenachádza, čo môže naznačovať smerovanie ku kolokvializácii politickej komunikácie. Líder politickej strany je takmer výlučne tematizovaný ako človek vstupujúci do verbálnych súbojov a zreteľne vyjadrujúci svoje postoje a názory – „igor“ debatuje, diskutuje, komentuje, vysvetľuje, opisuje či „nakladá“ svojim politickým oponentom (*Pozrite si, ako Igor včera naložil Pellegrinimu za obhajovanie zlodejín Smeru*).

Lema „slovensko“ sa s výnimkou politickej strany ĽSNS nachádza v zoznamoch tematických slov všetkých politických strán. V komunikácii strany OĽaNO sa ňou síce jednoducho referuje na geografický priestor s istou kultúrou a legislatívou, no oveľa častejšie sa tematizujú jeho negatívne stránky (*korupcia je alfou a omegou všetkých problémov, ktoré na Slovensku máme*). Tie však slúžia ako podklad vykreslenia pozitívnych možností, ktoré „slovensko“ má (*keď korupciu vykynožíme, Slovensko má šancu rásť*), akurát je potrebná „zmena“, „nádej“ či „statočnosť“. „Slovensko“ má navyše aj hodnotový rozmer – je to motivácia, pre ktorú sa oplatí byť aktívny (*Príďte nám 29. februára pomôcť vyhrať kľúčový zápas pre Slovensko.*). Vzhľadom na zvýšenú mieru obraznosti, ktorú pri analýze tejto lemy pozorujeme, by pri výskume predvolebnej komunikácie ďalej mohlo byť prínosné zameranie na metafory.

Prítomnosť lemy „voľba“, ktorá sa používa predovšetkým v podobe plurálového tvaru „voľby“ referujúceho na inštitucionalizovaný spôsob obsadzovania verejných funkcií v demokratických krajinách, je v skupine tematických slov predvolebnej komunikácie pochopiteľná, hoci so začleňovaním sociálnych médií do predvolebnej kampane sa hovorí o posilňovaní tzv. permanentnej kampane (Jaworowicz, 2016). Voľby sú prezentované ako súťaž, boj, zápas (*Predstavenie plánu, ako demokratická opozícia môže vyhrať voľby*), no zároveň ako časový bod, ktorý rozdeľuje politické dianie na predvolebné, plné negatívne hodnoteného konania (*Ďalší biznis za 60 miliónov EUR 2 týždne pred voľbami.*), a povolebné, ktoré má byť priestorom pre všetko pozitívne (*Sociálne výhody po voľbách zachováme!*).

b) SMER – SD

V tematizácii Slovenska, ktoré je najfrekventovanejším tematickým slovom v komunikácii politickej strany SMER – SD, nachádzame v porovnaní s predchádzajúcou stranou podobné aj odlišné body. Podobnou je referencia na konkrétny geografický priestor nachádzajúci sa v Európe alebo Európskej únii a vnímanie Slovenska ako niečoho zraniteľného, ako hodnoty, o ktorú je potrebné sa starať a venovať jej pozornosť (*„Vždy budeme chrániť Slovensko a našich občanov.“ – Peter Pellegrini*). Špecifické predvolebné postavenie tejto politickej strany a vtedajšie volebné preferencie však vplývajú na reflektovanie Slovenska ako krajiny, ktorá stojí na rázcestí a jej budúcnosť je nejasná (*Či Slovensko bude napredovať ako prosperujúca krajina, ktorá prinesie svojim občanom dlhodobú perspektívu a stabilitu alebo sa posunie do neistoty, hádok a sebadeštrukčnej rivality.*).

Vzhľadom na oficiálne deklarovanú sociálnodemokratickú profiláciu politickej strany, ktorá predpokladá prvky širšie chápaného populizmu, teda politickej koncepcie zameranej na bežného človeka, možno prítomnosť lemy „človek“ v skupine tematických slov považovať za prirodzené. Ľuďmi sa takisto rozumejú najmä voliči, občania, obyvatelia Slovenska, ktorí sú pôvodcami politického diania (*Rozhodnú o tom ľudia, či to bude zodpovedná zmena so skúsenosťami, schopnosťami a víziou, ako ďalej posunúť Slovensko dopredu, alebo títo páni, ktorí sa nevedia dohodnúť ani na dátume*

spoločného obedu.), ale aj motiváciou, pre ktorú strana realizuje svoje aktivity (*My bojujeme v prospech ľudí!*). Z uvedených príkladov takisto možno trochu prekvapivo vyčítať antielitárske smerovania, no pri hlbšej analýze je zrejmé, že tie sú vedené voči opozičným politikom. V prípade komunikácie tejto politickej strany však použitie lemy „človek" môže byť tiež kontextovo neutrálne. S politickou profiláciou strany sa viaže aj adjektívum „sociálny", ktoré v príslušných kolokáciách slúži ako podklad prezentácie dosiahnutých výsledkov – „sociálny balíček" a „sociálne opatrenia".

Sebaprezentácia politickej strany je postavená na rovnakých princípoch ako v prípade OĽaNO. Politická strana ako celok, tím a kolektív sa tematizuje najmä v kontexte svojich mediálnych aktivít (*Tlačová beseda strany SMER – SD k aktuálnej situácii v NR SR*), v kontexte skutočnosti, že predvolebne zastáva vládnuce postavenie (*Kým som premiérom Slovenska a vládne SMER – SD, Istanbulský dohovor nebude ratifikovaný proti vôli občanov SR.*) a v kontexte výzvy voličom, aby práve jej odovzdali svoj volebný hlas (*Dnes načúvajte znameniam a voľte č. 19 SMER – SD!*). Tematizácia strany s využitím tematizácie svojho člena, sa značne opiera o prezentáciu úspešných a už známych politických osobností (*PODPREDSEDA SMERU-SD JURAJ BLANÁR V DISKUSII NOVÉHO ČASU*). S tematickým slovom „smer – sd" tvorí kolokačne silný vzťah tematické slovo „strana", ktoré vytvára aj egoprezentačnú kolokáciu „naša strana" či kolokáciu „opozičné strany" posilňujúcu konflikt na predvolebnom politickom spektre.

Konštatovanie o kolokvializácii politickej komunikácie vyvodené z antroponým v zoznamoch tematických slov pri politickej strane SMER – SD neplatí; tu totiž nachádzame obe lemy, z ktorých pozostáva meno volebného lídra – „peter" aj „pellegrini", čo môže byť ovplyvnené frekventovaným tagovaním jeho virtuálneho profilu. Tematizácia tejto osobnosti je obyčajne kolokačne spojená s prestížnou funkciou predsedu vlády SR, ktorá znamená istú autoritu a môže teda niečomu alebo niekomu zvyšovať kredit (*PREMIÉR PETER PELLEGRINI SÚHLASÍ S VYJADRENÍM MINISTERKY DENISY SAKOVEJ*). Táto osobnosť je navyše vykresľovaná ako niekto, kto je schopný priniesť zmenu, kto drží slovo, kto je zodpovedný, čestný a férový.

c) Sme rodina

Tematizácia človeka ani v tomto prípade neobchádza možnosti jeho interpretácie v spojení s populistickým spôsobom komunikácie, ba možno povedať, že tu výrazne dominuje. Okrem spôsobov tematizácie zmienených pri predošlých politických stranách je v komunikácii strany Sme rodina „človek" niekto, kto má v živote ťažkosti alebo sa nachádza v nedobrom postavení (*Bojujeme za ľudí, počúvame čo ich trápi a chceme im pomôcť!*), niekto, koho záujmy dlhodobo nie sú zohľadňované a hájené (*SME RODINA je stranou, ktorá dokáže priniesť zmenu a konečne myslieť aj na ľudí.*) alebo niekto, kto si zasluhuje citlivý a solidárny prístup (*Keď má prísť zmena, je dôležité aby prišla strana, ktorá bude na ľudí myslieť srdcom a bude za nich bojovať.*). V týchto intenciách by mohlo byť zaujímavé zameranie na ovplyvňovacie techniky založené na emocionálnom podklade.

Vysoká frekvencia lemy „rodina“ je v prípade tejto politickej strany daná tým, že je súčasťou oficiálneho názvu, ktorý ale naznačuje aj jej programovú prioritu. Názov strany je v korpuse prítomný vďaka plateným obsahom a ich administratívnym častiam a v jadre príspevkov v kolokácii „sme rodina“ plní egoprezentačnú funkciu. Lema „rodina“ je takisto obsiahnutá v názve pravidelne prezentovaného *Programu pomoci rodinám*, teda dokumentu, v ktorom politická strana svoje priority konkretizuje. Ak sa však lemou nereferuje na vlastnú politickú stranu ale na spoločenskú jednotku, spravidla sa pozornosť upriamuje na jej finančné ťažkosti (*Mnoho Slovákov spolu so svojimi rodinami opúšťa rodné Slovensko s vidinou lepšieho finančného zabezpečenia a života v zahraničí.*).

Osamotené tematické antroponymum „boris“ aj v tomto prípade naznačuje kolokvializáciu komunikácie politickej strany. Spôsobov tematizovania líderskej osobnosti je niekoľko; zmieniť sa môžeme napríklad o pripomínaní jej účinkovania v médiách (*Už o chvíľu o 20:30 bude Boris diskutovať na Markíze*) neraz spojenom s výzvou sympatizantom, aby ho sledovali. „Boris“ je prezentovaný ako spoločenský a obľúbený človek (*Pri príležitosti profesionálneho sviatku Dňa diplomata Ruskej federácie Boris prijal pozvanie od veľvyslanca Ruskej federácie J. E. pána Alexeja L. Fedotova na spoločnú recepciu.*) s množstvom podporovateľov (*Dnes fandíme Borisovi v RTVS*). Pripomíname, že frekvenčné postavenie slova „boris“ je ovplyvnené tým, že aj táto lema je súčasťou oficiálneho názvu strany a teda aj súčasťou administratívnych zložiek mikroblogov.

Tematické slovo „slovensko“ sa rozvíja v podobných líniách ako v komunikácii predvolebne takisto opozičnej strany OĽaNO. Okrem priamej referencie na konkrétny stredoeurópsky región má „slovensko“ svoje ťažkosti (*Uvedomujeme si, že Slovensko má v súčasnosti množstvo problémov.*), avšak aj potenciál vyriešiť ich a zlepšiť svoj stav (*Posuňme Slovensko tým správnym smerom!*). To je možné dosiahnuť zmenou, na ktorej sa aktívne podieľa práve politická strana Sme rodina (*29. februára prinesieme Slovensku skutočnú zmenu.)*.

d) ĽSNS

Výrazne nízky počet komunikátov strany ĽSNS je spôsobený tým, že spoločnosť Facebook v roku 2017 zablokovala hlavnú stánku politickej strany s viac ako 80 000 sledovateľmi pre šírenie nevhodného obsahu.[7] Politický subjekt síce svoju oficiálnu stránku vytvoril opäť, no komunikácia so sympatizantmi už skôr prebieha prostredníctvom desiatok fanúšikovských stránok a stránok regionálnych organizácií. Následkom nízkeho počtu tokenov vo výskumnom korpuse tak získavame iba jedno tematické slovo, ktorým je „kotleba“. Interpretácia toho, prečo pri prezentácii

[7] Túto politickú stranu napriek nepomerne nízkemu počtu sledovaných tokenov v komunikácii v porovnaní s komunikáciou ostatných politických strán z analýzy nevylučujeme; spĺňa totiž uvedené metodologické kritérium, podľa ktorého je perspektívne skúmať tematické slová v textoch s rozsahom od 200 do 6500 tokenov (Čech – Kubát, 2016).

lídra dochádza k opačnej situácii ako v prípade politických strán OĽaNO a Sme rodina je komunikačne sťažená, pretože podkladom je iba päť výpovedí. Na ich základe je možné konštatovať iba to, že líder sa tematizuje v spojitosti s kritickým hodnotením istej skutočnosti (*Pohoršený Kotleba: BBSK chystá kšeft pre Poliakov!*).

e) SaS

Jediným prostriedkom egoprezentácie politickej strany Sloboda a Solidarita je skrátená podoba jej názvu – „SaS" –, ktorá je zároveň jej najfrekventovanejším tematickým slovom. Politická strana je, podobne ako v predchádzajúcich prípadoch, kolektívom, celkom alebo tímom deklarujúcim istý postoj, zámer či orientáciu, a to buď vo formálne aktívnej pozícii (*Každý už dnes vie, že SaS chce NIŽŠIE DANE a menej štátu pre každého.*), alebo s využitím inkluzívnej funkcie predložky „v" (*V SaS neprepadnú ani hlasy maďarských voličov*). Výrazná sebaprezentácia strany je ešte podporená tým, že v skupine tematických slov nenachádzame žiadne antroponymum odkazujúce na jej predsedu alebo lídra. K tematizácii strany tematizáciou kandidáta dochádza len zriedka; vtedy, keď je tento kandidát odborníkom v nejakej oblasti (*Aj o tom sme sa rozprávali s expertom SaS pre dane a dôchodkový systém Petrom Cmorejom.*). Tematizácia strany je namiesto toho pevne spojená s jej volebným programom (*SaS má najlepší program pre živnostníkov a podnikateľov a to je fakt*).

„Rok" sa tematizuje viacerými spôsobmi, ktoré sa, rešpektujúc sémantický význam slova, opierajú o istú faktografickosť – konkrétny rok je začiatkom alebo koncom niečoho alebo niektorý rok je dôležitý v histórii politickej strany. O niečo frekventovanejšie sa však lema „rok" spája s kritickým poukazovaním na dlhé obdobie vlády predvolebne koaličných strán (*Dvanásť rokov Smeru, dvanásť rokov korupcie, úpadku a rozkrádačiek za viac ako ŠESŤ MILIÁRD EUR*) a s konkrétnym momentom z minulosti, ktorý znamenal zmenu dovtedajšej spoločensko-politickej situácie na Slovensku (*Už tento piatok to budú presne dva roky od vraždy, ktorá otriasla Slovenskom.*).

Tematizácia Slovenska sa v komunikácii politickej strany SaS zásadne nelíši od jeho tematizácie v komunikácii predchádzajúcich predvolebne opozičných politických strán, možno iba pozorovať väčšiu frekvenčnú vyrovnanosť medzi jednotlivými tematizačnými líniami – „slovensko" referuje na presný geografický priestor, má svoje problémy, ale aj možnosti, ako ich vyriešiť. Tie sú opísané v dokumente *Návod na lepšie Slovensko*, ktorého prezentácia takisto zvyšuje početnosť tohto tematického slova.

Okrem toho, že „bratislava" je súčasťou administratívnych zložiek mikroblogov, politická strana SaS ju tematizuje aj v jadre svojich príspevkov. „Bratislava" je najmä miestom realizácie kontaktovej kampane (*Stretneme sa na Námestí SNP v Bratislave od štvrtej.*), no pozornosť púta kritika jej centralistického postavenia (*Úradníci v Bratislave by nemali rozhodovať o tom, z ktorých učební sa budú učiť deti v regiónoch po celom Slovensku*) a poukazovanie na konflikt medzi hlavným mestom, ktoré síce ponúka isté možnosti, no nie je „domovom", a ostatnými časťami Slovenska (*Urážajú ľudí, ktorí museli odísť od rodiny a hľadať dôstojný život v Bratislave či v zahraničí.).*

f) Za ľudí

Už príslušný gramatický tvar tematického slova „človek“ v názve politickej strany Za ľudí umožňuje uvažovať nad chápaním ľudí ako motivácie, pre ktorú strana vyvíja svoje aktivity. Aj v tomto prípade identifikujeme použitie tematického slova v populistických líniách všeobecne označujúcich obyvateľov Slovenska (*Andrej Kiska: „Bol by som veľmi rád, aby všetci ľudia boli hrdí na naše Slovensko.“*), no jeho miera je porovnateľná s použitím lemy v aspoň čiastočnej konkretizácii na istú skupinu (*Mnohí ľudia s ťažkosťami s čítaním nie sú schopní si klasickú formu programu prečítať.*). Politická strana Za ľudí však lemou „človek“ ako jediná nereferuje iba na sekundárnych, ale takisto na primárnych účastníkov politického diskurzu – s jej použitím tematizuje vlastných kandidátov (*Sme veľmi radi, že aj na nižších miestach kandidátky máme ľudí, ktorí majú za sebou príbeh.*). Tí sú sprítomňovaní aj v spojitosti s tematickým slovom „program“, ktoréhо časti majú medzi kandidátmi svojich špecialistov a odborníkov (*Viac už garantka programu pre zdravotníctvo @Andrea Letanovská – ZA ĽUDÍ.*).

Úvahy o obmedzovaní podprahovo pôsobiacej komunikácie strany Za ľudí môžeme rozvíjať aj s oporou o tematické slovo „slovensko“, ktoré je väčšinovo použité v priamej referencii na presný geografický areál. Zriedka tiež identifikujeme tematizačné línie už opísané pri ostatných predvolebne opozičných politických stranách, no aj jednu takú, ktorá je príznačná výlučne pre tento politický subjekt – „slovensko“ je prezentované ako niečo odcudzené slušným ľuďom, ako niečo, čo im imanentne patrí a čo by sa po voľbách mohlo dostať späť do ich vlastníctva (*Vráťme 29. februára Slovensko späť všetkým slušným ľuďom.*).

Pokiaľ ide o tematické antroponymá, v tomto prípade pozorujeme situáciu podobnú skôr predvolebne vládnej strane SMER – SD – identifikujeme meno aj priezvisko lídra. Kolokácia „andrej kiska“ je tematizovaná v spojitosti s mediálnym pôsobením politika, na ktorého referuje (*Už dnes 11:55 na RTVS Andrej Kiska a Peter Pellegrini*), a v spojitosti s pozíciou zakladateľa príslušnej politickej strany (*Andrej Kiska dal dokopy tím špičkových odborníkov.*). Neraz sa prezentuje ako názorovo vedúca osobnosť, ktorej vyjadrenia sú sprítomňované využitím technologických možností sociálnych médií (*@Andrej Kiska: Veľa som premýšľal o tom, čo sa teraz deje v Národnej rade. Sú situácie, kedy treba konať neštandardne a preto rozumiem Mirovi Beblavému. Všetko však má mať svoju mieru a myslím si, že je čas riešiť veci politicky.*). Špecifické je tematizovanie toho, že voči líderskej osobnosti je zo strany oponentov vedená antikampaň (*Špinavá antikampaň s cieľom diskreditácie osoby Andreja Kisku pokračuje.*), s čím takisto súvisí prítomnosť tematického slova „smer“, ktorý je za antikampaň zodpovedný (*Dnes som podal trestné oznámenie za krivé obvinenie mojej osoby v trápnom zinscenovanom videu, podľa všetkého v réžii Smeru.*). Za ľudí ako jediná politická strana aj tematizačne posilňuje konflikt s inou stranou.

Politická strana Za ľudí vo svojej komunikácii takisto pracuje s faktografickou oporou. Tematizačné línie slova „rok“ sú podobné tým, ktoré boli opísané pri politickej strane SaS. Najfrekventovanejšie sa kriticky poukazuje na obdobie dlhodobej

nečinnosti predvolebne vládnej strany (*Smer mal 12 rokov na to, aby pomáhal rodinám a dôchodcom.*), alebo pochvalne na obdobie aktivity člena svojej politickej strany (*VERONIKA REMIŠOVÁ – 4 ROKY V PARLAMENTE*).

Z analýzy tematických slov komunikácií volebne úspešných politických strán možno vyvodiť aj isté všeobecné špecifiká slovenskej sociálnomediálnej predvolebnej komunikácie. Typickým znakom predvolebného diskurzu je sebaprezentácia, ktorá sa profiluje ako prezentácia politickej strany, jej lídra alebo kandidátov. Predovšetkým politická strana SaS čerpá z prezentácie strany ako kolektívu, konkrétne osobnosti sú v jej komunikácii podľa analýzy tematických slov v pozadí, čo však neplatí pri ostatných politických stranách. Tematizácia líderskej osobnosti v niektorých prípadoch naznačuje kolokvializáciu politickej komunikácie (OĽaNO, Sme rodina), v niektorých sa vyznačuje charakteristikami príznačnými pre oficiálnu a verejnú komunikáciu (SMER – SD, Za ľudí). Lídri sú názorovými vodcami a kritikmi aktuálneho stavu, no tiež zodpovednými, schopnými, spoločenskými a férovými ľuďmi. Z viacerých analýz vyplýva posilňovanie konfliktu vládnych a opozičných politických strán v predvolebnom období, no iba strana Za ľudí vládnu politickú stranu aj tematizuje. Slovensko, ktoré je s výnimkou ĽSNS tematickým slovom všetkých skúmaných politických strán, je vnímané nielen ako geografický priestor, ale takisto ako objekt s hodnotovým atribútom; v tematických slovách politických strán sa objavuje preto, lebo strany nezriedka tematizujú záležitosti, ktoré vnímajú ako problémové. S frekventovaným tematizovaním ľudí sa spájajú aj populistické tendencie, ktoré sú príznačné najmä pre tri volebne najúspešnejšie politické strany (OĽaNO, SMER – SD, Sme rodina). Voľby ako hlavný faktor ovplyvňujúci charakter komunikácie tematizuje iba volebne najúspešnejšia politická strana, no všeobecne tematizované sú s politickým životom súvisiace reálie ako napríklad strana alebo program. Na podobe diskurzu sa prejavujú isté legislatívne okolnosti, ktoré sa ako prvky administratívneho štýlu prejavujú na textovej podobe mikroblogov. Každá z politických strán aspoň v obmedzenej miere využíva technologické možnosti sociálnych médií a pri viacerých z nich možno v priebehu predvolebnej kampane pozorovať repetitívnosť obsahu alebo modelovosť a schematickosť na formálnej úrovni mikroblogu. Hoci je potrebné konštatovať, že opísané črty predvolebného diskurzu nie sú veľmi prekvapivé, v aktuálnom príspevku sme im, minimálne v kvantitatívne ladených častiach, poskytli exaktnú metodologickú oporu.

Využitie kvantitatívneho podkladu pri skúmaní politického diskurzu sa v tejto štúdii teda ukazuje ako perspektívne, no nemenej dôležitá je následná kvalitatívna interpretácia získaných dát. Analýza zároveň naznačila niekoľko ďalších ciest, ktorými je možné viesť výskum predvolebnej sociálnomediálnej komunikácie – manipulačné a persuazívne techniky, metafory v politickom diskurze či kolokácie tematických slov. Nadstavbou by takisto mohlo byť porovnanie tematických slov celej sociálnomediálnej predvolebnej kampane s jej kľúčovými slovami. Takýmto spôsobom by sa totiž ešte spriezračnili jej tematicky najrelevantnejšie aspekty.

Bibliografia

BREZINA, Václav – McENERY, Antony – WATTAM, Steve: Collocations in context: A new perspective on collocation networks. In: International Journal of Corpus Linguistics, 2015, roč. 20, č. 2, s. 139 – 173.

BREZINA, Václav – WEILL-TESSIER, Pierre – McENERY, Antony: #LancsBox v. 5.x. [softvér]. 2020. Dostupný na: http://corpora.lancs.ac.uk/lancsbox.

ČECH, Radek: Language and Ideology: quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949-2011). In: Quality and Quantity, 2014, roč. 48, č. 2, s. 899 – 910.

ČECH, Radek – KUBÁT, Miroslav: Text length and the thematic concentration of text. In: Mathematical Linguistics, 2016, č. 2, s. 5 – 13.

DAI, Zheyuan – LIU, Haitao: Quantitative Analysis of Qeen Elizabeth II and American Presidents' Christmas Messages Over 50 Years (1967-2018). In: Glottometrics, 2019, roč. 19, č. 2, s. 63 – 88.

HOFFMANNOVÁ, Jana: Stylistika a ..... Praha: Trizonia 1997. 200 s.

JAGERS, Jan – WALGRAVE, Stefaan: Populism as political communication style: An empirical study of political parties´ discourse in Belgium. In: European Journal of Political Research, 2007, roč. 46, č. 3, s. 323 – 353.

JAWOROWICZ, Piotr: Wideokomunikowanie polityczne w Internecie. Youtube i polskie partie polityczne w latach 2011–2014. Warszawa: Difin 2016. 223 s.

KOLENČÍKOVÁ, Natália. Mikroblog jako gatunek w słowackich mediach społecznościowych. In: Zeszyty Prasoznawcze, 2018, roč. 61, č. 3, s. 460 – 475.

KUBÁT, Miroslav – ČECH, Radek. Quantitative Analysis of US Presidental Inaugural Adresses. In: Glottometrics, 2016, roč. 16, č. 2, s. 14 – 27.

McQUAIL, Denis: Úvod do teorie masové komunikace. Praha: Portál 2007. 447 s.

MACHO, Marián: Terminologické a metodologické východiská skúmania jazyka politického diskurzu. In: XLinguae, 2012, roč. 5, č. 1, s. 14 – 22.

MANOVICH, Lev: The Language of New Media. Cambridge/London: The MIT Press 2001. 354 s.

MAYFIELD, Antony: What is Social Media? London: iCrossing 2008. 36 s.

MOLNÁROVÁ, Patrícia: Kognitívno-sémantická interpretácia metafory v kontexte politicko-ideologického vývinu v rokoch 1965 – 1970. In: Slovenská reč, 2019, č. 2, s. 185 – 204.

LISTER, Martin – DOVEY, Jon – GIDDINGS, Seth – GRANT, Iain – KELLY, Kieran: New Media: a critical introduction. London/New York: Routledge 2009. 446 s.

ODALOŠ, Pavol: Charakteristiky a techniky slovenskej politickej komunikácie. In: Jazyk, média, politika. Eds. S. Čmejrková – J. Hoffmannová. Praha: Academia 2003. s. 217 – 243.

ORGOŇOVÁ, Oľga – DOLNÍK, Juraj: Používanie jazyka. Bratislava: Univerzita Komenského 2010. 229 s.

PAGE, Ruth – BARTON, David – UNGER, Johann Wolfgang – ZAPPAVIGNA, Michele: Researching Language and Social Media: A Student Guide. London/New York: Routledge 2014. 202 s.

PATRÁŠ, Vladimír: Politická komunikácia v slovenských mediálnych podmienkach na konci 90. rokov. In: Jazyk, média, politika. Eds. S. Čmejrková – J. Hoffmannová. Praha: Academia 2003. s. 174 – 216.

POPESCU, Ian-Iowitz: Text ranking by the weight of highly frequent words. In: Methods in the Study of Language and Text. Eds. P. Grzybek – R. Köhler. Berlín/New York: Mounton de Gruyter 2007. s. 555 – 566.

RAŠOVÁ, Dominika: Pragmatika jazykových javov v masmediálnej komunikácii. Kontrastívna štúdia na materiáli v slovenčine a v nemčine. Kraków: Spolok Slovákov v Poľsku 2013. 170 s.

ŠTEFANČÍK, Radoslav – DULEBOVÁ, Irina: Jazyk a politika. Jazyk politiky v konfliktnej štruktúre spoločnosti. Bratislava: Ekonóm 2017. 193 s.

sciendo
DOI 10.2478/jazcas-2022-0014

BASE DE DONNÉES NUMÉRIQUE DES CORPUS KABYLES ET EXPLOITATION. ESSAI D'ANALYSE LEXICOMÉTRIQUE DE LA DIMENSION IDENTITAIRE DANS LE DISCOURS ROMANESQUE

AREZKI IKHERBANE[1] – RAMDANE BOUKHERROUF[2] – NOURA TIGZIRI[3]
[123]Laboratoire d'Aménagement et d'Enseignement de la Langue Amazighe
[123] Université Mouloud Mammeri de Tizi-Ouzu, Tizi-Ouzu, Algérie

IKHERBANE, Arezki – BOUKHERROUF, Ramdane – TIGZIRI, Noura: Kabyle corpus digital database and exploitation. Test of lexicometric analysis of the identity dimension in the romanesque discourse. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 894 – 905.

Abstract: The purpose of this contribution is to show, through a preliminary analysis of a corpus sample composed of the first five kabyle novels (1963-1990), the contribution of lexicometry as a new method based on statistics, in the treatment of large corpora and the establishment of databases. The aim is to describe all the phases intrinsic to the preliminary processing of a corpus (transcription, tagging and lemmatization) before submitting them to the various stages of its exploitation. Thus, in our corpus, we have opted to deal with the theme of identity induced by the five works by highlighting both the overused vocabulary and the singularity of each work in relation to the corpus as a whole. But before moving on to the quantitative analysis of the vocabulary, a work of data preparation is necessary. We intend to focus on the orthographic choices to be adopted by removing all ambiguities, the marking out and the lemmatization of the corpus. In order to do this, we have resorted to Lexico5 computer tool.

Key words: corpus, kabyle, identity, novel, lexicometry, databases

0. DONNÉES INTRODUCTIVES

La langue amazighe (berbère), était une langue minorée, réduite à l'usage quotidien, parlée uniquement dans les régions amazighophones. Elle fait partie des langues chamito-sémitiques. Elle couvre une vaste aire géographique : de l'Afrique du Nord comprenant le Maroc jusqu'en Egypte en passant par l'Algérie, la Tunisie et la Lybie sans oublier le Sahara et une partie du Sahel Ouest-africain avec de nombreux locuteurs au Mali et au Niger (Chaker, 1983).

La réalité linguistique du domaine berbère (amazigh) montre que la langue connaît une fragmentation et une dialectalisation parfois importante d'un point à un autre. À l'intérieur du dialecte kabyle, des divergences se manifestent particulièrement en phonétique et dans le lexique. L'intercompréhension est parfois difficile avec les autres variantes de l'amazigh.

Dans le présent article, il est question de présenter quelques hypothèses préliminaires en phase expérimentale, d'un grand projet qui s'inscrit dans le cadre d'un

axe de recherche[1] portant sur les technologies linguistiques. Il s'agit d'un projet qui a pour objectif de mettre en place une grande base de données numérique des corpus amazighs[2] (oraux et écrits), accessible en ligne, dans la perspective de les exploiter dans le cadre du traitement automatique de l'amazighe (dictionnaires, atlas linguistiques, analyse du discours, grammaires, syntaxe, traduction automatique, enseignement, etc.) afin de les mettre à la disposition des chercheurs berbérisants qui veulent les exploiter.

Par ailleurs, loin de faire une analyse quantitative d'un grand corpus, l'objectif central de notre première exploitation consiste en une tentative d'application lexicométrique[3] sur un échantillon de corpus en mettant en exergue les traitements préalables de préparation adoptés : la transcription, le balisage et la lemmatisation.

Ainsi, nous nous appuyons sur un corpus des cinq premiers romans kabyles, formé d'une panoplie d'œuvres qui s'étale sur une période de trois décennies : Belaid Ait Ali (1963), Rachid Aliche (1981, 1986), Said Sadi (1983) et Amar Mezdad (1990).

L'intérêt de ce corpus réside dans le fait qu'il a suscité l'intérêt non seulement des études littéraires, mais aussi des études de statistique textuelle. C'est à ce titre que les travaux de la critique littéraire berbérisante montrent que la thématique identitaire est portée dans l'émergence de l'écriture romanesque berbère (Salhi – Sadi, 2016) puisque ce nouveau genre est né dans un contexte particulier, caractérisé par un mouvement de revendication pour la reconnaissance du berbère en tant que langue et identité. Ainsi l'étude statistique (Loikkanen, 1998) portant sur le vocabulaire du roman kabyle, montre que le thème le plus récurent est celui de la revendication culturelle et identitaire. C'est pourquoi nous proposons d'inscrire notre recherche dans ce sillage, mais avec une démarche qui diffère quelque peu, en vue d'aborder cet aspect d'identité tant de point de vue quantitatif que qualitatif.

Par ailleurs, en plus du statut relevant du sème **identité** caractérisant les conditions de production et de l'émergence du discours romanesque kabyle, notre choix en faveur de la lexicométrie, comme domaine issu de la statistique lexicale et des derniers développements en traitement automatique des langues naturelles (TALN), est motivé par le fait que l'application de ces méthodes sur un corpus littéraire, aussi vaste soit-il, permet d'obtenir des résultats d'une manière rapide et quasi systématique.

---

[1] Le Projet mené au niveau du laboratoire d'Aménagement et Enseignement de la Langue Amazighe de Université Mouloud Mammeri de Tizi-Ouzou, Algérie.

[2] Le projet s'inscrit dans le cadre des projets du comité d'évaluation de la recherche universitaire de l'enseignement supérieur et de la recherche scientifique dont il a fait l'objet. Il comprend trois axes : Base de données kabyle (2010), La transcription synchronisée des corpus oraux (2015) et les écrits anciens en tamazight : recueil, numérisation, récriture et géolocalisation (2018). Ainsi, ces derniers, ont fait l'objet de plusieurs communications et publications : Tigziri (2014), Boukherrouf et Tigziri (2015, 2016) et Tigziri, Jolivet et Boukherrouf (2017).

[3] Deux analyses lexicométriques ont été réalisées à l'aide du logiciel AntConc des corpus kabyles Tigziri (2016, 2017).

Le projet auquel nous nous attelons, compte tenu du corpus cité plus haut, sera axé sur la problématique du discours identitaire et la manière dont il est organisé à travers les œuvres précédemment mentionnées. Il s'agit d'examiner la dimension identitaire en tant qu'entité concrète, inhérente à l'ensemble des œuvres du corpus choisi, afin de dégager ses modalités de fonctionnement ainsi que sa structure interne, à partir de la qualification du mot et le sens du texte. Pour ce faire, nous ferons appel aux travaux de Labbé et Labbé (2013 ; 2019) qui ont bien décrit l'objet et la méthode de la lexicométrie:

> Elle permet de traiter de vastes ensembles de textes (corpus), d'établir leur vocabulaire, de classer les vocables en fonction de leur fréquence, de leur répartition, de leurs catégories grammaticales. Elle établit les contextes d'emploi d'un vocable et les combinaisons les plus fréquentes dans lesquelles il entre, ce qui permet de déterminer le ou les sens de ce vocable. Elle retrouve les principaux thèmes présents dans un corpus, son genre et son style. Elle segmente ce corpus en fonction des ruptures thématiques ou stylistiques. Pour obtenir ces résultats, des traitements préalables sont nécessaires : balisage des textes, correction et standardisation orthographiques, étiquetage des mots. Le texte peut alors entrer dans une bibliothèque électronique à la disposition des chercheurs. (2013, p.1).

Ceci étant, nous allons associer à l'étude du contexte qui tient compte des mots et de leurs dispositions, une analyse sémantique, qui cherche à identifier à l'intérieur de ces mots un sens au-delà des termes stricto sensu (Ada, 2004). Pour ce faire, nous tenterons d'abord de faire le relevé de tous les éléments de notre corpus (Maingueneau, 1991) avant de les exploiter pour caractériser l'œuvre dans son ensemble ou pour traiter les données lexicales par rapport à l'ensemble du vocabulaire de l'œuvre (Muller, 1982).

Le présent essai est composé de deux parties. Après une phase préparatoire et l'établissement des données, qui présentent le corpus, sa transcription, son balisage, sa lemmatisation et les différentes fonctionnalités prises en charge par notre outil d'analyse Lexico 5, nous passerons à la description progressivement les différentes phases de l'exploitation des données préparées.

1. PRÉPARATION ET ÉTABLISSEMENT DES DONNÉES

1.1 Présentation des données

Le corpus que nous soumettons à l'étude est constitué du discours romanesque produit en kabyle. Ces œuvres s'étendent sur une période allant de (1963 à 1990). Le premier roman intitulé *Lwali n udrar « Le saint de la montagne »* a été publié en 1963 par Dallet dans le Fichier de Documentation Berbère (FDB). L'ensemble de ces textes qui constituent notre corpus représentent, comme déjà cité, cinq romans, à savoir : *Lwali n udrar « Le saint de la montagne »* de Belaid Ait Ali, *Asfel « le sacrifice »* et

Faffa « diminutif de la France» de Rachid Aliche, *Askuti « Le scout»* de Said Sadi et *Iḍ d wass « nuit et jour »* de Amar Mezdad. Mis à part Rachid Aliche qui compte deux œuvres, ces textes qui représentent une œuvre de chaque auteur, ont été classés selon l'ordre chronologique des dates de publication pour pouvoir étudier l'évolution du vocabulaire, abstraction faite de la date de rédaction car dans le contenu des ouvrages, il y a presque aucune symétrie entre la date de publication et celle de sa rédaction ; l'écart est souvent conséquent. C'est le cas de « *Lwali n udrar* » de Belaid Ait Ali, écrit entre 1940 et 1945, mais qui n'est publié qu'en 1963 à titre posthume ; idem pour *« iḍ d wass »* de Amar Mezdad, écrit à partir des années 80 et non publié qu'en 1990.

1.2 L'outil d'analyse : Aperçu sur ses fonctionnalités

L'analyse des données est menée à l'aide du logiciel **Lexico 5**[4], développé depuis le milieu des années 2010. Il est le prolongement de Lexico 3 développé à partir de 2003. Lexico 3 est lui-même une suite logique de ses deux versions précédentes (Lexico 1 puis Lexico 2). L'une des nouveautés de la version Lexico 5 par rapport à la précédente est la prise en charge de l'encodage des corpus en Unicode, offrant une possibilité à un plus grand nombre de langues, entre autre le berbère, de bénéficier de ce traitement lexicométrique. Les fonctions documentaires (concordances, contextes), statistiques (spécificités), analyses multidimensionnelles (analyses factorielles des correspondances, arborées), sont l'essentiel de ce que renferme cet outil en termes de fonctionnalités.

Le dictionnaire est une donnée que le logiciel génère automatiquement sous forme de deux colonnes. Cette liste des formes présente dans le texte s'affiche toujours dans la colonne de gauche du logiciel, sous l'onglet dictionnaire, à côté des onglets navigation et rapport. Dans le dictionnaire de gauche (cf. Figure 1) les occurrences sont affichées dans un ordre lexicométrique (classement décroissant). Si on clique sur l'onglet Formes (ordre lexicographique), le classement des formes du texte s'affiche alors par ordre lexicographique (ordre alphabétique).

1.3 Etablissement et standardisation orthographique des textes

En dépit des disproportions constatées entre les romans proposés à l'étude, en termes de volume, qui diffèrent tant par le nombre de pages que de mots, nous les avons établis sous forme d'un seul et unique corpus. Ainsi, de tout le texte narratif des romans, nous avons supprimé uniquement les éléments qui renvoient au péritexte, autrement dit, les numéros et rubriques éventuels des chapitres, les dédicaces, les numéros des pages, les titres, les notes en annexe qui expliquent les mots, les lexiques, les chiffres, les dates, les présentations biographiques, les préfaces ainsi que les passages écrits en entier dans la langue étrangère.

[4] Téléchargeable sur http://www.lexi-co.com/Produits.html

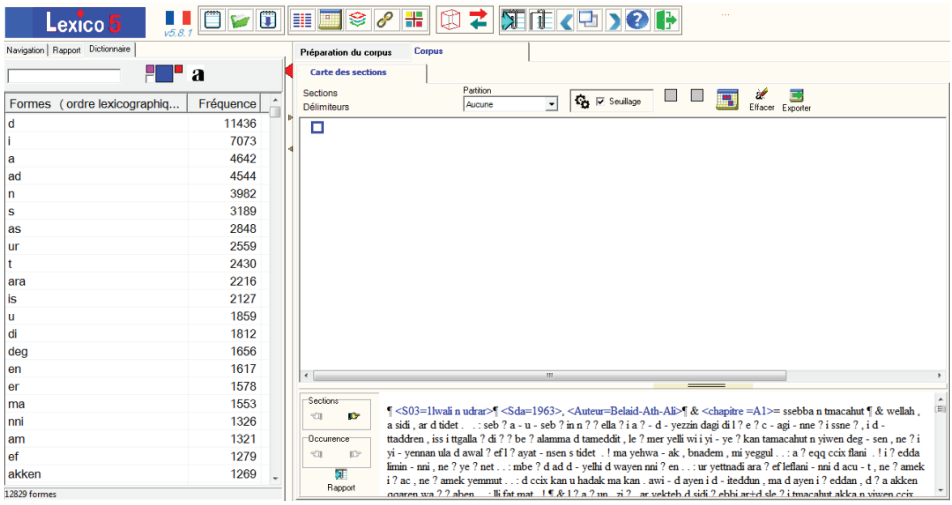


Fig. 1. Index alphabétique des données textuelles

Aussi, à cause de la notation que chaque auteur a choisi d'utiliser dans son texte, le découpage du texte en mots à base d'unités graphiques parait délicat. On a souvent écrit des éléments différents ensembles, comme c'est le cas pour *les affixes personnels* et *les prépositions.* Pour les y homogénéiser, nous avons dû procéder à la séparation entre les mots outils à fonction grammatique des autres lexèmes auxquels ils se rattachent. Mais malgré l'apparente ambiguïté de certaines classifications, nous avons dû s'en tenir, outre les choix adoptés par (Loikkanen, 1998), aux différentes propositions des manuels du kabyle pour déterminer certaines unités mal identifiées.

