ČASOPIS PRO MODERNÍ FILOLOGII 103, 2021, Č. 1, S. 36–59

OPEN ACCESS

Kolokační grafy a sítě s použitím nástroje #LancsBox: aplikace v angličtině a češtině*

Václav Březina (Lancaster) – Petr Pořízka (Olomouc)

COLLOCATION GRAPHS AND NETWORKS USING #LANCSBOX: APPLICATIONS IN ENGLISH AND CZECH

This article deals with the notion of collocation graphs and lexical networks, which not only represent the visualization of the collocational relationship between linguistic units — these have been traditionally displayed in a tabular form with frequency distributions and association measure values — but also an important analytical method in its own right. We illustrate the use of collocation graphs and networks with two case studies as examples demonstrating the use of this technique in lexicography and discourse analysis. The examples are based on both English and Czech corpora, which we analysed using #LancsBox, a free tool which can build collocation graphs and networks on the fly.

KEYWORDS

association measures, collocations, collocation graphs and networks, corpus, quantitative analysis, #LancsBox, conceptual metaphor, contrastive analysis, MI-score

KLÍČOVÁ SLOVA

asociační míry, kolokace, kolokační grafy a sítě, korpus, kvantitativní analýza, #LancsBox, konceptuální metafora, kontrastivní analýza, MI-score

DOI

https://doi.org/10.14712/23366591.2021.1.2

ÚVOD

Bezprostřední kontext, který v textech sdílejí lexikální výrazy, nám poskytuje jedinečnou možnost analyzovat základních struktury jazyka i interakci mezi slovy. V zásadě lze říci, že kolokace je jev spojený s opakovaným výskytem dvou (či více) slov v textu. V české terminologii tak hovoříme o souvýskytu slov. Záměrně zde používáme velice širokou definici kolokace, kterou různí autoři pojímají z různých

* Poděkování: Autoři by chtěli poděkovat dvěma anonymním recenzentům za vnímavé a faktické komentáře, jež bezesporu pomohly tento text vylepšit. Poděkování patří i redakci časopisu, zejména profesoru Klégrovi, za podporu při psaní a revizi článku.
Příspěvek vznikl za podpory MŠMT ČR udělené UP v Olomouci (IGA_FF_2020_021 „Bohemistika: literárněvědné a lingvistické přesahy a interpretace“). Prezentovaný výzkum byl dále podpořen následujícími granty: Granty ESRC č. EP/P001559/1,ES/K002155/1 a ES/R008906/1.

hledisek. Nemáme tedy na mysli jen lexikální jednotky a ustálená spojení typu *cestovní ruch* a *tratoliště krve*, ale též volnější asociace mezi slovy v diskurzu. Z tohoto důvodu také v případových studiích používáme poněkud široký kolokační rozsah pět slov vlevo a pět slov vpravo od nodu (viz kapitola 1). Kolokaci zde tak chápeme jako široký pojem s řadou definic (např. Firth, 1957; Sinclair, 1991; Čermák, 2006; Wray, 2008; Paquot — Granger, 2012; Gries 2013). V tomto textu, jenž vychází z předchozích výzkumů a studií publikovaných na toto téma v angličtině (Brezina, 2018b) a doplňuje je kontrastivní rovinou mezi angličtinou a češtinou, se tak navracíme k jednoduchému firthovskému pojmu kolokace definovanému jako „obvyklý souvýskyt slov" („the habitual co-occurrence of words") (Firth, 1957, s. 2). Tento souvýskyt je přitom statisticky identifikován užitím celé řady asociačních měr (blíže Evert, 2008; Pecina 2010; Gablasova et al., 2017b). Tyto asociační míry jsou nástroje, které operacionalizují konkrétní definici nebo teoretické pojetí kolokace. Předkládaná studie rovněž navazuje na metodologii korpusové lingvistiky, jež poskytuje empirický základ pro kvantitativní identifikaci a následnou extrakci kolokací. Například slovní spojení *v zásadě* z úvodu tohoto článku tvoří kolokaci,[1] jež se v reprezentativním korpusu češtiny SYN2015 vyskytuje 2041krát (průměrně 16,9krát v 1 milionu slov).[2]

Při hledání kolokací se však nemusíme omezovat jen na pouhou frekvenci souvýskytu dvou slov, příp. na výrazy s vysokou četností — v této souvislosti jsme dosud hovořili o obvyklém nebo opakovaném souvýskytu. Často je vhodnější podívat se na exkluzivní vztah mezi dvěma výrazy, jejichž spojení nemusí být tak frekventované. K tomuto účelu slouží kolokační míry jako MI-score (Pořízka, 2014, s. 40). Tabulka 1 uvádí výrazy identifikované právě prostřednictvím MI-score, vyskytující se v jedinečném (exkluzivním) vztahu se slovním tvarem *zásadě*. Jak je vidět, nejedná se jen o ustálená slovní spojení (*zásada rovnoprávnosti*), ale i o obecnější asociace se samotným výrazem (*v*) *zásadě* (např. *dvojím, irelevantní, neliší)*. MI-score (Church — Hanks, 1990) je jednou z nejvyužívanějších běžných asociačních měr (statistických metod) používaných v korpusové lingvistice právě k detekci a extrakci kolokátů, má ale i své nevýhody, s nimiž je třeba při analýze počítat — vliv výrazů s nízkou četností, detekce spíše příznakovějších, neobvyklejších kolokátů ad. (srov. Pořízka, 2014, s. 39n.).

Ke kvantitativní informaci — číselné hodnotě vyjadřující míru asociace mezi dvěma slovy — lze rovněž vygenerovat graf (obr. 1), tj. vizuální reprezentaci zkoumaných slovních spojení. Právě tato vizualizace kolokátů je užitečnou interpretační technikou vhodnou k analýze komplexnějších jazykových vztahů v korpusových datech. Tradičně jsou totiž kolokace chápány jako diskrétní jevy prezentované ve formě tabulkových seznamů kolokátů (tab. 1). Naší snahou je mj. ukázat, že kolokační spojení jsou nicméně propojenými entitami, jež mohou být zobrazeny v podobě kolokačních grafů a sítí.

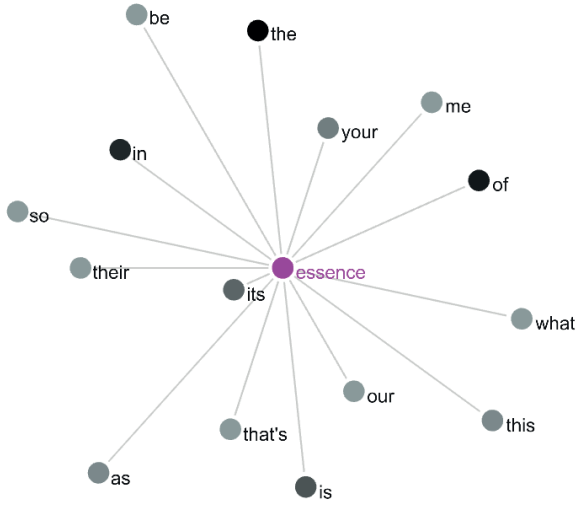
1 V určitém pojetí může být tato vazba chápána jako koligace, společný výskyt slov tvořící gramatickou vazbu. My se však zde přidržíme obecnějšího pojetí kolokace zahrnující jak lexikální, tak gramatické výrazy, a termín koligace ponecháme pro abstraktnější gramatické vztahy (např. sloveso + předložka), srov. Hunston, 2001.

2 Blíže viz https://wiki.korpus.cz/doku.php/cnk:syn2015 (poslední přístup: březen 2021).

	kolokát	MI-score
1.	rovnoprávnosti	10.695
2.	dvojím	10.367
3.	didaktické	10.235
4.	irelevantní	10.024
5.	dvojího	9.837
6.	rozlišujeme	9.782
7.	dokončeny	9.668
8.	namítat	9.457
9.	restituce	9.289
10.	neliší	9.104

TABULKA 1. Kolokáty slovního tvaru (typu) *zásadě* v korpusu SYN2015 (CPN: 03 – MI-score (3), L3–R3, C:5-NC:5)[3]

Pro ilustraci uvádíme kolokační graf vytvořený na základě BNC2014 Baby +, korpusu psané a mluvené angličtiny o pěti milionech slov (Brezina, 2019).



OBRÁZEK 1. Kolokační graf slovního tvaru *essence* v BNC2014 Baby + (CPN: 03 – MI-score (3), L3–R3, C:3-NC:3)

3 V závorce používáme u všech příkladů v textu tzv. kolokační notaci — CPN = Collocation Parameters Notation (blíže Brezina et al., 2015, s. 144n.). Formalizovaný zápis z tab. 1 má následující význam: ve třetím sloupci hodnoty MI-score s prahovou hodnotou 3 (03 – MI-score (3)); rozsah kontextu 3 výrazy vlevo i vpravo (L3–R3); minimální četnost kolokátu (C) i kolokace (NC) 5 výskytů. Blíže vysvětlujeme CPN zde v kap. 2.

Graf na obr. 1 vykresluje slovní tvary v blízkosti výrazu *essence* (esence, zásada) v kolokačním okně tří slov vlevo i vpravo. Výraz *essence* zde terminologicky označujeme jako **uzel**, tedy **(základové) slovo, o které se zajímáme**. Délka spojnic (hran) v grafu vyjadřuje sílu kolokace (i zde měřenou prostřednictvím MI-score): **čím je kolokace blíže k uzlu, tím silnější vztah vyjadřuje** (detailněji viz kapitola 2). Tento poměrně jednoduchý kolokační graf lze rozšířit do složitější kolokační sítě, což je komplexnější typ grafu, k němuž přináleží více uzlů s jejich jedinečnými a sdílenými kolokáty. Jednoduché kolokační grafy i komplexní kolokační sítě proto mají potenciál nejen účinně vytvářet přehled o datech, ale také, jak chceme ukázat i v této studii, přinést nové možnosti jazykové analýzy. Zatímco kolokační grafy poskytují užitečnou vizuální reprezentaci nejdůležitějších kolokátů kolem základového uzlu, a představují tak exploratorní alternativu k tradičnímu seznamu s nalezenými kolokáty, kolokační sítě jdou ještě o krok dále, naznačujíce složitější či komplexnější vztahy širší skupiny výrazů mezi sebou (uzlů). Kolokační sítě ukazují nejen sílu asociace, ale i její komplexnost vykreslenou jedinečnými i sdílenými kolokáty. To vše jsou informace o užití či jazykovém chování výrazů, jež nejsou z tradiční formy evidence kolokací zřejmé či snadno dostupné.

Myšlenka kolokačních sítí se opírá o původní výzkum Phillipse (1985) a byla použita například ve studiích terminologie (Williams, 1998), historického a sociálního vývoje jazyka (McEnery, 2006), ale i online diskurzu (Brezina, 2016). Donedávna však analýza kolokačních sítí vyžadovala značnou manuální práci. Se zveřejněním softwarového nástroje #LancsBox (Brezina et al., 2015), který kolokace automaticky identifikuje a jejich sítě dynamicky generuje, se tento analytický úkol stal pro lingvisty mnohem lépe zvládnutelným a přístupnějším.[4]

Tento článek nejprve pojednává o konceptu kolokačních grafů a sítí. Poté následují dvě případové studie, které ukazují použití tohoto konceptu jak v angličtině, tak i v češtině. Tyto studie jsou z oblasti lexikografie a analýzy diskurzu a využívají právě nástroj #LancsBox.[5] Co se týče samotného zaměření tohoto článku, jedná se o konceptuální studii ilustrující a kriticky hodnotící vizualizační techniku kolokačních grafů a sítí používaných v celé řadě oblastí korpusového výzkumu. Příklady prezentované