Par ailleurs, le mot est avant tout une unité graphique, séparé des autres par des blancs ou par des signes de ponctuation. Mais d'autres cas, ils sont constitués de deux ou plus de morphèmes composés qui rendent ambigu la définition que nous pourrions donner du mot *mulac « sinon »*, *bu qerru « grosse tête »* et *adrar ufud « le tibia »*. Ainsi, une unité graphique peut correspondre à plusieurs unités du lexique ou inversement une unité du lexique peut se composer de plusieurs unités graphiques. Pour effectuer le décompte des occurrences afin de regrouper les vocables, la méthode lexicométrique recommande d'élaborer les critères à base desquels on traite le texte et on reconstitue le lexique. Nous avons constaté, qu'hormis quelques compositions synaptiques, dont le processus de lexicalisation n'est pas encore achevé, il est rare de trouver beaucoup de mots composés écrits ensemble Dallet (1982). Ainsi, nous écrivons *ma ulac « si non », yal wa « chacun »*, au lieu de les écrire en un seul mot dont les parties sont groupées ou liées par un tiret. S'agissant du trait d'union justement qui, en kabyle, peut parfois si-

gnifier une seule unité de sens et d'autre fois deux voir plus, nous avons opté de le considérer comme un caractère délimiteur. Par ailleurs, dans le cas où l'unité est considérée comme une seule forme, nous avons introduit à la place du tiret le caractère « + » non délimiteur pour voir l'expression comptée comme une seule occurrence. Concernant le cas des particules de possession *bu « celui », u « de »*, nous les avons tantôt considérées séparément et tantôt rattachées. La particule est ainsi rattachée pour le cas des noms propres, du lexique des maladies et des plantes mais souvent détachée pour le cas des numéraux ou encore de ceux dont la formation n'est pas totalement grammaticalisée.

1.4 Balisage du texte

Avant de commencer à travailler statistiquement sur le texte, il est nécessaire de définir les normes lui permettant d'être segmenté en formes (unités textuelles) et afin qu'il soit reconnaissable par le logiciel, et ce conformément aux caractéristiques textuelles de chaque genre du discours. Ainsi, pour que la machine puisse traduire le texte dans un code qu'elle comprend, il suffit d'enregistrer le texte en format brute (.txt). Pour ce qui est de la séparation des mots, le caractère est soit délimiteur, c'est le cas de l'espace et plus généralement des signes de ponctuations, soit il ne l'est pas. En conséquence, une suite de caractères non délimiteurs contenue entre deux caractères délimiteurs est considéré comme une occurrence analysable. En raison de la spécificité de la langue amazighe, nous avons considéré le trait d'union comme délimiteur, car dans la plupart des cas celui-ci relie différents affixes aux verbes, noms ou adjectifs, etc.

Par ailleurs, pour considérer une unité composite comme une seule occurrence, il suffit de les relier à l'aide du signe « + » non délimiteur. De même, pour distinguer les unités homonymiques entre elles, notamment un nom propre et un adjectif, nous avons précédé l'ensemble des noms propres du corpus d'un astérix, et ce dernier, est considéré comme un marqueur des noms propres de l'ensemble du corpus.

L'autre étape importante réside dans le codage du texte à l'aide de certaines clés informatives codées, qui caractérisent les métadonnées et les données du corpus. Ainsi, la clé <Sda=1963> désigne l'année de publication, la clé <Auteur=Belaid Ait Ali> la partition par auteur, <S03=1> les titres d'un texte qui réunissent en une seule base textuelle différents sous-corpus délimitant (une division logique, thématique ou chronologique), nous pouvons comparer plusieurs parties du corpus et également naviguer dans le corpus plus facilement. Concernant notre corpus, il était nécessaire de répartir les textes selon les titres des romans pour pouvoir les comparer entre eux (<S03=1lwali n udrar>, <S03=2 asfel>, <S03=3 askuti>, <S03=4 faffa>, <S03=5 iḍ d wass>). C'est la même chose pour obtenir une séparation par date d'édition <Sda=1963>, par auteur <Auteur=Belaid Ait Ali>, ou bien encore par chapitre <chapitre =A1>=*ssebba n tmacahut « l'argument de l'histoire »*>. Aussi, pour qu'une

seule forme ne soit pas traitée deux fois, les majuscules devront être supprimées. Ainsi, la balise <S03=titre d'un texte> par exemple, nous permet de générer les données de chaque corpus (cf. Tableau 1).

Partie	Occurrences	Formes	Hapax	Fmax	Forme
1lwali n udrar	33898	4258	2251	2606	d
2 asfel	14789	3739	2281	1195	d
3 askuti	32626	5095	2882	2355	d
4 faffa	23614	5692	3472	1426	d
5 iḍ dwass	47957	6507	3323	3460	d
T corpus	**152884**	**15071**	**7444**	**11042**	**d**

Tab. 1. Partition par la balise du titre de l'œuvre

1.5 La lemmatisation des données du corpus

Après dépouillement et comptage des différents mots, pour des besoins d'analyse, nous avons regroupé un certain nombre de mots en un seul lemme. A l'aide de ce procédé, nous avons réuni en une seule classe quelques lemmes en suremploi. Ainsi, la lemmatisation est caractérisée par le regroupement en une seule forme canonique de toutes les flexions appartenant à une seule classe et catégorie grammaticale. Ainsi, nous avons représenté sous un seul lemme toutes les déclinaisons avec lesquelles nous avons pu écrire les mots dans le corpus étudié. Le lemme *tawwurt « la porte »,* nom féminin singulier, regroupe l'ensemble de ses variantes et leur flexion comme, *tiwwura, tewwura, tabburt, tebburt, tibbura, tebbura*. Le lemme préposition *« avec »* englobe les formes *s, es, yis, yiss*, et le lemme *ad*, particule de l'aoriste, englobe les formes variantes, comme *a, an, at.* Pour la forme verbale, comme langue aspectuelle, nous avons réduit l'ensemble des flexions à la forme de l'aoriste simple.

2. EXPLOITATION DES DONNÉES ET INTERPRÉTATION

Après avoir présenté les étapes principales inhérentes à la préparation et l'établissement du corpus, nous tenterons de présenter un essai de son exploitation en mettant en exergue l'index alphabétique de quelques données lemmatisées et la présentation des résultats des données des cinq œuvres par partition AFC.

2.1 Index alphabétique des données

La première information sur notre collection de textes apparaît sous forme d'une liste (d'index lexicographique) de toutes les unités du lexique ou vocables (entrées du dictionnaire) avec leurs effectifs (c'est-à-dire le nombre de leurs occurrences). Sous cette liste, qui donne, pour chaque mot, les formes graphiques sous lesquelles il s'actualise, nous avons résumé les plus importantes (cf. Tableau 2).

ini («dire, aoriste simple »)	1418
yenna	320
nniɣ	123
yini	112
yeqqar	105
tiniḍ	95
tenna	89
iniɣ	70
nnan	62
tenniḍ	34
yinin	29
inin	28
teqqar	27
innan	25
yeqqaren	23
qqareɣ	21
teqqareḍ	18
nini	15
nenna	13
neqqar	09
qqarent	08
yenni	07
teqqarem	07
ini	04
init	03
tinim	02
nnin	01
tinem	01
(…)	
ili («être», aoriste simple)	999
yella	278
yelli	96
tella	122
llan	120
yili	88
tili	78
lliɣ	33
nella	27
llant	26
telliḍ	21
llin	04
tiliḍ	09
ttilin	06
tettili	05
llin	04
ttilint	03
uxxam	156

(…)	
tamurt « pays »	408
tmurt	127
tmura	15
timura	04
(…)	
taddart « village »	282
tddart	249
tudrin	25
tuddar	08
leqbayel	34
aqbayli « kabyle »	79
leqbayel	34
teqbaylit	22
taqbaylit	15
aqbayli	06
uqbayli	02
(…)	
ṛebbi « dieu »	376
(…)	
Ccix « prêcheur »	259
(…)	
Netta/nettat « il/elle	483/236
(…)	
iw/ik « ma /ta »	393/336
(…)	
nekk/keč « moi/toi »	341/183
(…)	
Nutni/nekkni « ils/nous »	107/82
(…)	
ameẓyan « petit »	188
(…)	
muḥend « nom d'un personnage	208
(…)	
malḥa « nom d'un personnage »	136
(…)	
axxam « maison »	360
uxxam	164
axxam	156
ixxamen	29
yixxamen	11
(…)	
ass « jour »	398
wass	224
ussan	104
wussan	70
(…)	

nelli	17	ddunit	257
yettili	17	(…)	
ylin	17	awal	230
yettilin	03	awalen	8
ilit	03		
iliɣ	05		
nili	04		
tellamt	04		
yilin	04		
yettilin	03		
ilint	01		
tilimt	01		

Tab. 2. Extraits de l'index alphabétique : lemmatisation semi - automatique

Les premiers tests montrent que le discours est caractérisé par des unités graphiques sur-employées telles que *ini « dire » et ili « être »*, utilisées successivement comme verbe et/ou auxiliaire. Le prétérit de la deuxième personne du singulier (*Yenna*, 320 fois et *yella* 278) semble être la classe grammaticale la plus dominante dans le discours. Le récit sur des événements passés, discours rapporté serait la raison principale.

Si les substantifs *ṛebbi « dieu »* 376 et *ccix « prêcheur ou vieux »* 259 occurrences, peuvent être des marqueurs discursifs, à cause d'un supposé lien entretenu avec l'existence et la religion, ce n'est peut-être pas le cas de ces mêmes lexies selon qu'elles soient mises ou non dans un rapport de contextualisation au même titre que le substantif *ddunit « vie »* cité 257 fois.

La déictisation par emploi de pronoms personnels indépendants ou affixes (*netta/nettat (lui/elle)* 483/236 fois, *iw/ik (mien/tien)* 393/336 fois, *nekk/kečč (moi/toi)* 341/183 fois, *nutni/nekkni (eux/nous)* 107/82 fois, réunis par paires oppositives, rendent compte d'une mise en contexte du discours sur soi et l'autre. Ce rapport à l'altérité qui est par conséquent un des indicateurs majeurs du discours sur l'identité, constitue un réactif essentiel grâce auquel il est possible de clarifier l'ethos de celui qui parle et de celui pour qui le message est supposé prédestiné.

Le recours à l'adjectif *amezyan « petit »,* cité 188 fois aux côtés du nom propre *Muḥend* 208 fois, dénoterait de cette volonté de produire un discours sur la jeunesse. De la même manière, *Malḥa* qui a été citée 136 fois, est un autre nom propre qui, en plus de la vivacité qui le caractérise, rend bien compte du sentiment qu'un auteur puisse éprouver à l'égard d'une femme.

L'emploi des termes comme *tamurt « pays », axxam « maison » et taddart « village »* respectivement répétés 408, 360 et 282 fois, en tant que des espaces - concrets ou abstraits - ne laisse aucun doute quant à la place que ces auteurs leur ont accordés grâce à des choix discursifs qui rendent ces espaces plus attrayants. Ainsi, l'espace local serait l'une des caractéristiques principales dans le roman identitaire.

Contrairement au suremploi d'unités comme *ass « jours »* 398 fois, *awal « parole »* 230 fois, dont on ignore le sens exact en dehors du contexte de leur mise en discours, il y a lieu de souligner le sous emploi des formes adjectivales comme *aqbayli « le kabyle »* 79 fois et *amaziɣ « l'amazigh/berbère »* 30 fois seulement.

2.2 Essai d'analyse factorielle des correspondances des cinq œuvres

Pour rendre compte approximativement des grandes oppositions sous-jacentes dans le corpus de textes, deux constats sont mis en exergue, d'une part, l'œuvre de Belaid Ait Ali apparaît excentrée toute à droite avec comme caractéristique des termes religieux (*ṛebbi « Dieu » et ccix « vieux-prêcheur »).* De l'autre, les quatre œuvres restantes qui se rapprochent toutes de l'axe du milieu avec des différences parfois remarquables sur le plan du lexique. Ainsi, au moment où Rachid Aliche se rapproche des termes comme *ddunit « la vie »,* rendu par le pronom affixe *iw « mien(ne) »*, Said Sadi en porte le souci de *tamurt « pays »,* de manière subjective grâce au pronom personnel *nekk « moi »*. Un peu plus en haut de l'axe vertical, le nom propre, la forme adjectivale ainsi que la forme du personnel *nutni « eux »* sont tous ce qui font de Amar Mezdad une œuvre à part (cf. Figure 2). Nous pouvons conclure que l'écart et la singularité mis en exergue dans les cinq œuvres est d'ordre chronologique. En effet, la singularité de l'œuvre de Belaid Ait Ali est un fait marquant caractérisant la vision du discours identitaire kabyle.

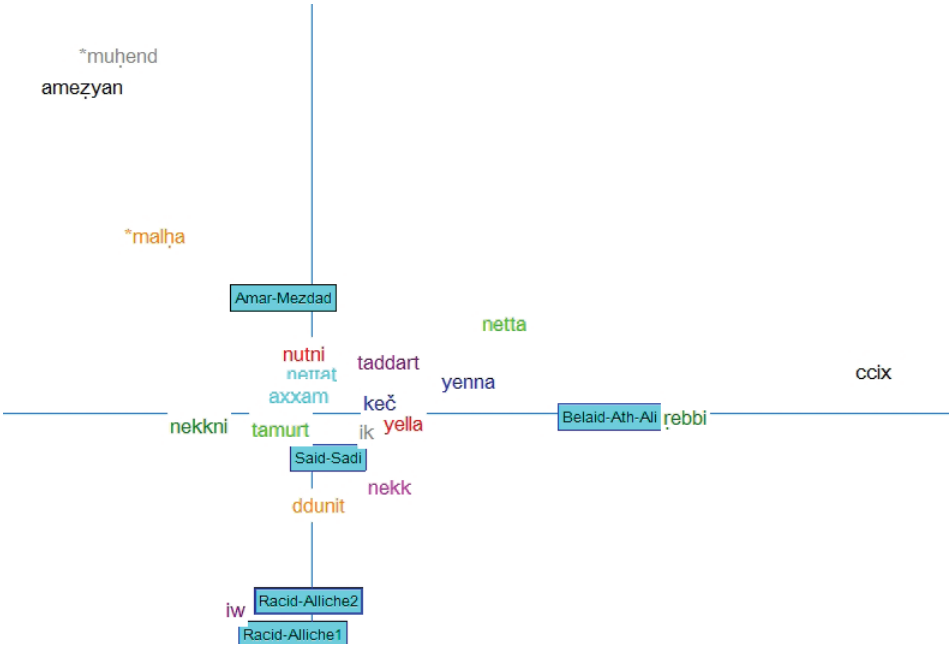


Fig. 2. Essai de présentation de partition par AFC

3. CONCLUSION ET PERSPECTIVES

Parmi les objectifs fixés dans notre projet, comme il a été signalé au début, il s'agira de mettre en place une base de données numérique des corpus amazighs, transcrits et balisés avant de les mettre à la disposition des chercheurs berbérisants pour des besoins d'analyse dans plusieurs domaines.

Ainsi, une fois toutes les données seront regroupées (orales et écrites), il conviendra de les classer en genres du discours et en domaines de traitement accompagnées de toutes les informations nécessaires à chaque corpus (les métadonnées).

Par ailleurs, avant de passer à l'étape de la transcription et de l'établissement des données, un travail de préparation des corpus en fonction des outils et des méthodologies à adopter est plus que nécessaire. Il permettra d'établir les données conformément aux perspectives théoriques et méthodologiques adoptées. Aussi, une phase d'expérimentation et de tests sur des choix de transcriptions à adopter seront d'une importance capitale à cette tâche de préparation des données à soumettre à l'exploitation.

C'est dans cette perspective, que nous avons choisi d'inscrire notre projet dans le domaine de la lexicométrie pour établir les données recueillies conformément aux choix méthodologiques de cette approche. Le choix de l'échantillon du corpus soumis aux premiers tests réside dans le fait que les cinq œuvres s'inscrivent dans un seul genre et partagent la thématique de l'identité.

Les résultats de ce test sont en phase expérimentale, d'autres étapes s'imposent afin d'établir les données avant de les généraliser pour l'ensemble des corpus.

Bibliographie

ADA, Myriam Scanu : Séminaire d'Analyse du texte fantastique : Littérature et informatique, Hyperbase logiciel pour l'analyse textuelle. Università degli Studi di Bologna 2004.

ALICHE, Rachid : Asfel. Mussidan : Fédérop 1981.

ALICHE, Rachid : Faffa. I yuɣen irgazen ur ttrun. Mussidan : Fédérop 1986.

BOUKHERROUF, Ramdane – TIGZIRI, Noura : Base de données kabyles : collectes de données et applications. Synchronisation texte / son. In : Iles d Imesli, 2015, No 7, pp. 193–206. Disponible sur : http://revue.ummto.dz/index.php/idi/issue/view/122

BOUKHERROUF, Ramdane – TIGZIRI, Noura : Etablissement des corpus et construction de l'objet pour l'analyse du discours. Transcription synchronisée à l'aide du logiciel Winpitch. TICAM'16, 28 et 29 novembre 2016, IRCAM, Rabat-Maroc.

CHAKER, Salem : Un parler berbère d'Algérie (Kabylie) : syntaxe. Aix-en-Provence : Publications de l'Université de Provence 1983.

DALLET, Jean-Marie : Dictionnaire kabyle-français. (Parler des Ait-Menguellat, Algérie). Paris : SELAF 1982.

DALLET, Jean-Marie – DEGEZELLE, Jules-Louis (Eds.): Les cahiers de Belaïd ou la Kabylie d'auton, T1 (Textes) et TII (traduction). Fort National, Fichier de Documentation Berbère (FDB) 1963. Disponible sur : http://dx.doi.org/10.1080/17409292.2016.1120548

LABBÉ, Cyril – LABBÉ, Dominique : Lexicométrie : quels outils pour les sciences humaines et sociales ? In : Usages de la lexicométrie en sociologie, France : Guyancourt 2013. Disponible sur : https://hal.archives-ouvertes.fr/hal-00834039

LABBÉ, Cyril – LABBÉ, Dominique : Humanités numériques. Données et méthodes. Marcel Proust. A la recherche du temps perdu. Semaine Data-SHS, 9-14 décembre, France : Université de Grenoble-Alpe 2019.

LOIKKANEN, Sinikka : Vocabulaire du roman kabyle (1981-1995). Mémoire de DEA, Paris : INALCO 1998.

MAINGUENEAU, Dominique : L'Analyse du discours. Paris : Hachette 1991.

MEZDAD, Amar : Iḍ d wass. Alger : Asalu 1990.

MULLER, Charles : Une nouvelle façon de voir le lexique : le Brunet. In : Le français moderne, 1982, No 4, Conseil International de la Langue Française, pp. 321–328.

SADI, Said : Askuti. Paris : Imedyazen 1983.

SALHI, Mohand Akli – SADI, Nabila : Le Roman Maghrébin en Berbère. In : Contemporary French and Franchophone Studies, 2016, Vol. 20, No 1, pp. 27–36.

TIGZIRI, Noura – JOLIVET, Remi – BOUKHERROUF, Ramdane : Base de données de corpus oraux : applications. The 3rd International Conference on Business Intelligence, march 29-31, 2017, Beni Mellal-Maroc. Disponible sur : https://www.cbi-bm.com/

TIGZIRI, Noura : La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique. In : Revue Asinag, 2014, Vol. 1, No 9, pp. 75–90.

TIGZIRI, Noura : Analyse textuelle à l'aide du concordancier AntConc d'une œuvre de Bélaid Ait Ali. In : Iles d Imesli, 2016, No 8, pp. 163–172.

TIGZIRI, Noura : Ait Menguellet ou l'éveilleur des consciences. In : Langue, Cultures, Communication, : L2C, 2017, Vol. 1, No 2, pp. 269–287.

sciendo
DOI 10.2478/jazcas-2022-0015

QUELQUES OBSERVATIONS SUR LA COMPOSITION PAR AMALGAME EN FRANÇAIS ACTUEL ISSUE DU PETIT ROBERT

RADKA MUDROCHOVÁ
Université Charles de Prague, Prague, Tchéquie

MUDROCHOVÁ, Radka: Some observations on the composition by blending in contemporary French from Petit Robert. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 3, pp. 906 – 915.

Abstract: The objective of this article is to analyze the composition by amalgam in current French by focusing on the one hand on the notion of amalgam in linguistics and on the other hand on the use and the frequency of use of the chosen amalgams in the diatopic variation of French. The notion of amalgam and / or of portmanteau word does not seem obvious and the explanations or definitions offered by dictionaries as well as by works on lexicology are not unanimous and differ from one another. Before presenting the results of a more detailed research, we therefore find it essential to frame the contribution in a theoretical context dealing with the notion of amalgam, or even of portmanteau word, which allows us to better understand the whole problem.

Key words: blending, French language, lexicography, loanword, diatopy

0. INTRODUCTION

La composition par amalgame représente un procédé de formation de mots qui couvre notamment l'ancienne dénomination de mot-valise (Sablayrolles, 2017, p. 56). Néanmoins, dans le titre du présent article, nous avons décidé d'employer le terme *amalgame* car il nous apparaît plus adéquat pour sa neutralité et son concept plus large, bien que le linguiste Vincent Renner (2015, p. 98), consacrant beaucoup de ses recherches au procédé en question, souligne « qu'en français le terme le plus communément utilisé pour dénommer le produit de l'opération morphologique est *mot-valise* » et que « le terme *mot-valise* a les faveurs des *francisants*, mais les anglicistes parlent, eux, plus volontiers *d'amalgame* » (Renner, 2006, p. 98).

1. NOTION D'AMALGAME

Les dictionnaires de langue générale ne comportent en aucun cas la notion d'amalgame et ne se limitent qu'aux mots-valises définis ainsi : (1) le ***Petit Robert*** : « mot composé de morceaux non signifiants de deux ou plusieurs mots (ex. *motel, cultivar, progiciel*) », le mot a été créé en 1956 par un calque de l'anglais *portmanteauword* ; (2) ***Larousse*** (en ligne) : « mot résultant de la réduction d'une suite de

mots à un seul mot, qui ne conserve que la partie initiale du premier mot et la partie finale du dernier (par exemple *franglais*) » ; ***Trésor de la langue française*** (TLFi) : une « création verbale formée par le télescopage de deux (ou trois) mots existant dans la langue » et cite un synonyme, celui de « mot porte-manteau ». Ces définitions sont souvent insuffisantes et ne permettent pas de discerner d'autres procédés de composition d'amalgames.

En revanche, parmi les différentes théories et études portant sur les amalgames lexicaux, nous pouvons observer une diversité terminologique ainsi que typologique. Pour ce qui est de la terminologie, elle oscille en français notamment entre *l'amalgame*, dont l'emploi est influencé par la terminologie anglaise, le *mot-valise*, celui-ci plus traditionnel, et nous trouvons aussi la notion de *télescopage*. À part celles-ci, il existe d'autres dénominations, désignant ce procédé de formation spécifique ou son fruit, certaines évoquées par J.-F. Sablayrolles (2000, p. 224) : *mot porte-manteau* (Carroll, Riffaterre), *mot-centaure* (Le Bidois), *croisement*, (Pei, Gaynor), *mot-tiroir, mot-gigogne, emboîtement* (Jakobson), *mot sandwich* (cité par Sablayrolles 2015, p. 187, employé par Ferdière en 1964, par ailleurs comme *mots-maux bile* dans le même article), d'autres complétées par A. Léturgie (2011, p. 77) ou V. Renner[1] (2006, p. 9 ; 2015, p. 98) : *mot-valisation/mot-valisage* (Fradin, Montermini et Plénat 2009), *valisage* (Bonhomme, 2009)[2], *amalgamation* (Renner, 2008, Léturgie, 2012), *processus de construction par association et troncation* (Bassac, 2004), *brachygraphie gigogne* (Clas[3], 1987), *compocation* (Cusin-Berche, 1999), *mixonymes/mixonymie* (Pottier, 1987, 1992, 2001), *imbrication* (Grésillon, 1984) et *la polygraphie des portmanteaus*. Certains linguistes parlent aussi d'*acronymie* (Guilbert, 1975 ; Koucourek, 1991 ; Cabré, 2006). J. Milner (1982) introduit le terme de *monstres de langue* en analysant des plaisanteries sur la langue, les mots-valises y compris. J. Chaurand (1977, p. 5) évoque, et critique d'ailleurs, le terme de *contamination*. En outre, la linguiste allemande, Cornelia Friedrich (2008, p. 21), propose, elle aussi, certains termes : *mélange, mot-tandem, mot-tiroir* (Morier, 1961), *compromot* (*compromis* + *mot*, dans l'étude de Dierickx, 1966), ou *bloconyme* (Dupriez, 1980). Madueke (2013, p. 45–46) et cite encore d'autres synonymes : *brunch-word, amalgrammes, mot articulé* (terme proposé par R. Galisson d'après la logique de l'existence des mots dérivés et des mots composés) ou *mots composés exocentriques*. La liste peut être achevée (?) par le répertoire cité par A. Grésillon (1984, p. 6) :

[1] V. Renner (2015, p. 99) résume également la terminologie employée en anglais, celle-ci peut être complétée par l'étude de C. Friedrich (2008, p. 21).

[2] Entre guillemets, M. Bonhomme (2009) emploie aussi la désignation : *contaminations* et dans l'article en question, on parle à plusieurs reprises de *mixage*.

[3] Dans l'introduction de son article, A. Clas (1987, p. 347) évoque, lui aussi, la riche terminologie de la langue française : *hapaxépie, haplolagie, haplologie, acronymie, crase, paronomase, croisement, amalgame, télescopage, emboîtement, mot valise, mot centaure, mot gigogne, mot contaminé, mot fusionné, mot portemanteau*.

druses (Stuchlik et Bobon, 1960), *mots fermentés* (Butor, 1962), *mots sauvages* (Rheims, 1969), ou par les créations de Moncelet : *bête-à-deux-mots, mots a(i) mants, mots croasés, mots-valistes* (1972, 1978, 1981).

Pour ce qui est de la typologie des amalgames, elle est riche, elle aussi. Néanmoins, pour les besoins de la présente contribution, nous aimerions souligner la dernière proposition de classification effectuée par Jean-François Sablayrolles (2019, p. 127). Il comprend la composition par amalgame comme une matrice externe faisant partie des procédés morpho-sémantiques, qui comporte trois sous-groupes distinctifs : factorisation, substitution et mot-valisation. D'après Sablayrolles, il y a d'autres procédés de composition qu'il faut distinguer de la composition dite « classique », il s'agit notamment de fractocomposition et de compocation. Pour expliquer, la **fractocomposition** se distingue de la composition « classique » par le fait qu'un « des éléments constitutifs n'est pas un mot complet, mais un fragment de celui-ci, qui vaut sémantique, pour l'ensemble » (*téléspectateur, écovignette*). En revanche, la **compocation** (terme né par compo[sé] + [tron]cation et créé par Fabienne Cusin-Berche), par ailleurs très répandu en français de nos jours, permet de créer « des composés avec deux fractolexèmes », le plus fréquemment l'apocope du premier membre à l'aphérèse du second (*tradismatique* = *tradition* + *charismatique*). Quant à la **factorisation**, une invention de Julie Makri-Morel, elle désire « restreindre la mot-valisation aux seuls cas où le segment homophone se situe à l'intersection des deux mots combinés » (*automo**débile*** = *automo**bile*** + *dé**bile***). Le dernier terme, le **mot-valise** qui recouvre traditionnellement tous les procédés d'amalgame, est selon Sablayrolles le cas où le mot se replie sur lui-même autour d'un axe : une ou des syllabes communes (au moins une voyelle) avec une superposition syllabique (*gangsterrorisme* = *gangs**ter*** + ***ter**rorisme*).

2. AMALGAMES ISSUS DU *PETIT ROBERT*

Même si nous avons expliqué et développé le concept d'amalgame dans le milieu linguistique français tout en expliquant le choix terminologique, pour une analyse du phénomène étudié dans le dictionnaire le *Petit Robert* (version électronique payante de 2021 disponible via https://www.lerobert.com), nous avons dû passé par la requête « mot-valise » dans la recherche avancée (recherche effectuée le 21.6.2020) pour recevoir l'échantillon de lexies souhaitées. Ainsi, nous avons obtenu 61 résultats que nous avons analysés plus en détail en nous concentrant sur : les parties du discours ; la datation, voire la première attestation des lexies dans le dictionnaire ; nombre d'anglicismes qui font partie de cette catégorie et des questions particulières comme le nombre de marques déposées, des recommandations officielles ou des régionalismes.

2.1 Parties du discours des amalgames

Premièrement, nous avons observé les parties du discours des amalgames issus du PR, les résultats sont résumés par le Tableau 1.

Partie du discours	Nombre	Pourcentage
adjectif	1	2 %
verbe	4	6 %
nom + adjectif	5	8 %
nom	6	10 %
nom féminin	16	26 %
nom masculin	29	48 %

Tab. 1. Partie du discours des amalgames issus du PR

En observant le tableau, nous remarquons que les noms sont en majorité, avec 92 % au total. Nous n'avons remarqué que quatre verbes et un seul adjectif si l'on ne prend pas en considération les cas où une lexie peut représenter deux parties du discours, un nom d'une part et un adjectif d'autre part, cette catégorie comprend 8 %.

2.2 Datations des amalgames dans le PR

Deuxièmement, pour ce qui est des datations (illustrées par la Figure 1 *infra*), la tranche la plus représentée est celle qui couvre les dates entre 1980 et 2000. La lexie amalgamée la plus ancienne est celle de *tangélo*, un « agrume de la taille d'une grosse orange, issu du croisement du pomélo et de la mandarine », datée de 1906. En revanche, la lexie *souplex* est la plus récente de notre corpus, créée en 2010 par l'assemblage des mots *sous-sol* et *duplex* désignant « appartement sur plusieurs niveaux dont un en sous-sol ».

2.3 Origine des amalgames

Troisièmement, nous avons analysé l'étymologie des amalgames en nous focalisant sur leur origine. Nous pouvons constater que 56 % de lexies de notre corpus comportent une note renvoyant à l'anglais. En majorité, avec 61 % de cas, il y a une simple remarque, « d'après l'anglais », « de l'anglais » ou « mot-valise anglais », à titre d'exemple : *chatbot, globish, flexitarien, crossfit, burkini, mook, hackathon, nomophobie*, etc. La tranche de 23 % est représentée par la mention « anglais américain » (p. ex. : *sitcom, podcast, télévangéliste, fanzine*), suivie par 13 % comprenant des calques de l'anglais (p. ex. : *permalien, permaculture, rançongiciel*). La dernière catégorie, moins importante, avec 3 %, concerne des traductions (terminologie employée par le PR), : *gratuiciel* : « mot-valise, de *gratuit* et *logiciel,* pour traduire l'anglais *freeware* ».

2.4 Questions particulières – variété diatopique, marque déposée, recommandations officielles

Finalement, nous avons traité d'autres questions plus particulières, notamment les observations concernant la variété diatopique, des recommandations officielles de certains termes empruntés à l'anglais et la représentation des marques déposées dans l'ensemble du corpus.

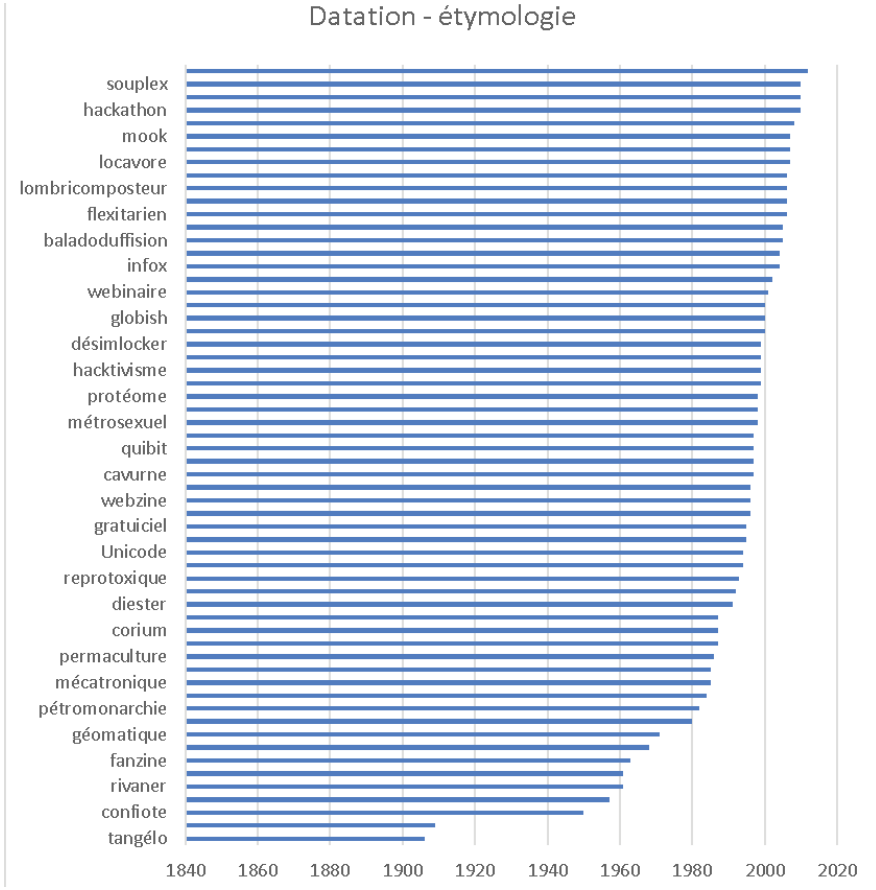


Fig. 1. Datation des amalgames dans le PR

Pour une définition de la marque déposée, le PR renvoie à l'Institut national de la propriété industrielle en ajoutant une simple définition : « Enregistrement, protection, propriété d'une marque. Marque de distributeur ». Néanmoins, la problématique du concept semble plus complexe notamment en l'occurrence avec d'autres termes qui y sont liés, à savoir : nom déposé, produit déposé, etc. qui ont fait, par ailleurs, l'objet de plusieurs études antérieures (cf. par exemple les travaux de Altmanova).

Pour ce qui est de notre corpus, les marques déposées, dans la ligne terminologique du PR, représentent 10 % : *globish, crossfit, burkini, diester, bisounours, bistronomie* et une lexie, celle de *mook*, est traitée par le dictionnaire comme nom déposé.

Quant aux questions diatopiques, le Canada a été évoqué par le dictionnaire dans 10 % de notre échantillon de lexies, à savoir : *gratuiciel* (« mot-valise, de gratuit et logiciel, pour traduire l'anglais freeware »), *enfirouaper* (« de enfifrer 'sodomiser, duper',

de fifre 'pénis', et rouâper, variante québécoise de râper »), *clavarder* (« de clavier et bavarder »), *pourriel* (« de poubelle et courriel »), *baladodiffusion* (« baladeur et diffusion, diffusion de fichiers audios et vidéos téléchargés sur un baladeur numérique à partir d'Internet »), *divulgâcher* (« recommandé en français du Canada » pour remplacer le verbe *spoiler*).

Les recommandations officielles, représentées par les amalgames, apparaissaient dans trois cas du corpus étudié, le premier lexème est celui de la recommandation officielle telle qu'elle est attestée par le PR (si elle provient des instances canadiennes, le terme est suivi par l'abréviation CA), le second représente le mot d'origine, un anglicisme pour les trois situations : *infox --- fake news, baladodiffusion (CA) --- podcast, divulgâcher --- spoiler.* En outre, la lexie *fake news* connaît deux autres formes recommandées, d'après les renseignements dans le *Grand dictionnaire terminologique* et sur *FranceTerme* (en ligne) : « En France, la Commission d'enrichissement de la langue française a publié une recommandation d'usage qui préconise l'emploi, entre autres, des termes *information fallacieuse*, *infox* ou *fausse nouvelle* pour désigner ce concept. ».[4] Pour ce qui est du verbe *divulgâcher*, il a été proposé en tant qu'équivalent officiel en 2016 par l'Office québécois de la langue française et a été validé par des instances françaises en 2020, publié dans le Journal officiel le 23.05.2020. Néanmoins, le dictionnaire PR marque toujours sa spécificité territoriale liée au Québec. Un autre équivalent qui n'est pas évoqué par le PR est celui de *audio à la demande* inclus dans la plateforme *FranceTerme* depuis le 23.5.2020.

L'objectif de l'analyse suivante a été de déterminer si les lexies d'origine et leurs équivalents officiels (souvent des amalgames) circulent de la même façon sur les deux territoires francophones, la France et le Canada francophone. C'est la raison pour laquelle nous avons effectué une recherche (27.1.2021) dans plusieurs outils linguistiques qui permettent de mesurer la fréquence d'emploi des lexies choisies. Ainsi, nous avons utilisé :

(1) des corpus *Aranea* (pour en savoir plus cf. notamment : Benko, 2018 ; Benko, Butašová, Puchovská, 2019 ; Nádvorníková, 2020), susceptibles de suivre les deux variétés (et même plus) du français, à savoir : Araneum Francogallicum Gallicum Maius (French French, 20.05) 1.25 G, abrégé par la suite en AFM, Araneum Francogallicum Canadiense II (Canadian French, 20.05) 406 M, abrégé en AFC ;
(2) les archives de presse d'un journal pour chaque territoire, disponibles gratuitement sur la Toile : 20minutes.fr, lapresse.ca ;
(3) la plateforme *Néoveille*, plateforme de repérage, analyse et suivi des néologismes en sept langues (cf. notamment : Cartier, 2016, 2018) qui permet de visualiser des emplois sur les différents territoires francophones, le Canada y compris.

Le Tableau 2 (*infra*) résume la recherche effectuée dans les trois ressources.

[4] http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26542775, (consulté le 31 décembre 2020).

Lexie	AFM	AFC	20minutes.fr	Lapresse.ca	Néoveille FR/CA
fake news	**1164**	**39**	**3240**	**33 600**	**889**/44
infox	57	3	310	34	65/3
fausse nouvelle	144	**37**	116	483	698/**208**
information fallacieuse	5	2	5	1	12/0
podcast	**3426**	381	**17 900**	**812 000**	**950**/32
baladodiffusion	378	**724**	10	242	6/**107**
audio à la demande	2	0	0	0	4/1
spoiler	**1730**	**165**	**3450**	6	**980**/3
divulgâcher	5	0	35	**517**	26/**6**

Tab. 2. Répartition des fréquences d'emploi des lexies étudiées dans les outils choisis

En observant le tableau, nous pouvons constater que le français de France affiche à chaque fois le nombre d'occurrences plus élevé pour les lexies d'origine, donc les anglicismes. En revanche, le français québécois diffère. Bien que l'emprunt *fake news* possède plus de résultats dans le corpus AFC ainsi que dans les archives de presse de *Lapresse.ca*, la plateforme *Néoveille* atteste plus d'occurrences pour l'équivalent *fausse nouvelle* (208) contre *fake news* où il y en a 44. En outre, le corpus AFC retient 39 occurrences de la lexie *fake news* contre 37 occ. pour la *fausse nouvelle*, donc une différence plutôt négligeable. Il est assez surprenant que la recommandation *information fallacieuse* circule très peu et le seul résultat retrouvé dans les archives *Lapresse.ca* renvoie à la recommandation elle-même (cf. l'extrait 1) :

> L'ex. 1 : « Au cœur de la polémique : la définition même d'une « fake news ». À cette expression, la Commission d'enrichissement de la langue française (CELF) préfère les termes « information fallacieuse » ou le néologisme « infox ». »[5]

Pour ce qui est du couple de lexies *podcast/baladodiffusion*, il suit en effet la même tendance. La lexie le *podcast* est majoritairement employée en France, tandis que la *baladodiffusion* montre plus d'occurrences dans les outils analysant le français canadien, sauf les archives de presse *Lapresse.ca* où le *podcast* semble, d'après les résultats de la recherche, plus fréquent. En effet, la *baladodiffusion* apparaît également dans des publicités en tant que terme officiel, cf. la Figure 2 *infra* reprise du site journaldequebec.com.

Fig. 2. Témoignage de l'emploi de la lexie baladodiffusion sur les sites officiels

[5] https://www.lapresse.ca/international/europe/201811/20/01-5204947-le-parlement-francais-adopte-les-lois-controversees-anti-fausses-informations.php, (consulté le 23 octobre 2020).

En revanche, l'emploi du terme *baladodiffusion* se limite en France aux cas où l'on évoque la situation au Québec, voire au Canada (cf. l'extrait 2), ou l'usage du terme est suivi de marques lexicographiques, des guillemets ou des italiques, ou les deux (cf. l'extrait 3). En plus, il faut souligner que le texte dans lequel apparaît la recommandation en question, opère avec le terme d'origine, le *podcast*, sans qu'il soit suivi de marquages, cf. l'extrait *infra*.

> L'ex. 2 : « Nous avons de bonnes surprises : "baladodiffusion" a remplacé "podcast", même chez Apple, mais pas chez vous. »[6]

> L'ex. 3 : « New York - Ils ne pesaient quasiment rien économiquement il y a encore trois ans, mais les **podcasts**, devenus un produit de consommation de masse aux Etats-Unis, génèrent désormais des centaines de millions de dollars de revenus, et ce n'est qu'un début […]. Créée il y a à peine 15 ans, la '***baladodiffusion***' est déjà entrée dans la vie de 124 millions d'Américains, […]. »[7]

Le dernier amalgame étudié, celui de *divulgâcher*, est de nouveau plus répandu, d'après la statistique évoquée par le Tableau 2, au Canada qu'en France, même si nous n'avons pas pris en compte l'emploi de la lexie *spoiler*, qui peut inclure non seulement un verbe mais également un nom. Tout comme dans le cas de l'équivalent *baladodiffusion*, la lexie *divulgâcher* apparaît dans le contexte français pour parler notamment de la situation sur le territoire francophone transatlantique comme en témoignent des exemples *infra* (cf. Figure 3) tirés de la plateforme *Néoveille*.