4 Je třeba poznamenat, že #LancsBox není jediným nástrojem vytvářejícím kolokační grafy a sítě, je ale jedním z nástrojů uživatelsky nejpřívětivějších (kolokační analýza je relativně velmi snadná) a volně dostupných, nekomerčních. Kolokační grafy produkuje např. i CONE (Gullick et al., 2010), nabízí však jen omezený počet možností identifikace kolokací. I korpusový nástroj Sketch Engine (dostupný z: https://www.sketchengine.eu/) umožňuje kromě tradičního způsobu vytěžování kolokátů generovat a vizualizovat vztahy mezi výrazy prostřednictvím tzv. word sketch diferencí (Kilgarriff et al., 2010; srov. též https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/) (cit. 2. 3. 2021). Jedná se ale bohužel o proprietární software, byť je v současnosti akademickým institucím v EU poskytován bezúplatně (grantová podpora projektu Elexis do března 2022). Kolokační grafy lze v neposlední řadě vytvářet i pomocí obecných (nelingvistických) softwarových nástrojů jako Gephi (dostupný z: https://gephi.org/).

5 Dostupný z www: http://corpora.lancs.ac.uk/lancsbox/ (cit. 2. 3. 2021), aktuálně ve verzi 6.0.

v tomto příspěvku a případové studie nemají za cíl poskytnout ucelený obraz o tématech, kterých se dotýkáme, ani prezentovat plně srovnatelnou diskusi o rozdílech mezi českými a anglickými kolokacemi. Namísto toho se zaměřujeme na ilustraci řady aspektů a metodických rozhodnutí, která musí výzkumník učinit při aplikaci dané techniky v konkrétním výzkumném kontextu.

1. KOLOKAČNÍ GRAFY A SÍTĚ: VYMEZENÍ POJMŮ A CHARAKTERISTIKA KONCEPTU

Přestože jsou kolokační grafy a sítě identifikovány a generovány pomocí nástroje #LancsBox automaticky, musíme vždy učinit řadu klíčových rozhodnutí, jež se týkají nastavení parametrů vytěžování kolokací: volba asociační míry, prahová hodnota (minimální hodnota asociační míry), minimální frekvence výrazů, kolokační rozsah atd. Podrobněji o těchto parametrech pojednává Gablasova a kol. (2017b). Pro standardizovaný zápis těchto parametrů navrhli Brezina et al. (2015) systém nazvaný Collocation Parameter Notation (CPN, zápis kolokačních parametrů) používaný i v tomto textu. Parametry a informace, jež je třeba uvést, shrnuje tabulka 2: identifikační číslo asociační míry,[6] název asociační míry (statistické metody) + její prahová hodnota, kolokační rozsah vlevo (L) a vpravo (R), minimální frekvence kolokátu + minimální frekvence kolokace a jakékoli další filtry použité během analýzy.

identifikační číslo asociační míry	název asociační míry	prahová hodnota	rozsah vlevo (L) a vpravo (R)	minimální frekvence kolokátu (C)	minimální frekvence kolokace (NC)	filtr
4b	MI2	3	L5–R5	5	1	funkční slova odstraněna

TABULKA 2. Notace kolokačních parametrů (Brezina et al.)[7]
4b-MI2 (3), L5–R5, C:5-NC:1; funkční slova odstraněna

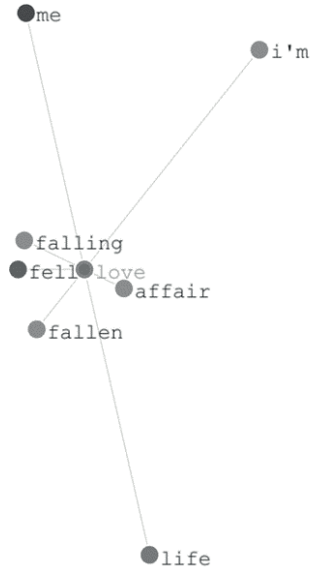
Souhrnně řečeno, aspekty kolokačního vztahu, jež lze prozkoumat metodou kolokačních grafů a sítí, jsou: i) **frekvence**, ii) **síla asociace**, iii) **poloha**, iv) **kolokační jednotka** a v) **kolokační propojení**. Všechny rysy lze v jazykové analýze využít, jak ilustrují případové studie níže v textu.

Ilustrujme nyní na konkrétním příkladu rozdíl (a) v tradiční, statické prezentaci kolokací a (b) dynamické reprezentaci prostřednictvím kolokačního grafu a sítě. Tabulka 3 ukazuje prvních sedm kolokátů substantiva *love* (láska) v kolokačním okně L3–R3 (tři slova vlevo i vpravo) z korpusu BE06 současné psané britské angličtiny (Baker, 2009). Kolokace jsou setříděny podle hodnoty MI-score.

6 S odkazem na seznam asociačních měr v publikacích Brezina et al., 2015, a Brezina, 2018.
7 V posledním řádku tab. 2 je uveden konkrétní příklad, navrhovaná forma zápisu.

kolokát	MI-score	AF (kolokace)	AF (korpus)
affair	8.86	5	37
fell	8.52	14	131
falling	8.52	5	47
fallen	8.37	5	52
me	5.57	23	1667
i'm	5.30	5	437
life	5.12	8	791

TABULKA 3. Kolokáty výrazu *love* v BE06 (CPN: 03 – MI-score (5), L3–R3, C:5-NC:5)



OBRÁZEK 2. Kolokační graf: *love* (přehledné zobrazení) v BE06 (CPN: 03 – MI (5), L3–R3, C:5-NC:5)

Tytéž informace mohou být vizualizovány jako kolokační graf (obr. 2) ukazující vztah mezi uzlem a jeho kolokáty, konkrétně vždy podle zvolené asociační míry. Délka spojnice (hrany) mezi uzlem a kolokáty je přitom nepřímo úměrná síle asociace: čím je asociace silnější, tím je spojnice kratší.

Jak je možno vidět na obr. 2, nejsilněji spojenými kolokáty k výrazu *love* podle MI-score jsou *affair, fell, falling* a *fallen*. Ve srovnání s tím představují výrazy *me, I'm* a *life* zřetelně méně těsné spojení. Graf dále zobrazuje i frekvenci jednotlivých kolokací (souvýskytů uzlu + kolokátu): tmavší barevný odstín je odrazem častějšího souvýskytu jednotek. Například ve vnějším kruhu kolokátů jsou *me* a *life* častějšími kolokáty než *I'm*.

V symetrickém okně (jako 3L–3R) je další kolokační dimenzí, již lze změřit, pozice kolokace v textu. Některé z kolokací se vyskytují v syntaktických (lineárních) pozicích, které v textu předcházejí uzlu, jiné v pozicích, jež následují. Například různé

formy slovesa *fall* vždy předcházejí výrazu *love*, aby vytvořily frázi *fall in love* (zamilovat se); ve srovnání s tím slovo *affair* vždy následuje za *love,* aby tak vytvořilo slovní spojení *love affair* (milostná aféra). Tato forma zobrazení kolokátů, která ilustruje jejich přesnou (syntaktickou) polohu v textu, ale vede k překrývání (jako na obrázku 3 níže), pokud se víceré kolokáty vyskytují v téže pozici. Obrázek 3 také ukazuje převládající tendenci jednotlivých kolokátů objevit se spíše vlevo či převážně vpravo od svého uzlu. Tato pozice je v grafu stanovena na základě výpočtu podílu případů vyskytujících se nalevo či napravo. Například substantivum *life* se obvykle objevuje až za výrazem *love*, jako v příkladu (1):

(1) *The BBC was clear that Mr Blunkett's* ***love life*** *was absolutely his own affair.* (BE06, B)
[BBC jasně uvedla, že **milostný život** pana Blunketta je jeho čistě soukromá záležitost.]

V některých případech však může *život* předcházet *lásce,* srov. níže:

(2) *Do You Believe In* ***Life*** *After* ***Love****?* (BE06, K)
[Věříte v **život** po **lásce**?]

Proto je na obrázku 3 kolokát *life* zobrazen s pravostrannou tendencí, ale nikoli zcela vpravo od uzlu (*love*). *I'm*, podobně jako *life,* má pravostrannou tendenci. Nutno upozornit, že rozhodnutí, zda je kolokát zobrazen nad uzlem či pod ním, je motivováno čitelností (přehledností zobrazení) a nevztahuje se k žádné z vlastností kolokace. Kromě uvedených dvou kolokátů s tendencí objevovat se spíše vpravo od uzlu ukazuje graf na obrázku 3 i kolokát *me,* gravitující mírně vlevo s převahou příkladů jako (3):

(3) *Dad needed to leave* ***me*** *to* ***love*** *the next child.* (BE06, K)
[Táta **mě** musel opustit, aby mohl **milovat** další dítě.]

Pozice v grafu nám tedy pomáhá interpretovat jazykové chování kolokací a poskytuje užitečné shrnutí typických syntaktických či textových pozic, v nichž se kolokáty vyskytují.

Doposud byly základními jednotkami kolokační analýzy slovní tvary (*wordforms*, *types*). Než budeme pokračovat, musíme učinit důležitou terminologickou poznámku. Terminologie se v odborné literatuře liší, pokud jde o užití výrazů jako *slovní tvar* (*wordform*), *lemma, typ* (*type*) a *slovníková forma* (*headword*). V tomto článku definujeme následující termíny v souladu s hlavními publikacemi v angličtině takto:

LEMMA: Skupina slovních tvarů, které jsou spojeny tím, že jsou flektivními tvary stejného základního slova. Lemma je obvykle označeno touto základní formou nebo kmenem slova.

TYP: (a) Jeden konkrétní slovní tvar. Jakýkoli rozdíl ve formě (například pravopis) dělá ze slova jiný typ.

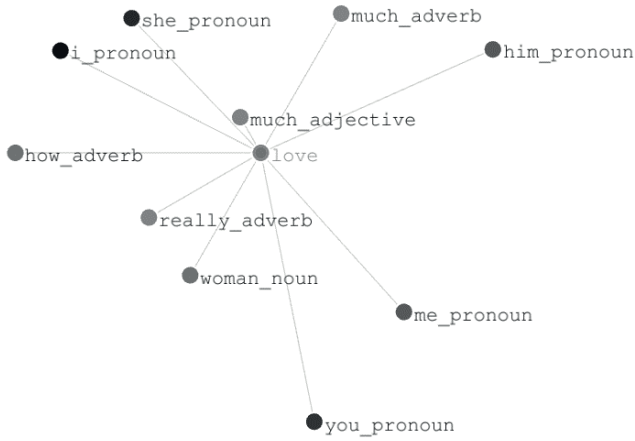
OBRÁZEK 3. Kolokační graf (poziční pohled): tvar *love* v BE06 (CPN: 03 – MI (5), L3–R3, C:5-NC:5)

Obě definice jsou převzaty z McEnery — Hardie (2011, s. 245, 252). Srovnej též podrobnou diskusi o tom, co je slovo v Brezina (2018, s. 38–41). V tomto úzu se termín *slovní tvar* používá synonymně s termínem *typ*. Naopak, *slovníková forma (headword)* se odlišuje od *lemmatu* v následujícím smyslu: Lemma je abstraktní třída zahrnující více slovních tvarů, zatímco slovníková forma je jeden tvar (základní forma slova nebo kmen), který se používá k reprezentaci celého lemmatu. Zdá se, že se česká terminologie[8] od toho úzu odlišuje v rozlišování typu jako nadřazeného termínu, který lze realizovat slovním tvarem nebo lemmatem např. ve frekvenčních seznamech slov a ve výpočtech *token–type ratio* atd. Zřetelně se zde též nerozlišuje mezi slovníkovou formou a lemmatem.