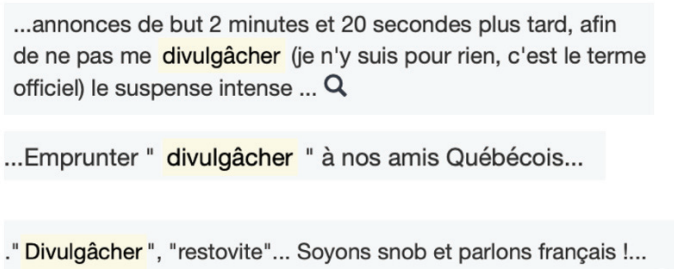


Fig. 3. Contexte du verbe divulgâcher (source : Néoveille)

3. CONCLUSION

Notre recherche a montré que le concept d'amalgame n'est pas unanime dans le milieu linguistique français. Les dictionnaires de langue générale du français préfèrent le terme de « mot-valise » pour y inclure tous les procédés de formation

[6] https://www.la-croix.com/Monde/Ameriques/LOQLF-gardien-gendarme-francais-Quebec-2017-08-03-1200867449, (consulté le 3 août 2020).

[7] https://lexpansion.lexpress.fr/actualites/1/actualite-economique/les-podcasts-nouvel-eldorado-avec-des-centaines-de-millions-de-dollars-a-la-clef_2030503.html, (consulté le 12 septembre 2020).

d'amalgames distingués par Sablayrolles. Néanmoins, pour les besoins de notre étude, nous avons emprunté le terme générique d'« amalgame » pour l'analyser plus en détail suite aux données reçues par le dictionnaire le *Petit Robert*.

Ainsi, nous pouvons constater que dans la formation d'amalgames, ce sont les noms qui prédominent. En outre, les anglicismes représentent également un groupe assez nombreux (56 % du corpus). Les marques déposées apparaissent parfois dans le contexte des amalgames et 10 % de lexies amalgamées provenaient de l'usage canadien. En comparant les recommandations officielles, représentées par des amalgames, avec leurs équivalents d'origine, des emprunts à l'anglais, nous notons que l'emploi des recommandations officielles est plus fréquemment attesté dans les corpus canadiens. Néanmoins, certaines recommandations sont intégrées plus tardivement dans la norme française, cf. le verbe *divulgâcher*, proposé en 2016 par l'Office québécois de la langue française et repris par le Journal officiel en 2020 en France.

Remerciements :
Le présent article s'inscrit dans le Projet Européen du Développement Régional « Créativité et adaptabilité comme conditions du succès de l'Europe dans un monde interconnecté » (No. CZ.02.1.01/0.0/0.0/16_019/0000734) et a bénéficié du soutien du projet de l'Université Charles « Progres Q10 : Le langage dans les changements de temps, d'espace et de culture ».

Bibliographie

ALTMANOVA, Jana : Du nom déposé au nom commun. Néologie et lexicologie en discours. Milano : EDUcatt, 2013.

BENKO, Vladimír – BUTAŠOVÁ, Anna – PUCHOVSKÁ Zuzana (Eds.): Webové korpusy Aranea. Učebnica pre učiteľov jazykov, prekladateľov, tlmočníkov, filológov a študentov filologických odborov. Bratislava : Univerzita Komenského, 2019.

BENKO, Vladimír : "Aranea: A Family of Comparable Gigaword Web Corpora". In : Web Corpora & Corpus Linguistics Portal. Bratislava, Slovak Academy of Scientces Ľ. Štúr Institute of Linguistics, 2018. Disponible sur : http://aranea.juls.savba.sk/aranea_about/index.html, (consulté le 21.01.2020).

BONHOMME, Marc : Mot-valise et remodelage des frontières lexicales. In : Cahiers de praxématique, 2009, No 53, pp. 99–120. Disponible sur : http://journals.openedition.org/praxematique/1091, (consulté le 27 mars 2020).

CARTIER, Emmanuel : Emprunts en français contemporain : étude linguistique et statistique à partir de la plateforme Néoveille. In : L'emprunt en question(s). Conceptions, réceptions, traitements lexicographiques. Eds. A. Kacprzak – R. Mudrochová – J.-F. Sablayrolles. Limoges : Lambert Lucas 2020, pp. 145–185.

CARTIER, Emmanuel : Neoveille, système de repérage et de suivi des néologismes en sept langues. In : Neologica, 2016, No 10, pp. 101–131.

CHAURAND, Jacques : Des croisements aux mots-valises. In : Le Français moderne, 1977, No 45, pp. 4–15.

CLAS, André : Une matrice terminologique universelle : la brachygraphie gigogne. In : Meta, 1987, Vol. 32, No 3, pp. 347–355.

FRIEDRICH, Cornelia : Kontamination – Zur Form und Funktion eines Wortbildungstyps im Deutschen. Thèse de doctorat, Erlangen-Nürnberg : Friedrich-Alexander-Universität, 2008.

GRÉSILLON, Almuth : La règle et le monstre : le mot-valise. Interrogations sur la langue, à partir d'un corpus de Heinrich Heine. Tübingen : Max Niemeyer Verlag, 1984.

LÉTURGIE, Antoine : Un cas d'extragrammaticalité particulier : les amalgames lexicaux fantaisistes. In : Linguistica, 2011, Vol. 51, No 1.

MADUEKE, Ijeoma Chidinma Sylvia : L'amalgamation lexicale dans un corpus spécialisé : analyse morphologique. Mémoire de Master, Saskatchewan : Université de Regina, 2013.

NÁDVORNÍKOVÁ, Olga : The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology. In : Linguistica Pragensia, 2020, Vol. 30, No 1, pp. 30–50. Disponible sur : https://dspace.cuni.cz/bitstream/handle/20.500.11956/117137/Olga_Nadvornikova_30-50.pdf?sequence=1&isAllowed=y, (consulté le 23 juillet 2020).

RENNER, Vincent : Les composés coordinatifs en anglais contemporain. Thèse de doctorat, Université Lumière-Lyon 2, 2006.

RENNER, Vincent : Panorama rétro-prospectif des études amalgamatives. In : Neologica, 2015, No 9, pp. 97–112.

REY, Alain : Le Petit Robert de la langue française. Paris : Éditions le Petit Robert, 2020. Disponible sur : https://petitrobert.lerobert.com/robert.asp, (consulté le 22 janvier 2021).

SABLAYROLLES, Jean-François : Comprendre la néologie. Conceptions, analyses, emplois. Limoges : Lambert Lucas, 2019.

SABLAYROLLES, Jean-François : La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes. Paris : Champion, 2000.

SABLAYROLLES, Jean-François : Quelques remarques sur une typologie des néologismes : Amalgamation ou téléscopage : un processus aux productions variées (mots valises, détournements…) et un tableau hiérarchisé des matrices. In : Neologia das linguas romanicas. Eds. I. M. Alves – E. S. Pereira. Sâo Paulo : Humanitas, 2015, pp. 187–218.

Sitographie

20minutes.fr, Disponible sur : https://www.20minutes.fr, (consulté le 2 novembre 2020).

Aranea, Disponible sur : http://unesco.uniba.sk/aranea/, (consulté le 2 novembre 2020).

FranceTerme, Disponible sur : http://www.culture.fr/franceterme, (consulté le 22 janvier 2021).

Grand dictionnaire terminologique, Disponible sur : http://gdt.oqlf.gouv.qc.ca, (consulté le 22 janvier 2021).

Lapresse.ca, Disponible sur : https://www.lapresse.ca, (consulté le 2 novembre 2020).

Larousse, Disponible sur : https://www.larousse.fr, (consulté le 21 octobre 2020).

Néoveille, Disponible sur : https://tal.lipn.univ-paris13.fr/neoveille/html/login.php?action=login, (consulté le 2 novembre 2020).

Trésor de la langue française informatisé, Disponible sur : http://atilf.atilf.fr, (consulté le 21 octobre 2020).

sciendo
DOI 10.2478/jazcas-2022-0016

L'EXTRACTION DE TERMES-CLÉS DE LA PÊCHE À L'AIDE D'OUTILS GNU/LINUX

AGNIESZKA K. KALISKA
Université Adam Mickiewicz, Poznań, Pologne

KALISKA Agnieszka K.: Extracting fishing terminology using GNU/Linux tools. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 916 – 926.

Abstract: The technological revolution that has occurred in recent decades has made accessible for researches large textual data collections. At the same time, the development of increasingly sophisticated computer tools provides them with new methods of analyzing texts. In the present study however we examine the functionalities offered by traditional tools, namely GNU/Linux tools, easily accessible via the command line but still unknown among linguists with little or no computer knowledge. Our goal is to show how using the web corpus on the one hand and the processing GNU/Linux tools on the other, we can extract key-terms of fishing jargon.

Key words: web corpus, GNU/Linux tools, key-term, fishing terminology.

1. INTRODUCTION

La révolution technologique qui s'est produite ces dernières décennies a rendu accessibles des collections de données textuelles de grande taille, parmi lesquelles le corpus web occupe une place spéciale. Parallèlement, le développement d'outils informatiques de plus en plus sophistiqués, souvent en accès libre, ouvre aux chercheurs de nouvelles voies d'exploitation des textes. Dans la présente étude nous nous proposons néanmoins d'examiner les fonctionnalités que nous offrent des outils traditionnels, à savoir les outils GNU/Linux accessibles pour tout utilisateur du système Linux via la ligne de commande. Accessibles, certes, ces outils restent cependant peu connus parmi les linguistes n'ayant aucune ou très peu de connaissance de l'informatique. Notre objectif est de montrer comment grâce au corpus web, d'un côté, et au traitement exécuté à l'aide d'outils GNU/Linux de l'autre, nous pouvons approfondir notre connaissance du vocabulaire spécialisé. Dans cette recherche nous utiliserons le corpus OSCAR (Open Super-large Crawled ALMAnaCH coRpus), le plus grand corpus multilingue disponible librement (Ortiz – Sagot – Romary, 2019).

Par *pêche* nous entendons la pêche *récréative*, dite aussi *pêche de loisir*, *récréationnelle* ou *plaisancière*. Nous commencerons par une brève caractéristique du jargon de la pêche et des méthodes de collecte de données lexicales, à commencer par des enquêtes traditionnelles, une recherche sur le terrain, jusqu'au traitement des données textuelles de grande taille. L'étape suivante de notre étude consistera à présenter les données lexicales collectées. Nous finirons par quelques conclusions concernant les avantages et les inconvénients de la méthode employée.

2. CARACTÉRISTIQUE DU JARGON DE LA PÊCHE

Nous nous souscrivons à la définition du *jargon* proposée par D. François-Geiger (1988) selon qui il s'agit d' « un parler technique qui peut être ésotérique pour le profane, mais dont la fin n'est pas de masquer l'objet du discours ; elle est, au contraire, d'en rendre l'expression plus rigoureuse, plus spécifique ». Les mots de la pêche servent ainsi à *nommer* – de manière juste, brève et précise (cf. aussi : Sourdot, 2011, p. 1) – les outils et méthodes employés pour attraper le poisson.

La deuxième fonction, non moins importante, est *communicative*. Déjà le fait que *pêcheur* ne désigne dans le cadre de la présente recherche ni un métier ni une formation diplômante (car il s'agit d'une activité récréative, non commerciale) rend la pêche accessible pour tous à toute étape de la vie. Le *savoir-pêcher* se transmet ainsi de manière informelle, d'une génération à l'autre, en famille et entre amis, quelque soient l'âge, le sexe, l'expérience professionnelle ou encore la situation sociale de nouveaux amateurs. Dans ces échanges des forums de discussion et des blogs sont une source d'informations précieuse, permettant à un débutant de retracer le parcours nécessaire pour atteindre son objectif. Ainsi le jargon de la pêche permet-il non seulement de transmettre les connaissances mais aussi d'entrer en contact, de dialoguer avec les autres et de nouer des liens durables. Il assume alors la troisième fonction – ie. *identitaire* – car l'échange est d'importance cruciale pour assurer le lien social.

Le dernier trait caractéristique du vocabulaire de la pêche réside dans la dimension ludique de celui-ci (cf. Sourdot, 2018 ; Kaliska, 2018). C'est sur le web qui est un lieu d'échange libre de connaissances que nous pouvons observer l'expressivité et la spontanéité du jargon de la pêche, et très probablement les premières traces écrites de désignations futures. Il ne s'agit donc pas d'un strict technolecte. Notons ce qu'annoncent les en-têtes de différents sites privés où on lit que la pêche est « un art » ou « une façon de vivre ». Ils témoignent qu'il s'agit pour leurs auteurs d'un véritable dévouement qui trouvera son reflet dans la langue utilisée : pleine d'humour, imagée et technique à la fois.

3. MÉTHODES DE COLLECTE DE DONNÉES LEXICALES

Pour un linguiste non pêcheur l'accès à ce vocabulaire n'est possible que par une des voies suivantes : la première étant un travail de terrain peut consister à distribuer, collecter et interpréter les enquêtes, ce qui est une méthode de travail plutôt coûteuse et insuffisante car, au final, les entretiens, voire les observations directes, permettant de résoudre des questions complexes peuvent s'avérer indispensables. L'efficacité de ces derniers réside justement dans la possibilité d'obtenir une image authentique du domaine dans son environnement naturel, sans que cela signifie pour autant qu'on n'aura pas de difficultés épistémologiques ou méthodologiques à surmonter.[1] Deux méthodes suivantes

[1] V. Labov, 1972 (pour le paradoxe de l'observateur) et Blanchet, 2012 (pour l'observation participante).

consisteraient à explorer soit les éditions papier (p.ex.: la presse de spécialité) soit le web (p.ex.: les blogs). Ici, comme pour les enquêtes, les observations de près des actions et des interactions *in situ* peuvent aider à comprendre mieux le matériel analysé.

Les méthodes de collecte de données citées ne s'excluent pas et peuvent être complémentaires. La question se pose néanmoins quel rôle attribuer à ces différentes sources ? Illustratif ou heuristique ? La question est d'autant plus complexe, si l'on se limite – comme c'est le cas de notre étude – à la dernière des sources citées, notamment le web, c.-à-d. l'ensemble des productions linguistiques sous forme de documents numériques librement consultables de chez soi par le biais d'un moteur de recherche, *en ligne*. Or, ce même ensemble devient – grâce aux projets tels que Common Crawl ou OSCAR (Ortiz – Sagot – Romary, 2019) – accessible aussi en forme brute (*plain text*).[2]

En même temps, le développement récent d'outils informatiques mis à la disposition de linguistes leur a permis d'accélérer dans leur travail : trouver et calculer des (co-) occurrences de mots sont-ils devenus plus faciles à exécuter dès que des outils spécifiquement développés pour ce type de tâches sont devenus disponibles (p.ex. : MeWeX, LEM et d'autres outils élaborés pour l'analyse des textes polonais par le consortium Clarin-PL, cf. p.ex. : Piasecki, 2014 ; cf. aussi : Drouin, 2015). Dans la présente étude nous nous proposons néanmoins d'examiner les fonctionnalités que nous offrent les outils traditionnels, à savoir les outils GNU/Linux, accessibles pour tout utilisateur du système Linux via la ligne de commande, tels que *cat* (et ses variantes, p.ex. : *zcat*), *grep* (et ses variantes, p.ex. : *egrep*), *sed*, *sort*, *tr*, *uniq*, *wc* (cf. p.ex.: Shotts, 2012), auxquelles s'ajoutent les expressions dites *rationnelles* (parfois appelées *normales*, *régulières* ou, tout court, les *motifs*, dites encore les *regex* par la compocation des mots anglais *regular* et *expression*). Les expressions régulières sont un outil très puissant nous permettant d'accéder à des ensembles de chaînes de caractères (séquences de texte) prédéfinies selon une connaissance préalable de structures dont on abstrait les caractéristiques les plus saillantes pour construire le corps du motif. À l'aide des programmes en ligne de commande nous pouvons ensuite extraire ce qui nous intéresse, le trier, dédupliquer, modifier, compter et numéroter, de manière à obtenir des résultats que l'on peut ensuite soumettre au dépouillement terminologique manuel.

4. WEB EN TANT QUE CORPUS

Le corpus web est parfois défini par la formule de trois *V* : *volume*, *vélocité* et *variété*. Ses données sont massives, hétérogènes (textes, images et sons) et se renouvellent en permanence (cf. Longhi, 2017 ; *cf*. aussi : Fouqueré – Isaac, 2003, p. 115 pour *volatilité* de données). La collecte automatisée de ces différentes données produites massivement est possible grâce aux pratiques de crawling (indexation des pages web) et de scraping (extraction d'informations issues d'Internet). Telle est, notons-le, la vocation de l'organisation Common Crawl : explorer les pages web et

[2] Un tel corpus fourni sans métadonnées est, en principe, destiné à être utilisé dans le TAL.

fournir au public, gratuitement, ses archives sous forme d'ensembles de textes. Le corpus OSCAR, coordonné par l'équipe ALMAnaCH dans le centre de recherche Inria de Paris, est un ensemble de textes multilingue obtenu par filtrage du corpus Common Crawl ; il est à télécharger librement sur le site du projet.

Les données issues d'Internet sont ainsi stockées sous forme de fichiers texte brut. La première caractéristique d'un tel ensemble est qu'il est gros, varié et bruité. On y trouvera non seulement toute une variété de textes et de registres mais aussi des pages entières de données qui pour un linguiste lexicographe sont plutôt de moindre qualité. Ce seront, par exemple, des énumérations de chiffres et de signes spéciaux qui, dans un ensemble de textes battus (*shuffled*), ne disent pratiquement rien.

D'autres caractéristiques peuvent découler de la spécificité du discours électronique médié qui a des traits à la fois oraux et écrits. Y sont présents à différents et pas toujours les mêmes degrés (p.ex. chat vs. blog ; cf. Panckhurst, 2007) des phénomènes comme écritures hors norme, absence ou emploi excessif de certaines formes grammaticales (cf. p.ex.: Amitay, 1999, cité après Fouqueré – Issac, 2003, p. 116), fautes de frappe, omission spontanée ou systématique de signes diacritiques, fautes d'orthographe dues à l'homophonie des formes ou à l'écriture phonétique.

Les traits cités posent des difficultés de traitement, certes, mais il paraît que la plus grande difficulté consiste à repérer dans un corpus aussi immense qu'OSCAR (23,206,776,649 mots graphiques) les structures utiles pour la recherche. Pour ce faire, il faut avoir une connaissance préalable de mots clés et de structures potentiellement intéressantes et sur cette base construire un ou des motifs utiles pour la recherche plus approfondie de termes-clés (termes-candidats).

5. DÉPOUILLEMENT TERMINOLOGIQUE

5.1 Sous-corpus de la pêche

La constitution d'un corpus cohérent à partir de données aussi variées que le web n'étant pas facile, nous nous sommes servis d'un mot clé *pêche* afin de construire sur sa base un motif – à savoir : `(^|\s)\b(p|P)êch\w+\b` – qui nous a permis d'accéder à 1,312,157 lignes contenant la séquence *pêch(e)-* (en début du mot pour distinguer entre *pêche*, *pêcheur* et *pêcher*..., d'un côté, et *empêcher* et *dépêcher*... de l'autre), ce qui constitue env. 0,28% du corpus entier (461,343,098 lignes).[3]

[3] Tout fichier texte comporte un certain nombre de caractères, à savoir : caractères imprimables, espaces et sauts de ligne. Ces derniers sont des caractères de contrôle indiquant le passage à la ligne suivante (*line feed*). Les limites d'une ligne ne sont donc pas ce qu'on voit aux frontières d'une fenêtre d'affichage. Dans notre sous-corpus français, la longueur moyenne d'une ligne égale à 656 caractères (sans compter les sauts de ligne), tandis que la ligne la plus longue s'y compose de 950,890 caractères et la plus courte en a une centaine. Encore que les lignes doubles aient été éliminées avant que le corpus eût été mis en ligne par ses créateurs, certaines se répètent – ce sont des lignes qui ne diffèrent, par exemple, que d'un caractère ou d'un mot, il fut donc impossible de les détecter à l'étape du filtrage.

Bien que le motif que nous avons proposé nous ait permis d'accéder relativement vite à un nombre considérable d'emplois, on n'a pas pu à cette étape-là contrer les difficultés qu'occasionne l'homonymie lexicale. En effet, il y a *pêche* qui signifie 'action ou manière de pêcher' (cf. TLFi) et *pêche*, son homographe, qui signifie 'fruit du pêcher'. Analogiquement, *spinning* désigne une technique de pêche ou une activité sportive de cardio-training. Ce type de difficultés ne peuvent être résolues autrement qu'au cours d'une lecture minutieuse des exemples extraits.

5.2 Extraction des termes-clés : expressions régulières

Les expressions régulières facilitent considérablement la fouille du corpus. Nous nous en sommes servis pour repérer, à l'aide de la commande `egrep`, des noms de méthodes de pêche. Les motifs ont été bâtis à la base de structures que nous avons considérées comme potentiellement utiles pour le repérage des termes recherchés (p.ex.: *X est une méthode*). Notons que le même mot peut apparaître dans différentes structures :

Structure > Motif > Exemples			
`(\ble\b|\bla\b)\s\b\w+\b(\s\b\w+\b){,2}\sest\s(une|la)\ s(méthode|technique)`			
Ex. : [L]*a pêche à* ***la dandinette est une technique*** *à utiliser lorsqu'on a repéré un groupe de dorés*.	la bolognaise la bouée la bouillette délavée la carpe à la batterie la carpe au coup la carpe au pellet waggler la charfia ou chrafi la cuiller la dandine la dandinette la ligne à la traine	la madrague la mouche au tenkara la mouche en nymphe la mouche noyée la mouche sèche la pêche de fond la pêche en float tube la pêche en nymphe la pêche en texan la pêche verticale le down shot le drop shot	le feeder à longue distance le filet barrière le filet maillant le fireball le jig le jigging lourd le tenkara le toc le vif le wacky
`\bla\b\spêche\s(à|au)\s\b\w+\b(\s\b\w+\b){,2}`			
Ex. : *Idéale pour* ***la pêche à la dandinette****, cette canne devient inexploitable au lancer*.	la pêche à la bombette la pêche à la carpe la pêche à la cuillère la pêche à la dandinette la pêche à la dérive la pêche à la ligne morte la pêche à la mouche noyée la pêche à la mouche sèche la pêche à la nymphe la pêche à la pirouette la pêche à la traîne la pêche à la verticale la pêche à la volée la pêche à rôder	la pêche à soutenir la pêche à vue la pêche au boobie la pêche au booby la pêche au buscle la pêche au coup la pêche au fire la pêche au feeder la pêche au flotteur la pêche au jig la pêche au lancer la pêche au léger la pêche au leurre lancé la pêche au leurre manié	la pêche au mort manié la pêche au mort posé la pêche au posé la pêche au poser la pêche au quiver la pêche au rubber jig la pêche au shad en linéaire la pêche au stream la pêche au streamer la pêche au swing tip la pêche au tenya la pêche au toc la pêche au vif

\bpêche\b\s\w+(ante|elle|que|ère|enne|one|ane|ée|euse|eure)(\s|\n|$)			
Ex. : *Pour la **pêche itinérante** le panier se porte en bandoulière*.	la pêche carnassière la pêche civelière la pêche civelle la pêche coquillère la pêche germonière	la pêche harengière la pêche itinérante la pêche journalière la pêche morutière la pêche maquereautière	la pêche piroguière la pêche planante la pêche supersonique la pêche téléscopique la pêche thonnière

Tab. 1. Extraction des noms de méthodes de pêche.

Certains syntagmes que nous avons pu extraire renvoient à la pêche non récréative, p.ex.: *pêche baleinière*, *pêche électrique*, *pêche halieutique*, *pêche haussière*, *pêche hauturière* (dite aussi *pêche au large*), *pêche nourricière*.

Certaines appellations sont des mots anglais (p.ex.: *surfcasting*, *drop shot*, *jig*). Celles-là et d'autres importations anglaises sont, en effet, assez fréquentes dans le jargon de la pêche, ce qui peut s'expliquer par le fait que les Anglais ont été les premiers à faire de la pêche une activité récréative (Sourdot, 2018, p. 276). Grâce aux motifs faisant référence à la structure interne de certains dérivés (p.ex.: *-ing*), nous avons pu extraire des noms de méthodes (p.ex.: *popping*, *spinning*, *trolling*), d'activités (p.ex.: *cofishing*) et de qualités (employés comme épithètes auprès des noms de leurres, p.ex.: *sinking* et *suspending* – en parlant de woblers). Pareillement pour les composés avec *bait* (fr. 'leurre'). Ces derniers n'ont pas, en général, d'orthographe stable, ce que confirment les exemples de corpus : ils s'écrivent soudés ou séparés par un espace, ou encore avec un tiret d'union – cf. Tableau 2.

5.3 Extraction des termes-clés : *n*-grammes

Dans le traitement automatique des langues, les *n*-grammes renvoient aux suites de *n* caractères (ce à quoi on recourt couramment dans l'apprentissage automatique, ang. *machine learning*) ou aux suites de *n* mots. Dans le cas des *bi*-grammes, le *n* égale à 2, dans le cas des *tri*-grammes, le *n* égale à 3, et ainsi de suite. L'idée est qu'à partir d'une séquence de *n* caractères ou de *n* mots il est possible d'évaluer la probabilité d'occurrence pour un caractère ou un mot suivant. Les méthodes statistiques sont, en effet, en usage également pour évaluer le potentiel terminologique des mots, encore que depuis un certain temps elles cèdent la place aux réseaux neuronaux.

Structure > Motif > Exemples			
\b\w+ing\b			
Ex. : *Toutes les techniques sont praticables ici mais nous privilégions la pêche aux leurres (lancer et **jigging**)*.	baitcasting casting cofishing cranking fluorocasting flyfishing jigging matching	pitching popping powerfishing rockfishing rolling slow-jigging spinning	sportfishing stalking streetfishing surfcasting trolling twitching urban fishing

\b\w+(-|\s)?bait\b		
Ex. : *Le **crankbait** peut également être coulant ou suspending permettant d'accéder aux couches d'eaux inférieures*.	bigbait (big-bait, big bait) buzzbait (buzz bait, buzz-bait) chatterbait (chatter bait) cran(k)bait (cran(k)-bait, cran(k) bait) glidebait (glide bait) groundbait hardbait (hard bait, hard-bait)	hookbait (hook bait, hook-bait) jerkbait powerbait (power bait) propbait (prop bait) softbait (soft bait) spinnerbait (spinner bait,spinner-bait) stickbait (stick bait)

Tab. 2. Extraction des importations anglaises en *-ing* et *bait*.

Or, les *n*-grammes peuvent servir de point de départ pour l'extraction des termes-clés syntagmes. Il ne suffit pas, certes, que les mots qui se suivent dans le texte soient deux, trois ou quatre, pour constituer un syntagme (p.ex. : *pêche à*, *pêche à la* et *la pêche à la* sont des *n*-grammes sans pour autant être syntagmes). Dans la suite nous expliquons les étapes de l'extraction de *n*-grammes avant qu'ils soient soumis à une lecture permettant d'en tirer ceux qui sont, en fait, des termes de pêche.

5.3.1 Termes-candidats avec les mots clés : *canne*, *flotteur*, *leurre* et *moulinet*

Ci-dessous nous présentons quelques exemples de termes-clés qui furent extraits à partir des *n*-grammes à deux, trois et quatre mots textuels.[4] Nous nous limitons ci-dessous aux syntagmes dans lesquels les substantifs *canne*, *moulinet* et *leurre* sont des noyaux.

Une première série d'exemples a pour le noyau le substantif *leurre* ('appât artificiel utilisé pour la capture des poissons' – cf. TLFi). Une fois les *bi*-grammes *leurre X* extraits, nous les avons passés au crible de la sélection automatique : nous avons commencé par l'élimination des mots vides (*stop words*) tels que les articles, les prépositions, les pronoms et – puisque nous avons décidé de nous limiter dans cet exercice aux syntagmes nominaux – nous avons passé à l'élimination des formes conjuguées de verbes (p.ex.: *a*, *est*, *dépend*). Nous avons ainsi obtenu env. 700 emplois uniques que nous avons soumis à un dépouillement manuel. Notons que déjà pour les *tri*-grammes la procédure était différente. Étant donné que trois mots peuvent correspondre à une synapsie et que les synapsies sont fréquemment utilisées dans les

[4] Voici le haut de la liste de 10,810,023 *bi*-grammes : *de la* (1,248,804 occurrences), *de l'* (764,699), *à la* (525,833), *la pêche* (381,253), *de pêche* (339,510). Notons qu'en haut de cette liste fréquentielle de séquences pour la plupart des cas inutiles (car ni syntagmes ni termes recherchés) le mot *pêche* apparaît dans deux séquences différentes (non chevauchantes). Les *tri*-grammes, *canne à pêche* et *bateau de pêche* occupent les positions soixante-onzième et cent-seizième sur la liste de 28,237,535 séquences, avec 10,291 et 8,147 occurrences (sans compter les variantes mal écrites, p.ex.: *pęche*, ou grammaticales, p.ex. : *bateaux de pêche*), tandis que les *quadri*-grammes *pêche à la mouche* et *pêche à la ligne* se trouvent déjà à la position quatrième et à la position septième du haut de la liste fréquentielle de 39,101,958 suites avec, respectivement, 12,363 et 10,492 occurrences. Ces données confirment la spécificité sémantique du corpus étudié.

nomenclatures techniques, ce que Benveniste constata déjà quand il introduisait la notion de synapsie en linguistique il y a 50 ans (Benveniste, 1966, p. 172), l'élimination des mots vides ne pouvait plus se faire sans avoir soumis ceux-là à une sélection, tenu compte de la position centrale qu'un mot *vide*, tel la préposition *à* ou la préposition *de*, occupe dans le composé synaptique (cf. Tableau 3).

n-grammes	**Exemples**		
bi-grammes : *leurre X* (syntagme nominal) Ex. : *Le **leurre lipless** ou lipless crankbait est un leurre coulant dépourvu de bavette.*	leurre anguille leurre artisanal leurre artificiel leurre brochet leurre casting leurre dur leurre écrevisse leurre fluo	leurre fouetté leurre grenouille leurre jig leurre lesté leurre libellule leurre lipless leurre lézard leurre maison	leurre ondulant leurre pencil leurre plombé leurre plongeant leurre souple leurre tournoyant leurre vairon leurre worm
tri-grammes : *leurre Prép N* (synapsie) Ex. : *Ce **leurre de suspension** combine le profile d'un poisson appât et l'action d'un poisson blessé.*	leurre à bavette leurre à brochet leurre à dandiner leurre à distance leurre à flapper leurre à hameçon	leurre à hélice leurre à jigger leurre à perche leurre à truite leurre à truites leurre de pêche	leurre de prospection leurre de spinning leurre de surface leurre de suspension leurre de traîne leurre de traque
quadri-grammes : *leurre Prép* (*N Adj* | *N N* | *Adj N*) (synapsie) Ex. : *La cuiller tournante (par opposition à cuiller ondulante) est un **leurre à palette tournante**.*	leurre à action moyenne leurre à collerette double leurre à double caudale leurre à hameçon simple leurre à la bombette leurre à palette tournante leurre à tête plombée leurre de grande prospection leurre de pêche wobbler leurre de pêche crankbait leurre de pêche cuillère leurre de pêche modulaire		leurre de pêche spinnerbait leurre de pêche profonde leurre de pêche télescopique leurre de pêche vibrant leurre de pêche swimbait leurre de pêche worm leurre de type finesse leurre de type crankbait leurre de type jerkbait leurre de type mouche leurre de type shad

Tab. 3. Termes-clés avec *leurre* comme noyau du syntagme

Parmi les *quadri*-grammes extraits certains sont des syntagmes nominaux libres, non terminologiques, dont les satellites fournissent des caractéristiques à propos des traits en quelque sorte accidentels de leurres tels que leur couleur (p.ex.: *leurre de couleur criarde*, *leurre de couleur discrète*, *leurre de couleur sombre*, *leurre de couleur jaune*), poids (p.ex.: *leurre de faible grammage*), forme (p.ex.: *leurre de forme longiligne*, *leurre de forme triangulaire*) et taille (p.ex.: *leurre de grande taille*, *leurre de grosse taille*, *leurre de petite taille*, *leurre de taille moyenne*).

Trois séries d'exemples qui suivent ont pour les noyaux les mots clés *canne*, *flotteur* et *moulinet* :

***n-grammes* – Exemples**		
bi-grammes : (*canne* | *flotteur* | *moulinet*) (*N* | *Adj*) (syntagme nominal)	*tri*-grammes : (*canne* | *flotteur* | *moulinet)* *Prép N* (synapsie)	*quadri*-grammes : (*canne* | *flotteur* | *moulinet*) *Prép* (*N Adj* | *N N* | *Adj N*) (synapsie)
canne anglaise canne baitcasting canne casting canne feeder canne silure canne spinning canne toc canne téléréglable canne télescopique	canne au posé canne à appâts canne à déboité canne à emboîtement canne à emmanchement canne à pêche canne de jig canne de spinning canne de traîne	canne à pêche anglaise canne à pêche blanche canne à pêche bolognaise canne à pêche jigging canne à pêche spinning canne à pêche surfcasting canne à pêche téléscopique canne à scion souple canne de surfcastiong léger
flotteur boule flotteur carotte flotteur glisseur flotteur lumineux flotteur poire flotteur sondeur flotteur tige	flotteur de cassant flotteur de pêche flotteur de sonde flotteur goutte d'eau flotteur pêche mer flotteur pour carpes flotteur sans antenne	flotteur de forme crayon flotteur de forme poire flotteur en forme de stick flotteur de pêche ascendant flotteur de pêche coulissant flotteur de pêche balle flotteur de pêche nuit
moulinet amovible moulinet baitfinesse moulinet débrayable moulinet casting moulinet cranking moulinet match moulinet spincasting moulinet spinning moulinet verrouillable	moulinet de pêche moulinet de surfcasting moulinet frein avant moulinet frein arrière moulinet grande capacité moulinet jigger électronique moulinet lancer lourd moulinet large arbor moulinet tambour tournant	moulinet à bobine fixe moulinet à bobine libre moulinet à frein avant moulinet à grande récupération moulinet à ratio lent moulinet à tambour tournant moulinet à tambour fixe moulinet de pêche centerpin moulinet de pêche rotatif

Tab. 4. Termes-clés avec *canne*, *flotteur* et *moulinet* comme noyaux des syntagmes.

À part les exemples cités, d'autres composés synaptiques furent extraits et soumis à la vérification manuelle, p.ex.: *bobine à tambour fixe*, *cuillère ondulante*, *cuillère tournante*, *cuillère vaironnée*, *émerillon à agrafe*, *émerillon à billes*, *émerillon baril*, *émerillon rolling*, *hameçon à œillet*, *œillet de guidage*, etc.

6. CONCLUSION

La brève présentation que nous venons de faire montre à quel point les outils GNU/Linux, et les expressions régulières, peuvent s'avérer utiles dans la collecte des candidats-termes nominaux. La connaissance des propriétés morphosyntaxiques des syntagmes, composés réguliers et synaptiques, permet au linguiste d'accéder relativement vite aux séquences qu'il peut ensuite soumettre à une vérification manuelle et en sélectionner celles pour lesquelles le potentiel terminologique n'est pas uniquement l'affaire de fréquence, surtout lorsqu'on travaille sur un corpus aussi

immense et bruité que le corpus web et que l'objet de notre recherche – en l'occurrence le jargon de la pêche – n'y est qu'un pour mille. C'est d'ailleurs pour cette raison que le corpus traité a, certes, rendu possible « la quête de l'inconnu », il « a montré », comme disait Scheer (2004), sans pour autant « démontrer » la totalité du parler étudié. Et pour qu'une vision ainsi bâtie soit plus complète, il faudrait – à part l'extraction des termes nominaux – prévoir un traitement qui aboutirait à la sélection des verbes et des constructions verbales typiques pour le jargon de la pêche, telles que *prospecter les zones profondes*, *scanner le fond* (en parlant des humains), et *se jerker*, *travailler* (en parlant des leurres).

Bibliographie

AMITAY, Einat : Anchors in context: A corpus analysis of web pages authoring conventions. In : Words on the Web – Computer Mediated Communication. Eds. L. Pemberton – S. Shurville. Éd. Intellect Books 1999.

BLANCHET, Philippe : La linguistique de terrain. Méthode et théorie. Rennes : Presses universitaires de Rennes 2012.

DROUIN, Patrick : Acquisition automatique de termes : simuler le travail du terminologue. In : Études de linguistique appliquée, 2015, Vol. 180, No 4, pp. 417–427.

FOUQUERÉ, Christophe – ISSAC, Fabrice : Corpus issus du Web : constitution et analyse informationnelle. In : Revue québécoise de linguistique, 2003, Vol. 32, No 1, pp. 111–134.

FRANÇOIS-GEIGER, Denise : 1988, Les paradoxes des argots. In : Actes du Colloque culture et pauvretés. Eds. A. Lion – P. de Meca. 1988, pp. 17–24, Tourette 13-15 décembre 1985, La Documentation française.

GANASCIA Jean-Gabriel : Les *big data* dans les humanités. In : Critique, 2015, No 819-820, pp. 627–636.

KALISKA, Agnieszka : Potoczność a terminologiczność socjolektu wędkarskiego. Polskie i francuskie nazwy technik wędkarskich. In : Orbis Linguarum, Vol. 49, Dresden-Wrocław : Neisse Verlag & Oficyna Wydawnicza ATUT 2018, pp. 109–126.

LABOV, William : Some Principles of Linguistic Methodology. In : Language in Society, 1972, Vol. 1, No 1, pp. 97–120.

LONGHI, Julien : Humanités, numérique : des corpus au sens, du sens aux corpus. In : Questions de communication, 2017, Vol. 31, pp. 7–17.

ORTIZ, Pedro J. – SAGOT, Benoît – ROMARY, Laurent : OSCAR – Open Super-large Crawled ALMAnaCH coRpus. Disponible sur : https://oscar-corpus.com/ux normes (accès : 12-14 octobre 2020).

PANCKHURST, Rachel : Discours électronique médié : quelle évolution depuis une décennie ? In : La langue du cyberespace : de la diversité aux normes. Ed. J. Gerbault. Paris : L'Harmattan 2007, pp. 121–136.

PIASECKI, Maciej : User-driven language technology infrastructure – the case of Clarin-PL. In : Proceedings of the 9th Language Tehnologies Conference. Ljubliana : Information Society 2014.

SCHEER, Tobias : Le corpus heuristique : un outil qui montre mais ne démontre pas. In : Corpus, 2004, No 3, pp. 153–192.

SHOTTS, William E., Jr. : The Linux Command Line. Éd. No Starch Press 2012.

SOURDOT, Marc : Les mots de la pêche ou « emprunter c'est jouer ». In : Journée d'hommage à A.-M. Houdebine. La Rochelle 17 juin 2011. Disponible sur : l.brunethunault.free.fr/17juin/marcSourdot.pdf].

SOURDOT, Marc : Il faut appeler un « shad » un shad. In : Le poids des mots. Hommage à Alicja Kacprzak. Eds. A. Konowska – A. Woch – A. Napieralski – A. Bobińska. Łódź : Wydawnictwo Uniwersytetu Łódzkiego 2018, pp. 275–281.

Sitographie

Mots vides français en ligne. Disponible sur : https://countwordsfree.com/stopwords/french

Ressources lexicales sur le site d'Yann van der Cruyssen. Disponible sur : http://www.nurykabe.com

TLFi : Trésor de la langue Française informatisé. Disponible sur : http://www.atilf.fr/tlfi, ATILF - CNRS & Université de Lorraine.

Remerciements :
Je tiens à remercier Karol Kaczmarek, doctorant en informatique à l'Université Adam Mickiewicz/ Applica.AI, pour son aide dans le téléchargement et le filtrage du corpus. Un grand merci aussi à tous les internautes, souvent anonymes, qui contribuent éminemment à l'échange de connaissances liées au traitement automatique du texte.

sciendo
DOI 10.2478/jazcas-2022-0017

COMMENT LES DIFFÉRENTS TYPES DE CORPUS LINGUISTIQUES ÉCLAIRENT (OU NON) LES DIFFÉRENTS TYPES DU LEXIQUE SUBSTANDARD : ANALYSE CONTRASTIVE À PARTIR DU VOCABULAIRE DE LA COMÉDIE « LES KAÏRA », EXEMPLE TYPIQUE DU GENRE FILMIQUE DIT « DE BANLIEUE »

ALENA PODHORNÁ-POLICKÁ[1] – ANNE-CAROLINE FIÉVET[2]
[1]Université Masaryk de Brno, Brno, Tchéquie
[2] L'École des hautes études en sciences sociales, Paris, France

PODHORNÁ-POLICKÁ, Alena – FIÉVET, Anne-Caroline: How different types of linguistic corpora shed light (or not) on various categories of substandard lexicon: contrastive analysis of vocabulary in the comedy "Les Kaïra" [Porn in the hood], a typical example of the hood film genre. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 3, pp. 927 – 941.

Abstract: The arrival of WaC corpora, including Aranea family corpora, with its "close-to-spoken language" writings from different non-formal web pages brought the new options to researchers of sociolects, mainly to those who were previously obliged to observe youth collectives in its spontaneous discourses with its consequent time-consuming transcripts. Non-spontaneous spoken language from rap songs or youth film dialogues also help researchers to describe the level of societal diffusion of some typical features of youth slang. In this paper, we focus on demonstration of these crossed approaches in order to describe three types of verbs, used in a successful comedy about Parisian peri-urban post-adolescents *Les Kaïra* (2012), representing different types of substandard lexicon.

Key words: substandard verbs, French, neology, film dialogues, corpus linguistics, hood films

0. INTRODUCTION

La dynamique langagière est traitée abondamment par les linguistes, surtout les lexicologues, qui s'appuient sur des corpus textuels de plus en plus larges et de plus en plus accessibles. Jusqu'à récemment, ces corpus textuels avaient pour inconvénient de présenter essentiellement les néologismes des journalistes (cf. Sablayrolles, 2013 ; Cartier, 2019) et de faire l'impasse sur la néologie créée, promue et partagée notamment par la jeune génération. En effet, pour ce type de néologisme, seule l'observation sur le terrain - aussi bien des réseaux concrets (Popovičová Sedláčková, 2012 pour le slovaque, à titre d'exemple) que des réseaux virtuels (Chovancová, 2009 pour le français, entre autres) - permettait d'observer les tendances de cette dynamique inhérente à chaque langue vivante.

Cette dichotomie était encore plus prononcée dans les pratiques lexicographiques. En effet, cette deuxième catégorie de néologismes identitaires à un moment donné pour la jeune génération en entier ou pour une partie d'entre elle ne rentre pas facilement dans les dictionnaires généraux, elle est au contraire retraçable à partir des différents dictionnaires spécialisés de jeunes amateurs de langue. Une récupération dictionnairique officielle, si elle a lieu, s'opère avec un décalage temporel qui pourrait être expliqué, parmi d'autres facteurs, par l'instabilité sémantique de nombreux nouveaux items, plus particulièrement de ceux qui comportent une forte valeur expressive et identitaire pour les jeunes. Malgré le caractère instable de tous les aspects variationnistes (Gadet, 2003), d'où l'impression de « futilité » de toute entreprise lexicographique sur le sujet, nous partageons l'idée de Gaétane Dostie que « s'il faut choisir [...], il est préférable d'avoir une vue figée de ce qui bouge, que de n'avoir aucune vue sur le sujet » (Dostie, 2004, p. 192). C'est dans cette perspective qu'Alena Podhorná-Polická a lancé, en 2009, le projet d'un corpus de chansons de rap francophone, le *RapCor*. Ce dernier permet de compléter les résultats de recherches qualitatives et quantitatives sur le terrain et aide à reconstituer l'histoire des néologismes identitaires pour les jeunes à l'époque de leur promotion via le rap (voir Podhorná-Polická – Fiévet, 2013). La version la plus actuelle qui comprend les textes de 1288 chansons de rap est disponible sur la plateforme SketchEngine (pour une description plus détaillée de cette version intitulée *RapCor 1288*, voir Podhorná-Polická, 2020).

La situation a considérablement évolué depuis l'arrivée des corpus « big data », gratuitement accessibles et compilables de manière répétitive et à faible coût à partir des pages web (web as corpus ou WaC). L'apparition des corpus de la famille WaCky (Baroni et al., 2010), TenTen (cf. Jakubicek et al., 2013) et d'Aranea (Benko, 2014) a ouvert et démocratisé l'accès à des productions du « parlécrit »[1] français : le *frWaC* de 2010 a été rendu disponible en 2013 avec une taille de plus d'un milliard trois cent millions de mots. La même année arrive aussi la version française de la famille TenTen : le *FrTenTen12*, compilé du web francophone en 2012 avec presque 10 milliards de mots ; suivi par la version compilée en 2017, qui fait la moitié de la taille de la version de 2012 mais s'avère intéressante du point de vue de la synchronie dynamique. Et, enfin, en 2014, arrive la famille de corpus Aranea (Benko, 2014) qui a l'avantage de montrer les contrastes, sous *Araneum Francogallicum*, entre les différentes parties de la francophonie grâce aux cinq sous-corpus (*Gallicum, Belgicum, Canadiense, Helveticum* et *Africanum*). En effet, leur taille inégalable (pour la dernière version d'*Araneum Franco-*

[1] Étudiant les spécificités de la communication sur Minitel à l'époque, Anne-Marie Jeay a proposé un mot-valise néologique qui n'a rien perdu de son ergonomie et adéquation quant à la description de ce qui se passe sur les réseaux sociaux trente ans plus tard : « **Sur les messageries télématiques les individus sont censés dialoguer, mais en fait ils se 'parlent' par écrit.** Le langage y est donc un 'parlécrit' » (Jeay, 1991, p. 31). C'est l'auteure qui souligne.

gallicum Maximum, compilée en mai 2020, par exemple, il s'agit de 10,9 G, soit presque 11 milliards de positions (tokens) / 9,3 milliards de mots) et leur constitution extrêmement rapide en comparaison avec des corpus partant de bases de données de textes soumis aux droits d'auteurs (cf. Cvrček et al., 2020) fait de ce type d'outil une méthode puissante de vérification des données de terrain, désormais incontournable pour les chercheurs variationnistes. Si nous n'avons trouvé aucune attestation pour les argotismes plutôt « jeunes » tels que, à titre d'exemple, *keum* (verlan de *mec*, au sens de « garçon, (jeune) homme ») dans le corpus littéraire qui était un des rares disponibles gratuitement entre 2008 et 2013, à savoir le sous-corpus français du corpus parallèle *InterCorp* (créé dans le cadre du Corpus national tchèque), la situation s'améliore un peu quantitativement mais pas qualitativement à partir de sa version 7, lancée en 2014, où nous avons pu découvrir les 11 premières occurrences de *keum*, ceci dans une collection *Open subtitles*, formée de traductions de films par des amateurs inconnus que le Corpus national tchèque a décidé d'abriter dans le souci de diversifier les textes parallèles et d'augmenter leur taille (Nádvorníková – Vavřín, 2014). Dans le *frWac*, qui est abrité par le même Corpus national tchèque dès 2013 mais qui n'a attiré notre attention qu'en 2015, 53 occurrences de *keum* sont désormais disponibles. Mais il faut attendre l'intégration d'une des versions du corpus *Aranea Francogallicum* (AF) sous le même « toit » tchéco-slovaque, à savoir le *AF Maius* (de mars 2015 ; 1,2 G) pour que la véritable puissance des corpus de type WaC devienne évidente (avec 400 occurrences de *keum*), ouvrant ainsi la porte à la « big-data argotologie ».

Notre attention dans ce travail portera sur un film racontant l'histoire de trois *keums* d'une banlieue sud de Paris, *Les Kaïra*, sorti en 2012, année où les informaticiens de la Faculté d'informatique de l'Université Masaryk ont pu compiler le susmentionné *FrTenTen12* qui comporte 4 538 occurrences du lexème *keum*. Les nouvelles versions du corpus *TenTen* et d'*AF Maximum* (respectivement de 2017 et de 2020) montrent bien, si l'on observe la chute des chiffres de fréquence relative (i.p.m.), que cet argotisme circule actuellement moins dans le parlécrit aspiré sur la toile et serait donc moins répandu qu'en 2012.

1. VERS UN CORPUS PARALLÈLE DE DIALOGUES ET DE SOUS-TITRES

En parallèle du RapCor, de manière plus ponctuelle mais depuis plus longtemps (2007), les deux auteures constituent une base de données portant sur les films qui mettent en scène des jeunes, vivant ou non en banlieue. Nous préférons le terme anglais de « hood films » (cf. Mével, 2008) à ses équivalents français « cinéma de banlieue » (Grodner, 2020) ou encore « cinéma sur les jeunes de la banlieue » car ces derniers ont pour inconvénient de véhiculer des connotations négatives autour des termes « banlieue » et « jeune de banlieue ».

Notre but scientifique est d'observer la dynamique de diffusion et la variation diachronique du lexique substandard dans les dialogues de plusieurs « hood films » (Podhorná-Polická – Fiévet, 2008 ; Fiévet – Podhorná-Polická, 2020). Il est également nécessaire de prendre en compte les défis traductologiques et épistémologiques qui émergent quand on compare les dialogues avec les sous-titres intra-lingues (pour les sourds et malentendants, SM) et inter-lingues, ceci avec une question primordiale : les traducteurs vers les langues dites mineures telles que le tchèque, traduisent-ils à partir des dialogues ou se simplifient-ils la tâche en s'appuyant sur des sous-titres SM et, surtout, anglais, qui sont souvent disponibles avec la copie du film ?

C'est pour cette raison que notre base de données interne, comportant des transcriptions incomplètes de dialogues, a été récemment revue et retravaillée avec l'objectif de rendre disponible un corpus parallèle, formé : 1) de dialogues de « hood films » (un corpus similaire à *The Movie corpus* (Davies, 2019–) étant indisponible pour le français jusqu'à présent, nous nous inspirons du corpus de Dekhissi (2013), formé de transcriptions sélectives de 38 « films de banlieue » français parus entre 1984 et 2011) et 2) de sous-titres officiels de différents types (faisant ainsi écho au corpus de Open subtitles[2] qui a l'inconvénient de ne pas fournir les métadonnées sur les protagonistes du film et sur les sous-titreurs – amateurs de fan-subbing). Le résultat de ce travail sera disponible sous le nom de *HoodFilmCor* sur la plateforme Sketch Engine courant 2022. Afin de présenter ce nouveau corpus parallèle de plus près, nous montrons ci-dessous (voir Tableaux 1 et 2) les extraits de deux films qui vont y apparaître prioritairement, *Les Kaïra* (2012) et *Bande de filles* (2014). Il s'agit de deux tableaux, pour l'instant sous Excel, parce que la mise en page dans un gestionnaire de corpus avec balisage n'est pas encore achevée (en partenariat avec Pavel Rychlý, co-auteur de Sketch Engine, de la Faculté d'informatique de l'Université Masaryk). Le balisage consiste à donner plusieurs informations : sur quelle ligne apparaît le texte en question, qui énonce la réplique, de quelle couleur est le sous-titre et comment il est positionné (la couleur et la position jouent un rôle important dans les sous-titres SM surtout). Le corpus apporte également des informations sur le minutage des sous-titres (colonne de gauche), sur les métadonnées, sur les protagonistes du film (tranche d'âge, sexe, apparence, etc.) et, au niveau textuel, apporte des informations sur l'alignement des répliques (pour la transcription fidèle des énoncés) et au niveau des images pour les différentes versions des sous-titres disponibles (SM, anglais et d'autres langues qui nous intéressent, notamment le tchèque, le slovaque et l'allemand). Par exemple, pour le cas de *Bande de filles*, deux versions de sous-titres tchèques du même film, créés pour deux festivals distincts par deux traducteurs professionnels, nous permettent de faire l'analyse des procédés de traduction audiovisuels détaillés.

[2] Corpus disponible dans le cadre du corpus parallèle *InterCorp* dès sa version 7 (publiée le 19 décembre 2014).

Tab. 1. Les Kaïra (extrait)

Minutage	Sous-titres EN	Sous-titres SM
		<\V><1>Musique classique</1></\V>
		<g><\V><1>...</1></\V></g>
		<R><1>Le vinyle grésille.</1></R>
		<\V><1>Musique hip hop</1></\V>
		<g><\V><1>...</1></\V></g>
0:00:28	<1>Listen, imagine</1><2>that if you could bang</2>	<J><1>- Regarde.</1><2>Imagine que si tu peux ken</2></J>
0:00:31	<1>Beyonce or Shakira,</1>	
0:00:33	<1>who's a better bang?</1>	<J><1>ou Beyoncé ou Shakira, laquelle</1><2>des deux tu préfères le plus la ken ?</2></J>
0:00:35	<1>Stop with the lame questions.</1><2>Jackass.</2>	<J><1>-Vas-y, dégage,</1><2>avec tes questions en carton.</2></J>
0:00:38	<1>Bone any babes lately?</1>	<J><1>- Depuis quand t'as pas niqué ?</1></J>
0:00:42	<1>You serious?</1><2>You think I don't fuck?</2>	<g><B><1>-T'es sérieux,</1><2>tu crois que je ken pas ?</2></B></g>
0:00:44	<1>I fuck every babe I want.</1><2>I know tons.</2>	<g><B><1>Moi je ken toutes les meufs</1><2>que je veux.</2></B></g>
0:00:47	<1>It's easy. They dig me.</1>	<g><B><1>J'ai pas de problème,</1><2>elles me kiffent. C'est toi qui ken pas.</2></B></g>
0:00:49	<1>You can't get laid, loser.</1>	
0:00:51	<1>Why are you lying, bro?</1><2></2>	<B><Abd><1>Il est là le doss.</1></Abd><Mou><2>-Pourquoi tu mens ?</2></Mou></B>
0:00:53	<1>We're always together.</1><2>If you fucked, I'd	<B><1>On est tous les jours ensemble,</1><2>je t'aurais cramé !</2></B>

Sous-titres SM	Qui parle FR	Transcription fidèle (par répliques)
<\V><1>Musique classique</1></\V>		
<g><\V><1>...</1></\V></g>		
<R><1>Le vinyle grésille.</1></R>		
<\V><1>Musique hip hop</1></\V>		
<g><\V><1>...</1></\V></g>		
<J><1>- Regarde.</1><2>Imagine que si tu peux ken</2></J>	Mousten	Tiens, regarde, imagine que si tu peux ken ou Beyoncé ou Shakira, laquelle des deux tu préfères le plus la ken ?
<J><1>ou Beyoncé ou Shakira, laquelle</1><2>des deux tu préfères le plus la ken ?</2></J>	Mousten	
<J><1>-Vas-y, dégage,</1><2>avec tes questions en carton.</2></J>	Abdelkrim	Vas-y, dégage avec tes questions en carton là. Super chelou, c'keum.
<J><1>- Depuis quand t'as pas niqué ?</1></J>	Mousten	Genre toi, ça fait combien d'temps qu't'as pas niqué une meuf, toi ?
<g><B><1>-T'es sérieux,</1><2>tu crois que je ken pas ?</2></B></g>	Abdelkrim	Eh, t'es sérieux là. Et mais, regarde-moi bien, tu crois qu'j'ken pas, moi ? Moi, j'ken toutes les meufs que j'veux. J'en connais trop des meufs. Moi j'ai pas de problème avec ça, les meufs, elles m'kiffent, mon gars. C'est toi l'chien d'la casse, qui ken pas. Il est là, l'doss.
<g><B><1>Moi je ken toutes les meufs</1><2>que je veux.</2></B></g>	Abdelkrim	
<g><B><1>J'ai pas de problème,</1><2>elles me kiffent. C'est toi qui ken pas.</2></B></g>	Abdelkrim	
<B><Abd><1>Il est là le doss.</1></Abd><Mou><2>-Pourquoi tu mens ?</2></Mou></B>	Abdelkrim, puis Mousten	Et mais, pourquoi tu mens, mon frère ? Eh, on est tous les jours ensemble. Si tu niquais des meufs, j't'aurais d'jà cramé, sale mytho.
<B><1>On est tous les jours ensemble,</1><2>je t'aurais cramé !</2></B>	Mousten	

Tab. 2. Bande de filles (extrait)

Minutage	Sous-titres EN	Sous-titres CZ1(JM)	Sous-titres CZ2
0:15:26	<B><1><Lad>- Always with your crew.<Lad></1><2><Abd>- You too.</Abd></2></B>	partě?</Abd></1><2><Lad>- Ty taky.</Lad></2></B>	<B><1><Abd>- Vždycky s partou.</Abd></1><2><Lad>- Ty taky.</Lad></2></B>
0:15:29	<B><1><Adi>- Right.</Adi></1><2><Lad>- Always together.</Lad></2></B>	<B><1><Lad>- Jo.</Lad></1><2><Adi>- Pořád spolu.</Adi></2></B>	<B><1><Lad>- Jasně.</Lad></1><2><Adi>- Vždycky spolu.</Adi></2></B>
0:15:32	<1><B>I heard you clashed with some girls.</B></1>	<B><1>Prej ste se srazily</1><2>s ňákejma holkama.</2></B>	<B><1>Slyšel jsem, že ste se pohádaly.</1><2>s nějakejma holkama.</2></B>
0:15:34	<1><B>Bullshit!</B></1>	<1><B>Kecy!</B></1>	<1><B>Kecy!</B></1>
0:15:37	<1><B>That's what they said.</B></1>	<1><B>Říkaly to.</B></1>	<B><1><Abd>- Říkaly to.</Abd></1><2><Lad>- Počkej.</Lad></2></B>
0:15:39	<1><B>Hold on a second.</B></1>	<1><B>Tak počkej.</B></1>	
0:15:40	<1><B>What did they say exactly?</B></1>	<1><B>Co přesně říkaly?</B></1>	<1><B>Co přesně říkaly?</B></1>
0:15:42	<1><B>That you guys chickened out.</B></1>	<1><B>Že jste zdrhly</B></1>	<1><B>Že jste zbabělci.</B></1>
0:15:44	<1><B>Chickened out?</B></1>	<1><B>Zdrhly?</B></1>	<1><B>Zbabělci?</B></1>
0:15:45	<1><B>We chickened out?!</B></1>	<1><B>My a zdrhnout?</B></1>	<1><B>My a zbabělci?</B></1>
0:15:48	<1><B>We'll waste them any time.</B></1>	<1><B>Rozmáznem je, kdykoli se nám zachce.</B></1>	<1><B>Zničíme je, kdykoli se nám zachce.</B></1>
0:15:51	<B><1>You got that from the chick.</1><2>always doing selfies?</2></B>	<B><1>To máš od tý kozy,</1><2>co furt fotí selfíčka?</2></B>	<B><1>To máš od tý holky,</1><2>co si pořád fotí selfie?</2></B>
0:15:54	<1><B>Doing duck-face in her bathroom.</B></1>	<B><1>Co špulí pusu v koupelně</1><2>jak topmodelka?</2></B>	<B><1>Jak na sebe neustále</1><2>špulí pusu v koupelně?</2></B>

Sous-titres CZ2	Sous-titres SM	Qui parle FR	Transcription fidèle (par répliques)
<B><1><Abd>- Vždycky s partou.</Abd></1><2><Lad>- Ty taky.</Lad></2></B>	<1><d><J>-Toujours en équipe.</J></d></1>	filles	Toujours en équipe, ce aue je vois.
	<1><ddc><B>-Vous aussi.</B></ddc></1>	Abdel	
<B><1><Lad>- Jasně.</Lad></1><2><Adi>- Vždycky spolu.</Adi></2></B>	<B><Lad><1><gcc><h>-On est là.</h></gcc></1></Lad><Adi><2><d>-Toujours ensemble.</d></2></Adi></B>	Lady, puis Adiatou	
<B><1>Slyšel jsem, že ste se pohádaly.</1><2>s nějakejma holkama.</2></B>	<1><g><J>-Ça a parlé sur vous.</J></g></1>	Abdel	Ça a parlé sur vous là si vous êtes embrouillé avec des filles.
	<1><Abd><dcc><h><B>-Ons embrouillé avec des filles.</B></h></dcc></Abd></1><2><Lad><d><J>-N'importe quoi !</J></d></Lad></2>		
<1><B>Kecy!</B></1>			
<B><1><Abd>- Říkaly to.</Abd></1><2><Lad>- Počkej.</Lad></2></B>	<J><g><1><Abd><h>-Ils ont pas dit ça ?</h></Abd></1><2><Lad>-Attends.</Lad></2></g></J>	Abdel, puis Lady	Wallah, ils ont pas dit ça ? Attends, attends
<1><B>Co přesně říkaly?</B></1>	<1><gcc><B>Elles ont dit quoi ?</B></gcc></1>	Lady	Elles ont dit quoi éxactement?
<1><B>Že jste zbabělci.</B></1>	<1><g><J>-Elles ont dit que</J></g></1> <1><gc><B>-vous avez peur d'elles.</B></gc></1>	Abdel	Moi j'étais pas là, mais elles ont dit que vous étiez des lâches et vous avez peur d'elles.
<1><B>Zbabělci?</B></1>	<B><1><Lad><dcc><h>-Genre, on a peur d'elles.</h></dcc></Lad></1><2><Adi><d>-	Lady, puis	Comment ça, des lâches ?
<1><B>My a zbabělci?</B></1>	Des lâches ?</d></Adi></2></B>	Adiatou	Genre. Genre, on a peur d'elles.
<1><B>Zničíme je, kdykoli se nám zachce.</B></1>	<1><B>-On va les exploser.</B></1>		Mais on les prend quand elles veulent, on va les exploser.
<B><1>To máš od tý holky,</1><2>co si pořád fotí selfie?</2></B>	<g><J><1>C'est pas celle qui se prend en photo</1><2>dans sa salle de bains ?</2></J></g>		Ça c'est pas un coup de la meuf là qui se prend toujours en photo dans sa salle de bains là ?
<B><1>Jak na sebe neustále</1><2>špulí pusu v koupelně?</2></B>	<1><g><J>Avec sa bouche de canard, là.</J></g></1>	Lady	Genre top-modèle avec la bouche de canard là

2. MÉTHODES CROISÉES POUR MIEUX CIRCONSCRIRE LA DIFFUSION DU LEXIQUE EXPRESSIF POUR LES JEUNES

Depuis les débuts de nos travaux sur les corpus filmiques, nous avons appliqué la méthode des filtres successifs qui consiste à rechercher les lexèmes dans des dictionnaires, du plus standard au plus argotique. Cette méthodologie a été utilisée dans le cadre de tous nos travaux mentionnés *supra*. De plus, depuis l'arrivée des WaC, les mots sont également recherchés dans les différents corpus en ligne, ce qui permet de valider ou d'invalider certaines hypothèses sur leur circulation. Il s'agira ici d'approfondir cette méthodologie que nous avons pour la première fois appliquée sur le corpus du film de Franck Gastambide sorti en 2012, *Les Kaïra* (Fiévet – Podhorná-Polická, 2020). *Kaïra* est le verlan de *racaille* (« personne peu recommandable » (PR), aujourd'hui plutôt avec le sens en usage de « délinquant juvénile » (*Dictionnaire de la zone*, DZ)), mais tandis que *racaille* est négativement connoté, *kaïra* (orthographié aussi *kaillera* ou *caillera*) apporte plutôt un effet laudatif, voire comique. Le film a connu un grand succès au cinéma lors de sa sortie en salles en 2012 avec 1 million d'entrées. *Les Kaïra* raconte l'histoire de trois amis, jeunes adultes d'une cité de la banlieue parisienne, plus exactement Melun en Seine-et-Marne. Dans l'espoir de faire des conquêtes féminines, ils vont essayer de percer dans le milieu du cinéma pornographique. Une grande partie du film retrace leurs tentatives de se procurer une sex-tape démo pour un producteur de films X, tentatives qui finissent la plupart du temps par des échecs.

Puisque le film *Les Kaïra* parle très ouvertement des relations entre filles et garçons, ceci nous a amenées, dans cette recherche précédente (Fiévet – Podhorná-Polická, 2020), à analyser ce qu'on peut appeler la dragolalie et la pornolalie : en effet, plus de 50 mots et expressions différents, pour plus de 150 occurrences, ont été relevés concernant les thématiques de la drague et de la sexualité. Nous appuyant sur *Le Petit Robert* (PR) comme dictionnaire d'exclusion (considérant que les mots qui y sont répertoriés avant la sortie du film sont déjà très connus, nous avons plus exactement exclu les mots présents dans le PR jusqu'en 2011, considérée comme l'année de tournage du film), nous avons analysé en détail les 81 occurrences restantes (31 lexèmes). Parmi eux, six expressions ont été ajoutées au PR après 2011 (comme *mettre/se prendre un vent* (PR2012), *pécho* (PR2015) ou *être en chien* (PR2017) pour les dragolaliques et *avoir la gaule* (PR2012), *film de boule* (PR2015) ou *défoncer* (PR2019) pour les pornolaliques). La plupart des lexèmes ont été trouvés dans les dictionnaires d'argot (19 sur 31, surtout dans le DZ) mais les six restants n'ont été trouvés nulle part (*dragon de Komodo, chacaler, surcheum* et *taper un blocage sur qqn* pour les dragolaliques et *anaconda* et *poutre de Bamako* pour les pornolaliques).

Ainsi, les résultats ont pu mettre au jour plusieurs niveaux de circulation, du plus évident (le lexème relevé est présent dans tous les dictionnaires d'argot des

jeunes consultés) au plus énigmatique (le lexème relevé dans le film n'est présent nulle part). Afin d'étudier notre hypothèse que le recours aux grands corpus web peut nous donner des indications supplémentaires sur ces différents niveaux de circulation ou encore spécifier les nuances sémantiques que les dialogues sous-entendent mais que les dictionnaires ne notent pas, nous avons décidé, pour cet article, de sélectionner trois verbes dont la circulation est *a priori* différente :

- un lexème à faible circulation, qu'on trouve dans aucun dictionnaire d'argot (pas d'indication sémantique, pas d'indication sur sa circulation et sur la période) : ***chacaler***
- deux lexèmes à large circulation, qui comportent plusieurs notations graphiques avec une faible dictionnairisation officielle: ***ken*** pour les pornolaliques et ***choper/pécho*** pour les dragolaliques (le verbe *choper* et sa verlanisation *pécho*, sachant qu'il existe des nuances sémantiques entre les deux).

2.1 Le lexème *chacaler* : hapax ou argotisme à définir ?

Dans le corpus de dialogues du film *Les Kaïra*, nous avons pu relever deux occurrences du verbe *chacaler*. La première est prononcée dans une phrase énoncée par un des trois principaux personnages qui s'appelle Mousten et qui est incarné par le scénariste Franck Gastambide (« et qu'y'avait qu'une seule meuf que tout l'monde a chacalée », 16.49) et la deuxième est prononcée par son acolyte, Abdelkrim (« J'ai vu quand t'es allé la chacaler à côté des chiottes » (36.04). Le lexème *chacaler* est intéressant à observer de plus près puisqu'il n'est présent dans aucun des dictionnaires consultés (PR, AFP, CTT, DZ, BOQ – voir la liste *infra*). On trouve seulement le lexème *chacal* dans le dictionnaire papier d'argot commun des jeunes, *Lexik des cités* (2007, p. 99, LC ; « 1) radin. Synonyme : crevard, pince 2) avide. Synonyme : crevard ») et dans le dictionnaire en ligne collaboratif « Wiktionnaire »[3] (« personne sournoise et opportuniste »).

Quant au lemme *chacaler* dans les différents corpus disponibles cités *supra*, il n'est présent ni dans les corpus WaC, ni dans le corpus RapCor1288. En revanche, une recherche plus complexe dans l'AF Maximum (AFM) grâce à la requête [lemma="chacal.*"&tag="V.*"] permet d'obtenir un seul résultat correspondant au sémantisme relevé dans le film : « Moi je refuse de penser que ce sont des gens frustrés et en manque de sensations sexuelles qui la ***chacalent*** » (saisi en 2015 à partir du site https://ntrjack.mondoblog.org/2013/08/29/les-camerounaises-sont-belles/). Cet exemple unique apporte pourtant des informations intéressantes : d'une part que le mot est bien annoté syntaxiquement mais mal lemmatisé (lemme *chacalent*) puisque les concepteurs de corpus WaC en général, y compris Benko (auteur d'AFM), préfèrent prendre les formes des mots pour lemmes si ces formes manquent dans le dictionnaire interne du *Tree tagger* (logiciel gratuit largement utilisé pour

[3] https://fr.wiktionary.org/wiki/chacal

l'annotation morphosyntaxique et la lemmatisation du français, cf. Stein et Schmid, 1995) à la place du « none » (absence de lemme) que propose *TreeTagger* dans de pareils cas. D'autre part, du fait que nous n'avions jamais entendu ce verbe en dehors de ce contexte filmique dans des discussions entre jeunes, cet exemple nous permet de répondre négativement à notre questionnement : est-ce que, en 2012, *chacaler* était un lexème identitaire pour le groupe de pairs autour du scénariste (voire sa création idiolectale), que ce dernier aurait essayé de faire connaître plus largement en le faisant répéter dans les dialogues ?

Dans les autres corpus WaC interrogés, à savoir *frWaC, FrTenTen12* et *FrTenTen17*, il est possible de trouver une série dérivationnelle basée sur ce verbe : *chacaliser, chacalisoter*, ce qui montre la vivacité de la métaphore animalière, importée fort probablement de l'habitat traditionnel des chacals au Maghreb, dans son acception verbale. Les informations recueillies autour de ce lexème nous permettent ainsi de proposer une définition de *chacaler* : « draguer avec avidité, de façon opportuniste (en approchant sa « proie », sans y mettre les formes) ».

2.2 Le lexème *choper* et sa verlanisation *pécho* : changements formels et sémantiques

Le verbe *pécho* a été choisi à partir de la liste des dragolaliques comme un exemple prototypique d'une circulation large dans le français hexagonal d'aujourd'hui mais aussi d'une dictionnarisation problématique qui se reflète dans le traitement de ces lemmes dans les grands corpus. Résultat d'une métathèse régulière sur le mot bisyllabique *choper* (absent des dialogues du film étudié), *pécho* a été énoncé trois fois dans *Les Kaïra*, chaque fois par Mousten dont deux fois dans la locution *pécho des meufs* (« on va pécho les meufs », 24.48 et « où est-ce qu'on va pouvoir pécho des meufs », 59.27) et la dernière fois, juste après cette dernière scène, dans le sens de « se procurer ; voler » (« tu lui demandes qu'il pécho l'enveloppe », 59.57). La polysémie de *pécho* reflétant la polysémie de son « verbe-miroir » *choper*, c'est paradoxalement seulement le sens de « draguer, séduire » qui est répertorié dans l'ouvrage lexicographique de référence, le PR. Au fait, il faut le chercher sous l'entrée *choper* qui comporte la marque lexicographique FAM. (« familier ») pour toutes ces acceptions dont les trois premières, anciennes : 1) Voler (vieilli) ; 2) Arrêter, prendre (qqn) ; 3) Attraper. La quatrième, « Parvenir à séduire qqn », toujours marquée comme FAM., n'est ajoutée qu'à partir de l'édition 2015, et c'est justement sous cette acception du champ sémantique de la dragolalie qu'apparaît *pécho* sous une entrée cachée, avec une limitation (erronée comme en témoignent les occurrences du film ainsi que de nombreux exemples dans les corpus WaC) de son emploi uniquement en tant que participe passé[4]. Absent du dictionnaire de référence en matière de français substandard, l'Argot &

[4] « Parvenir à séduire (qqn). ABSOLT *Il a chopé !* - VERLAN au p.p. *pécho*. 'le fils d'un chanteur très connu que j'ai pécho' (L. Pille) ».

français populaire (AFP), *pécho* figure, en revanche, dans le DZ, dictionnaire de référence pour l'argot commun des jeunes, en ligne, comme entrée distincte et autonome par rapport à *choper*. Il est à noter qu'ici, la suite d'acceptions légèrement différente témoigne d'une spécialisation de *pécho* par rapport à *choper.* Pour ce dernier, le lien paronymique avec *chiper* (« voler »[5]) est plus marquant mais, à la différence de nos observations sur l'usage du verbe *pécho* dans les corpus WaC, la prédominance du sens « draguer, séduire » pour le verbe verlanisé n'est pas pris en compte dans l'ordre des acceptions. De la même manière que pour le rapport *racaille – kaïra* cité *supra*, on assiste ici aussi à un effet d'allégement du poids référentiel que constatait Goudaillier en 2002 déjà : « le verlan est une pratique langagière qui vise à établir une distanciation effective par rapport à la dure réalité du quotidien […]. Le lien au référent serait plus lâche et la prégnance de celui-ci moins forte, lorsque le signifiant est inversé, verlanisé » (2002, p. 18).

Il est intéressant de noter que le dictionnaire *Comment tu tchatches !* de Goudaillier (CTT) a introduit, dès sa première édition parue en 1997, la variante phonétique *peucho* [pøʃo] juste en dessous de l'entrée *pécho* [peʃo]. Dans le DZ, en revanche, l'entrée *pécho* renvoie encore à sa variante où la fin du mot subissait une reverlanisation monosyllabique [peoʃ], orthographiée *péauche* et *péoch.* Les corpus du type WaC peuvent venir sur ce point éclairer le niveau de circulation de ces variantes phonétiques (et graphiques). Comme en témoignent les résultats du Tableau 3, les trois variantes n'ont été trouvées nulle part, ce qui renforce notre hypothèse qu'il s'agit de variantes créées et diffusées de manière locale qui tendent à disparaître avec la perte d'expressivité du verbe *pécho*.

La variation graphique reste néanmoins une caractéristique typique des créations *a priori* orales qui surgissent à partir des interactions dans un groupe en quête d'identité (générationnelle, socio-ethno-géographique ou autre). Quant à *pécho*, sa graphie a été longtemps variable (les mots verlanisés étant transcrits soit avec un tiret, soit sans, en gardant le digramme *-er* de l'infinitif ou non, avec une hésitation pour l'accent au-dessus du *e*, etc.), mais comme pour d'autres verlanisations à haute circulation (*tess* pour *cité*, par exemple ou *ken*, voir *infra*), elle tend à se stabiliser dans le parlécrit des usagers des réseaux sociaux, tout en respectant les règles de l'économie. Le fait de faire entrer la graphie *pécho* dans le PR ne fera certainement qu'accélérer cette stabilisation. Pour faire ressortir l'oscillation graphique telle que nous avons pu la voir dans nos questionnaires auprès de jeunes dans nos travaux précédents, une série d'autres graphies (*pecho, per-cho, pé-cho, pêcho*) a été évaluée dans le Tableau n°3.

Même pour le verbe *choper*, dont l'orthographe semble être stabilisée, à en croire le PR, l'oscillation entre un ou deux P (*choper* ou *chopper*) est lisible à partir des citations données dans l'AFP, par exemple. Ce phénomène, ainsi que l'homony-

[5] Cf. article *choper* dans le PR et celui d'AFP, où il serait dérivé de *coper* (« faire un faux pas »).

mie du verbe *chopper* avec le déonyme *chopper* (anglicisme, « moto de sport avec les guidons très haut », prononcé et souvent aussi orthographié comme *choppeur*) est reflété également dans le Tableau 3.

	frWaC (2010)	Araneum Francogallicum III – Maximum (2013-2019) AFMaxi	French Web 2012 (FrTenTen12)	French Web 2017 (FrTenTen17)	RapCor1288 (2020)
DOCs	2 268 304	17 767 539	20 400 411	14 088 683	1288
TOKENS	1 613 814 684	10 881 222 203	11 444 973 582	6 845 630 573	767 483
MOTS	5 911 017	9 327 453 482	9 889 689 889	5 752 261 039	709 057
choper	2 243	25 893	41 490	11 858	11
chopper	1 245 (dont 1075 pour verbe)	17 336	30 512	5 861	3
pécho	128	2 013	1 697	1 203	6
pecho	95	547	310	185	0
pêcho	6	100	0	52	0
pé-cho	1	14	0	10	0
per-cho	0	8	2	3	0
peucho, péoch, péauche	0	0	0	0	0

Tab. 3. Résultats des requêtes pour lemmes *choper, pécho* et leurs variantes graphiques dans le *frWaC*, *AFM, FrTenTen12, FrTenTen17* et dans le *RapCor*1288.

Le Tableau 3 apporte un témoignage intéressant de la distribution de différentes variantes graphiques pour les verbes *choper* et *pécho* dans différents corpus WaC (frWaC de 2010, AFM de 2013-2019, les deux FrTenTen de 2012 et de 2017) et dans un corpus annoté manuellement, le RapCor. La requête simple permet de regrouper la flexion verbale pour *choper* mais le taggeur automatique n'arrive pas à gérer sans faute l'homonymie entre *chopper* verbe et *chopper* nom, ce qui est sous-entendu pour les corpus de cette taille. Or, pour *pécho*, invariable au niveau des désinences mais variable au niveau de ses éléments vocaliques, la lemmatisation automatique est extrêmement erronée. Si l'on prend pour exemple les 128 occurrences du frWaC pour *pécho* (requête simple, sans sensitivité à la majuscule), 5 d'entre eux ont pour lemme « Pécho » puisque la majuscule a fait que le TreeTagger l'a interprété comme un nom propre. Or, lorsqu'on regarde la distribution de 128 occurrences de *pécho* en catégories grammaticales, on a affaire à seulement 51 verbes (soit un taux de 39,8 % d'annotation correcte), les autres cas ont été interprétés comme NOM (nom commun ; 45,3 %), ADJ (adjectif ; 2,3 %) ou encore comme NAM (nom propre ; d'autres

11 cas ayant la minuscule, soit 12,5 %). Les défauts d'annotation automatique par TreeTagger se font encore plus sentir sur les 2 013 résultats pour *pécho* dans l'AFM où les tags comportent 0 verbe, 1 650 noms, 302 adjectifs et 57 noms propres – le taux d'annotation correcte est alors de 0 %. Quant au FrTenTen17, annoté par un autre outil, le FreeLing, la totalité des occurrences de *pécho* est annotée comme nom (propre ou commun, en fonction de l'initiale).

Le *RapCor* a été ajouté aux côtés de ces grands corpus web pour deux raisons : d'une part pour montrer que la lemmatisation y est semi-automatique et les variantes graphiques sont prises en considération (le résultat de TreeTagger est revu chanson par chanson et les variantes graphiques sont intégrées dans notre dictionnaire interne sous le même lemme fédérateur afin de simplifier les requêtes dans un avenir proche). D'autre part, si l'on regarde la fréquence relative de *pécho* par rapport à la taille du corpus, on constate une spécialisation du RapCor par rapport au français substandard (frWaC – 0,08 ; AFM – 0,22 ; FrTenTen12 – 0,17, FrTenTen17 – 0,21 et RapCor – 7,96). Pour *choper*, ce rapport (les argotismes y sont 7 fois plus fréquents) se confirme.

De plus, toujours en chiffres relatifs (i.p.m.), il est intéressant du point de vue de la synchronie dynamique de comparer FrTenTen12 (2012 : année de sortie du film) d'un côté et AFM et FrTenTen17 (postérieur au film) de l'autre, pour constater que la circulation est en baisse pour *choper* et en légère hausse pour *pécho*.

2.3 Le lexème *ken* : un usage ergonomique mais une lexicographie complexe

Notre troisième exemple tiré du film *Les Kaïra* est le verbe le plus fréquent de la catégorie des pornolaliques. La verlanisation apocopée formée à partir de l'arabisme entièrement intégré au français, *niquer* [nike] > *[keni] > [kɛn], *ken* est apparue 25 fois dans les dialogues du film, dans le sens de « posséder sexuellement ».

La recherche dans les différents corpus nous sera utile, ici aussi, pour mieux circonscrire : 1) la distribution réelle des variantes graphiques qui ont été notées dans les différents dictionnaires d'argot et 2) le problème de la lemmatisation défaillante des verbes invariables dans les corpus annotés automatiquement.

De même que pour l'homographie entre le verbe et le nom *chopper*, la graphie la plus fréquente pour la suite phonique [kɛn], à savoir *ken*, apporte une homographie avec le nom propre d'origine celte, assez fréquent, à savoir Ken. Cette forme abrégée du prénom masculin Kenneth est connue surtout comme compagnon de la poupée Barbie. Dans le discours (entre filles notamment), il peut fonctionner également comme déonyme pour désigner un synonyme de *beau gosse* (« bel homme/garçon »), où l'initiale peut s'orthographier avec une minuscule. La négligence du style, typique pour le parlécrit des sms d'abord, puis pour le parlécrit sur les réseaux sociaux, et très présente dans les corpus web, complique les requêtes. Par exemple, parmi les 32 706 occurrences de *ken* (requête simple) dans l'AFM, on peut trouver, en dehors des emplois nominaux susmentionnés, à la fois les formes phonétisantes de *qu'en = ken*, les noms de produits divers importés du Japon (*ken* étant un homonyme très fréquent en japo-

nais, ayant une trentaine de sens différents dont 7 extrêmement fréquents, p. ex. le fameux manga *Hotaru no ken*, le nom du département + *ken* « département », etc.). Avec la lemmatisation erronée (0 % de verbes, comme *supra* pour *pécho*), il est difficile de trouver des occurrences de *ken* avec le sens de « posséder sexuellement » ou « abîmer, détruire ». En utilisant la requête CQL complexe : [tag="PRO.*"][word="ken"] qui donne 83 occurrences, on enlève à la fois les prénoms Ken mais on limite la recherche uniquement aux pronoms dont certains renvoient à la personne qui fait l'action (*je ken, j'ken, on ken*) et d'autres renvoient toujours au Ken (déonyme ou prénom) : *ce ken*, *moi ken*. D'autres contextes verbaux peuvent être retrouvés avec la requête [lemma="avoir|faire"] [word="ken"] – 71 occurrences, notamment les participiales (*je l'ai ken, on a ken*) et les pronominaux passifs (*se faire ken*).

Le chercheur peut alors accéder (même si c'est de façon peu aisée) aux attestations de circulation fréquente de ce verbe et apprendre, par une requête rapide, que *ken* est la variante graphique privilégiée par les scripteurs sur la toile : dans l'AFM, *kène* renvoie au verbe étudié moins de 30 fois parmi les nombreux africanismes orthographiés ainsi, *kèn* une seule fois, et aucune preuve n'est trouvée de l'utilisation des orthographes *quène*, *kéne* ou *kén*. Or, si un étudiant en FLE ou un traducteur sont à la recherche de ce verbe dans les dictionnaires, ils ne trouveront *ken* ni dans le PR, ni dans l'AFP. Pour le CTT, ils devront aller le chercher sous la lettre Q (entrée *quène*). En ce qui concerne d'autres dictionnaires d'argot, seul *Bien ou quoi* (BOQ) privilégie la graphie *ken*. Quant au DZ en ligne, l'entrée principale y est bizarrement *kéner*, avec la remarque « s'emploie aussi sous sa forme invariable *kène* ». L'adverbe « aussi » est inapproprié parce que nous avons observé l'emploi exclusif de cette forme invariable, aussi bien dans les corpus web que dans le RapCor (23 occurrences dont une seule avec la graphie originelle du rappeur Ol'Kainry « elles kennent toutes », chanson *En attendant* de 2001, qui ne prouve pourtant pas l'existence de la forme régularisée en *-er* à l'infinitif). Pour être précis, l'AFM apporte une occurrence de *kéner* (taggué en tant que NOM) dans un contexte métalinguistique : « ce qui le rend ultra 'kénable' (du verbe 'kéner'). ».