V tabulce 3 a na obrázku 2 se různé gramatické (flektivní) tvary slovesa *fall* objevily jako tři různé kolokáty (*fallen, falling* a *fell*). Pro některé typy analýz může být naopak užitečné či vhodné pracovat s lemmaty (viz i pozn. níže o důležitosti volby základní jednotky v souvislosti s typologicky různými jazyky). Lze například nastavit *love* jako lemma slovesa *to love*. Tímto způsobem vyloučíme všechna nominální užití výrazu *love*, ale zahrneme do analýzy všechny slovesné tvary, včetně *loves, loving, loved*. Budeme-li tedy pracovat s lemmaty, výsledný graf bude odlišný, jak je vidět na obrázku 4. Uzly i všechny kolokáty v grafu představují slovníkovou formu výrazů a slovnědruhovou příslušnost a zastupují všechny slovní tvary přináležející k danému lemmatu (paradigmatu).

Volba základní jazykové jednotky kolokační analýzy je metodologicky velmi důležitá, tradičně si konkurují **type** (slovní tvar) a **lemma**. Tato volba nabývá důležitosti obzvláště u tak typologicky rozdílných jazyků, jako je angličtina (analytický typ) a čeština (flexivní typ). Zatímco v případě angličtiny nemusí být rozdíl v užití type/lemma během analýzy týchž textů nijak zásadní (srov. ale Sinclair, 1991, kde se dis-

8 Přehled základních pojmů korpusové lingvistiky, dostupný z https://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu (cit. 2. 3. 2021).



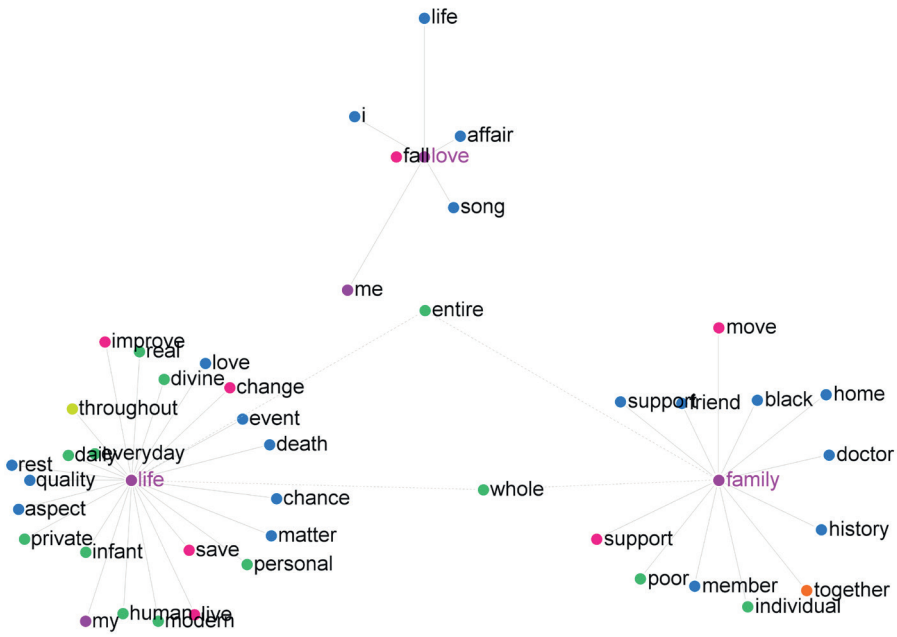
OBRÁZEK 4. Kolokační graf — poziční pohled, lemmatizovaný: lemma *love* v BE06 (CPN: 03 – MI (5), L3–R3, C:5-NC:5)

kutuje o důležitosti jednotlivých slovních tvarů v kolokacích pro angličtinu) a běžně je základní jednotkou typ, u češtiny může být v mnohých situacích vhodnější lemma vzhledem k flektivnímu charakteru jazykového systému. Lemma je totiž jednotka zastupující všechny textové tvary vážící se k jednomu paradigmatu. Jinak řečeno: výběr jednotky (slovní tvar vs. lemma) by mohl být pro flektivní jazyky důležitější než pro jazyky neflektivní povahy, aby byly do kolokační analýzy zahrnuty všechny typy daného lemmatu (paradigmatu).

Nakonec musíme zvážit ještě jednu důležitou vlastnost kolokací — **konektivitu** (Phillips, 1985; Brezina et al., 2015). Kolokace v jazyce i diskurzu vstupují do bohaté **sítě významových a křížových asociací**. Významové asociace představují jednoduchý vztah mezi dvěma výrazy (uzlem a jeho kolokátem), zatímco křížové asociace zahrnují skupinu širších vztahů mezi více slovy v rámci určitého textu/diskurzu. Analýza a vizualizace těchto sítí nám tak může pomoci odkrýt důležité vztahy v jazyce (textu či diskurzu). Kolokační konektivitu nelze snadno a efektivně zachytit tradičním způsobem (kolokačním seznamem); vhodnou formou zobrazení této konektivity je komplexní graf nazývaný **kolokační síť** (viz obrázek níže).

Obrázek 5 reprezentuje kolokační síť, jež ukazuje spojení mezi třemi různými substantivními lemmaty: *love* (láska), *life* (život) a *family* (rodina). Barvy kolokátů v grafu označují jejich slovnědruhovou příslušnost, např. modrá značí substantiva, zelená adjektiva a červená slovesa. Jak můžeme vidět, *love* je spojeno s *family* nepřímo, prostřednictvím kolokátu *life*. *Family* a *life* navíc sdílejí dva společné kolokáty: *whole* a *entire*. Tyto kolokáty se vyskytují v anglických spojeních *whole life/family* a *entire life/family* (celý život / celá rodina). Prozkoumat ale můžeme i kolokace, jež jsou jedinečné (nesdílené) pro každý ze tří uzlů.[9]

9 Pro další teoretickou diskusi o konceptu kolokačních sítí a jeho využití v analýze viz Phillips, 1985; Brezina et al., 2015; Baker, 2016, a Brezina, 2016.



OBRÁZEK 5. Kolokační síť: *love*, *live* a *family* v BE06 (CPN: 03 – MI (5), L3–R3, C:5,0-NC:5,0)

2. PŘÍPADOVÁ STUDIE 1: KOLOKAČNÍ SÍTĚ V LEXIKOGRAFII

Jazykové korpusy jsou pro lexikografické účely využívány velmi často (Granger — Paquot, 2014). Typický slovníkový popis lexému obsahuje definici slova (Hanks, 2016) a některé další související informace (např. výslovnost, etymologii, morfologická specifika, příklady použití atd.). Jak již bylo zdůrazněno jinde (Bejoint, 2016, s. 21), slovníky pro běžného uživatele jen zřídka obsahují slova sémanticky související; a pokud jsou zahrnuty, často se omezují jen na základní sémantické vztahy jako synonymie a antonymie, hyponymie a meronymie (Murphy, 2015). Kromě konkordanční analýzy lexikografové stále častěji používají k zachycení významových nuancí slova právě kolokace. Z korpusových analýz je zřejmé, že nad rámec těchto lexikálních vztahů jsou slova v jazyce i diskurzu propojena prostřednictvím bohaté sítě konceptuálních vazeb (Cope et al., 2011).

Ještě dále jdou v dnes již klasické knize Lakoff a Johnson (1980) uvedením tzv. konceptuálních metafor; tedy metafor obsažených v našem myšlení či jeho strukturaci. Otázkou je, zda lze prostřednictvím korpusových dat a pomocí kolokačních analýz (sítí) najít pro konceptuální metafory, jak je definují právě Lakoff a Johnson, empirický podklad. Pokud ano, kolokační sítě by nám mohly pomoci při lexikografickém popisu slov nad rámec obvyklých analýz používaných v elektronické korpusové lexikografii. V této studii jsme pro naše analýzy využili konceptuálně metaforické dvojice pojmů (i) *čas* a *peníze* a (ii) *láska* a *cesta*.

2.1 METODA A DATA

Tato případová studie využívá dva korpusy psaného jazyka: (1) BE06 — korpus psané britské angličtiny o rozsahu jeden milion slov (Baker, 2009), (2) KFS — český Korpus funkčních stylů,[10] 4milionovou databázi obsahující 6 tradičních funkčních stylů. BE06 obsahuje 15 hlavních žánrů (registrů) psané angličtiny, včetně publicistiky, akademických textů a beletrie; každý text v korpusu obsahuje vzorek daného žánru (registru) o velikosti 2 000 slov.

Pro tuto případovou studii jsme si vybrali dva konceptuálně metaforické páry pojmů Johnsona a Lakoffa (1980, 2002), první pro angličtinu, druhý pro češtinu. Záměrně jsme vybrali různé konceptuální metafory pro ilustraci rozličných sémantických domén. Záměrem tedy není komparace obou jazyků z hlediska konceptuálních metafor — to by vyžadovalo samostatnou detailní studii.

— TIME a MONEY
— LÁSKA a CESTA

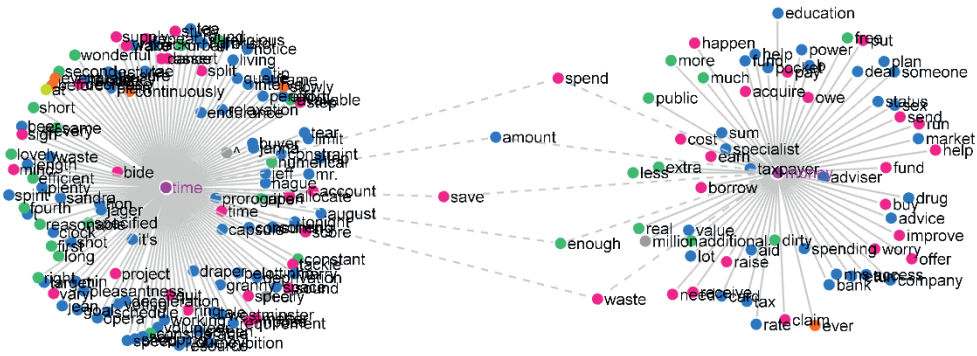
Pomocí nástroje #LancsBox jsme vytvořili kolokační sítě kolem klíčových výrazů (viz výše), abychom tak prozkoumali lexikální (konceptuální) síť vztahující se k těmto klíčovým lexémům. Nastavení pro angličtinu i češtinu bylo totožné: základní jednotkou bylo lemma, aby byly do analýzy zahrnuty všechny tvary daného morfologického paradigmatu; asociační mírou bylo MI-score, s prahovou hodnotou 5 pro statistiku a 3 pro četnost kolokace.

2.2 VÝSLEDKY A DISKUSE

Obr. 6 prezentuje výsledky analýzy z anglického korpusu BE06. Vidíme, že každý z klíčových pojmů má své vlastní, jedinečné kolokáty (TIME: 133 a MONEY: 62); uprostřed se nacházejí kolokáty sdílené oběma lexémy (5). V tomto případě ale není analýza zaměřena na jedinečné kolokace kolem každého z uzlů (nadto většinou z grafu bezprostředně nečitelné),[11] ale právě na sdílené kolokáty, s nimiž jsou tyto hlavní uzly, tj. dva klíčové pojmy, spojeny.