Cet exemple du verbe *ken* renforce alors notre conviction qu'il y a matière à creuser aussi bien du côté des lexicographes traditionnels (ajustement des graphies des entrées principales en fonction des usages réels) que du côté des chercheurs en linguistique de corpus (élargissement des dictionnaires intégrés aux taggers par le lexique substandard).

3. CONCLUSION

Nos trois exemples ont permis de confirmer notre hypothèse de travail. En effet, les corpus linguistiques du type WaC (Web as Corpus) apportent des informations précieuses sur le niveau de diffusion des argotismes (souvent des néologismes) qui sont identitaires pour les jeunes à une époque donnée, ils servent de **banques d'attestation des usages dans le passé proche**. La riche palette de **notations gra-**

phiques qui les accompagne (notamment s'il s'agit d'emprunts ou de verlanisations) complique les requêtes. Et ce sera donc un défi pour celui qui aura comme mission de revoir les lemmes de façon manuelle (ex.: *ken*/ *kène*/*quène,* verlan de *niquer*, « posséder sexuellement »). Les WaC sont un outil intéressant pour ceux qui s'intéressent au français substandard mais l'accès à certaines formes, en particulier aux formes verbales invariables, reste difficile face à la **complexité des variantes formelles** et encore plus aux **nuances sémantiques**, faute d'annotation morphosyntaxique et de lemmatisation automatiques sans révision manuelle.

On arrive alors à l'époque où les pratiques orales de jeunes qui n'ont eu que peu d'impact dans les corpus écrits lors de la décennie précédente deviennent scripturalisées grâce à l'essor des réseaux sociaux et peuvent être étudiées par les chercheurs grâce à l'arrivée des corpus WaC, ce qui ouvre une brèche pour la « big-data argotologie ».

Acknowledgements: The research has been supported by the Masaryk University Development Fund (project MUNI/FR/1366/2019).

Bibliographie

BARONI, Marco – BERNARDINI, Silvia – FERRARESI, Adriano ZANCHETTA, Eros : The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. In : Language Resources and Evaluation, 2009, Vol. 43, No 3, pp. 209–226. Disponible sur: https://doi.org/10.1007/s10579-009-9081-4

BENKO, Vladimír : Aranea: Yet Another Family of (Comparable) Web Corpora. In : Text, Speech and Dialogue. Eds. P. Sojka *et al*. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, Springer International Publishing Switzerland, pp. 257–264.

CARTIER, Emmanuel : Néoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. In : Neologica, 2019, No 13, pp. 23–54.

CHOVANCOVÁ, Katarína : Pour une pragmatique de l'écriture interactive en ligne : le statut de l'énoncé dans le chat ». In : La langue en contexte, Helsinki : Université d'Helsinki 2009, pp. 199–211.

CVRČEK, Václav – KOMRSKOVÁ, Zuzana – LUKEŠ, David – POUKAROVÁ, Petra – ŘEHOŘKOVÁ, Anna – ZASINA, Adrian Jan – BENKO, Vladimír : Comparing web-crawled and traditional corpora. In : Language Resources and Evaluation, 2020, No 54, pp. 713–745. Disponible sur : https://doi.org/10.1007/s10579-020-09487-4

DAVIES, Marc : The Movie Corpus. (2019–). Disponible sur : https://www.english-corpora.org/movies/

DEKHISSI, Laurie : Variation syntaxique dans le français multiculturel du cinéma de banlieue. Thèse de doctorat sous la direction d'Aidan Coveney et Zoë Boughton, Exeter : Université 2013.

DOSTIE, Gaétane : Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique. Bruxelles : De Boeck/Duculot 2004.

FERRARESI, Adriano – BERNARDINI, Silvia – PICCI, Giovanni – BARONI, Marco : frWaC. Ústav Českého národního korpusu FF UK 2013, Praha. Disponible sur : http://www.korpus.cz

FIÉVET, Anne-Caroline – PODHORNÁ-POLICKÁ, Alena : La variation du lexique substandard dans le cinéma sur la banlieue : analyse argotologique du champ lexical des relations garçons-filles dans le film « Les Kaïra ». In : Diversité et variations de la langue française au XXIe siècle. Eds. R. Mudrochová – B. Courbon. Plzeň : Nakladatelsví Nava 2020, pp. 183–224.

GADET, Françoise : La variation sociale en français. Paris : Ophrys 2003.

GOUDAILLIER, Jean-Pierre : De l'argot traditionnel au français contemporain des cités. In : La linguistique, 2002, Vol. 38, No 1, pp. 5–23.

GRODNER, Manon : Le « cinéma de banlieue » : représentation des quartiers populaires ? Enjeu d'un cinéma entre réalité et fantasme. Paris : L'Harmattan 2020.

JAKUBÍČEK, Miloš – KILGARRIFF, Adam – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel – SUCHOMEL, Vít : The tenten corpus family. In : 7th International Corpus Linguistics Conference CL, 2013, pp. 125–127.

JEAY, Anne Marie : Les messageries télématiques. Une communication paradoxale. Paris : Eyrolles 1991.

MÉVEL, Pierre-Alexis : Traduire La haine : banlieues et sous-titrage. In : Glottopol, revue sociolinguistique en ligne, 2008, No 12, pp. 161–181.

NÁDVORNÍKOVÁ, Olga – VAVŘÍN, Martin : Korpus InterCorp – francouzština, verze 7 z 19. 12. 2014. Ústav Českého národního korpusu FF UK 2014, Praha. Disponible sur : http://www.korpus.cz

PADRÓ, Lluís – STANILOVSKY, Evgeny : FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul 2012. Disponible sur : http://nlp.lsi.upc.edu/publications/papers/padro12.pdf.

PODHORNÁ-POLICKÁ, Alena – FIÉVET Anne-Caroline : Argot commun des jeunes et français contemporain des cités dans le cinéma français depuis 1995 : entre pratiques des jeunes et reprises cinématographiques. In : Glottopol, revue sociolinguistique en ligne, 2008, No 12, pp. 212–240.

PODHORNÁ-POLICKÁ, Alena – FIÉVET Anne-Caroline : Le rap en tant que vecteur des innovations lexicales : circulation médiatique et comportement des locuteurs. In : Écarts et apports des médias francophones. Eds. M. Abecassis – G. Ledegen. Oxford, Bern, Berlin, Bruxelles, Frankfurt : Peter Lang 2013, pp. 113–139.

PODHORNÁ-POLICKÁ, Alena : RapCor, Francophone Rap Songs Text Corpus. In : Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020. Eds. A. Horák et al. Brno : Tribun EU 2020, pp. 95–102.

POPOVIČOVÁ SEDLÁČKOVÁ, Zuzana : Slang v mládežnickom diskurze. Bratislava : Univerzita Komenského v Bratislave 2013.

RYCHLÝ, Pavel : Manatee/Bonito – A Modular Corpus Manager. In : 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University 2007, pp. 65–70.

SABLAYROLLES, Jean-François : D'où viennent les mots nouveaux ?. In : Sciences humaines, Le langage en 12 questions, 2013, Vol. 3, No 246, p. 14.

STEIN, Achim – SCHMID, Helmut : Étiquetage morphologique de textes français avec un arbre de décisions. In : Traitement automatique des langues, 1995, Vol. 36, No 1-2, pp. 23–35.

Dictionnaires (avec les abréviations utilisées) :

DZ : Cobra le Cynique [Abdelkarim Tengour] (2000-2020). Le Dictionnaire de la Zone. Disponible sur : https://www.dictionnairedelazone.fr

AFP : COLIN, Jean-Paul – MÉVEL, Jean-Pierre – LECLÈRE, Christian : Argot et français populaire. (1ère édition sous le titre « Dictionnaire de l'argot », 1990), Paris : Éditions Larousse 1990-2008.

LC : Collectif Permis de vivre la ville : Lexik des cités. Paris : Fleuve Noir 2007.

CTT : GOUDAILLIER, Jean-Pierre : Comment tu tchatches! Dictionnaire du français contemporain des cités. Paris : Maisonneuve & Larose (4ème éd. 2019 ; 1ère éd. 1997).

BOQ : LAFFITTE, Roland – YOUNSI, Karima : Bien ou quoi? La langue des jeunes à Ivry et Vitry-sur-Seine. Paris : SELEFA 2004.

PR : *Le Petit Robert*. Paris : Dictionnaires Le Robert 1997–2021.

sciendo
DOI 10.2478/jazcas-2022-0018

DIDACTISER LES CORPUS PARALLÈLES SPÉCIALISÉS : LE CAS DES DIRECTIVES EUROPÉENNES

ELEFTHÉRIA DOGORITI – THÉODORE VYZAS
Université d'Ioannina, Ioannina, Grèce

DOGORITI, Elefthéria – VYZAS, Théodore: Didactising specialised parallel corpora: the case of European directives. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 942 – 950.

Abstract: Within the framework of a didactic proposal, this article proposes to present a preliminary step to the specialized translation French-Greek. It will attempt to highlight the benefits of autonomous learning through the consultation of a corpus of specialized parallel texts established by the EU institutions. The use of concordancers will provide solutions to students wishing to study the variability of terminology and specialized vocabulary at monolingual and bilingual levels.

Key words: specialized translation French-Greek, parallel corpora, variability, terminology, concordancer

0. INTRODUCTION

L'enseignement de deux langues de spécialité ainsi que de la traduction spécialisée dans un département de Langues Étrangères Appliquées (LEA), tel que le nôtre, fonctionnent en complémentarité et visent la maîtrise des propriétés intrinsèques de la langue et du discours spécialisés des domaines d'application dans lesquels les étudiants exerceront leurs compétences. À cet effet, les étudiants doivent être sensibilisés tant aux spécificités de différents types de genres, de registres ou de textes, qu'aux difficultés liées au passage en français spécialisé (langue d'enseignement pour notre recherche).

Dans le cadre d'une proposition d'exploitation didactique, cette étude, en tant que démarche préalable à la traduction spécialisée français-grec, tentera de présenter les aspects positifs d'un apprentissage autonome que la consultation d'un corpus de textes parallèles spécialisé établi par les instances européennes peut offrir aux étudiants. Grâce au corpus, les étudiants pourront maîtriser la terminologie et le lexique spécialisé dans les deux langues dans un ou plusieurs domaines d'application au moyen d'un concordancier.

Il faut noter que les corpus parallèles ont en effet déjà trouvé une large application dans l'enseignement des langues, l'enseignement de la traduction, dans la recherche contrastive, aussi bien que dans la lexicographie bilingue (Loock, 2016; Schaeffer-Lacroix, 2009 ; Zanettin, 2009).

Notre démarche s'appuie pour la première étape sur la présentation d'un genre textuel spécifique tel qu'une directive européenne, à savoir juridique, qui foisonne en terminologie. Plus concrètement, la directive dont les versions française et grecque seront étudiées se caractérise par des termes de spécialité et des phraséologies qui sont spécifiques : a) au domaine scientifique du droit qui constitue le support « naturel » du texte et b) vu sa thématique, aux sciences biomédicales et au domaine de l'économie. Elle comporte aussi du lexique scientifique transdisciplinaire et de la phraséologie de langue générale.

La deuxième étape théorique se fondera sur le corpus de textes parallèles spécialisés, étudié à la lumière de la théorie fonctionnaliste (Nord, 2005) conjointement avec la terminologie textuelle (Bourigault – Slodzian, 1999).

Dans ce contexte, nous proposerons le parcours que les étudiants seront invités à suivre : repérage des unités lexicales ayant trait aux domaines respectifs qui se croisent au sein du texte tel qu'il s'affiche dans les deux langues, emploi des concordanciers, vérification des termes et de la phraséologie en matière d'équivalences terminologiques et lexicales entre les deux langues. Nous tenterons d'expliquer pourquoi nous considérons qu'une telle exploitation du corpus favoriserait une approche que Johns (1991) a baptisée le « data-driven learning », c'est-à-dire un apprentissage sur corpus.

1. LES TEXTES EUROPÉENS COMME CORPUS PARALLÈLES

1.1 Un bref tour d'horizon

L'Union européenne (UE) constitue une source inépuisable de législation pour tous les États membres. Les textes européens sont caractérisés par une pluralité thématique qui, au niveau linguistique, se concrétise au moyen du lexique spécialisé de tous les domaines de la vie quotidienne (Cao, 2007, p. 14). Depuis 2006, le Centre Commun de Recherche (JRC) en collaboration avec d'autres organisations de l'UE, en application du principe du multilinguisme, ont mis à disposition des ressources parallèles multilingues dans les 24 langues officielles (multilingual parallel language resources) (Steinberger *et al*., 2014) afin de rendre la législation accessible non seulement aux experts mais aussi au grand public (Michel, 2018).

Si cet effort est dans la bonne direction, la bibliographie internationale fait état d'une certaine opacité au niveau du sens des textes européens qui est due à plusieurs raisons : pluralité thématique, rédaction par des personnes dont la langue maternelle n'est pas celle du texte original, négociations constantes entre juristes et linguistes, absence de révision systématique (Sosoni, 2011 ; Mac Aodha, 2018). Le Guide pratique commun (2015, p. 10) préconise ainsi que la rédaction d'un acte juridique, indépendamment du type de chaque texte législatif (Directive, Règlement, Rapport etc.), soit sans équivoque, dépourvue d'éléments superflus et précise.

Le site EUR-Lex, qui nous sert de support pour la combinaison linguistique français-grec, donne accès à la législation de l'UE et propose un affichage multilingue simultané jusqu'à trois versions linguistiques d'un même document. Il s'agit, donc, des corpus parallèles téléchargeables et exploitables offrant des textes alignés au niveau du paragraphe dans des fichiers distincts (Loock, 2016, p. 164).

1.1.1 Les particularités de la traduction des textes parallèles européens

Pour passer à la traduction proprement dite, celle-ci constitue le véhicule de publication dans toutes les langues de l'UE. Le stéréotype veut que la traduction des textes européens soit considérée comme traduction juridique, puisque, pour la plupart, elle concerne des textes législatifs qui, à leur tour, donnent lieu à de nouveaux textes législatifs (Biel, 2017, p. 32).

Si les textes juridiques nationaux ont des exigences strictes aux niveaux du lexique et de la phraséologie notamment (Griebel, 2016, p. 207), mais aussi de la syntaxe et du registre afin d'atteindre une cohérence qui évite les malentendus, tel n'est pas le cas des textes européens. Cela explique pourquoi l'étude comparative de ces textes a été un défi pour nous.

Plus spécifiquement, en matière de lexique spécialisé et de terminologie, les traducteurs se voient parfois obligés d'inventer des termes – Goffin (1994) les appelle *eurolexies* – afin d'éviter la polysémie. Il s'ensuit que, les néologismes n'étant pas rares, il ne faut pas passer sous silence les interférences lexicales et sémantiques entre langues officielles au travers de la traduction (Trebits, 2008, p. 41). Cela car notamment les termes juridiques et dans un moindre degré les expressions et tournures employés ne doivent pas être trop étroitement liés à une langue ou à un système juridique national (Mac Aodha, 2018, p. 63).

Pour ce qui est de la syntaxe, les traducteurs sont parfois forcés « de tordre l'ordre logique et les règles grammaticales » de la langue cible afin d'observer les consignes d'alignement des paragraphes entre « original » et traduction (Mac Aodha, 2018, p. 41).

1.2 Les particularités des directives

Sur le plan de la transposition au niveau national, le législateur national a pour mission d'adapter les termes juridiques et techniques au contexte juridique qui est le sien. C'est la raison pour laquelle les directives doivent être formulées d'une façon moins détaillée que les autres types d'actes « pour laisser aux États membres une marge suffisante d'appréciation lors de la transposition » (Guide pratique commun, 2015, p. 12).

Il en découle que toute directive a une fonction à deux étapes à remplir, la première étant sa réception par le législateur national et la seconde la réception par le public de chaque État membre, donc un public bien ciblé. La première étape, celle qui nous intéresse, passe par la traduction, une traduction qui ne doit pas perdre de vue le public cible.

2. CADRE THÉORIQUE : L'APPROCHE FONCTIONNALISTE ET LA TERMINOLOGIE TEXTUELLE

Notre matériau consiste en la Directive européenne 2001/83/CE (consolidée en 2019), instituant un code communautaire relatif aux médicaments à usage humain, versions française et grecque.

Nous considérons que le caractère de traduction des actes juridiques européens se prête au modèle fonctionnaliste (Nord, 2005). Selon cette approche, le traducteur doit tenir compte de la fonction juridique du texte cible et de l'aspect pragmatique de la communication par le biais de la traduction (Šarčević, 2003).

La terminologie textuelle vient en complément de notre cadre théorique comme approche descriptive qui focalise sur le fonctionnement réel des unités lexicales en contexte. Se basant sur la variabilité syntaxique des termes repérée dans divers corpus, cette théorie pose que le texte joue un rôle décisif, car grâce à sa composante syntaxique, il crée un réseau d'éléments sémantiques qui contribuent à la construction du sens. De plus, l'exploration terminologique consiste à appliquer les termes extraits d'un texte en les réutilisant dans un autre texte (Bourigault – Slodzian, 1999), ce qui met en valeur la démarche préalable que nous adoptons.

3. MÉTHODOLOGIE ET ÉTUDE DU MATÉRIAU

3.1 Principes méthodologiques

Notre petit corpus garantit des conditions pragmatiques d'usage communes telles que le contexte juridique de référence des textes, le contexte d'usage sociolinguistique des deux langues juridiques ainsi que le statut respectif des deux langues considérées dans le contexte de production.

Pour mener à bien cette étude, nous mettons à profit la double visualisation du texte en français et en grec. Grâce à l'alignement des deux versions, celle-ci « représente une ressource linguistique privilégiée pour l'observation, la description et la comparaison » (Escoubas-Benveniste, 2013, p. 148) des structures lexico-syntaxiques, à savoir termes juridiques, biomédicaux et économiques ainsi que les phraséologies respectives, dans ces deux versions équivalentes dans le cadre du caractère interdisciplinaire des deux textes.

Pour étudier de manière efficace le comportement des structures lexico-syntaxiques, nous aurons recours au concordancier AntConc (version 3.4.3W). Cela afin de créer des concordanciers dans les deux langues qui nous aideront à procéder à des comparaisons pour voir d'abord s'il y a conséquence et à quel degré quant à la terminologie employée au sein de chaque texte (comportement des termes monolexicaux/polylexicaux) et par la suite vérifier la conséquence des équivalences entre les deux versions (Figure 1).

Fig. 1. Concordances du terme *mise*

3.2 Étude du matériau et résultats

L'emploi de concordanciers nous a donné la possibilité de procéder à des constatations qui sont présentées ci-après de façon concise, suivies d'exemples à titre indicatif.

3.2.1 Texte français

De nombreux termes utilisés dans des domaines différents sont sémantiquement clarifiés, ce qui est très important pour les étudiants, qui, pour la plupart, n'ont pas de connaissances thématiques.

Les concordanciers nous permettent de distinguer le sens de chaque structure tant au niveau grammatico-syntaxique qu'au niveau lexico-syntaxique.

a) En matière de termes monolexicaux :

Le mot *rapport* est rencontré aussi bien au singulier dans *par rapport à* qu'au pluriel *rapports* désignant les documents européens. *Administration* est employé uniquement en tant que terme médical.

b) En matière de termes polylexicaux :

L'usage fréquent de *mise* montre que, à plus de 80% des occurrences, ce mot participe à *mise sur le marché*, ce qui témoigne de la fréquence élevée de ce terme polylexical ; à son tour, *mise sur le marché* fait partie de *autorisation de mise sur le marché* et de *autorisation communautaire de mise sur le marché*. Le terme *effets* – toujours au pluriel – est le plus souvent suivi de *indésirables* qui est parfois accompagné de *graves* (*effets indésirables graves*). Parfois nous trouvons *nocifs* au lieu de *indésirables* sans aucun autre complément.

3.2.2 La recherche de l'équivalent en langue cible

Pour mieux aborder les termes polylexicaux français et grecs au niveau de la traduction, nous présentons les modèles de base respectifs.

a) Les termes polylexicaux français ont comme base une structure bilexicale qui peut comprendre :

i) nom + adjectif : *autorité compétente, globuline antivariolique, données expérimentales*

ii) nom + préposition (+ article) + nom : *distribution en gros, mise sur le marché*

iii) nom + nom : *État membre* ou rarement :

iv) nom + verbe : *précautions à prendre* (Scarpa, 2010, p. 193-199).

Les termes polylexicaux (à trois mots et plus) proviennent des combinaisons variées des séquences mentionnées ci-dessus: *médicaments à usage humain, concentration en principes actifs, produits dérivés du sang humain.*

b) La typologie de base pour les termes polylexicaux grecs étudiés est la suivante (Anastassiadis-Symeonidis, 1986, p. 147) :

1) adjectif + nom : *φαρμακευτικά ιδιοσκευάσματα, ψυχοτρόπα φάρμακα, ραδιενεργά ισότοπα, πειραματικά δεδομένα*

2) nom + Ø / article défini + nom en cas génitif : *γεννήτρια ραδιονουκλεϊδίων, εμβόλιο της ευλογιάς.* Le type nom + nom tel que *Κράτος μέλος*, plutôt rare, est emprunté au français.

Les termes polylexicaux (à trois mots et plus) proviennent des combinaisons variées des séquences mentionnées ci-dessus parfois avec l'usage d'un verbe ou d'une préposition : *φάρμακα που προορίζονται για ανθρώπινη χρήση, περιεκτικότητα σε δραστικές ουσίες, παράγωγα του ανθρωπίνου αίματος, ανοσοσφαιρίνες ανθρώπινης προέλευσης*.

Les exemples qui suivent, obtenus au moyen de concordanciers de façon comparative, illustrent la multiplicité ou la stabilité du rendement aussi bien des termes que des locutions.

Même s'il n'est pas facile de séparer l'aspect grammatico-syntaxique de l'aspect lexical, pour des raisons de clarté quant aux particularités des deux textes, nous avons essayé de regrouper nos remarques en deux catégories. Les exemples présentés sont exclusivement à titre indicatif.

I) au niveau grammatico-syntaxique notamment :

Variabilité de rendement de certaines structures avec ou sans changement de catégorie grammaticale, de nombre, de voix ou même avec une traduction libre :

visant à garantir le…	για τη διασφάλιση του…
visant à établir la sécurité…	που αποβλέπουν στην εξακρίβωση της ασφάλειας…

qui fait l'objet d'une demande…	για το οποίο έχει επίσης υποβληθεί αίτηση…
a fait l'objet, dans l'État membre importateur, d'une analyse qualitative complète…	έχει υποστεί στο κράτος μέλος εισαγωγής πλήρη ποιοτική ανάλυση…
de faire l'objet de risques importants d'abus médicamenteux…	να αποτελεί σημαντικό κίνδυνο φαρμακευτικών καταχρήσεων…

De plus, il y a eu des exemples de non-traduction en grec :

à l'égard de	---
à cet égard	---

II) au niveau lexical notamment :

Variabilité de rendement de certains termes monolexicaux ainsi que de certains termes polylexicaux en partie ou en totalité :

médicaments radiopharmaceutiques	ραδιοφαρμακευτικά προϊόντα / ραδιοφάρμακα
échanges de médicaments	εμπορία των φαρμάκων / συναλλαγών επί των φαρμάκων
échanges de substances thérapeutiques d'origine humaine	εμπόριο θεραπευτικών ουσιών ανθρώπινης προέλευσης

ou même une traduction libre :

Les États membres prennent toutes dispositions utiles pour que les…	Τα κράτη μέλη εξασφαλίζουν ότι οι…

Par ailleurs, nous avons relevé un cas de néologisme en grec :

…firmes innovatrices…	…καινοτομουσών εταιρειών…

Il est à signaler que seuls les termes monolexicaux et polylexicaux désignant les institutions et les documents juridiques européens demeurent stables tant dans le texte français que dans le texte grec :

Conseil de l'Europe	Συμβούλιο της Ευρώπης
…directive 80/836/Euratom du Conseil du 15 juillet 1980	…οδηγίας 80/836/Ευρατόμ του Συμβουλίου της 15ης Ιουλίου 1980

Les exemples présentés montrent la multiplicité et la complexité des cas, qui se répercutent sur la prise de décision quand il s'agit du choix de l'équivalent pertinent en langue cible. En outre, il est évident que les concordanciers mettent en exergue le rôle du contexte.

4. DISCUSSION ET PROPOSITIONS

Ayant mis en valeur des textes parallèles pour un apprentissage sur corpus et ayant employé des concordanciers afin d'étudier la terminologie et la phraséologie en contexte, nous avons constaté que celles-ci sont sujettes à des variations et à des combinaisons aussi bien entres elles qu'avec la langue générale.

Au niveau monolingue, l'emploi de concordanciers au sein du texte français, nous a donné la possibilité, en premier lieu, d'étudier aisément les termes et la phraséologie dans un environnement où se rencontrent plusieurs lexiques spécialisés reflétant ainsi leur caractère interdisciplinaire et de porter l'accent sur l'axe paradigmatique de chaque terme. Nous avons ainsi eu l'occasion d'observer la stabilité, la variabilité et l'alternance éventuelles des structures en question et de relever par la suite la terminologie et le lexique spécialisé de chaque domaine.

Pour passer au niveau contrastif, l'alignement des deux versions nous a facilité la tâche du repérage des équivalents grecs. L'utilisation comparative de concordanciers nous a permis d'identifier à chaque fois la stratégie de traduction (Loock, 2016, p. 165), ce qui pourrait s'avérer précieux pour les cours de langue de spécialité et de traduction spécialisée. Malgré la grande complexité des cas étudiés, nous n'avons pas repéré, à une exception près, de choix malheureux qui auraient embrouillé les étudiants.

De plus, du point de vue pédagogique, l'intérêt d'un apprentissage sur corpus réside dans la concentration artificielle de structures cibles pour faciliter l'analyse et l'apprentissage de celles-ci. Grâce à la version systématisée par le concordancier des structures grammatico-syntaxiques et lexico-syntaxiques, les étudiants en traduction spécialisée auront la possibilité de réfléchir sur la complexité du fonctionnement des termes et de la phraséologie monolingues ou multilingues. Ajoutons ici que l'exploitation pédagogique de la triple visualisation comportant l'anglais, du fait que cette langue fait partie de la majorité des programmes LEA, serait fructueuse, car la plupart des textes européens sont d'abord rédigés en anglais.

Enfin, les étudiants profiteront d'une approche très efficace grâce à l'utilisation des corpus parallèles et des concordanciers qui favorisent l'apprentissage autonome.

Bibliographie

BIEL, Łucja : Quality in institutional EU translation: Parameters, policies and practices. In : Quality aspects in institutional translation. Eds. T. Svoboda – Ł. Biel – K. Łoboda. Berlin : Language Science Press 2017, pp. 31–57.

BOURIGAULT, Didier – SLODZIAN, Monique : Pour une terminologie textuelle. In : Terminologies nouvelles, 1999, No 19, pp. 29–32.

CAO, Deborah : Translating Law. Clevedon/Buffalo/Toronto : Multilingual Matters 2007.

Concordancier AntConc3.4.3w. Disponible sur : https://www.laurenceanthony.net/software/antconc/

ESCOUBAS-BENVENISTE, Marie-Pierre : Prédicats juridiques et schémas d'arguments dans les textes des arrêts de la Cour. Approche bilingue français-italien. In : La traduction juridique : Points de vue didactiques et linguistiques. Eds. M. Meunier – M. Charret-Del Bove – E. Damette. Publications du CEL, 2013, pp. 141–166.

GOFFIN, Roger : L'eurolecte : oui, jargon communautaire: non. In : Meta, 1994, Vol. 39, No 4, pp. 636–642. Disponible sur : https://www.erudit.org/en/journals/meta/1994-v39-n4-meta185/002930ar.pdf

GRIEBEL, Cornelia : Textes et discours juridiques: aspects cognitifs et traductologiques. In : Manuel des langues de spécialité. Eds. W. Forner – B. Thörle. Berlin/Boston : Walter de Gruyter 2016, pp. 205–226.

Guide pratique commun du Parlement européen, du Conseil et de la Commission à l'intention des personnes qui contribuent à la rédaction des textes législatifs de l'Union européenne. Union européenne. Disponible sur : https://eur-lex.europa.eu/content/techleg/FR-guide-de-redaction-legislative.pdf

JOHNS, Tim : Should you be persuaded: two examples of data-driven learning. In : Classroom Concordancing. Eds. T. Johns – P. King. In : English Language Research Journal, 1991, No 4, pp. 1–16.

LOOCK, Rudy : La traductologie de corpus. Université de Lille 2016.

MAC AODHA, Máirtín : Lexicographie, traduction et langues minoritaires : le cas de l'irlandais au sein de l'Union européenne. Thèse de doctorat. Université de Strasbourg 2018. Disponible sur : https://tel.archives-ouvertes.fr/tel-02081060/document

MICHEL, Hélène : La transparence dans l'Union européenne : réalisation de la bonne gouvernance et redéfinition de la démocratie. In : Revue française d'administration publique, 2018, Vol. 1, No 165, pp. 109–126.

NORD, Christiane : Text Analysis in Translation. Theory, Methodology and Didactic Application of a Model for Translation-Oriented Analysis. Second Edition. Amsterdam & New York : Rodopi 2005.

ŠARČEVIĆ, Susan : Legal Translation and Translation Theory: A Receiver-oriented Approach. In : La traduction juridique, Histoire, théorie(s) et pratique. Ed. J.-C. Gémar. Université de Genève 2003, pp. 329–347.

SCARPA, Federica : La traduction spécialisée, Une approche professionnelle à l'enseignement de la traduction. Ottawa : Les Presses Universitaires d'Ottawa 2010.

SCHAEFFER-LACROIX, Eva : Corpus numériques et production écrite en langue étrangère. Une recherche avec des apprenants d'allemand. Thèse de doctorat. Université de la Sorbonne nouvelle – Paris III 2009. Disponible sur : https://tel.archives-ouvertes.fr/tel-00439095/document.

SOSONI, Vilelmini : Training translators to work for the EU institutions: luxury or necessity? In : The Journal of Specialised Translation, 2011, No 16, pp. 77–108. Disponible sur : https://www.jostrans.org/issue16/art_sosoni.pdf

STEINBERGER, Ralf – EBRAHIM, Mohamed – POULIS, Alexandros et al. : An overview of the European Union's highly multilingual parallel corpora. In : Language Resources & Evaluation, 2014, Vol. 48, No 4, pp. 679–707. Disponible sur : https://doi.org/10.1007/s10579-014-9277-0

TREBITS, Anna : English lexis in the documents of the European Union : A corpus-based exploratory study. In : Working Papers in Language Pedagogy. Eotvos Lorand University, 2008, Vol. 2, pp. 38–54.

ZANETTIN, Federico : Corpus-based translation activities for language learners. In : The Interpreter and Translator Trainer (ITT), 2009, Vol. 3, No 2, pp. 209–224.

ΑΝΑΣΤΑΣΙΑΔΗ-ΣΥΜΕΩΝΙΔΗ, Άννα : Η νεολογία στην κοινή νεοελληνική. Διδακτορική διατριβή. Θεσσαλονίκη: ΑΠΘ, Επιστημονική επετηρίδα της Φιλοσοφικής Σχολής, παρ. 1986, αρ. 65. [Anastassiadis-Symeonidis, Anna : La néologie en grec standard. Thèse de doctorat. Université Aristote de Thessalonique 1986].

sciendo
DOI 10.2478/jazcas-2022-0019

PROPOSITION D'EXPLOITATION DU CORPUS D'ÉTUDE POUR LE FRANÇAIS CONTEMPORAIN EN DIDACTIQUE DU FLE

FANNY LAFONTAINE
Université Palacký d'Olomouc, Olomouc, Tchéquie[1]

LAFONTAINE, Fanny: Proposal to use the study corpus for contemporary French in Didactics of French as a Foreign Language. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 3, pp. 951 – 966.

Abstract: The ORFÉO platform (Tools and Research on Written and Oral French) has been making available to users since 2018 a Study Corpus for sampled Contemporary French as well as operating tools. Although this resource is intended for an audience of researchers and students in the fields of linguistics and automatic language processing, we endeavor in this article to report on the didactic potential that it offers within the framework of a Licensing Syntax course treating "subordination" and intended for Czech and Slovak students at levels B1 to C1 in French. We propose a didactic sequence composed of four activities and pursuing three objectives: consolidation of the mastery of the basic functions of *dont* («which») from a corpus of friendly conversations; the use of simple query interface tools and the introduction of certain principles of corpus sociolinguistics. The corpus-based approach, by confronting learners with authentic contextualized data, helps to redefine the teaching-learning priorities of a language by giving primacy not to respect for grammatical norms but to genre norms.

Keywords: Didactics of French as a Foreign Language; Data-driven learning; corpus linguistics; "dont"; sociolinguistics.

0. INTRODUCTION

Le français institutionnellement enseigné/appris en tant que langue étrangère (mais aussi L1) correspond à une de ses variétés, la variété cultivée, qui tire sa reconnaissance du « traitement socio-culturel qui en a été fait par une certaine élite des voix » au cours des siècles (Besse, 2001, p. 47). Bien que les manuels d'inspiration communicative donnent à apprendre un éventail des variétés de la langue française, ils ne font que remplacer la référence à la variété écrite par une référence à une oralité franco-normée : « ce qui est donné à apprendre reflète moins la diversité des usages réellement pratiqués par les francophones qu'une certaine représentation de cette diversité » (ibid.).

Afin d'exposer les apprenants à un input authentique et varié en langue cible, émergent depuis les années 1990, en complément à l'utilisation des manuels de

[1] La présente publication a été financée par le ministère tchèque de l'Éducation, de la Jeunesse et des Sports (IGA_FF_2021_022).

langue, des propositions d'exploitation des outils de la linguistique de corpus à des fins didactiques. Les corpus sont, en premier lieu, exploités pour prélever des données audio et/ou textuelles utilisées comme support authentique, à partir desquelles l'enseignant construit des activités de compréhension et de production visant à l'acquisition de compétences communicatives, linguistiques et socio-culturelles, à l'image des ressources produites par le projet CLAPI-FLE[2]. L'autre mode d'exploitation des corpus, désigné « approche sur corpus » (désormais ASC), consiste à recourir au concordancier (cf. §1.3.) pour enseigner/apprendre un fait langagier précis à partir de l'observation et de l'analyse de son fonctionnement en contexte (Boulton, 2018 ; Di Vito, 2018).

Cette nouvelle méthodologie améliorerait la compétence sociolinguistique des apprenants, composante essentielle de la compétence de communication, en favorisant, par rapport à un support traditionnel, le « développement d'une production plus naturelle » (Boulton, 2018, p. 76), c'est-à-dire plus conforme aux pratiques langagières réelles des locuteurs/scripteurs de la langue cible en fonction des situations extralinguistiques.

Avant de proposer une illustration des potentialités de l'ASC en milieu universitaire, nous présentons brièvement la plateforme ORFÉO[3] (Outils et Recherches sur le Français Écrit et Oral) qui permet d'interroger le Corpus d'Étude pour le Français Contemporain (CEFC), que nous avons détourné à des fins didactiques.

1. PRÉSENTATION DE LA PLATEFORME ORFÉO

L'objectif du projet ORFÉO est de permettre aux chercheurs dans les domaines de la linguistique et du traitement automatique des langues de mener différentes études comparatives sur des données de genres variés, de façon à élaborer une grammaire des usages du français à partir d'une approche *corpus driven*[4]. En accès libre depuis 2018, la plateforme ORFÉO offre un corpus de référence de français parlé et écrit – composé respectivement de quatre et six millions de mots, échantillonné et enrichi par des données secondaires de diverses natures – ainsi que des outils d'interrogation.[5]

[2] Cet usage des corpus est utile notamment pour répondre à des besoins langagiers précis de certains types d'apprenants, besoins non représentés dans les méthodes de langue, tels que la maîtrise de l'oral et de l'écrit académiques pour un public d'étudiants.

[3] Disponible sur : https://www.projet-orfeo.fr/

[4] Comme l'indiquent Deulofeu et Debaisieux (2012), la linguistique de corpus « aboutit à substituer à la conception d'une grammaire unique pour une langue celle de grammaires multiples rendant compte des faits observés dans des usages écrits et oraux diversifiés en fonction de situations de production, allant des plus spontanées aux plus élaborées » (p. 33).

[5] Se reporter à Benzitoun *et al.* (2016) pour une présentation détaillée du projet.

1.1 Échantillonnage du corpus

Les ressources orales sont définies selon cinq variables de la situation de communication :

- type (entretien, réunion, transaction, conversation, narration, médias oraux, cours, discours, présentation publique, repas, explication, consultation, activité) ;
- secteur (professionnel, privé) ;
- milieu (amical, académique, scolaire, commercial, familial, associatif, politique, médical, religieux, sportif) ;
- nombre de locuteurs ;
- situation de l'enregistrement (face à face, en public, téléphone, télévision, radio).

Quant au corpus écrit, quatre genres discursifs (composés de 1,5 million de mots chacun) sont représentés, littérature, presse, écrit scientifique et écrit non planifié, lesquels donnent lieu à une répartition en dix sous-genres : presse quotidienne nationale, presse quotidienne régionale, communication, article, thèse, HDR, actes congrès, chapitre d'ouvrage, clavardage et ouvrage.

1.2 Données secondaires

Chaque échantillon de langue orale et écrite est accompagné de données secondaires pouvant, dans l'analyse, être mises en relation avec les formes linguistiques produites. Celles-ci comprennent :

- une identification du corpus source ;
- des métadonnées générales (soit des informations relatives à l'échantillonnage des données) ;
- le texte ou une transcription alignée texte/son ;
- une représentation de chaque unité « élémentaire » de texte sous la forme d'un arbre syntaxique (comprenant un étiquetage en catégories morphosyntaxiques et des annotations en dépendances syntaxiques) ;
- des fichiers texte/son téléchargeables dans un fichier. zip.

Les métadonnées générales des données orales incluent en outre un résumé du contenu de l'échantillon, la date et la durée de l'enregistrement, ainsi que des indications sur la qualité du son. Enfin, chaque locuteur est identifié à l'aide de métadonnées sociologiques indiquant son âge, son sexe, son niveau d'études, son lieu de naissance et sa profession.

1.3 Outils d'exploitation du corpus

Deux outils d'exploitation du corpus sont proposés : la recherche simple et la recherche avancée[6]. Nous n'aborderons ici que la première, celle utilisée lors de notre séquence didactique.

[6] Cette interface permet d'effectuer des requêtes sur une catégorie grammaticale ou sur une fonction syntaxique, dans l'ensemble du corpus ou sur un corpus spécifique.

L'interface de recherche simple offre la possibilité de constituer un corpus de travail à partir d'un des 18 corpus sources ou des métadonnées générales. Le corpus ainsi défini peut être exploité en recourant au logiciel de concordance, qui permet de rechercher une chaîne de caractères (mot ou expression). Le résultat de la requête est affiché sous la forme de lignes de concordance[7], où le mot-clé, au centre, est entouré de son cotexte linguistique (jusqu'à 20 mots de part et d'autre). La sélection du mot-clé donne accès à l'affichage, dans une nouvelle fenêtre, des métadonnées et du texte suivi, à l'intérieur duquel le mot recherché est aisément identifiable. La recherche par concordance permet également la visualisation d'un diagramme circulaire présentant des indications statistiques sur la fréquence d'utilisation du mot-clé dans les différents corpus source.

2. ILLUSTRATION DE L'ASC

La séquence présentée ici s'adresse à un public d'étudiants tchèques et slovaques en formation de philologie française et prend place dans le programme d'un cours de syntaxe de Licence portant sur la notion de « subordination ». Elle s'intéresse aux usages de « dont » en français spontané, dont l'observation et l'analyse ont été circonscrites à ses attestations dans un corpus de conversations amicales[8] composé de 44 occurrences[9]. Aucune occurrence de ce sous-corpus n'a été écartée, qu'il s'agisse des occurrences non normatives de « dont » ou de celles d'hésitation et d'alternance avec d'autres pronoms relatifs. Ces données fournissent aux étudiants une représentation fidèle de la variété de ses usages selon cette situation de parole particulière, qui se caractérise notamment par l'émergence d'un discours moins soumis à la *pression normative*.

Nous abordons à la suite les finalités et le déroulement de cette séquence[10].

2.1 Objectifs didactiques

La séquence proposée réunit les trois points de vue didactiques mentionnés par Fligelstone (1993) concernant l'utilisation des corpus : *exploiting to teach* (le corpus comme support d'enseignement), *teaching to exploit* (le corpus pour enseigner la langue) et *teaching about* (le corpus comme objet d'enseignement).

[7] Les concordances peuvent être téléchargées.

[8] Le postulat sous-jacent qui préside au choix de ce sous-corpus est que les structures les plus représentatives du système de la langue s'actualisent en français non planifié : celles-ci constituent « une base commune à tous les locuteurs élaborée dans la phase d'acquisition en milieu naturel de la langue maternelle » (Présentation du Projet Orféo).

[9] Les 44 occurrences de « dont » se distribuent comme suit : 36 occurrences d'emploi normatif décrites au § 2.2.2 ; cinq occurrences non normatives décrites au § 2.2.3 ; une occurrence d'hapax non standard (« les États-Unis c'est côté super artificiel dont avec les chirurgies esthétiques ») ; et deux occurrences d'alternance/hésitation avec un autre pronom relatif (« c'est la seule île qui qui dont la sur laquelle il y a il y a rien quoi »).

[10] En raison de la situation sanitaire, cette séquence n'a pas encore été mise en place.

S'agissant d'une forme qui pose des difficultés d'apprentissage tant en FLM[11] qu'en FLE, il est question tout d'abord que les étudiants consolident leur maîtrise des emplois les plus représentatifs de « dont », en adoptant une démarche inductive partant de l'interprétation des lignes de concordance pour faire ressortir ses régularités de fonctionnements.

Cette séquence vise, d'autre part, à former les étudiants à manier l'interface de requête simple de la plateforme ORFÉO (par l'interrogation de la fréquence d'une forme sur l'ensemble des données orales et écrites, la constitution d'un sous-corpus d'étude, l'utilisation du logiciel de concordance et des métadonnées des locuteurs) pour que ceux-ci soient en mesure de consulter librement les corpus sur d'autres faits langagiers.

Le dernier objectif didactique, qui consiste à enseigner certains principes de la (socio) linguistique de corpus, cherche en premier lieu à faire « émerger une conception non normative de la grammaire qui devient description des faits de langue en contexte » (Di Vito, 2018, p. 53), et ce en s'appuyant sur la partition fondamentale pour une langue opposant les « énoncés grammaticaux », qu'ils soient standard ou non, aux « énoncés agrammaticaux ». Les premiers se caractérisent par leur fréquence et leur régularité syntaxique, tandis que les seconds constituent une erreur de performance dont la reproductibilité est nulle[12] (Deulofeu – Valli, 2007). En second lieu, il s'agit d'aborder les phénomènes de variation, en termes non pas uniquement de variantes de la langue normalisée, mais d'exploitation différenciée des ressources linguistiques selon les divers genres oraux et écrits[13].

2.2 Scénario didactique

La séquence, d'une durée d'environ trois heures, s'organise autour de quatre activités déclinées à la suite.

2.2.1 Analyse quantitative de données

Lors d'une première activité, les étudiants sont invités à relever la fréquence de « dont » et « duquel » (et de ses variantes morphologiques) à l'oral et dans les quatre sous-genres de l'écrit, de façon à produire une analyse quantitative comparative.

Voici le relevé de ces formes dans l'ensemble du corpus :

[11] « Les difficultés que rencontrent les maîtres pour […] enseigner l'usage [de ''dont'' et ''lequel''] à l'école primaire montrent bien que ces formes ne sont pas intégrées dans la connaissance première que les enfants ont de la grammaire. » (Blanche-Benveniste, 2010, p. 101)

[12] Le corpus présente une occurrence d'hapax non standard : « les États-Unis c'est côté super artificiel dont avec les chirurgies esthétiques ».

[13] Pour quelques exemples de descriptions fines de ces variations, voir notamment Benzitoun, Corminboeuf et Cappeau, 2017 ; Blanche-Benveniste, 1997 ; Cappeau, 2016.

	Oral	Écrits non planifiés	Littérature	Presse	Écrits scientifiques	Total
« Dont »	766	596	2 447	1 694	1 396	6 899
« Duquel »	8	11	49	37	35	140
« De laquelle »	9	5	40	39	36	129
« Desquels »	5	3	31	11	19	69
« Desquelles »	3	1	24	12	20	60
Total	25	20	144	99	110	398

Tab.1. Répartition de « dont » et « duquel » (et ses variantes morphologiques) dans l'ensemble du corpus

On observe ainsi que « dont » est en moyenne cinq fois plus fréquent à l'écrit qu'à l'oral (pour six millions de mots, 1 149 occurrences orales contre 6 133 occurrences écrites). À l'écrit, c'est dans la littérature que « dont » est le plus fréquent (40 % des occurrences), suivie de la presse (27 %), de l'écrit scientifique (23 %) et de l'écrit non planifié (10 %).

« Duquel » (et ses variantes) est dix fois plus utilisé à l'écrit qu'à l'oral (pour six millions de mots, 37,5 occurrences orales contre 373 occurrences écrites). À l'écrit, des variations d'utilisation de « duquel » se dessinent selon les genres discursifs : littérature (38,6 %), écrit scientifique (29,5 %), presse (26,5 %) et écrits non planifiés (5 %).

Il ressort de cela que ces deux formes apparaissent majoritairement dans les écrits planifiés (75 % de l'ensemble des emplois de « dont » et 85 % de ceux de « duquel »). Toutefois, « dont » est 30 fois plus fréquent que « duquel » à l'oral, et seize fois plus à l'écrit. Sa plus forte fréquence s'explique par sa plus grande diversité de fonctionnements. « Dont » appartient au registre courant, bien qu'en français parlé et dans l'écrit non planifié, les locuteurs et scripteurs l'utilisent peu (et principalement dans des tournures formulaires, comme nous le verrons). « Duquel », du fait de sa rareté à l'écrit et de sa quasi-inexistence dans les données orales et écrites non planifiées, relève quant à lui du discours élaboré.

L'analyse de la répartition de ces deux formes peut être complétée par d'autres faits d'observation, de nature qualitative, étayant leur caractère non interchangeable : Blanche-Benveniste (2010) remarque – sur la base d'un corpus de deux millions de mots de français parlé composé de conversations familières et de situations de parole publique – que les occurrences de « duquel » (et ses variantes morphologiques) « ne font pas concurrence à ''dont'' et semblent privilégier le domaine sémantique de l'intériorité, ''au sein duquel'', ''à l'intérieur de laquelle'' : ''le cabinet au sein duquel on travaille'' » (p. 103)[14].

[14] Nous constatons que les étudiants tchèques et slovaques, dans des exercices de langue traitant des éléments relatifs, notamment ceux « à trous », ont tendance à produire « duquel » dans des contextes où l'on attendrait *naturellement* « dont », comme dans « l'animal duquel je t'ai parlé ».

2.2.2 Analyse des usages normatifs de « dont » en français spontané

Durant la seconde activité, les étudiants proposeront un classement ordonné des occurrences d'emplois normatifs de « dont » représentées dans le corpus et rendront compte des éléments lexicaux les plus fréquents dans l'environnement de cette forme. L'analyse est basée sur 36 lignes de concordances préalablement distinguées par l'enseignant[15].

Le classement de ces occurrences montre que cette forme se partage entre deux types d'emplois, d'inégale importance : un emploi de relatif (où « dont » exerce différentes fonctions de complément d'un syntagme prépositionnel à l'intérieur d'une relative) et un emploi d'adverbe :

Usages de « dont »	Numéro des occurrences	Exemples et/ou faits remarquables	Nombre d'occurrences
Emploi relatif			
- **Dépendance verbale**			
1) Objet indirect	(5), (9), (14), (15), (16), (19), (21), (22), (25), (27), (30), (31), (34), (37), (40), (41)	10 occ. « parler de » 2 occ. « s'occuper de » 1 occ. « rêver de », « être issu de », « se souvenir de », « avoir horreur de »	16 (44,4 %)
2) Complément de manière	(8), (17), (18), (28), (29), (43)	« La façon/manière dont : « la façon/manière de »	6 (16,7 %)
- **Dépendance nominale**			
3) Complément du nom	(7), (10), (12), (13), (24), (32), (33), (39)	« un Méditerranéen dont la femme s'en va » : « la femme de »	8 (22,2 %)
Emploi adverbial Syntagme verbal *dont* syntagme nominal	(1), (2), (3), (4), (11), (23)	« Il a encore plein de contacts dont le toubib »	6 (16,7 %)

Tab.2. Classement des emplois normatifs de « dont » dans le sous-corpus de conversations amicales

Lorsque « dont » est sous la dépendance d'un verbe, il est remarquable qu'il se situe dans presque la moitié des cas dans la valence du verbe « parler de » (ce qui représente par ailleurs le quart de l'ensemble des occurrences du corpus). Le paradigme de verbes peut être complété par les observations de Blanche-Benveniste

[15] Se reporter à l'annexe pour visualiser l'ensemble de la liste des résultats.

(2010) : les verbes les plus fréquents admettant un Objet indirect, outre « parler de », sont « avoir besoin de », « se passer de », « faire partie de », « se souvenir de », « s'occuper de » et « se servir de » (p. 102).

En outre, dans un peu plus de 15 % des cas, « dont » est restreint à des tournures stéréotypées faisant intervenir les deux lexèmes « façon » et « manière ». Il se combine alors avec n'importe quel verbe admettant un complément de manière, qu'il s'agisse d'un complément circonstanciel (« la façon dont tous les personnages sont en lien ») ou d'un Objet (« la façon dont ça se termine »).

Lorsque « dont » est dépendant d'un nom, certaines tendances se dégagent également : le nom désigne une relation de parenté (comme dans l'exemple ci-dessus) ; on relève d'autre part une occurrence de la tournure formulaire « dont je tairai le nom ».

En ce qui concerne les emplois tels que « il a encore plein de contacts dont le toubib »[16], qui totalisent un peu plus de 15 % de l'ensemble des occurrences normatives, « dont » ne peut être considéré comme un élément relatif introduisant une proposition subordonnée, mais plutôt comme un adverbe, commutable avec « notamment ». Dans cet emploi, on observe que le syntagme nominal qui précède « dont » est pourvu d'un déterminant (simple ou complexe) indéfini de quantité (comme « quelques », « plein de », « plusieurs » ou un numéral) et que « dont » sert à prélever une certaine quantité de ce syntagme nominal.

Il apparait ainsi que dans le discours non planifié les emplois normatifs les plus fréquents de « dont » se concentrent principalement sur quelques éléments de lexique[17] : on remarque que tantôt c'est le lexique de l'élément recteur qui est déterminant (« parler » du côté des verbes ; indication d'un lien de parenté, du côté des noms) ; tantôt c'est celui du terme régi (« de cette manière »/« de cette façon »).

2.2.3 Analyse des usages non normatifs de « dont » en français spontané

À partir des lignes de concordance mises de côté, les étudiants mettront en évidence les régularités de fonctionnement des usages non normatifs de « dont », qui constituent un peu plus de 10 % de l'ensemble des occurrences du corpus. Comme l'indique, en effet, Delabre (1995), « l'utilisation de "dont", au-delà de la règle de base, n'est pas un phénomène anarchique, mais elle est à la fois la manifestation d'un changement linguistique et le signe de l'émergence de nouvelles régularités » (p. 7).

L'analyse de ces occurrences laisse apparaître deux cas[18].

---

[16] Cet emploi est rarement mentionné dans les grammaires de référence. Sa description serait à affiner sur la base d'un plus grand nombre d'occurrences.

[17] La fréquence des phénomènes relevés ici est relativement semblable à celle observée sur de plus vastes corpus de français parlé (Blanche-Benveniste, 2010).

[18] Toutes les possibilités d'emplois non normatifs de « dont » ne sont pas représentées par ce corpus. Se reporter à Delabre (1995) pour d'autres exemples, notamment ceux où « dont » est complément d'un syntagme prépositionnel : « à noter aussi un bel hommage à R. Barthes, *dont* c'est le dixième anniversaire *de la mort* » (p. 5).

Soit « dont » est complément d'un verbe qui ne construit pas de complément en « de »[19] :

(1) C'est ça ce dont j'étais en train de *penser*
(2) le truc dont *je me rappelle* c'est qu'il a le ventre qui gonfle[20]
(3) il nous vient plein d'idées dont tu peux en *choisir* une

Soit la relation d'appartenance qu'indique « dont » est exprimée une deuxième fois dans la proposition relative, que ce soit avec le clitique « en » (4) ou avec un syntagme prépositionnel auquel « dont » est coréférent (5), raison pour laquelle ces relatives sont dites « pléonastiques » :

(4) un support dont on peut pas complètement s'*en* passer
(5) on a mis en place des permanences psychologiques avec une association qui s'appelle NNAAMMEE et dont elle est donc stagiaire *de cette association*

2.2.4 Identification des facteurs influençant la variation des usages

Les étudiants, après avoir pris connaissance des caractéristiques sociologiques des locuteurs ayant produit les occurrences non normatives de « dont », formuleront leurs hypothèses sur les facteurs externes qui organisent cette variation[21].

Métadonnées : locuteur MD	
Identiant du locuteur	MD
Âge du locuteur	21-60
Sexe du locuteur	F
Profession du locuteur	étudiant
Niveau d'études du locuteur	études supérieures
Lieu de naissance du locuteur	France, PACA, Marseille

Fig.1. Exemple de métadonnées d'un locuteur dans le corpus

Au regard de l'ensemble de ces informations, il apparait que, dans le corpus proposé, les formes produites ne peuvent être mises en relation avec les usagers, mais avec la spécificité de la situation de communication : les locuteurs ont tous un niveau d'études supérieures − ce qui va à l'encontre des représentations courantes où les *infractions* ou *déviances* à la norme sont fréquemment associées à un manque de scolarisation des locuteurs – et ils ne présentent pas de caractéristiques homogènes quant à leur milieu socio-économique, leur âge et leur situation géographique.

[19] Parallèlement à cela, « que » peut être employé là où la norme attendrait « dont », même avec des verbes fréquents comme « parler de » ou « avoir besoin de » : « je voulais faire un stage de formation que j'avais besoin » (Blanche-Benveniste, 2010).

[20] Notons que la construction « se rappeler de » semble devenir l'usage dominant en français contemporain.

[21] Ces paramètres extralinguistiques distinguent entre les variations de la langue selon les différents usagers (variation diachronique, diatopique et diastratique) et selon l'usage qu'en fait chacun (variation diaphasique) (Gadet, 1989).

3. CONCLUSION

Nous nous sommes attachée, dans cet article, à donner une illustration dans le milieu universitaire du potentiel didactique de l'exploitation des outils de la linguistique de corpus, en l'occurrence des ressources du projet ORFÉO, qui donnent accès à un corpus échantillonné de dix millions de mots représentatifs pour l'observation de la langue des locuteurs natifs, de la langue d'une époque et de la langue d'un genre. Le recours à de telles données semble aujourd'hui indispensable afin de *désartificialiser* l'enseignement/apprentissage du français.

Dans la séquence pédagogique proposée, la consultation de cette ressource par l'apprenant remplit trois objectifs : l'apprentissage d'un fait langagier – dont la maitrise est par ailleurs perçue comme difficile, l'appropriation des principaux outils de l'interface de requête simple et l'acquisition de connaissances sur les méthodes de la sociolinguistique de corpus. Ces deux derniers points, s'ils peuvent concourir à modifier les représentations des étudiants sur la façon d'apprendre et de concevoir une langue, peuvent en outre intéresser ceux s'orientant vers un travail de recherche dans le domaine de la linguistique de corpus. D'autre part, un des intérêts de cette séquence est que son architecture peut être réemployée pour décrire d'autres faits linguistiques s'attachant à concilier les approches syntaxique et sociolinguistique.

Enfin, l'ASC, en confrontant les apprenants à des pratiques langagières contextualisées qui, bien que fréquentes, sont souvent ignorées dans le cadre de l'apprentissage du français, comme la variation des éléments relatifs en français non planifié, redéfinit les priorités d'enseignement/apprentissage d'une langue : il s'agit de mettre l'accent sur l'acquisition des outils linguistiques pour correspondre au genre visé, c'est-à-dire sur les normes de genres, plutôt que sur les normes grammaticales, qui restreignent par ailleurs les usages possibles en système. Cette acquisition passe en outre par l'observation et l'apprentissage des combinaisons lexicales dans l'environnement de la forme étudiée.

Bibliographie

BENZITOUN, Christophe – DEBAISIEUX, Jeanne-Marie – DEULOFEU, Henri-José : Le projet ORFÉO : un corpus d'étude pour le français contemporain. In : Corpus, 2016, No 15, pp. 1–19.

BENZITOUN, Christophe – CORMINBOEUF, Gilles – CAPPEAU, Paul : Réflexions sur les exploitations différenciées de la grammaire. In : Revue de Sémantique et Pragmatique, 2017, No 41, pp. 135–154.

BESSE, Henri : Peut-on naturaliser l'enseignement des langues en général, et celui du français en particulier ?. In : Le français dans le monde, 2001, numéro spécial, pp. 29–57.

BLANCHE-BENVENISTE, Claire : La notion de variation syntaxique dans la langue parlée. In : Langue française, 1997, No 115, pp. 19–29.

BLANCHE-BENVENISTE, Claire : Le français. Usages de la langue parlée. Leuven : Peeters 2010.

BOULTON, Alex : Les corpus et les TIC comme aide à la découverte des langues. In : Le français moderne, 2018, No 3, pp. 71–84.

CAPPEAU, Paul : Questions sur l'oral : médium, syntaxe, genre. In : Le français aujourd'hui, 2016, Vol. 4, No 195, pp. 23–36.

DELABRE, Michel : *Dont* en français contemporain : norme, grammaire et théorie linguistique. In : L'Information Grammaticale, 1995, No 64, pp. 3–8.

DEULOFEU, Henri-José – VALLI, André : Quels faits faut-il retenir pour une description grammaticale satisfaisante ?. In : Les linguistes et la norme, aspects normatifs du discours linguistique, 2007, Neuchâtel : Peter Lang, pp. 87–110.

DEULOFEU, Henri-José – DEBAISIEUX, Jeanne-Marie : Une tâche à accomplir pour la linguistique française du XXIe siècle : élaborer une grammaire des usages du français. In : Langue française, 2012, Vol. 4, No 176, pp. 27–46.

DI VITO, Sonia : Apprendre le FLE en modalité data-driven. In : Le français moderne, 2018, No 3, pp. 51–60.

FLIGELSTONE, Steve : Some reflections on the question of teaching, from a corpus linguistics perspective. In : ICAME journal 17, 1993, pp. 97–109.

GADET, Françoise : Le français ordinaire. Paris : Armand Colin 1989.

Annexe

№	Nom de fichier	Corpus	Contexte gauche	Résultat	Contexte droit
1	unine11a04...	OFROM (O)	a fait presque deux semaines euh avec moi et tout il m'a fait visiter euh euh quelques villes euh	dont	euh ouais dont Tokyo bien sûr et puis justement là j'étais ouais donc il était là il me faisait
2	unine11a04...	OFROM (O)	et puis sinon aussi un inter-rail euh euh en inter-rail euh donc là on c'était pendant dix jours euh	dont	cinq jours où on voyageait en fait et puis là euh je suis allée euh enfin euh à Lubljiana en
3	ffamcv08	CORALROM (O)	que par exemple mes parents bon mon père travaille plus dans ce régiment, mais il a encore plein de contacts	dont	le toubib dont le chauffeur du de du successeur et caetera et donc ils se font inviter à manger ils
4	ffamcv08	CORALROM (O)	mes parents bon mon père travaille plus dans ce régiment, mais il a encore plein de contacts dont le toubib	dont	le chauffeur du de du successeur et caetera et donc ils se font inviter à manger ils invitent des gens
5	28_JD_AL_J...	TUFS (O)	rapport au au à un à un divorce un peu douloureux d'accord donc ça c'est c'est ça	dont	il t'a parlé alors est-ce que tu choi voilà parce qu'il y a eu un conflit avec l'

6	28_JD_AL_J...	TUFS (O)	de ce programme on a mis en place des permanences psychologiques voilà avec une association qui s'appelle NNAAMMEE et	dont	elle est donc stagiaire de cette association donc ouais effectivement on rencontre des gens qui ont à voir avec ce
7	28_JD_AL_J...	TUFS (O)	ouais ouais ouais c'est puis je pense un peu vexé dans son honneur tu sais euh un un méditerranéen	dont	la femme s'en va en lui disant c'est fini je pense qu'il a un peu de mal
8	28_JD_AL_J...	TUFS (O)	tout de suite en libéral quoi ouais et les les associations il y a enfin c'est organisé la façon	dont	des psychologues peuvent intervenir association ouais ouais ouais besoin ben là l'exemple d'une association comme laquelle c'est
9	28_JD_AL_J...	TUFS (O)	c'est le c'-~ moi je bon par profession je m'intéresse au langage euh visiblement la p-~ la personne	dont	on parlait là c'était quelqu'un qui avait des de des des diplômes universitaires si j'ai bien compris
10	28_JD_AL_J...	TUFS (O)	de s'en sortir par rapport à tous les autres hein donc il y en a bien un échec social	dont	on fait porter la culpabilité à l'individu quoi et ça ça ça amène ou la dépression ou la violence
11	PRI-BAY-2	CRFP (O)	cours d'espagnol avec lui parce qu'il en donnait à Bayonne c'était rigolo alors ces ces réfugiés espagnols	dont	beaucoup des basques nous ont amené euh la danse les chants euh moi je me suis inscrit parce que les
12	Raei_leh_s...	TCOF (O)	groupes de combien alors j'ai eu des groupes en fait j'ai fait plusieurs cours, mais euh la classe	dont	j'ai été ouais le plus responsable en fait donc j'avais les cours à vraiment à préparer euh ils
13	ffamcv01	CORALROM (O)	dépendance et l'indépendance synergie et désynchronisation et on peut se sentir comme livrée au hasard victime d'une destinée	dont	l'essentiel nous échappe et donc l'amour ben ça serait une question de de dosage entre intimité distance quoi

14	26_FA_SR_C...	TUFS (O)	qu'on appelle nous le club junior et sur NNAAMMEE euh euh plus euh enfin a-~ appelé euh secteur jeune	dont	NNAAMMEE s'occupe voilà donc donc les jeunes ensuite on travaille aussi beaucoup avec les adultes nous plus les mamans
15	fr12_2005_...	TUFS (O)	a toujours une part de de philosophie dans cette littérature, mais non, mais je vois pas ap-~ après le cours	dont	tu parles bien ouais bon c'est pas grave et puis sinon en littérature comparée au premier semestre j'avais
16	styBM1s	VALIBEL (O)	ne nous envoie rien d'autre hum et alors on on a quand même euh je pense les les endroits	dont	elle parle là pour Naples hum bon bé justement bon écoute moi j'ai oui parce que on j'ai
17	12VLHW1109...	TUFS (O)	là elle se cale comme ça debout et elle se met à man-~ à manger sa carotte, mais la manière	dont	elle était debout et dont elle fixait euh le vide tu sais on aurait dit qu'elle était en train
18	12VLHW1109...	TUFS (O)	ça debout et elle se met à man-~ à manger sa carotte, mais la manière dont elle était debout et	dont	elle fixait euh le vide tu sais on aurait dit qu'elle était en train de bouffer en regardant la
19	32_XC_MB_1...	TUFS (O)	important d'aller voir ses amis que finalement visiter le Japon visiter le Japon pour moi c'est quelque chose	dont dont	je rêve depuis très longtemps et ben que j'ai enfin pu j'ai pu me permettre de le faire
20	10_MD_EM_1...	TUFS (O)	le coup est de savoir si, mais le pire c'est pour bouger les pions ouais c'est ça ce	dont	j'étais en train de penser est-ce que parce que parce que là on va prendre une photo statique ouais
21	10_MD_EM_1...	TUFS (O)	à deux jeux tout d'abord le Mastermind euh, mais on n'est pas sûres d'abord le le Mastermind	dont	on vous a parlé en cours euh et et ou euh le Triomino le Triomino tatatan il y a s
22	10_MD_EM_1...	TUFS (O)	voir pour le projet du second semestre NNAAMMEE et moi avons pensé à deux jeux tout d'abord le Mastermind	dont	on vous a parlé en cours ou le Triomino avec trois grilles différentes prédéfinies et qui représentent trois niveaux d'

23	05_SB_LZ_1...	TUFS (O)	origine c'est devenu connu et maintenant effectivement donc au Canada et en Europe aussi j'ai vu plusieurs versions	dont	la version anglaise évidemment euh le au niveau États-Unis je sais pas en dehors d-~ de ces autres pays je
24	05_SB_LZ_1...	TUFS (O)	mangas euh parfois l'animé le manga sont un peu différents et je crois qu'il y avait un manga	dont	le l'histoire était un peu plus sombre que la version animé ça arrive de temps en temps ça oui
25	05_SB_LZ_1...	TUFS (O)	est pas forcément évident pour quelqu'un qui n'a pas grandi avec euh avec la technologie et les jeux	dont	on parlait tout à l'heure qui qui pouvaient euh trouver ça trop difficile à comprendre, mais qu'est ce
26	Christine_...	TCOF (O)	y a tous les trucs modernes quoi hum ah oui d'accord ouais c'est la seule île qui qui	dont	la sur laquelle il y a il y a rien quoi ouais c'est cool quoi et tu as un
27	10CJTD1109...	TUFS (O)	taxi euh elle existe ouais c'est ça je me suis posé exactement la même question est-ce que la femme	dont	il parle existe ou pas c'est hm pff parce que à la f-~ fait la première chose que j'
28	18_AB_LS_1...	TUFS (O)	c'est d'accord bah histoire d'amour à peu près ce qui est assez intéressant c'est la façon	dont	tous les personnages sont ont en lien sans même le savoir ouais enfin comme quoi euh l'action de quelqu'un
29	18_AB_LS_1...	TUFS (O)	ont dit qu'ils allaient en faire un deuxième ou hum hum non ils ont pas dit, mais la façon	dont	ça se termine tu le sais que il y aura une suite puis généralement maintenant c'est ça quand un
30	Martha	TUFS (O)	qu il y a des chansons d'amour très euh ouais ouais ouais euh très très complètes comme les bagenatos	dont	a parlé tout à l'heure les vieux bagenatos tout ça la salsa il y a pas mal de mm
31	Martha	TUFS (O)	oui j'aime tout le rock euh il y a aussi des choses à une époque dans le secteur populaire	dont	nous sommes issus par exemple quand nous étions petits on on écoutait beaucoup de Rancheras ah oui oui rancheras c'

32	01BHGM1109...	TUFS (O)	une autre elle s'est fait une entorse au pouce parce que elle est tombée je crois c'est elle	dont	la selle s'est renversée elle est tombée elle s'est fait une entorse au pouce donc en plus on
33	01BHGM1109...	TUFS (O)	ben un carton non pas qu'un carton j'avais euh deux trois sacs à emmener plus l'imprimante ouais	dont	j'ai oublié le câble ça va être très très utile maintenant ben ça va vachement m'embêter
34	16KSLR1109...	TUFS (O)	elle va faire ses courses Casino un Casino un Casino ah oui, mais faut passer par le par le chemin	dont	je t'ai parlé qui est un peu plus court, mais c'est un petit Casino hein c'est pas
35	Prov_pin_8...	TCOF (O)	est quand même un support ouais, mais non c'est sûr ouais oui non, mais c'est pas un support	dont	on peut pas complètement complètement s'en passer on peut pas s'en passer quand même ouais je crois pas
36	01_OG_NH_1...	TUFS (O)	de quelque chose oui c'est après tu rebondis après nous dans notre esprit il nous vient plein d'idées	dont	tu peux en choisir une déjà déjà c'est déjà un choix difficile et puis après parce que une conversation
37	12_JG_AI_1...	TUFS (O)	tous les albums et il y en a pas mal sur chaque album ah oui ben il y en a	dont	se souvient, mais euh à part euh les deux premiers je crois sinon euh mh moi hier j'ai écouté
38	15_LW_MG_1...	TUFS (O)	Japon a quand même une meilleure ima~ une meilleure image parce que euh les États-Unis c'est côté super artificiel	dont	avec euh avec les chirurgies esthétiques les bimbos et tout ça hum ah c'est vrai euh le côté euh
39	19LCGK1109...	TUFS (O)	a vu ses deux posters là n-~ ah oui je m'en rappelle ouais la brigade des ouais machin oui	dont	on taira le nom ici pour les âmes sensibles oh là là ouais et du coup il est avec sa
40	19_CB_CV_1...	TUFS (O)	peut-être à à la BU hum euh tu sais sur la métamorphose des dieux hum le texte bah celui de	dont	je vais m'occuper faudrait peut-être que j'aille lire le petit passage sur Weber quand même ouais, mais j'

41	20_FD_CB_1...	TUFS (O)	avec des coquillettes des coquillettes des coquillettes oh c'est super bon rigole pas c'est bon non le truc	dont	j'ai horreur non ce que j'adore c'est la sauce au poivre je trouve ça très bon avec
42	08LFBM1109...	TUFS (O)	mais euh, mais ça va pas devenir tes meilleurs amis quoi oui les les g-~ les gens que les gens	dont	je fais à qui je parle pendant mon boulot en général c'est bonjour euh remplis la feuille et au
43	16_FB_EL_1...	TUFS (O)	ai l'air en fait toi parles familier tu es avec les profs tu les vouvoies, mais de la manière	dont	tu parles on a l'impression que c'est euh que c'est familier oui je sais pas si tu
44	25_CC_SL_1...	TUFS (O)	savais pas qu-~ enfin j'avais oublié ouais voilà parce que ça fait longtemps moi c'est vraiment le truc	dont	je me rappelle c'est qu'il a le ventre qui gonfle parce qu'il veut pas se faire sangler

Tab.3. Liste des résultats de la recherche sur « dont » dans le sous-corpus de conversations amicales d'ORFÉO

sciendo
DOI 10.2478/jazcas-2022-0020

COMPARATIVE CORPUS-DRIVEN STUDY OF PREPOSITIONAL SEMANTICS IN RUSSIAN AND CZECH

VICTOR ZAKHAROV
Saint Petersburg University, Saint Petersburg, Russian Federation

ZAKHAROV, Victor: Comparative corpus-driven study of prepositional semantics in Russian and Czech. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 967 – 976.

Abstract: This paper deals with prepositions with causal meaning in Russian and Czech. In Slavic languages prepositions are closely connected to cases. Russian and Czech prepositions have many common features. Prepositions show a relation in space or time or a special relationship between two or more people, places, things or situations. In the current paper we are dealing with causal relations. There are different ways to express them. Among these means, the most common are prepositional-case forms and complex sentences with a subordinate causal part. We analyze the repertoire of causal prepositions in both languages and describe their statistical representation in corpora. Another task is to reveal translation equivalents between two languages.

Key words: preposition, causal meaning, Russian language, Czech language, corpus statistics, parallel corpora

1. INTRODUCTION

The preposition is perhaps the most mysterious part of speech in all languages. Its frequency is extraordinarily high. In Russian and Czech, more than 10% of tokens in a given text are prepositions (Lyashevskaya – Sharov, 2009; Bartoň et al., 2009). Prepositions are heterogeneous in both languages: there is a small group of primary prepositions (about 30) and a few hundreds of secondary ones. The latter are motivated by content words (nouns, adverbs, verbs), which may be combined with primary prepositions, thus forming complex multiword expressions. A strict division between secondary multiword prepositions and prepositional phrases, however, is not specified. This is a task for a special corpus-based research.

The semantics of prepositions is a special topic of interest. Primary prepositions tend to be highly polysemous. For instance, the preposition *в* in Russian (‘in’) has as much as 18 meanings (BAS, 2010), whereas the Czech preposition *v* (‘in’) has 16 meanings (SSJČ, 1979). The majority of these meanings are, nevertheless, quite rare. There are also many unique or “empty” meanings when the preposition constitutes a part of an idiom. Prepositional ambiguity is manifested in the complex nature of the prepositional meaning and in selective preferences of certain prepositions, depending on the context.

Prepositions are characterized as function words used to express various relationships between main (governor) and dependent (governee) members of a phrase. The difficulty is that the relations expressed by prepositions are multi-sided, grammatical, lexical and extralinguistic. Prepositions are often said to have no real lexical meaning. At the same time, they express different semantic relations between words and their meanings must naturally directly correspond to these relations.

We believe that the meaning of prepositions is realized in prepositional constructions. In order to describe prepositional meaning, it is necessary to describe the meaning of a prepositional construction. It is important to provide a special metalanguage for such a description. For Russian, we use the syntaxeme classification of G. Zolotova (2011). The syntaxeme combines the governor, preposition and the governee in a case form, while it represents a minimal semantic-grammatical unit. The classification of syntaxemes in Zolotova (2011) is made according to the pattern of semantic roles: directive, destinative, correlative, quantitative, qualitative, locative, mediative, temporative; 27 syntaxemes in total (ibid., p. 383). All grammars of Czech describe prepositional meanings autonomously in terms of adverbial modifiers. As a rule, 7 types of such meanings are distinguished (Štícha, 2018).

2. PREPOSITIONS WITH CAUSAL MEANING IN RUSSIAN AND CZECH

For this study we chose Czech and Russian prepositions with similar meanings, i.e., causal relationship (‘příčinný vztah’ in Czech). This is the meaning of constructions where the prepositional group indicates the cause of an action or the influencing factor. The word “cause” is the basic term used to interpret the whole lexical composition, which is associated with the category of determining the cause. Dictionaries of a language divide two meanings of this word: 1) cause as a phenomenon that inadvertently causes another phenomenon, ontological cause; 2) cause as a basis, precondition for the realization of an event, action, i.e., subjective, explainable cause. Among ways to express causal relations, the most common are prepositional-case forms and complex sentences with a subordinate causal part, and most of the causal conjunctions come from prepositions.

A lot of studies are devoted to this type of relation (Křížková, 1967; Vsevolodova – Yascenko, 1988; Horák, 1989; Diessel – Hetterle, 2011) and only several studies deal with causal prepositions (Kroupová, 1980; Levontina, 1997; Iordanskaya – Melchuk, 1996; Luraghi, 2005). However, these studies were not based on corpora. Our study relies on statistics describing the use of these prepositions in large text material.

Causal relationships can be expressed in Russian and Czech by some primary and by a large number of secondary prepositions. They can be used in combination with nouns and pronouns in genitive, dative, accusative and instrumental. The lists of causal prepositions differ in different sources. For the Czech language, we used the prepositions listed in Štícha (2018) with a few deviations. The main part of this preposition list is as

follows: *z*, *za*, *pro*, *skrz*, *od*, *na*, *v důsledku, následkem* ‘due to’, *za příčinou* ‘for the reason of’, *vinou* ‘due to’, *díky* ‘thanks to’, *kvůli* ‘due to’, *z důvodu* ‘for reasons of’, *u příležitosti* ‘on the occasion of’, *v souvislosti s* ‘in relation to’, *v závislosti na* ‘depending on’, *vlivem* ‘owing to’, *vycházeje/jíc z* ‘coming from’. In Russian, according to our analysis, the list is as follows: *за*, *из*, *из-за*, *на*, *от*, *по*, *под*, *после*, *при*, *с*, *через*, *благодаря* ‘thanks to’, *в зависимости от* ‘depending on’, *в ответ на* ‘in response to’, *в результате* ‘as a result of’, *в свете* ‘in light of’, *в связи с* ‘due to’, *в силу* ‘by force of’, *за счёт* ‘on account of’, *исходя из* ‘drawing from’, *на основании* ‘on the basis of’, *на основе* ‘based on’, *на почве* ‘on the ground of’, *по причине* ‘because of, for the reason of’. (Here we provide English equivalents only for the secondary prepositions because the meanings of the primary ones are highly context-dependent).

These prepositions form clusters of intra- and interlanguage synonymy. Different prepositions can express the same meanings and grammatical relations when used in the same phrases. In the sentences *Он не пришел по причине болезни – Он не пришел из-за болезни – Он не пришел вследствие болезни* (‘He did not come because of the disease – He did not come due to the disease – He did not come on account of the disease’) it is possible to interpret the prepositions as synonyms. In Czech, equivalent synonymous groups are formed by groups ‘*v důsledku/kvůli/pro*’ or ‘*kvůli/v důsledku/následkem/z důvodu*’.

Consequently, our tasks in the current study were as follows:

1) to obtain statistical characteristics of prepositions from corpora;
2) to show differences in use of prepositions in various genres;
3) to find main translation equivalents for the basic set of prepositions with the causal meaning.

3. RESULTS AND DISCUSSION

The results of the statistical analysis presented in this article were obtained from the Araneum Russicum III Maius and Araneum Bohemicum IV corpora (1.25 billion tokens each, www.unesco.uniba.sk), the Russian National Corpus (RNC) (321 million tokens, www.ruscorpora.ru), the Czech National Corpus (ČNK, www.korpus.cz) – syn v8 corpus, 5,4 billion tokens, and InterCorp v13 (20 million tokens in Russian subcorpus) (including parallel corpora for 40 languages). We used the Russian-Czech and Czech-Russian parallel subcorpus which consists of both Russian texts translated into Czech and Czech texts translated into Russian. Moreover, it also includes texts translated into Czech and Russian from other languages, mostly from English (Rosen et al., 2020).

Prepositional causal constructions occupy a rather modest place among all prepositional constructions. In a large experiment on the extraction and annotation of Russian prepositional constructions, only 349 causal constructions were identified among 10047 constructions, i.e., 3.47%. Their analysis showed that in Russian, most prepositional constructions with the causal meaning are those with the following

prepositions: *по* (24% of all occurrences inside causal constructions), *за* (19%), *из-за* (16%), *от* (12%), *в связи с* (6%), *на* (5%), *на основании/на основе* (4%), *по причине* (3%). For Czech, such an extensive experiment is yet to be undertaken, and so more or less accurate statistics cannot be provided currently.

However, for most of the **primary** polysemous prepositions that make up the majority of causal constructions, the causal meaning is not the primary one. In order to identify the prepositions for which the causal meaning is the only or the primary meaning, we conducted another experiment. As a rule, the experiment was aimed at secondary multiword prepositions. 12 Russian and 11 Czech prepositions mostly from the top part of the frequency list were selected and constructions with these prepositions were extracted from two Russian and two Czech corpora. Then the first 50 randomly selected constructions were annotated by hand according to the realized meaning. Usually, in those cases where this meaning differed from the causal one we found a free combination of a preposition and a common word. The assignment of these combinations to the category of prepositions is sometimes complicated by the fact that the common word has to some extent retained its original lexical meaning.

The obtained results are shown in Tables 1 and 2. They demonstrate relative frequencies of each preposition and the percentage of constructions with causal meaning.

Preposition	**Frequency (ipm) in Araneum**	**% of prepositional use, Araneum**	**Frequency (ipm) in RNC**	**% of prepositional use, RNC**
в зависимости от (ото)	111,60	100	27,72	98
исходя из (изо)	44,70	100	14,95	100
за счет (за счёт)	142,57	88	41,70	100
в связи с (со)	119,10	100	57,93	100
на основе	105,40	82	35,46	82
на основании	57,30	100	31,78	100
в результате	173,80	100	81,26	54
по причине	19,70	86	12,34	96
в ответ на	12,20	100	18,09	98
на почве	1,10	100	3,59	96
в свете	2,00	84	3,68	78
в силу	10,70	90	10,54	80

Table 1. Relative frequency (ipm) and percentage of prepositional use of Russian multiword prepositions in the Araneum Russicum III Maius and in the RNC.

Preposition	**Frequency (ipm) in Araneum**	**% of prepositional use, Araneum**	**Frequency (ipm) in ČNK**	**% of prepositional use, ČNK**
díky	313,30	100	186,20	100
kvůli	225,60	100	215,40	100
následkem	5,50	100	5,70	100

v důsledku	26,50	100	26,60	100
v souvislosti	41,30	92	28,80	92
v závislosti na	23,00	100	16,80	100
vinou	3,10	96	8,40	98
vlivem	17,00	66	17,40	80
z důvodu	43,80	96	15,70	100
za příčinou	0,10	100	0,10	100
vycházeje/jíc z	only 3 items	66	-	-

Table 2. Relative frequency (ipm) and percentage of prepositional use of Czech secondary prepositions in the Araneum Bohemicum IV Maius and in the ČNK.

Different corpora for both languages, one a well-balanced national corpus and the second a web-based corpus, were selected on purpose. We aimed to make sure that the choice of corpora has no significant effect on the final results. The result is ambiguous. We see that prepositions in national corpora tend to have generally lower frequency than those in web-corpora. This is quite clear, since the percentage of journalistic and business texts – for which the secondary prepositions are more characteristic – is smaller.

Causal prepositions, especially the secondary ones, occur more frequently in journalistic and theoretical writing. We have observed the distribution of causal prepositions in texts of various genres and types. Our data, as well as data from Czech sources, indicate that the relative frequency of causal prepositions is highest in the texts from the domain of legislation and journalism. From there, however, they also permeate into colloquial language or fiction.

4. TRANSLATION EQUIVALENTS

In most cases, the Russian primary causal prepositions do not correspond to their similar Czech counterparts. On the one hand, the repertoire of prepositions is different in both languages, and on the other hand, they are used differently. These are either differences in the semantic structure of the prepositions or differences in individual use cases.

Actually, only Russian *из* 'out of' and *за* 'for'[1] coincide with the Czech *z* and *za*: *сделать что-нибудь* ***из*** *расчета – udělat něco* ***z*** *vypočítavosti* ('to do something out of self-seeking'); *наказывать* ***за*** *проступок – trestat* ***za*** *provinění* ('to be punished for a wrongdoing'). Such coincidence, however, does not always take place, e.g.: *сердиться* ***за*** *упрек – zlobit se* ***pro*** *výtku* ('to be angry for a reproach').

The preposition *от* 'from' is highly polysemous in Russian and occurs very often in various meanings. Therefore, translation of constructions with this preposition may differ: ***от*** *болезни –* ***на*** *нетос* 'from illness', ***от*** *наводнения –* ***následkem*** *povodně*

[1] Only a rough (context-free) translation for prepositions mentioned in the paper henceforth can be provided out of context.

‘due to flood’, ***от*** *этого слова* – ***z*** *tohoto slova* ‘from this word’, *произошло не* ***от*** *хорошей жизни* – *nedošlo* ***v důsledku*** *dobrého života* ‘did not come from a good life’.

In Russian, the preposition *от* is used in conjunction with the name of a disease: *умер* ***от*** *рака*, ***от*** *сердечного приступа*. In such cases, the preposition *na* with the accusative is used in Czech: *zemřel* ***na*** *rakovinu,* ***na*** *infarkt* ‘he died of cancer, of a heart attack’. It could be translated as *v důsledku* or *následkem* ‘as a result, as a consequence’ when the cause is not directly related to a consequence that has already occurred in the meantime.

The Czech preposition depends often on **the type of the noun** entering the prepositional phrase and sometimes on the type of the verb. If a given noun denotes the meaning of an inner feeling or state of spirit, the Russian expression corresponds not to the prepositional construction in Czech, but rather to the corresponding noun in the instrumental: *плакать* ***от*** *радости* – *plakat* ***radostí*** ‘weep for joy’. However, constructions containing nouns with other meaning are translated sometimes in a similar way, too (cf. *заболеть* ***от*** *шума* – *onemocnět hlukem* ‘to get sick from noise’, *погибнуть* ***от*** *рук* – *zahynout rukou* ‘to die at the hands of’).

The choice of the Czech equivalent can also depend on the meaning of the governing verb. The construction *зависеть* ***от*** *чего-нибудь* ‘depend on something’ is used always with the genitive of the noun. The governee is then a proverbial determination of the cause. In Czech, the verb *záviset* ‘depend’ is associated with the preposition ***na*** ‘on’ and the noun in the locative: *зависеть* ***от*** *укрепления демократии* – *záviset* ***na*** *upevnění demokracie* ‘depend on the strengthening of democracy’. The rules of both languages therefore do not allow for other variants of translation of this construction.

However, sometimes different translations of the same prepositional construction are possible. They are determined by the style, preference of a translator and the prevailing practice. We attempted to identify some Russian-Czech prepositional equivalents in causal constructions. Sometimes these correspondences are quite frequent, sometimes not.

As the first step of the study we analyzed the translation of a few Russian preposition on the base of the InterCorp Russian-Czech corpus (Rajnochová et al., 2020).

The frequency of prepositions in InterCorp is high, and in the absence of semantic annotation, the task of selecting causal prepositional constructions is not trivial. For this purpose, the main lexical markers (nouns) were manually identified in constructions containing the analyzed prepositions with a causal meaning. Then these words were included in the CQL query, taking into account the required dependent noun case. Following this step, a manual cleaning of the list of extracted constructions was performed.

It should also be emphasized that InterCorp is mainly composed of literary texts and film transcripts. Translators often took more creative liberty and did not always translate Russian prepositional constructions by using Czech prepositional constructions.

The following preliminary results of translation equivalents analysis were obtained. As previously mentioned, the preposition *из* (*изо*) is one of the few that has a direct Czech equivalent (the preposition *z* (*ze*)). From the selected 212 *из*-constructions with a total frequency of 1127, some constructions with a frequency greater than 12 were selected. The results are shown in Table 3.

Construction	**Frequency in corpus**	**The number of causal meanings**	**Translation using the preposition z(ze)**	**Another causal preposition[2] or periphrasis[3]**	**Translation by other means[4]**
из вежливости 'out of courtesy'	40	39	24	2	15
из безопасности 'out of safety'	37	15	12	1	2
из жалости 'out of pity'	24	24	14	6	4
из ненависти 'out of hatred'	17	16	9	5	2
из необходимости 'out of necessity'	16	10	7	1	2
из мести 'out of revenge'	13	13	7	3	3
из гордости 'out of pride'	13	12	6	6	0

Table 3. Translation equivalents for Russian preposition *из* (*изо*)

Thus, the analysis showed that indeed the preposition *из* in causal meaning is mainly translated by the preposition *z* and only occasionally by other prepositions.

The preposition *из-за* shows a noticeably different picture. In the 280 analysed causal constructions, it is translated by a variety of prepositions: *kvůli*, *díky*, *pro*, *z* (*ze*), *vzhledem*, *výsledkem*, *důsledkem* (in descending order of the frequency of the equivalents).

Then, the secondary multiword prepositions containing the primary preposition *в* 'in' were analyzed. The preliminary analysis results are shown in Table 4. Numbers in cells show the number of translation equivalents for the corresponding Russian preposition in the column.

Czech preposition	**в зависимости от(о)**	**в ответ на**	**в результате**	**в связи с (со)**	**в силу**
	the number of constructions				
	260	201	947	489	145

[2] **из** предосторожности – **kvůli** vší opatrnosti 'for all caution'

[3] хранитель печати взглянул на меня **из вежливости** 'the seal superior looked at me out of **politeness**' – představený pečeti se ohlédl **zdvořile** po mně 'the seal superior looked at me **politely**'

[4] **из жалости** и сострадания – **máme soucit**, slitovali jsme se máme soucit, slitovali jsme se 'we have compassion, we have pity'

	The number of translations				
v závislosti na	11				
podle (dle)	123 (5)				
v (jako) odpověď na		14 (3)			
výsledkem			13		
následkem			10		
v důsledku (-em)			162 (18)		1 (3)
díky			31		14
po			35		
při			34	7	
pro				5	
v souvislosti s				63	
vzhledem k				17	4
ve vztahu				5	
ve spojitosti (ve spojení)				8 (2)	
kvůli				57	
z					19
podle					7

Table 4. Translation equivalents for Russian multiword prepositions with causal meaning

The table does not include prepositional equivalents with frequencies of less than 5: *vinou, u příležitosti, na základě.* Let us note that the preposition *в зависимости от* 'depending on' was often translated by the construction *záleží na* (14 times).

We see that certain prepositions are sometimes translated by many variants. Due to this variance of translations, they can be perceived as synonyms. It should be also emphasized that Russian prepositional constructions were often translated to Czech by subordinate clauses or other phrases. However, this is a research task for a separate study.

In a number of synonymic rows, the asymmetry of translation equivalents should be noted when translating from Russian into Czech and from Czech into Russian. It is not clear whether these results are a matter of languages (linguistics), or rather a matter of translation practice (translatology). A more detailed and in-depth analysis is yet to come.

5. CONCLUSION

Causal meanings in prepositional constructions in both Russian and Czech are expressed by a few primary prepositions and by a large number of secondary prepositions. The use of secondary multiword prepositions expressing the determination of the cause is very varied in Russian, as evidenced by the frequent and diverse occurrence of these prepositions in journalistic texts. In Czech, the number of secondary prepositions is not as high. In addition, they are often synonymous with each other and do not have the same semantic variety as Russian

prepositions. For this reason, it is often difficult to find an equal counterpart to the Russian prepositional construction.

We would like to emphasize that it is important to deal with the task of comparing not just individual prepositions in two languages, but the prepositional systems as a whole. Such an analysis would make it possible to establish a study of prepositional meanings as a systemic contribution to linguistic theory, and this analysis would also prove to be useful for practical tasks.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grant No. 17-29-09159 "Quantitative grammar of Russian prepositional constructions"). We express our sincere gratitude to the 2nd year students of the SPbU Mathematical Linguistics Department for their valuable help in annotating the data. Our cordial thanks belong to Vladimir Benko and Alexandr Rosen for their help in work with corpora. We also thank Anastasia Golovina for her help in checking the English text of the paper.

Bibliography

BARTOŇ, Tomáš – CVRČEK, Václav – ČERMÁK, František – JELÍNEK, Tomáš – PETKEVIČ, Vladimír: Statistiky češtiny. Praha: Nakladatelství Lidové noviny 2009. 215 p.

BAS: Большой академический словарь русского языка. Т. 14. Гл. ред. К. С. Горбачевич – А. С. Герд. Москва – Санкт-Петербург: Наука 2010. 656 с.

DIESSEL, Holger – HETTERLE, Katja: Causal clauses: A crosslinguistic investigation of their structure, meaning, and use. In: Linguistic Universals and Language Variation, Ed. P. Siemund. Berlin: De Gruyter Mouton 2011. pp. 21–52.

HAVRÁNEK, Bohuslav et al.: Slovník spisovného jazyka českého. Praha: Academia 1960–1971.

HORÁK, Emil: Východiská pre konfrontáciu predložkového systému slovenčiny s inými jazykmi. In: Studia Academica Slovaca 18. Ed. J. Mistrík. Bratislava: Alfa 1989, pp. 167–183.

IORDANSKAYA – MELCHUK: ИОРДАНСКАЯ, Лидия Н. – МЕЛЬЧУК, Игорь А.: К семантике русских причинных предлогов (ИЗ-ЗА любви ~ ОТ любви ~ ИЗ любви ~ *С любви ~ ПО любви). In: Московский лингвистический журнал, 1966, Т. 2. с. 162–211.

KŘÍŽKOVÁ, Helena: Adverbiální determinace s významem časovým a příčinným. Slavia, 1967, Vol. 36, pp. 507–531.

KROUPOVÁ, Libuše: Vztah významu gramatického a lexikálního u předložek. Slovo a slovesnost, 1980, Vol. 41, No 1, pp. 49–52.

LEVONTINA: ЛЕВОНТИНА, Ирина Б.: ИЗ-ЗА, ИЗ, ОТ, ПО, С, ЗА, БЛАГОДАРЯ, ПО ПРИЧИНЕ, ВСЛЕДСТВИЕ, В РЕЗУЛЬТАТЕ, ВВИДУ, А СИЛУ. In: Новый объяснительный словарь синонимов русского языка. Под общ. рук. Ю. Д. Апресяна. Москва: Школа «Языки русской культуры» 1997, с. 144–152.

LURAGHI, Silvia: Prepositions in Cause expressions. Papers on grammar, 2005, Vol. 12, No 2, pp. 609–619.

LYASHEVSKAYA – SHAROV: ЛЯШЕВСКАЯ, Ольга Н. – ШАРОВ, Сергей А.: Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). Москва: Азбуковник 2009. 1087 с.

RAJNOCHOVÁ, Natálie – RUNŠTUKOVÁ, Naděžda – VAVŘÍN, Martin: Korpus InterCorp – ruština, verze 13 z 1. 11. 2020. Ústav Českého národního korpusu FF UK, Praha 2020. Available at: http://www.korpus.cz [cit. 29. 01. 2021].

ROSEN Alexandr – VAVŘÍN, Martin – ZASINA, Adrian: Korpus InterCorp – čeština, verze 13 z 1. 11. 2020. Ústav Českého národního korpusu FF UK, Praha 2020. Available at: http://www.korpus.cz [cit. 29. 01. 2021]

SSJČ: Slovník spisovného jazyka českého. Zv. VII. Hl. red. B. Havránek. Praha: Academia 1989. 442 p.

ŠTÍCHA, František et al.: Velká akademická gramatika spisovné češtiny. I. Morfologie. Praha: Academie 2018. 1148 p.

VSEVOLODOVA – YASHCHENKO: ВСЕВОЛОДОВА, Майя В. – ЯЩЕНКО, Татьяна А.: Причинно-следственные отношения в современном русском языке. Москва: Русский язык 1988. 208 с.

ZOLOTOVA: ЗОЛОТОВА, Галина А.: Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. 4-е изд. Москва: Editorial URSS 2011. 440 с.

sciendo
DOI 10.2478/jazcas-2022-0021

IDENTIFYING ERRORS IN RUSSIAN WEB CORPORA

MARIA KHOKHLOVA
St Petersburg State University, St Petersburg, Russian Federation

KHOKHLOVA, Maria: Identifying errors in Russian web corpora. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 977 – 985.

Abstract: The explosion of the Web leads to the production of large amounts of texts and inevitably influences their quality. Errors that tend to occur more often can distort results, especially when texts are used for scientific purposes, in language teaching or learning. Hence, there is a need to examine the existing corpora based on web texts and to clean up the data, which may contain such "noisy" fragments. In our study, we deal with the problem of errors and analyze the Aranea Russicum Maximum corpus. Among such errors, we can name, above all, encoding errors, incorrect font types, as well as segments written in other languages. These phenomena result in incorrect morphological analysis and lemmatization, frequency distortion, as well as the fact that lexical units cannot be found and therefore displayed to corpus users. The paper focuses on the errors, describes their types and outlines possible ways to eliminate them.

Key words: corpora, web texts, errors, typos, orthography, typography, Russian language

1. INTRODUCTION

The technologies for corpus building have evolved rapidly over the last twenty years. Currently, there is no doubt that corpora have become more and more focused on extensive text collections. The largest and most significant projects in corpus linguistics are based on web texts, and it is crucial to pay attention to the quality of these texts. Methods of lemmatization and of morphological and syntactic analysis were originally developed for a standard (literary) language. Therefore, their application to web texts can give erroneous results which can be even worse in the case of "noisy" texts.

When searching on the Internet for factual information, erroneous web pages may not have such a negative impact on the result[1], as it may be the case when using corpora for linguistic purposes. Errors and misprints can even lead to unexpected results[2], as their repeated duplication increases frequency of these phenomena. The

[1] This can be explained by two reasons. First, search engines like Google or Yandex correct typos (often being overzealous in this regard). Secondly, it can be important for a user to get the results quickly, so some of these errors will be overlooked.

[2] As a fairly well-known example in Russian literature, we can recall the poem by S. Yesenin "The Black Man", which was being published for a long time with the incorrect line *на шее ноги* 'on the neck of the leg' instead of the correct one *на шее ночи* 'on the neck of the night', thus, the misprint concerned two Russian letters *г* and *ч* (that are similar in their form in handwritten text) and led to misinterpretation.

accuracy of automatic applications based on web corpora can also suffer from their possible poor quality and erroneous nature. Furthermore, a certain number of Web texts is obtained via optical character recognition which may again cause text distortion. The objectives of our research can be summarized as follows: 1) to create a typology of the encountered errors; 2) to determine possible ways to correct errors; 3) estimate their approximate number and the degree of their influence on the results.

1.1 Background

Web corpora gradually supersede traditional (or classical) corpora, as they contain even larger amounts of data and allow linguists to observe unique language phenomena. Studies of new large corpora and their characteristics have become popular recently. Authors examine various frequency distributions in corpora of various sizes (Khokhlova, 2016), the possibilities of using large corpora to study collocations (Khokhlova – Benko, 2020) or to explore the impact of corpus size on the quality of embeddings generated from the corpora (Kutuzov – Kunilovskaya, 2018). Special attention is paid to the process of corpus building: selection of texts, their crawling, subsequent processing and linguistic annotation (Jakubíček – Kovář – Rychlý – Suchomel, 2020). Corpora users can give feedback on possible errors. However, the question of how clean the corpus texts are also requires a deep analysis.

When creating corpora, one should deal with the question whether they can contain texts written in other languages. The answer to it is not so obvious and causes further discussion: how long can be the texts in foreign languages to be included in a corpus? For example, authors of the British National Corpus claim that "foreign language words do occur in the corpus" (British National Corpus).

The issue of typos in texts has been raised by a number of authors. Automatic spelling correction systems should be mentioned in this respect in the first place (Shavrina, 2017). The authors T. O. Shavrina and A. A. Sorokin (2015) use a probabilistic model based on the weighted Levenshtein distance to correct typos in social media texts. The normalization of this type of texts is considered by E. Clark and K. Araki (2011). The distribution of orthographic errors of various types in English and German web pages was analyzed in Ch. Ringlstetter, K. Schulzand and M. Mihov (2006). R. Baeza-Yates and L. Rello (2012) discuss the lexical quality of the web pages written in English and Spanish and propose a measure for their evaluation. The work by V. V. Shapoval (2009) raises a rather interesting question about the literacy of schoolchildren and makes a distinction between errors in a text written by hand and in a printed one, which shows errors that appeared relatively recently due to the massive use of computers.[3]

[3] Although typos could be found in earlier printed texts, they have become widespread to this extent recently.

1.2 Methodology and data

In this paper, we will dwell on several types of errors found in a large corpus and try to describe their most common features. As a material, we consider the Aranea Russicum Maximum corpus (Benko, 2014) (19.8 billion tokens equal to 16.0 billion words). We used a language identification guide by R. S. Gilyarevskiy and V. S. Grivnin (1965) in order to analyze text fragments written in other languages and to investigate whether letters inherent to alphabets are used in Russian texts. By using queries with regular expressions containing non-Russian symbols, we detected some possible errors. The examples of errors listed in T. O. Shavrina and A. A. Sorokin (2015) were also inspected in the corpus.

1.3 Classification of errors

The corpus texts contain various kinds of errors. These vary in nature: they could have been made during automatic text processing (for example, incorrect lemmatization or morphological analysis) or can result from the initial quality of texts.

We can list following errors found in the web corpus:

- orthographic (spelling) errors (caused either by users' intention or lack of language competence);
- typographical (graphemic) errors;
- encoding errors;
- errors in lemmatization (morphological errors);
- errors of incorrect OCR procedures.

In our study, we will focus on three groups: graphemic, encoding errors and errors that may be associated with incorrect recognition of scanned texts. Typos appear during the actual generation of texts. These errors occur, for example, on the internet discussion forums where a person writes quickly which inevitably leads to mistakes. The third group relates mostly with fiction and scientific literature that is uploaded to the Web after OCR procedures. In this case, misprints are caused by the similarity between letters, for example, between the Russian *п* and *и*. Here, we will not be dealing with spelling errors related to the literacy level. We also do not consider intentional distortion, i.e. spelling mistakes made for some purpose, as in, for example, the notorious "olbanskiy language", popular in Russia at the beginning of the 21st century. Another problematic aspect of text corpora is represented by the duplicate content of web pages, which also deserves a separate comprehensive analysis.

1.4 Errors: A case study

1.4.1 Wrong key

The first kind of such errors appears users when typing their texts use the wrong key for a given letter, which affects the graphemes. The letters on a keyboard are located next to each other, so it can lead to pressing an incorrect letter or even an

additional letter, that can be pressed accidentally. For example, the preposition *для* 'for' can be mistyped as *ддя* the letters *д* and *л* are located next to each other in the standard Russian keyboard layout. The corpus shows 619 occurrences of such misprinted word *для* (0.03 ipm) which received false POS tags of a noun or a verb, but not of a preposition, which would be correct. A user can repeat the letter by pressing the same key by mistake. The verb *говорить* 'to speak' is represented by the form *гооворить* resulting from double *о* (the corpus shows 18 examples). The letters can be even misplaced, such as in the same verb *говорить* and its incorrect variant *гооврить* (75 occurrences in the corpus) or the preposition *вместо* 'instead of' and *вмсето* (54 hits). Such errors are difficult to find because they can occur in any word, but one can also see the regularity of certain cases. Additional letters appended to words can be a further obstacle for correct normalization. The letter *ю* tends to appear at the end of sentences instead of an adequate punctuation mark, be it period or comma as the next key (for example, *сказатью* instead of *сказать* 'to say' has 6 hits in the corpus). This problem can be solved by using spell checkers or other software for correcting spelling of the words not found in the dictionaries.

1.4.2 Wrong keyboard layout

The second type of error is related to typing on the wrong keyboard layout (it's the Latin one in the case of Russian), that is, when a user forgets to switch between the languages. Apparently, these errors are more common in chats or comments where a short answer is required, or at the beginning of a text. However, they are recently becoming less common since special programs identify the language and switch the layout when a user enters incorrect text. It is interesting that the encountered examples of such wrong spelling refer mostly to advertising texts about this software. The most common cases are the following: *ghbdtn* instead of the Russian *привет* 'hi' (65 hits), *rfr* instead of the Russian *как* 'how' (120 hits), and *ltkf* instead of the Russian *дела* 'case, as part of the construction with the meaning: how are you doing'. The mentioned words were annotated as punctuation marks with the tag Z. by combining the given tag with regular expressions for lemmata written with Latin characters, one can indicate errors or "noisy" texts. Altogether we managed to find more than 153 mln tokens matching the pattern *[atag="Z.*"&lemma="[A-Za-z]*"]*. These tokens correspond both to foreign (for example, *Yahoo*, *Corporation*, *Michael*, *email* etc.) and misprinted words like the examples mentioned above.

1.4.3 Combination of upper- and lowercase letters

A combination of upper- and lowercase letters can constitute a certain problem for processing data that affects the results of lemmatization and morphological annotation. Although this type of errors deserves a deeper analysis which would help us to understand whether a given example is misprinted. Thus, capital letters used

inside a word can stand for stress, e.g. *бОльший* 'big', pronunciation peculiarities, e.g. *сердеШный* instead of the correct form *сердечный* 'hearty/cordial', abbreviations like *мАч* 'mAh' or *кГц* 'kHz' etc. In order to eliminate errors in these cases, there is a need to convert words to lowercase, so that they will be lemmatized properly. Errors can occur between sentence boundaries, resulting in joining of words without spaces, like in *посмотретьДепортация* instead of *посмотреть. Депортация* 'look. Deportation'.

1.4.4 Misuse of hyphens

Misuse of hyphens can also result from an incorrect OCR when words are hyphenated at the end of lines, like in the incorrect *загадоч-ный* instead of the correct *загадочный* 'mysterious'. The corpus shows 6,986 examples (0.35 ipm) of such an error (e.g. words with hyphen ending *-ный*). Although incorrectly lemmatized with hyphens, such words have nevertheless received proper morphological tags A, corresponding to adjectives. At the same time, the misuse of hyphens or dashes can cause not only errors in lemmatization or morphological analysis but also incorrect tokenization. This type of errors, however, seems to quite solvable. Russian spelling suggests a closed list of items that should be written with hyphenation (like *-таки*, *-ка*, *-нибудь*, etc)[4] and the ones that are not found in the list can thus be detected as possible errors occurring at line breaks and fixed as such.

1.4.5 Special characters

Further inconvenience of 'noisy' texts consists in special characters. For example, the symbol ¶ shows 11,212 occurrences (0.60 ipm) in the corpus. One also finds numerous examples of HTML encoding fragments, such as non-breaking space ( ) or different types of dashes (&ndash or &dash). Thus,   has 195,266 hits (9.9 ipm), while &ndash / &mdash account for 7,044 occurrences (0.36 ipm). These characters can be combined with other errors and in this way used to reveal irrelevant texts, for example, they can signal incorrect encoding of a given document.

1.4.6 Combinations of Latin and Cyrillic characters

There are words that contain characters from the two alphabets. Cyrillic characters can be replaced with similar Latin letters in order to 'cheat' search engines. The following letters of the Cyrillic alphabet have their Latin lowercase and uppercase counterparts: *a*, *e*, *o*, *p*, *c*, *y* (the uppercase of the last letter is different for the two alphabets) and *x*. The uppercase is the same for the following letters: *B*, *K*, *H* and *T*, such as in *скромный* 'modest'. Its correct form in Cyrillic (*скромный*)

[4] The well-known guide is the "Handbook of spelling and literary editing" by D. E. Rosenthal (2001). A list of hyphenated compounds is presented in the dictionary "Together or separately" (Bukchina, Kalakutskaya, 2006).

graphically coincides with the variant *скромный* (found in the corpus), where the Latin *c* is used. Although annotated properly as an adjective, all 9 occurrences of this lexical item had the same misspelled lemma. When searching for such words copied from corpus output, engines like Yandex and Google correct typos for queries and warn the users about the errors. However, such a procedure does not always lead to the desired result. For example, the Russian abbreviation *ДНК* 'DNA' mistyped with the Latin *H* instead of the Cyrillic *H* will be transformed into ДРК. Search engines assume that the false letter should be substituted for the correct one located on the same key (that is the Russian *Р*). Hence, the results will be related not only to ДНК but also to ДРК, standing for 'the Democratic Republic of the Congo'. According to preliminary estimation, the corpus contains about 2 mln examples written both in Cyrillic letters and with at least one of their above mentioned Latin counterparts, e.g. *серебро* 'silver', *роса* 'dew', *свергнуть* 'overthrow', *ожидать* 'expect', *верхом* 'on horseback, astride', *Некоторые* 'some' etc. They can be detected by using regular expressions with both Russian characters and a certain set of Latin ones, e.g. *[а-я]+[аоерсух]+*, which will find 311,263 words ending with non-Russian characters, many of them having wrong lemmata and morphological tags. However, these examples require a further analysis.

1.4.7 Combinations with extended Cyrillic graphics

A similar type of typographic errors results from the mixture of Cyrillic characters and characters from the Russian alphabet. In such cases, it is either foreign words or mistakes in Russian words, made during automatic text recognition. Such additional characters are used in languages other than Russian, namely, Abkhazian, Azerbaijani, Belorussian, Bulgarian, Ukrainian etc. Nevertheless, in a large number of cases, these letters are misrepresented in Russian words (see Table 1). For example, the character *р̣* is used instead of the letter *р,* or the character *ҷ* substitutes for the letter *ч*.

Table 1. The misuse of the characters in Russian words

Wrong character	Correct letter	Absolute frequency	Frequency in ipm	Examples written with wrong characters	Examples written with correct letters
ґ	г	6,725	0.34	ґраницы, временныґе, виґдение	границы, временные, видение
ҽ	е	49	0.00	жизнедеятҽльности, экстрҽнной, внутрҽнних	жизнедеятельности, экстренной, внутренних
р̣	р	704	0.04	приватизации, адресата, словаре	приватизации, адресата, словаре
т̣	т	1,496	0.08	размышляют̣, незамет̣но, культ̣уры	размышляют, незаметно, культуры
ҷ	ч	1,089	0.06	аллегорических, ҷтобы, мальҷик	аллегорических, чтобы, мальчик
ԯ	л	19	0.00	ԯалґеко, озԯа́чаԯо, ԯа́чаԯось	далеко, означало, началось

ӈ	н	1,275	0.07	ӈᴇ, сӈа́чала	не, сначала
ᴇ	е	11,042	0.60	ничᴇго, нᴇсчастных, слᴇдующее	ничего, несчастных, следующее
г	г	1,814	0.09	критического, детского, гармонического	критического, детского, гармонического
ς	с	90	0.00	повторноςтью, доςтаточная, инςтанции	повторностью, достаточная, инстанции
Є	Э, е, ё	11,995	0.61	Єто Єлизавета тЄмные	Это Елизавета тёмные (темные)
S	Б	4,803	0.24	Sолее, Sыл Sыстро, неверная кодировка	Более Был, Быстро
s	different	8,483	0.43	культурыs spи sочему	культуры при почему
J	ё	10,755	0.54	*Jлочка* *серьJзным* тJмными	*Ёлочка* серьёзным тёмными
ль	none, ль	3,270	0.17	неделиль жизнедᴇкҭᴇльности самостоятᴇльно	недели жизнедеятельности самостоятельно
Ъ	none, dash	19,968	1.01	знакомство вЪ „ отношение между людьми	знакомство — отношение между людьми

Letters from other alphabets may indicate foreign words and thus can be used for filtering out non-Russian lexis. For example, the character "*ґ*" can either be misused in Russian words or can identify tokens written in Ukrainian (for example, *ґрунтуватися*, *ґатунку*, *ґвалтують*). The above-mentioned errors differ in their frequencies: some of them are widespread in texts, e.g. the negation *не* 'not' incorrectly written as *ӈᴇ*, with 1,255 occurrences (about 98% of all found words with the character *ӈ*). All the mistyped words were lemmatized incorrectly, which resulted in frequency distortions. Certain letters can indicate mostly irrelevant meaningless and "noisy" texts, e.g. the letter È.

2. DISCUSSION

T. O. Shavrina and A. A. Sorokin (2015) use Levenshtein distance only for words that differ in two characters at most. Nevertheless, we managed to find examples containing more than two misprints, such as the noun *жизнедеятельность* 'life activity' with five typos in its variant *жизнедᴇкҭᴇльность*. As the authors rightly point out, Levenshtein distance can help to get rid of some misprints.

Based on the examples we have reviewed and analyzed, we can conclude that, for the most part, texts that contain errors described above are characterized by low quality and contain a large number of other errors. One can assume that the frequency of these units for such a large Russian corpus is insignificant; however, when considering peculiar linguistic phenomena, we often deal with small frequencies, so it is important to obtain and study "clean" data.

The analysis suggests that in a number of cases, mistyped letters can be replaced by usual (traditional) letters, that are inherent to the Russian language. Apart from words that erroneously mix characters from two alphabets (Latin and Russian, extended Cyrillic and Russian), there are encoding errors, foreign words, meaningless sequences etc. Moreover, it is interesting to mention that the texts written in different languages tend to be of different lengths. For example, fragments in Udmurt are short whereas Ukrainian or Serbian words are usually contained in larger texts.

3. CONCLUSION

In our study, we examined in detail the most common errors in Russian web corpora and outlined several ways of how to solve them. "Noisy" texts can be cleaned either by deleting irrelevant or large mistyped text fragments or by rewriting them with traditional Russian characters. The proposed list is by no means exhaustive. It can be tricky to get false negative results in a corpus, as these are not included in the output and hence are not presented to users. The described examples seem to be quite interesting, although not so numerous. According to preliminary estimates, they account for no more than 1.5% of the total volume of the Aranea Russicum Maximum corpus. Nevertheless, we consider it important to pay attention to the issue of cleaning up data, as well as removing irrelevant information from the corpora. The resulting findings can be used for building new corpora.

In the future, it would also be interesting to see if there is a correlation between errors of a certain type and the type of the text, its genre, the topic of the websites or other characteristics. We also believe that our work will allow us to eliminate errors and, therefore, to produce corpora of high quality.

Bibliography

BAEZA-YATES, Ricardo – RELLO, Luz: On measuring the lexical quality of the web. In: Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. Eds. C. Castillo – Z. Gyongyi – A. Jatowt – K. Tanaka. Lyon, France 2012, pp. 1–6. Available at: https://dl.acm.org/doi/pdf/10.1145/2184305.2184307

BENKO, Vladimír: Aranea: Yet another family of (comparable) web corpora. In: International Conference on Text, Speech, and Dialogue. Eds. P. Sojka – A. Horák – I. Kopeček – K. Pala. Cham: Springer 2014, pp. 247–256.

British National Corpus. Available at: http://www.natcorp.ox.ac.uk/corpus/

BUKCHINA – KALAKUTSKAYA: БУКЧИНА, Бронислава З. – КАЛАКУЦКАЯ, Лариса П.: Слитно или раздельно. Москва: Дрофа 2006. 936 с.

CLARK, Eleanor – ARAKI, Kenji: Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. In: Procedia — Social and Behavioral Sciences. Eds. N. A. Aziz – K. Hasida – A. W. A. Rahman – H. Saito. 2011, 27, pp. 2–11.

GILYAREVSKIY – GRIVNIN: ГИЛЯРЕВСКИЙ, Руджеро С. – ГРИВНИН, Владимир С.: Определитель языков мира по письменностям. Москва: Наука 1965. 376 с.

JAKUBÍČEK, Miloš – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel– SUCHOMEL, Vít: Current Challenges in Web Corpus Building. In: Proceedings of the 12th Web as Corpus Workshop. Language Resources and Evaluation Conference (LREC 2020). Eds. A. Barbaresi – F. Bildhauer – R. Schäfer – E. Stemle. Marseille, 11–16 May 2020, 2020, pp. 1–4.

KHOKHLOVA, Maria: Large Corpora and Frequency Nouns. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". Ed. V. P. Selegey, Vol. 15(22). Moscow: RSUH 2016, pp. 224–238.

KHOKHLOVA, Maria – BENKO, Vladimír: Size of corpora and collocations: the case of Russian. In: Slovenščina 2.0, 2020, Vol. 8, No 2, pp. 58–77.

KUTUZOV, Andrey– KUNILOVSKAYA, Maria: Size vs. structure in training corpora for word embedding models: Araneum Russicum maximum and Russian national corpus. In: Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science. Eds. W. M. P. van der Aalst et al. 10716 LNCS. Cham: Springer 2018. https://doi.org/10.1007/978-3-319-73013-4_5

RINGLSTETTER, Christoph – SCHULZ, Klaus – MIHOV, Stoyan: Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. Computational Linguistics, 2006, 32(3), pp. 295–340.

ROSENTHAL: РОЗЕНТАЛЬ, Дитмар Э.: Справочник по правописанию и литературной правке. Москва: Айрис-пресс 2016. 368 с.

SHAPOVAL: ШАПОВАЛ, Виктор В.: Новые типы ошибок в письменной речи. In: Русский язык в школе, 2009, № 9, с. 76–83.

SHAVRINA – SOROKIN: ШАВРИНА, Татьяна О. – СОРОКИН, Алексей А.: Моделирование расширенной лемматизации для русского языка на основе морфологического парсера TnT-Russian. In: Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Ред. В. П. Селегей. Москва: Российский государственный гуманитарный университет 2015. URL: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ShavrinaTOSorokinAA.pdf.

SHAVRINA: ШАВРИНА, Татьяна Олеговна: Методы обнаружения и исправления опечаток: исторический обзор. In: Вопросы языкознания, 2017, № 4, с. 115–134.

sciendo
DOI 10.2478/jazcas-2022-0022

A PROJECT WORK AS A WAY OF BRINGING CORPORA TO SECONDARY SCHOOL

MARINA KOGAN[1] – VICTOR ZAKHAROV[2]
[1] Peter the Great St. Petersburg Polytechnic University
[2] Saint Petersburg University

KOGAN, Marina – ZAKHAROV, Victor: A project work as a way of bringing corpora to secondary school. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 986 – 995.

Abstract: Corpus linguistics is one of the most dynamic and rapidly developing areas of modern linguistics. It affects all areas of linguistics, including methodology of teaching foreign languages, translation and other linguistic disciplines. Corpus linguistics has had a direct impact on teaching foreign languages. However, in general, it remains a marginal method in teaching. Analysis of publications on the subject allows us to conclude that very few studies are long-term and aimed at working with schoolchildren. This article proposes a model for the development of sustainable interest among high school students in online corpora as sources of linguistic information, including the initiation stage in the form of project work in mini-groups to study well-known sayings with the consequent stage aiming at completing tasks supplementing the main textbook on a regular basis. The organization of project work addressing the corps of 11th grade students of the Natural Science Lyceum at Peter the Great St. Petersburg Polytechnic University is described. The paper outlines further research.

Key words: corpus linguistics, language pedagogy, longitudinal studies, method of projects/project work, proverbs, sayings

1. INTRODUCTION

Corpus linguistics (CL) is a relatively new direction in linguistics, which is engaged in the creation and use of corpora for solving various linguistic problems. Today's corpora have become an integral part of linguistics and one of the methodological cornerstones used for research in vocabulary, grammar, discourse. Corpora are increasingly used as the basis for dictionaries, course books, and grammars (Chambers, 2019, p. 461). CL's impact on English lexicography scholars compare with a "revolution" (Hanks, 2012). After the advent of corpora, all linguistic science including applied linguistics became different.

Its role in teaching foreign languages is also very important and since the early 1990s, enthusiasts have been talking about the revolution that CL approaches will bring to the field of teaching foreign languages. However, almost 30 years later, researchers are forced to admit that this revolution has not happened yet, with the only important exception of the "indirect application" of CL approaches in teaching through the use of

modern dictionaries and textbooks, which all are prepared on the basis of corpora. However, direct access to corpora by both foreign language (FL) teachers and students is so limited globally that Bolton and Cobb called it a marginal activity on the map of Language pedagogy, despite compelling evidence from a meta-analysis of 64 experimental papers on DDL efficiency and effectiveness in teaching English to different categories of students, with different levels of FL competence (Boulton – Cobb, 2017).

2. BACKGROUND. LITERATURE REVIEW

2.1 Problems hindering the widespread implementation of CL approaches in teaching foreign languages

Challenges hindering the widespread adoption of CL approaches in everyday teaching practice include:

1) insufficient ICT competence of FL teachers and insufficient training in the field of corpus linguodidactics;
2) an unusual format for presenting query results in the form of truncated concordance lines;
3) insufficient level of students' FL competence to analyze authentic examples contained in the corpus;
4) lack of FL teachers' readiness of preliminary preparing students to perform such tasks;
5) periodical changes of the resource interface and query syntax, which makes detailed manuals from previous years outdated (e.g. (Shaw, 2011);
6) lack of corpora created with students' needs (Cobb, Boulton, 2015, p. 4) and pedagogical goals in mind (Braun, 2007);
7) gap between the results of corpus studies and their applicability in teaching foreign languages (Charles, 2007, Chambers,2019);
8) adherence of language teachers to the priority of grammatical rules described in textbooks over language patterns fixed in corpora;
9) technical problems and problems associated with lack of time and financing (Chambers, 2019).

2.2 Analysis of longitudinal studies on using corpora by language learners

There are very few studies examining how deeply rooted the skill of using corpora is in students after the end of the experimental corpus course. To date, attention has focused primarily on student achievements and evaluation of corpus work immediately after a corpus course has been completed (Charles, 2014, p. 31).

Pérez-Paredes and Sánchez-Hernández (2018) found out that the use of corpora had limited or no impact on the writing practices of Spanish researchers two years after they had received in-service training on using corpora as a resource in writing scientific articles in their field.

The most consistent longitudinal research has been carried out by M. Charles. In one of her articles (Charles, 2014), she showed that out of 40 subjects 70% continued to use their created corpus a year later, of which 38% were active users who accessed the corpus every week and 32% were inactive users who accessed the corpus once a month on average. In her plenary talk at TaLC 2020 (Charles, 2020), M. Charles summarized the accumulated data for a longer research period from 2009 to 2017: she collected feedback from 221 participants one year after completing the course. During this period, the number of postgraduate students who has heard about corpora before the start of the course increased from 50 to 65%, but the number of those who use them has not changed (about 23%), with frequent users using corpora every week in their writing practice accounting for 10% of users. The results of the postponed survey one year after the end of the course show that the total number of users remains approximately stable at 62%, with the number of active users remaining at about 40%. These findings lead us to put forward the following hypothesis: the earlier students become familiar with corpora, the more chances there are that they will develop a sustainable interest in this type of linguistic resources and that they will be using them when necessary in their future studies of the foreign language. In the context of this hypothesis, two questions arise: what is the best students' age and material/activity for introducing corpora to them?

2.3 Corpora as a resource for organizing students' project activities

The documents regulating the goals and objectives of modern education in Russia pay great attention to the involvement of students in project activities. A special journal for school teachers of the foreign language, named "Inostrannye Jazyki v Shkole[1] [Foreign languages at school], regularly publishes articles on the theory and practice of using the project method in teaching foreign languages. However, as noted by Morozov – Urazayeva (2018), teachers very rarely turn to work with corpora although corpus linguistics provides a valuable tool for carrying out project activities in FL classes.

Here, one can mention the A. Boulton's paper (2011) on MA students being familiarized with corpus linguistics through the project activities aiming at using corpora and CL approaches for conducting research in students' fields of interest. This is another example of a long-term study, the results of which were presented at EuroCALL 2019 (Boulton, 2019) and TaLC 2020 conferences. In the reports from these gathering, it was noted that the students in question 'gradually got involved in the game' after overcoming initial difficulties. Obviously, the assignments requiring MA students to address the corpora described in this paper are not suitable for high school students. The direct transfer of corpus linguistic methodology onto FL instruction is not really appropriate in a secondary school context and it needs to be

[1] Web-site of the journal *Inostrannye Jazyki v Shkole*: https://iyash.ru/

harmonized with theories of FL instruction, i.e. with didactics (Wicher, 2020, p. 32). This implies that the assigned corpus tasks should be designed in close relation to the main coursebook used by learners.

With the overall aim to have the students explore on-line corpora from as many different angles and in as many different ways as possible, we decided to organize the introductory phase of corpus work in a hands-on mode with the COCA[2] and NOW[3] corpora in the format of a project work.

2.4 The selection of the research problem

The research problem for 11th grade students at the Natural Science Lyceum (NSL) in Petersburg should be designed in a way that that by solving them, the students will become familiar with corpus search techniques, the analysis of the results, with the possibility of testing their own hypotheses etc. The tasks should be relevant to language level and interests of students, they should take into account their strengths, go in line with the subject (discipline) syllabus i.e. be compatible with the course book prescribed for teaching and learning English in NSL, as well as to allow the teacher to organize team work. We think that the task to conduct corpus search and analysis of proverbs and sayings meet all the above mentioned requirements. Such a task will allow students to 'safely' start an independent corpus search avoiding most difficulties and 'dangers' described in literature on the one hand and to get familiar with the most important features of modern big online linguistic corpora on the other.

Proverbs and sayings are included in most English textbooks, so high school students are familiar with at least the most common ones. Researchers and EFL teachers see a great potential in use of proverbs and sayings to activate grammatical constructions, enrich vocabulary, develop communicative competence, play a relaxation game, etc. (Pavlova, 2010).

The results of the experiment with the first-year linguistic students to study proverbs and idioms using corpus resources and tools described by I. Komarova and M. S. Kogan (2019) are very promising: A survey conducted at the end of the course showed that the first year students sincerely liked the hands-on corpus tasks and found this method of learning idioms effective. We believe that the analysis of proverbs and sayings and how they are used in the corpus is an interesting project assignment for high school students, which will learn them corpora skills. Besides, proverbs and sayings, like many other phraseological units, are fixed in the corpora not only in their standard form, but also in different modifications. Revealing this fact contributes to development of traits crucial for true researchers/scientists: the

[2] Davies, M.: The Corpus of Contemporary American English (COCA). Available online at: https://www.english-corpora.org/coca/ (2008-).

[3] Davies, M.: Corpus of News on the Web (NOW). Available online at: https://www.english-corpora.org/now (2016-).

ability to discover and interpret new unexpected facts. In the context of our hypothesis, we consider this task to be a good starting point for involving students into independent corpus work in the future.

3. THE ORGANIZATION OF AND TASKS FOR THE PROJECT WORK

The study was conducted at NSL in Autumn 2020. The participants were 16 students of the 11th grade (aged 16-17). They are very good at math, physics, and informatics and have diverse interests. They have different levels of command of English because they came from different schools. Their interest in English classes in NSL vary from very dedicated to rather superficial.

In NSL, *Forward* textbook is used at English classes as the main coursebook. Each Unit has a vocabulary section with a number of collocations and fixed expressions to learn. Some Units focus on idioms and present at least some proverbs and sayings, including *To Err is Human…,* which is used as the Heading of Unit 3 (Grade 10) (Verbitskaya, 2016). The number of exercises for training the proverbs and idioms is insufficient, which is typical of most English language textbooks (Cobb, 2019). Each unit includes a so-called *Project ideas* section. According to the syllabus there are two 45 minutes long classes a week in the autumn term.

The main part of the 3-week Project work was performed outside the classroom, with a 30 min introductory lecture and one 45 min class allocated for mini-teams reports about their findings at the end of the project. During the introductory presentation the teacher set the goals of the project and did the following:

- Gave a brief but carefully prepared talk about corpus linguistics;
- Introduced briefly COCA and NOW corpora;
- Specified the corpus-based task for project work for each mini-team;
- Supplemented students with detailed written instructions on conducting corpus and dictionary research and requirements for the final Report;
- Formed 6 mini-teams of two to three students to deal with a particular proverb/saying.

Then, each team worked independently outside the classroom for two weeks. Finally, each team presented their findings in front of the class answering their classmates and teacher's questions.

Each mini-team was given one item from the following list of sayings' parts for qualitative and quantitative analysis in the COCA and NOW corpora during the project work.

1 To err is human…
2 The grass is always greener…
3 Necessity is the mother of…

4 who laughs last
5 Early to bed and early to rise…
6 Birds of a feather.

The students had to conduct quantitative research in the COCA in the mode *List* and both quantitative and qualitative research in the modes *Chart* and *KWIC*. From the *List* mode, they had to obtain the total frequency and number of occurrences of the proverbs/sayings in the canonic form and its modifications containing the given part. In the *Chart* mode, they had to visualize and comment the results by using the bar chart, with the aim of evaluating how the phrase has been used in different genres and time periods, of becoming familiar with a list of examples of the phrase (the given part of the saying) used in a particular genre (each member of the team could choose a genre independently). Finally, they had to make a conclusion about the canonic form of the saying and confirm the conclusion using Cambridge English Dictionary (https://dictionary.cambridge.org/).

In the *KWIC* mode with the Right end alignment function, the students had to study the proverb endings; calculate the number of instances of the standard/canonic and modified forms of the saying and provide examples of modified forms. In case of recurrent modifications, the students were asked to specify these modifications. Also, they had to provide examples of the saying endings which they found unusual, striking, or original. Another task drew students' attention to a sentence type: they had to notice if the sentence containing the saying or part of it is affirmative, negative, interrogative, an exclamation or an unfinished thought (marked with a series of 3 dots at the end of each sentence.)

In the *KWIC* mode with the Left end alignment function, the students had to calculate the number of instances of the saying at the beginning of a sentence; as a part of a sentence (the first word starts with a small letter); as a quotation used in the inverted commas. Finally, students had to learn the frequency of the canonic form of their saying in the COCA corpus.

The students had then to become familiar with the NOW corpus. The NOW corpus-related tasks were included into the project work because

1 We wished to provide a first-hand experience of work in more than one corpus to the students;
2 In this way, the students could see the influence the corpus size had on their search results;
3 The NOW corpus enabled us to familiarize students with some functions that are difficult to access in COCA, for example, with arriving at the source text following a hyperlink, a randomized sampling function that allows one to get 100 or 200 examples of the saying use in random, rather than in reverse chronological order. The students had finally to submit a Report on the results of the research.

4. RESULTS AND DISCUSSION

The participants showed a sincere interest in the CL introduction lecture. They asked some technical questions on corpus markup which is due to their general high level of computer science and programming knowledge. At the end of the lecture, there was a lively discussion on the issue of how often native English speakers use proverbs in their everyday speech. The discussion allowed the teacher to emphasize that this is a problem which needs exploration and the corpus research can help answer this question.

At the final lesson, all the students noted that they spent a lot of time on the task of multidimensional analysis of the saying part in the two corpora, which on average amounted to 5-6 hours per group, or about 2 hours per person. There were no complaints about the difficulties associated with the analysis of the concordance lines, which is in line with our expectations, as in the COCA there were fewer than 100 examples of the use of 5 out of 6 sayings (see Table 1).

Saying's part	COCA	NOW
To err is human	87	737
The grass is always greener	79	359
Necessity is the mother of	78	1043
who laughs last	26	115
Early to bed and early to rise	0	84
Birds of a feather	230	1678

Tab. 1. Search results for parts of proverbs/sayings in COCA and NOW

The selection of a randomized sample of 100 examples in the NOW corpus also allowed students to analyse them quite comfortably. The *KWIC* view mode with a given phrase in the centre, and the closest words coloured in different colours, with the total number of no more than 200 examples can be called a user-friendly interface, because it allows users to perform a convenient qualitative and quantitative analysis of the proverb use contexts both in the canonical form and in its various modifications.

The students conscientiously analysed the concordance lines and also described the trends in the use of their proverb over time, using the results obtained in the *Chart* view. They admitted the usefulness of the corpora for linguistic research, but complained that it was tiresome to analyse concordance lines. They noticed that although the modifications of sayings are multiple, the number of really unusual, striking examples of sayings' endings is very small.

The students enjoyed the opportunity of changing different modes / formats of visualization in *Chart* view mode (frequency by year (Fig. 1), all subsections at once (Fig. 2), frequency by country (Fig. 3).

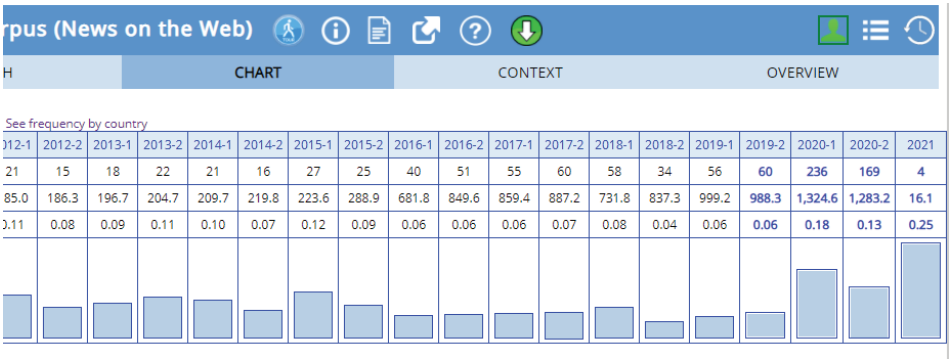


Fig.1. Frequency of *Necessity is the mother of* by year in the NOW corpus

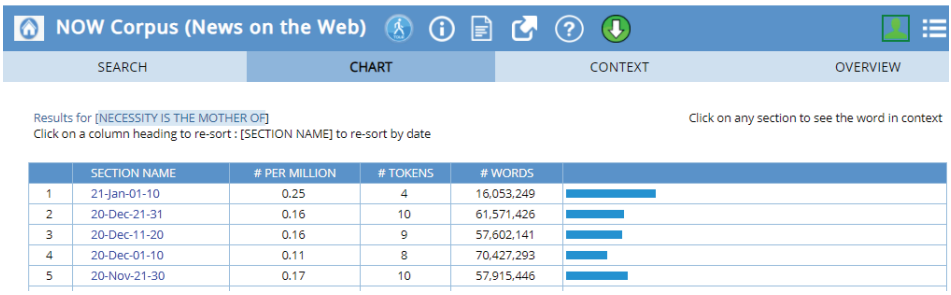


Fig.2. Results for *Necessity is the mother of* in the NOW corpus (all sections at once)

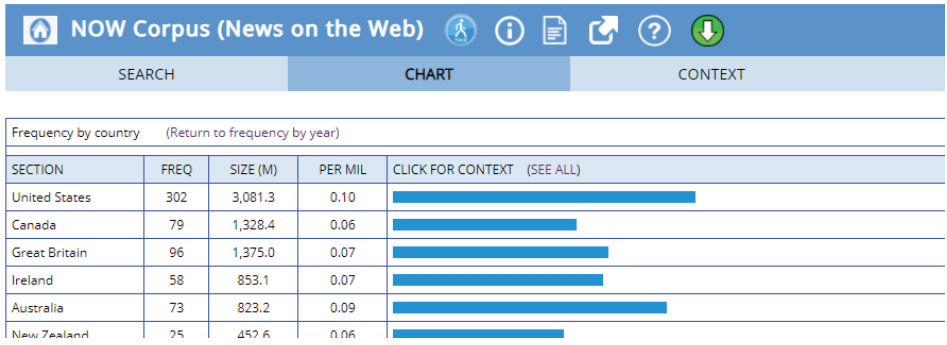


Fig.3. Results for *Necessity is the mother of* in the NOW corpus (frequency by country)

As an additional assignment, the students had to get familiarized with the extended usage context of their favourite modification of the saying searching in the COCA corpus. They were also asked to go to an external site in order to see the full text of the article searching in the NOW corpus. In the Reports, they noted that this was a good way to find out exactly how the unusual modification of the saying they liked is used.

While doing the project work, students faced some technical problems. When asked to elaborate what went wrong, they admitted that they often preferred a trial and error method of getting familiar with a new resource/software to following the instructions (some problems could have been avoided if the instructions had been followed). Another reported technical problem was loosing connection with the website when accessing it from mobile devices/smartphones. This might be attributed to the fact that the mobile version of the COCA website is less stable, because students who experienced it could not cope with it. Those students were unwilling to continue to use the corpus until the connection becomes more robust.

5. CONCLUSION

We believe that:

1. The selected material for the corpus analysis and the proposed format of work (mini-team project work) are effective for introducing high school students to modern online corpora.

2. This task contains elements of both true research and a computer game (the latter due to the unpredictability of answers and multiple searches).

3. We hope that our subjects have acquired basic skills of conducting the corpus research thanks to repeated use of the resource during the project work.

4. Some problems, typical for novice users of the COCA/NOW corpora, result from the complex nature of this resource.

5. Thus, despite some controversy in learners' responses to online corpora, we think that the suggested way of familiarizing learners with this linguistic resource (by means of a project work performed in mini-teams with hands-on the corpora as the initiation stage) is the correct first step on developing sustainable interest in corpora in high school students.

6. Further research is required to clarify the following issues:

1) how to incorporate corpus-based tasks into the learning process, with the aim of in-depth studying language phenomena, that is, as naturally as possible and in accordance with the main textbook content;

2) how to provide a support during students' individual and independent work with corpora while doing the assigned corpus-based tasks;

3) how to measure the impact of corpus-based work on students' language and technical skills development;

4) how to evaluate students' interest in and readiness/willingness to use this linguistic resource in the future.

Bibliography

BOULTON, Alex: Bringing corpora to the masses: Free and easy tools for interdisciplinary language studies. In: Corpora, Language, Teaching, and Resources: From Theory to Practice. Ed. N. Kübler. Bern: Peter Lang 2011, pp. 69–96.

BOULTON, Alex: First contact with language corpora: Perspectives from students. In: CALL and complexity. Eds. F. Meunier – J. Van de Vyver – L. Bradley – S. Thouësny. 2019, pp. 51–56. DOI : 1014705/rpnet201938985

BOULTON, Alex – COBB, Tom: Corpus use in language learning: a meta-analysis. In: Language Learning, 2017, Vol. 67, No 2, pp. 348–393.

BRAUN, Sabine: Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. In: ReCALL, 2007, Vol. 19, No 3, pp. 307–328.

CHAMBERS, Angela: Towards the corpus revolution? Bridging the research–practice gap. In: Language Teaching, 2019, Vol. 52, No 4, pp. 460–475. DOI:10.1017/S0261444819000089

CHARLES, Maggie: Reconciling top-down and bottom-up approaches to graduate writing: using a corpus to teach rhetorical functions. In: Journal of English for Academic Purposes, 2007, Vol. 6, No 4, pp. 289–302.

CHARLES, Maggie: Getting the corpus habit: EAP students' long-term use of personal corpora. In: English for Specific Purposes, 2014, Vol. 35, No 1, pp. 30–40.

CHARLES, Maggie: From take-up to take-off? A longitudinal view of data-driven learning in English for academic purposes. In: TALC 2020. Book of Abstracts. Ed. H. Tyne Perpignan 2020, pp. 7–8. Available at: https://langident.hypotheses.org/files/2020/07/Abstracts140720b.pdf

COBB, Tom: From Corpus to CALL: The use of technology in teaching and learning formulaic language. In: Understanding formulaic language: A second language acquisition perspective. Eds. A. Siyanova-Chanturia – A. Pellicer-Sanchez. London-New York: Routledge 2019, pp. 192–210.

COBB, Tom – BOULTON, Alex: Classroom applications of corpus analysis. In: Cambridge Handbook of English Corpus Linguistics. Eds. D. Biber – R. Reppen. Cambridge: Cambridge University Press 2015, pp. 478 – 497.

HANKS, Patrick: The Corpus Revolution in Lexicography. In: International Journal of Lexicography, 2012, Vol. 25, No 4, pp. 398–436. DOI:10.1093/ijl/ecs026

KOMAROVA – KOGAN: КОМАРОВА, Ирина А. – КОГАН, Марина С.: Исследование английской фразеологии с помощью подходов корпусной лингвистики. In: Компьютерная лингвистика и вычислительные онтологии: Труды XXII Международной объединенной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. СПб: Университет ИТМО 2019, с. 40–49. URL: https://ojs.itmo.ru/index.php/CLCO/issue/view/56

MOROZOV, Evgenii A. – URAZAYEVA, Nailya R.: The use of linguistic corpora in design activities on German language lessons at the university. In: Perspektivy nauki i obrazovania – [Perspectives of Science and Education], 2018, Vol. 36, No 6, pp. 187–195. DOI: 10.32744/pse.2018.6.21

PAVLOVA: ПАВЛОВА Елена А.: Приемы работы с пословицами и поговорками на уроках английского языка. In: Иностранные языки в школе, 2010, №5, с. 37–44.

PÉREZ-PAREDES, Pascual – SÁNCHEZ-HERNÁNDEZ, Purificación: Uptake of corpus tools in the Spanish Higher Education context: A mixed-methods study. In: Research in Corpus Linguistics, 2018, Vol. 6, pp. 51–66.

SHAW, Erin Margaret: Teaching Vocabulary through Data-driven Learning. Brigham Young University 2011, 130 p. Available at: https://scholarsarchive.byu.edu/etd/3024/.

VERBITSKAYA, Maria (ed.): Forward: English Student's Book. (Grades 10, 11). Moscow: Ventana-Graph. Person Education Limited 2016.

WICHER, Oliver: Data-driven learning in the secondary classroom: a critical evaluation from the perspective of foreign language didactics. In: Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners. Ed. P. Crosthwaite. London: Routledge 2020, pp. 31–46.

sciendo
DOI 10.2478/jazcas-2022-0023

CHINESE LANGUAGE WORD EMBEDDINGS BASED ON THE CORPUS HANKU

RADOVAN GARABÍK
Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

GARABÍK, Radovan: Chinese language word embeddings based on the corpus Hanku. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 996 – 1004.

Abstract: Vector models based on word embeddings are an indispensable part of advanced Natural Language Processing research and language analysis. We describe several Chinese language (Pǔtōnghuà) word embeddings, the differences from "western" language models caused by specific orthographic and linguistic features of the written Chinese language, and introduce a publicly available web interface for querying the vector models, aimed at linguistically or pedagogically oriented users.

Key words: word embeddings, Chinese, Pǔtōnghuà, corpus, NLP

1. INTRODUCTION

Recently, vector models based on word embeddings (Mikolov et al., 2013) became an indispensable part of advanced Natural Language Processing (NLP) research and language analysis. Originally conceived as a method working on raw, linguistically unannotated corpus (on the surface level of word forms), it has been often used in other configurations, e.g. on the space of lemmas, in order to better capture semantic values of the language, or on substring of words in the form of the fastText algorithm (Bojanowski et al., 2017), improving the analysis of inflected languages, without the need of "traditional" lemmatization and related NLP processes.

A vector space obtained by word embeddings is a very good model of semantic relations (compare Şenel et al., 2018); spatial relations between vectors correspond to semantic relations (similarities, differences, semantic categories, semantic clusters) between words. The models also extend into proper names; informally, we will speak about the "semantic closeness" and "synonyms" also for proper names, by which we mean the closeness of vectors in our models.

1.1 Chinese Language

Chinese as a macrolanguage is a group of language varieties of the Sinitic branch of the Sino-Tibetan languages. The modern prestigious and official variety (*Pǔtōnghuà* 普通话) is the common national speech of the Han nationality, using Beijing pronunciation as the standard pronunciation, Beijing speech as the basic dialect, and the model writing

of the modern vernacular prose as the norm for the grammar. It is based on northern dialects, in particular the standard written language is based on Beijing variant of Mandarin Chinese; and this is generally understood nowadays by the term "Chinese language". Modern Chinese language is in many respects, both inherently linguistic and sociolinguistic, quite different from other widespread languages:

- specific writing system, based on morphosyllabic script (*Hànzì* 汉字), where the basic units of the script – graphemes ("characters", *zì* 字) correspond to morphemes and syllables (with exceptions)[1] (Gajdoš, 2012)
- the language is almost completely isolating, words never change their form
- words are mostly bisyllabic
- the discrepancy between the spoken and written forms (Gajdoš, 2014)
- no space between words in writing
- in fact, the very notion of "word" is rather in flux; in Chinese corpus linguistics and NLP, word segmentation is a nontrivial challenge; the concept of "word" is even more weakened by the absence of a word stress and there is a significant disagreement among literate native speakers about the "correct" word segmentation (Sproat et al., 1996)
- significant amount of homophones

In the past, (Mandarin) Chinese has been marked by stark diglossia and stratification, with formal written texts being in Literary Chinese (*wényánwén* 文言文); in some aspects, this has been carried into contemporary language. A decisive factor for the discrepancy between the spoken and written forms, among other things, is the intellectualization of a language – the Literary Chinese is still one of the essential sources that affect the current (written) language in lexis and syntax. A consequence of these trends is the written language, which although based on the spoken language includes such "foreign" elements – the residue of the literary language *wényánwén* (Gajdoš, 2011). One important aspect of Literary Chinese is that words are mostly monosyllabic; later we discuss a vector model where this feature could be relevant.

2. CHINESE WORD EMBEDDINGS

There are some specifics when trying to make word embedding models of Chinese. Given the fact that most words are two characters long (corresponding to two syllables), the fastText algorithm would not be suitable for Chinese, either as written or even in some romanization. In many other languages (especially those using Latin/Greek/Cyrillic scripts) we can easily consider word embeddings to reflect a raw language, escaping the eventual trap of pre-existing linguistic bias, since the only

[1] Contemporary written Chinese often incorporates Latin script (Roman alphabet) elements, either as foreign (or even domestic) proper names (e.g. CNN, CCTV, QQ), abbreviations, or internet slang (e.g. CNM), often in combination with Arabic digits or Hànzì characters (2B, A片); this phenomenon is noticeably present already for some time (Hansell, 1994).

necessary prerequisite is tokenization, which can be performed quite efficiently and even universally (see e.g. Michelfeit et al., 2014). In Chinese, tokenization into words requires either statistical or rule based methods, introducing some amount of errors, and the exact way of segmenting text into words (or, looking from the opposite side, grouping individual characters into words) is subject to interpretation.

We compiled three models of simplified Chinese, based on the same source, the Chinese web corpus Hanku (Gajdoš et al., 2016) and the Chinese literature subcorpus.[2] The size of the web corpus is 1215 480 206 unicode characters; tokenized into words (*cí* 词), the size is 744 709 741 tokens. As expected from a web corpus, it contains a significant number of repeated texts – after deduplicating (on the paragraph level), the size of the corpus is 819 793 592 unicode characters, 501 782 955 tokens. The size of the whole corpus (deduplicated web corpus and the Chinese literature subcorpus) the word embeddings are trained on is 949 902 689 unicode characters, 594 461 715 tokens. The vectors are trained using skip-gram models, with 200 dimensions and a context window of 7 tokens (slight variations in these hyperparameters, as well as switching the model to Continuous Bag of Words do not change the overall results much). The models are downloadable from our webpage[3] in text Gensim format.

2.1 Model trained on the level of individual words

This model, labelled cí 词 is the closest to the usual web embedding usage. Basic units of the text are words, composed of one or several graphemes (characters). Tokenization is performed by ZPar (Zhang – Clark, 2011), with several enhancements – non-Hànzì elements in the text are separated from Hànzì characters, punctuation characters are tokenized individually, sequences of digits forming numerals are grouped together and tokenized as single tokens, similarly sequences of Roman letters are uppercased, grouped and tokenized as single tokens corresponding to words written in Roman alphabet. Roman characters used in conjunction with Hànzì are treated as parts of the word – thus A片 and 二B would be one token each, not two.

2.2 Model trained on the level of individual characters

This model, called *zì* 字 is compiled at the level of characters – basic units of the text are individual Hànzì characters. Almost identically to the *cí* 词 model, Roman alphabet elements are still uppercased and tokenized as separate tokens[4]. Combinations of Hànzì characters and Latin letters or digits are split into individual Hànzì characters and non-Hànzì remains (e.g. A片 will be tokenized as two tokens, A and 片, but 2B will be one token, unlike its variant 二B that is tokenized as two tokens). In this way, we hope to uncover semantic relations of Hànzì characters, if there are any.

[2] The Hanku corpus contains three subcorpora – the web subcorpus, the subcorpus of literary Chinese and the subcorpus of legal Chinese.

[3] https://www.juls.savba.sk/data.html

[4] We forgo the discreteness characterizing Roman letters in Chinese texts.

2.3 Model trained on Hànyǔ pīnyīn representation of words

There is a rather straightforward, though not completely unambiguous, one way transformation of Hànzì characters into their Hànyǔ pīnyīn transliteration (the opposite way is much more ambiguous). We included an automated transcription into Hànyǔ pīnyīn in our source corpus; the transcription was performed by the *xpinyin* package[5], however, no disambiguation of characters with multiple readings has been performed.

Building a special mode of the web interface that translates characters on the fly into Hànyǔ pīnyīn would be rather simple, but there would be no additional linguistic value in such an endeavour – one can always use an on-line transliteration service (see e.g. DZ Translit[6]) to obtain the same results.

Then there is the possibility to compile a vector model directly on the transliterated words (where the syllables within one word are concatenated together). Tokens transcribed in this way correspond to the 词 model, the transcription is a surjective function (each character in our transliteration is assigned only one reading). The model therefore mirrors the semantic relations of the 词 model, with the exception of relations of homophones (multiple characters with identical pronunciation), where we expect the corresponding vectors to fall to a different region of the semantic space, roughly between the expected meanings of the homophone original words in Hànzì (something we are used to when dealing with homonyms in word embeddings in other languages). To facilitate using the model, we mark the tones using digits 1 to 5 (neutral tone has the number 5), not the usual diacritical marks.

3. WEB INTERFACE

3.1 Modes of Operation

Word embeddings are quite easy to use; there are several mature OpenSource software frameworks, libraries and packages in major programming languages, providing both training and querying the models; or the vectors themselves can be imported into a mathematical/statistical software of one's choice. Nevertheless, this approach is somewhat cumbersome for casual users (such as teachers or learners of the language), or in linguistic research. We built a web interface to the models, with the intention to be used by both experienced linguists (or lexicographers) and laymen. The interface and some of the possibilities it offers has already been described (Garabík, 2020) and we just summarize the main points here (focusing on the Chinese language models[7]):

[5] https://lxneng.com/posts/70

[6] http://quest.ms.mff.cuni.cz/cgi-bin/zeman/translit/translit.pl

[7] We build several models per language; all the other (non-Chinese) languages use common methodology and model types (based on lemma, word form and word form using the fastText algorithm), given specific features of the Chinese language and writing system outlined above, this methodology is neither completely applicable nor optimal for Chinese. This is the main reason we treat the Chinese models separately, taking advantage of the features of the writing system to arrive at better results.

– Any (syntactically correct and using words existing in the corpus, i.e. a query that results in a valid vector) query will display a table of nearest words from the embedding model and a visualization graph, displaying the surroundings of the result, in either 2D, 3D or 4D projection, using ISOMAP dimensionality reduction.
– At the most basic usage, the portal works as a souped-up thesaurus. Querying a single word displays a table (see Table 1 for an example) containing words semantically close to the searched term, with a numeric value quantifying the "closeness" (defined as $\sqrt{1-\cos^2\varphi}$, where φ is the angle between the vectors corresponding to the two words). Note that the closeness need not be directly comparable across different models. We also point out that word embeddings do not deal with homonymy/polysemy well – if the same word has two different meanings, its vector will be roughly a mixture of both vectors, i.e. not corresponding to any of them; or, more realistically, one of the meanings dominates and the vector points to this meaning's region of semantic space.
– Querying two or more words displays similar table, showing the vectors close to all of the words (i.e. a normalized sum of their vectors), which reflects words that are semantically similar to all of the input words; the interface additionally shows the semantic closeness ($\sqrt{1-\cos^2\varphi}$) of the first two words, as a simple number from the interval [0, 1] to give the user a hint about the level of their synonymy.
– Simple vector arithmetic, consisting of addition and subtraction, is supported. The result of the expression is used as a vector around which we look for semantically close words and display the table of them in a similar manner to the previous usage cases.
– It is possible to query (uppercased) non-standard words in Roman alphabet or combinations of Hànzì characters and Roman letters or digits; these are treated as bona fide words in the *cí* 词 model and give valuable insight into modern Internet slang, a subset of lexicon that is often not covered by existing dictionaries.

3.2 Usage and Examples

The models can be used as a substitution of a thesaurus; for a given query, we get not only the semantically closest words, but also their semantic closeness – "true" synonyms have the value close to zero.

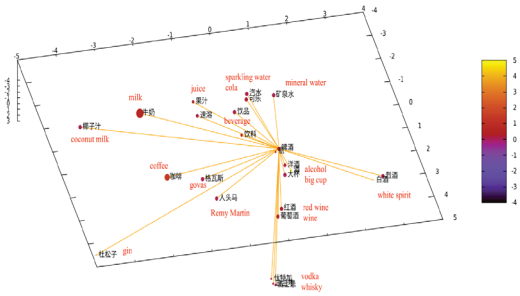
⊠	word	count	
0.000	龙	39537	G 百度 W
0.557	蛇	26177	G 百度 W
0.573	虎	13446	G 百度 W
0.574	争虎斗	47	G 百度 W
0.576	飞龙	1945	G 百度 W
0.598	凤	2769	G 百度 W

Table 1: Semantically closest words to the word lóng 龙 [dragon]. The second column is the word close to the query, the first column is the semantic closeness of the word, the third one number of

occurrences in the corpus, the fourth column contains links to external sources (Google search[8], Baidu search[9], English language Wiktionary[10]). Note that *zhēng hǔ dòu 争虎斗[11] is a phantom word, a relic of incorrect tokenization, as indicated by the low number of occurrences in the corpus (47). The translations of the words from top to bottom are: dragon; snake; tiger; *zhēng hǔ dòu fight against each other; flying dragon; phoenix (Chinese mythological bird).

	word	count	
0.000	TMD	1524	G 百度 W
0.313	他妈	4146	G 百度 W
0.341	煞笔	1043	G 百度 W
0.380	真尼玛	144	G 百度 W
0.387	它妈	209	G 百度 W
0.387	尼玛	4942	G 百度 W
0.397	妈逼	1638	G 百度 W
0.403	脑残	3800	G 百度 W
0.403	特么	2618	G 百度 W
0.404	傻B	805	G 百度 W
0.408	狗日	4421	G 百度 W
0.422	喷子	4146	G 百度 W
0.429	畜生	6882	G 百度 W

Table 2: Semantically closest words to the word (token) TMD – an example of using Roman letters as "native" parts of Chinese texts. The closest word tāmā 他妈 [damn it] with the semantic closeness of 0.313 is almost a synonym. We refrain from providing translations of the table, since we would have to include content warning for the benefit of our more sensitive readers.



Picture 1: 4D visualization of the word query píjiǔ 啤酒 [beer]. The fourth dimension is represented by different colours (probably not visible in the printed version of this article). We can see several semantic clusters around the term. We included translations of the words in the visualization.

[8] https://google.com

[9] https://www.baidu.com

[10] https://en.wiktionary.com

[11] The combination of characters is a part of the idiom *lóng zhēng hǔ dòu* 龙争虎斗 [fierce struggle between two evenly-matched opponents].

If there are at least two query terms (separated by either space or the plus sign), the interface calculates their semantic closeness and displays the value directly. For example, the closeness of the words *Sīluòfákè* 斯洛伐克 [Slovakia] and *Jiékè* 捷克 [Czech(ia)] is 0.312, much closer than *Rìběn* 日本 [Japan] and *Cháoxiǎn* 朝鲜 [North Korea] (0.638), which in our interpretation of the semantic model means that in a typical Chinese text, 日本 and 朝鲜 are perceived as rather different, but 斯洛伐克 and 捷克 are somewhat indistinguishable.

We noticed an interesting result – desemantized single characters in the *zì* 字 model are grouped together. This is somewhat surprising, because in Pǔtōnghuà these characters are almost exclusively used only for their phonetic value (e.g. in foreign language transcriptions) and not their original meaning, and the *zì* 字 model does not otherwise exhibit semantic properties of the individual characters, neither any obvious closeness of other classes of characters.

☒	word	count	
0.000	斯	723111	G 百度 W
0.322	尼	263755	G 百度 W
0.395	帕	43605	G 百度 W
0.397	尔	570967	G 百度 W
0.414	姆	67231	G 百度 W
0.456	迪	110605	G 百度 W
0.463	拉	482471	G 百度 W
0.483	弗	49477	G 百度 W
0.507	蒂	68818	G 百度 W

Table 3: Querying the 字 model for the character sī 斯. The original meaning of the character is desemantized and it is used only for its phonetic value sī. The whole region of our vector space (i.e. the semantic space) around this character is devoid of meaning – all the "semantically close" characters returned by our vector model (in the table) are used only for their phonetic values. From top to bottom: sī, ní, pà, ěr, mǔ, dí, lā, fú, dì.

3.3 Vector Arithmetic

One of the distinguishing, powerful and somewhat surprising features of word embedding models is working vector arithmetic – subtraction and addition of words has straightforward semantic interpretation, as a transfer to a different place in the multidimensional semantic space. Our web interface supports simple vector arithmetcs, consisting of addition and subtraction of (arbitrary number of) vectors.

The prototypical example used to demonstrate vector arithmetic in word embeddings is the "equation" *king* − *man* + *woman* = *queen* (or a local language equivalent), and we would like to use an appropriate Chinese language equivalent for demonstration purposes. The Chinese term for *king*: *guówáng* 国王 is an unassuming word not deeply connected with Chinese history, thus we use *huángdì*

皇帝 [emperor] instead. The equation 皇帝 − 男人 + 妇女 (i.e. *huángdì* 皇帝 [emperor] − *nánrén* 男人 [man] + *fùnǚ* 妇女 [woman]) gives *tàihòu* 太后 [empress dowager or the mother of an emperor] as the semantically closest frequent word (almost as expected; although not what we usually get for the query in "European" languages, it is quite understandable given Chinese history[12]). On the other hand, 皇帝 − 他 + 她 (*huángdì* 皇帝 [emperor] − *tā* 他 [he] + *tā* 她 [she]) gives *huánghòu* 皇后 [empress consort or wife of a ruling emperor[13]] and 皇帝 − 他 + 它 (*huángdì* 皇帝 [emperor] − *tā* 他 [he] + *tā* 它 [it]) gives *huángquán* 皇权 [imperial power]. Let's recall that Chinese does not use gendered personal pronouns and the distinction in writing between masculine, feminine and neutrum 3rd person pronouns has been introduced at the beginning of 20th century under the influence of "modern and progressive" western languages; nevertheless, the vector transfer clearly reflects semantic properties of these pronouns as written in modern Chinese.

Demonstrating a geographical example, we know the traditional Chinese drink is tea – what would, in the eyes of the word embeddings model, be the French equivalent? The query 茶叶 + 法国 − 中国 (i.e. *cháyè* 茶叶 [tea leaves] + *Fǎguó* 法国 [France] - *Zhōngguó* 中国 [China]) gives *hóngjiǔ* 红酒 [red wine] as the semantically closest word. We can interpret it as the typical product corresponding to tea leaves, if we make a transfer from the Chinese region of the semantic space to the "French" one (that is, France as written about in the Chinese language corpus). Similarly, 茶 + 法国 − 中国 (i.e. *chá* 茶 [tea] + *Fǎguó* 法国 [France] - *Zhōngguó* 中国 [China]) gives the result *kāfēi* 咖啡 [coffee] as the (whether right or wrong) typical French beverage corresponding to the *tea* in China in the mental image of an average(d) Chinese speaker.

For comparison, 茶 + 日本 − 中国 (i.e. *chá* 茶 [tea] + *Rìběn* 日本 [Japan] − *Zhōngguó* 中国 [China]) gives *qīngjiǔ* 清酒 [sake] as the Japanese semantic equivalent of Chinese *tea* (again, from Chinese perspective).

CONCLUSION

Word embeddings in modern written Chinese benefit from a specific approach, compared to naïve straightforward application of existing algorithms and software tools and packages. Models based on words (*cí* 词) give expected results, conditioned on word segmentation of adequate quality. By tokenizing sequences of Roman letters

[12] In ancient China, empresses were unheard of – there was only one ruling empress, *Wǔ Zétiān* 武则天 of the Zhōu (late Táng) dynasty (and the wife of a ruling emperor was usually not politically significant). Since many emperors ascended the throne as children, the emperor's mother would often possess notable political power. Perhaps the best known example is Empress Dowager *Cíxī* 慈禧 of the Qīng dynasty.

[13] As opposed to the female ruling monarch; both of these roles are covered by the English term *queen*. This is sometimes disambiguated in a European context by two two-word terms *queen regnant* and *queen consort*.

and combinations of non-Hànzì and Hànzì characters we obtain information of semantic relations of these unconventional words, often used in online Chinese slang, a register seldom covered in existing dictionaries.

We provide a web interface for casual or less technically oriented users that provides basic query methods within the word embedding models, returning a list of semantically related results, allowing quantifying semantic relatedness, and providing several visualization methods.

Bibliography

BOJANOWSKI, Piotr – GRAVE, Edouard – JOULIN, Armand – MIKOLOV, Tomáš: Enriching word vectors with subword information. In: Transactions of the Association for Computational Linguistics, 2017, No. 5, pp. 135–146.

GAJDOŠ, Ľuboš – GARABÍK, Radovan – BENICKÁ, Jana: The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. In: Studia Orientalia Slovaca, 2016, Vol. 15, No. 1, pp. 21–33.

GAJDOŠ, Ľuboš: The discrepancy between spoken and written Chinese methodological notes on linguistics. In: Studia Orientalia Slovaca, 2011, Vol. 10, No. 1, pp. 155–159.

GAJDOŠ, Ľuboš: Čínsky jazyk a čínske písmo. In: Historická revue, 2012, Vol. 23, No. 7, pp. 47–50.

GAJDOŠ, Ľuboš: Synsémantické slová v rámci stratifikácie čínskeho jazyka. In: Miscellanea Asiae Orientalis Slovaca. Bratislava: Univerzita Komenského 2014, pp. 121–131.

GARABÍK, Radovan: Word Embedding Based on Large-Scale Web Corpora as a Powerful Lexicographic Tool. In: Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 2020, Vol. 46, No. 2, pp. 603–618.

中华人民共和国中央人民政府: 国务院关于推广普通话的指示, 1956. Available online: http://www.gov.cn/test/2005-08/02/content_19132.htm

HANSELL, Mark: The Sino-Alphabet: The Assimilation of Roman Letters into the Chinese Writing System. In: Sino-Platonic Papers, 1994, Vol. 45, pp. 1–28.

MICHELFEIT, Jan – POMIKÁLEK, Jan – SUCHOMEL, Vít: Text Tokenisation Using unitok. In: 8th Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU 2014, pp. 71–75.

MIKOLOV, Tomáš – CHEN, Kai – CORRADO, Greg – JEFFREY, Dean: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR 2013.

ŘEHŮŘEK, Radim – SOJKA, Petr: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.

ŞENEL, Lutfi Kerem – UTLU, İhsan. – YÜCESOY, Veysel – KOÇ, Aykut. – ÇUKUR, Tolga: Semantic structure and interpretability of word embeddings. In: EEE/ACM Transactions on Audio, Speech and Language Processing, 2018, Vol. 26, No. 10, pp. 1769–1779.

SPROAT, Richard W. – SHIH, Chilin – GALE, William – CHANG, Nancy:. A stochastic finite-state word-segmentation algorithm for Chinese. In: Computational Linguistics, 1996, Vol. 22, No. 3, pp. 377–404.

ZHANG, Yue – CLARK, Stephen: Syntactic Processing Using the Generalized Perceptron and Beam Search. In: Computational Linguistics, 2011, Vol. 37, No. 1, pp. 105–151.

POKYNY PRE AUTOROV

Redakcia JAZYKOVEDNÉHO ČASOPISU uverejňuje príspevky **bez poplatku** za publikovanie.
Akceptované jazyky: všetky slovanské jazyky, angličtina, nemčina. Súčasťou vedeckej štúdie a odborného príspevku je abstrakt v angličtine (100 – 200 slov) a zoznam kľúčových slov v angličtine (3 – 8 slov).
Súčasťou vedeckej štúdie a odborného príspevku v inom ako slovenskom alebo českom jazyku je zhrnutie v slovenčine (400 – 600 slov) – preklad do slovenčiny zabezpečí redakcia.
Posudzovanie príspevkov: vedecké príspevky sú posudzované anonymne dvoma posudzovateľmi, ostatné príspevky jedným posudzovateľom. Autori dostávajú znenie posudkov bez mena posudzovateľa.
Technické a formálne zásady:

- Príspevky musia byť v elektronickej podobe (textový editor Microsoft Word, font Times New Roman, veľkosť písma 12 a riadkovanie 1,5). V prípade, že sa v texte vyskytujú zvláštne znaky, tabuľky, grafy a pod., je potrebné odovzdať príspevok aj vo verzii pdf alebo vytlačený.
- Pri mene a priezvisku autora je potrebné uviesť pracovisko.
- Text príspevku má byť zarovnaný len z ľavej strany, slová na konci riadku sa nerozdeľujú, tvrdý koniec riadku sa používa len na konci odseku.
- Odseky sa začínajú zarážkou.
- Kurzíva sa spravidla používa pri názvoch prác a pri uvádzaní príkladov.
- Polotučné písmo sa spravidla používa pri podnadpisoch a kľúčových pojmoch.
- Na literatúru sa v texte odkazuje priezviskom autora, rokom vydania a číslom strany (Horecký, 1956, s. 95).
- Zoznam použitej literatúry sa uvádza na konci príspevku (nie v poznámkovom aparáte) v abecednom poradí. Ak obsahuje viac položiek jedného autora, tie sa radia chronologicky.

Bibliografické odkazy:

- knižná publikácia: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda, vydavateľstvo SAV 2008. 204 s.
- slovník: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV 2011. 1088 s.
- štúdia v zborníku: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda, vydavateľstvo SAV 2000, s. 111 – 117.
- štúdia v časopise: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, roč. 60, č. 1, s. 3 – 12.
- internetový zdroj: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 2010. Dostupný na: http://korpus.juls.savba.sk [cit. DD. MM. RRRR].

INSTRUCTION FOR AUTHORS

JOURNAL OF LINGUISTICS publishes articles **free of publication charges**.
Accepted languages: all Slavic languages, English, German. Scientific submissions should include a 100-200 word abstract in English and a list of key words in English (3-8 words).
Scientific articles in a language other than Slovak or Czech should contain a summary in Slovak (400-600 words) – translation into Slovak will be provided by the editor.
Reviewing process: scientific articles undergo a double-blind peer-review process and are reviewed by two reviewers, other articles by one reviewer. The authors are provided with the reviews without the name of the reviewer.
Technical and formal directions:

- Articles must be submitted in an electronic form (text editor Microsoft Word, 12-point Times New Roman font, and 1.5 line spacing). If the text contains special symbols, tables, diagrams, pictures etc. it is also necessary to submit a pdf or printed version.
- Contributions should contain the full name of the author(s), as well as his/her institutional affiliation(s).
- The text of the contribution should be flush left; words at the end of a line are not hyphenated; a hard return is used only at the end of a paragraph.
- Paragraphs should be indented.
- Italics is usually used for titles of works and for linguistic examples.
- Boldface is usually used for subtitles and key terms.
- References in the text (in parentheses) contain the surname of the author, the year of publication and the number(s) of the page(s): (Horecký, 1956, s. 95).
- The list of references is placed at the end of the text (not in the notes) in alphabetical order. If there are several works by the same author, they are listed chronologically.

References:

- Monograph: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda, vydavateľstvo SAV 2008. 204 p.
- Dictionary: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV 2011. 1088 p.
- Article in a collection: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda, vydavateľstvo SAV 2000, pp. 111 – 117.
- Article in a journal: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, Vol. 60, No 1, pp. 3 – 12.
- Internet source: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 2010. Dostupný na: http://korpus.juls.savba.sk [cit. DD. MM. YEAR].

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Objednávky a predplatné prijíma/Orders and subscriptions are processed by:
SAP – Slovak Academic Press, s. r. o., Bazová 2, 821 08 Bratislava
e-mail: sap@sappress.sk

Registračné číslo 7044

Evidenčné číslo 3697/09

IČO vydavateľa 00 167 088

Ročné predplatné pre Slovensko/Annual subscription for Slovakia: 12 €, jednotlivé číslo 4 €
Časopis je v predaji v kníhkupectve Veda, Štefánikova 3, 811 06 Bratislava 1