10 Je důležité si uvědomit, že pojem funkční styl není v současné české literatuře nekontroverzní. Pro podrobnější diskusi viz Hoffmannová et al. (2016) a Cvrček et al. (2020). Při výběru korpusů jsme se zaměřili na srovnatelnost ve velikosti a struktuře mezi BE06 a KFS. Místo toho, abychom využili velké korpusy, jako je BNC pro angličtinu a ČNK pro češtinu, jsme záměrně zvolili menší databáze. Tím jsme chtěli ilustrovat funkčnost nástroje #LancsBox, který umí importovat a zpracovat (lemmatizovat a přiřadit morfologické značky) korpusy, které má uživatel k dispozici. V mnoha výzkumných kontextech tyto funkce nástroje #LancsBox představují analytickou výhodu.

11 #LancsBox také zobrazuje seznam všech kolokátů ke každému z uzlů v přehledové tabulce.



OBRÁZEK 6. Konceptuální metafora: TIME IS MONEY v BE06 (CPN: 03 – MI (5), L5–R5, C:3,0-NC:3,0)

	lemma	frekvence v korpusu	frekvence ve spojení s time	frekvence ve spojení s money
1	amount_n	124	9	4
2	enough_adj	112	6	3
3	save_v	106	6	3
4	spend_v	241	52	21
5	waste_v	37	14	7

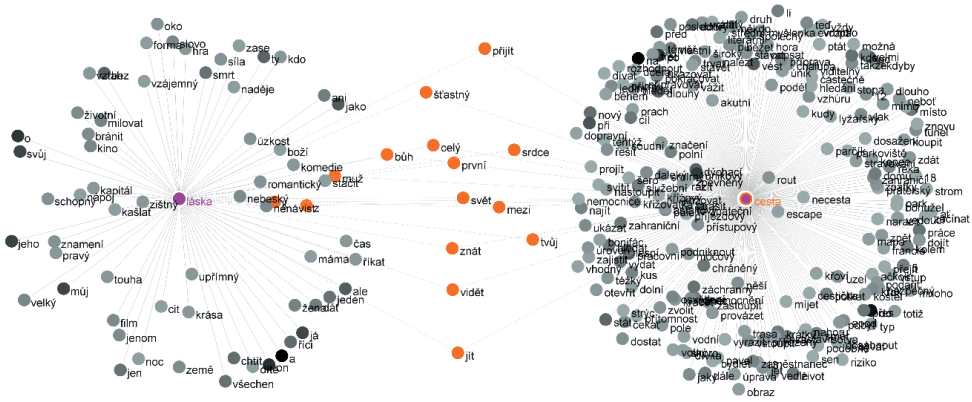
TABULKA 4. Sdílené kolokáty lexémů TIME a MONEY (data BE06)

Tato analýza TIME a MONEY (ČAS a PENÍZE) se sítí sdílených kolokací ukazuje, že průnik obou pojmů charakterizují slova *amount* (množství) a *enough* (dostatek), výrazy kvantifikační povahy, a dále se vyskytují ve spojení se slovesy *save* (šetřit), *spend* (trávit/utrácet) a *waste* (plýtvat), s nimiž se, jak dokládá frekvenční seznam v tabulce 4., klíčové výrazy pojí (nej)častěji.

Použití sdílených kolokátů lze ilustrovat na následujících příkladech:

(3) *You'll be **spending** a significant **amount** of **time** together on the day and...* (BE06, E) [V ten den spolu **strávíte** značné **množství času** a...]
(4) *FOREIGN AID — **time** to **spend** our **money** on our own people!* (BE06, F) [ZAHRANIČNÍ POMOC — **čas utratit peníze** za naše vlastní lidi!]
(5) *...in the race and **save wasted time** veering off-course.* (BE06, E) [... v závodě a **ušetřit promarněný čas** vybočením z kurzu.]
(6) *Regardless of how much **time**, effort, or **money** you've **spent** building an iPhone application, Apple...* (BE06, E) [Bez ohledu na to, kolik **času**, úsilí nebo **peněz** jste **strávili/utratili** vytvářením aplikace pro iPhone, Apple...]

Druhou dvojicí našich analýz konceptuálních pojmů je LÁSKA a CESTA v českém korpusu funkčních stylů (KFS; Pořízka, 2021). Pro vytvoření kolokační sítě (obrázek 7) jsme využili celý korpus o velikosti 3,3 mil. slov (s interpunkcí 4 mil.) obsahující styl



OBRÁZEK 7. Konceptuálně metaforické pojmy: LÁSKA a CESTA v KFS (CPN: 03 – MI (5), L5–R5, C:3-NC:3)

umělecký (U), odborný (O), publicistický (P), administrativní (A), řečnický (R) a hovorový (H, někdy též: prostě sdělovací).

Každý z pojmů má opět své jedinečné kolokáty (LÁSKA — 77; CESTA — 230) a celkem 15 kolokátů sdílených, mezi něž patří substantiva *bůh*, *svět*, *srdce*, *muž*; slovesa *jít*, *přijít*, *znát*, *vidět*; výrazy kvantifikační povahy *první* a *celý*; dále adjektivum *šťastný*, posesivní zájmeno *tvůj* a nakonec předložky *mezi*, *z* a *k* vyjadřující prostorovost či směrovost (srov. tabulka 5).

	lemma	frekvence v korpusu	frekvence ve spojení s láska	frekvence ve spojení s cesta
1	bůh	706	8	8
2	svět	2099	11	24
3	srdce	908	11	8
4	muž	1716	5	10
5	jít	5386	10	45
6	přijít	2517	7	17
7	znát	1293	6	15
8	vidět	3077	6	20
9	první	3361	8	20
10	celý	3692	7	31
11	šťastný	513	8	9
12	tvůj	1990	15	11
13	mezi	3487	9	37
14	z	28168	58	175
15	k	20141	75	187

TABULKA 5. Sdílené kolokáty lexémů LÁSKA a CESTA (data KFS)

Níže opět uvádíme z našich dat několik konkordancí sdílených kolokátů:

(7) Začíná zamotaný příběh o **lásce**, smrti a širém **světě**, příběh, který diváka nepustí od začátku do konce. (KFS, P)
(8) Poesie nemá jiný cíl nežli konstatovat a dále rozvíjet formy a kreace úděsnosti, lhostejnosti, zběsilosti, **lásky** a smrti v konkrétním **světě**. (KFS, U)
(9) Jediný platný vztah je láska. Platí, že **Bůh je láska**. (KFS, U)
(10) Ale to jsem neměl říkat, poněvadž Julda to **srdce** chtěl pořád **vidět**, jestli na tom **srdci** je ta **láska vidět**. (KFS, U)
(11) To je ta **cesta**, kterou **ze** sna **znám**. (KFS, U)
(12) Jeho hlas dozněl na chodbě venku a v sále se chýlila ke konci píseň lidového dua: Havaj, ty země květů, lžeš **lásku celému světu** havajských kytar sněníím… (KFS, U)

V některých dokladech nalezneme více než jeden sdílený kolokát — srov. př. 10: *srdce*, *vidět* + uzel LÁSKA; př. 11: *znát*, *z/ze* + uzel CESTA a př. 12: *svět*, *celý* + uzel LÁSKA. Př. 9 zase jako by skutečně naplňoval či demonstroval konceptuální metaforu NĚCO je NĚCO, zde BŮH *je* LÁSKA.

V obou výše uvedených analýzách (*time*/*money*, *láska*/*cesta*) byly detekovány jako sdílené kolokáty (dále též SK) především autosémantické slovní druhy (výrazy významově samostatné). U pojmů TIME a MONEY v anglickém korpusu BE06 jsou všechny sdílené výrazy plnovýznamové; v českém korpusu KFS u klíčových výrazů LÁSKA a CESTA jsou autosémantika, pokud k nim připojíme i pronominální výraz *tvůj*,[12] k synsématikům v procentuálním poměru 80 : 20.

Zda lze použít techniku kolokačních grafů a sítí pro komplexnější analýzu konceptuálních metafor, je třeba v budoucnu prověřit dalšími analýzami, neboť na datech KFS se ukazuje, že jsou výsledky ovlivňovány například typem textu. Tabulka 6 prezentuje dílčí subanalýzy provedené na pojmech LÁSKA a CESTA, jež tento vliv dokládají.

V tabulce 6 nás zajímají především poslední dva sloupce, jež uvádějí jednak počet sdílených výrazů nalezených v různých textových subkorpusech a jednak poměr autosémantických slovních druhů na jedné straně ke slovním druhům ostatním (synsémantickým, specifickým).[13] Největší vliv na výsledek subanalýz vykazuje beletrie (data s texty B), neboť v subkorpusech bez těchto textů buď výrazně klesá počet detekovaných sdílených kolokátů, nebo je nulový (0–3). V datech s literárními texty

12 Zájmena nejsou autosémantika, ale mají jejich distribuci a kopírují jejich vlastnosti (Karlík et al., *CzechEncy*, heslo *Slovní druh*, dostupné online: https://www.czechency.org/slovnik/SLOVN%C3%8D%20DRUH).

13 Klasifikací slovních druhů pro češtinu je několik. Nejužívanější je dělení z akademické *Mluvnice češtiny* (1986, s. 16n.), jež dělí nezákladní slovní druhy na nástavbové a nesamostatné, příp. klasifikace z *Příruční mluvnice češtiny*, kterou používáme v tomto textu (Karlík et al., 1997, s. 73–75). Mezi autosémantika tak řadíme i numeralia a mezi specifické slovní druhy pronomina a interjekce.

data — (sub)korpus	velikost (mil.)	kolokáty k láska	kolokáty k cesta	sdílené kolokáty (SK)	SK — poměr autoS / synS+spec
KFS-all	3,3	77	230	15	12:3
KFS-B+P	2,0	67	155	13	10:3
KFS-B+O	2,0	60	172	14	10:4
KFS-P+O	1,4	18	87	3	0:3
KFS-B	1,4	51	122	12	8:4
KFS-P	0,7	14	36	3	0:3
KFS-O	0,7	3	49	0	null
KFS-A	0,5	0	26	0	null

TABULKA 6. Kolokační profil výrazů LÁSKA a CESTA v rozdílných typech textů (korpus KFS)[14] — čísla vyjadřují četnost výskytu (absolutní frekvenci)
Vysvětlivky: KFS = korpus funkčních stylů; all = všechny funkční styly, B = beletrie, P = publicistika, O = odborná lit., A = administrativa; SK = sdílené kolokáty, autoS = autosémantika, synS = synsémantika, spec = specifické slovní druhy

bylo k uzlům *láska* a *cesta* nalezeno 12 až 15 sdílených výrazů (sloupec 5), vždy přitom dominantně převažují kolokáty plnovýznamové (sloupec 6). Naopak v datech „nebeletristických" šlo vždy o sdílené kolokáty synsémantické, příp. slovnědruhově specifické. V textech odborných a administrativních nebyly nalezeny žádné sdílené kolokáty, v subkorpusech s publicistikou pouhé tři sdílené synsémantické výrazy: v KFS-P+O předložky *k*, *mezi*, *z* a v KFS-P předložky *k*, *na*, *z*.

Seznamy sdílených kolokátů z dat obsahujících literární texty se příliš neliší, což může ukazovat na dominanci beletrie pro kolokační analýzu, nebo obráceně řečeno: jiné texty než beletristické měly na výslednou podobu extrahovaných výrazů jen minimální vliv — viz přehled níže (abecedně tříděno):

— KFS-all: *bůh, celý, jít, k, mezi, muž, první, přijít, srdce, svět, šťastný, tvůj, vidět, z, znát*
— KFS-B+P: *bez, bůh, celý, k, mezi, muž, první, přijít, srdce, svět, šťastný, z, znát*
— KFS-B+O: *bez, bůh, celý, jít, k, mezi, první, přijít, srdce, svět, šťastný, vidět, z, znát*
— KFS-B: *bez, bůh, celý, k, pro, první, přijít, srdce, svět, šťastný, z, znát*

Toto porovnání ukazuje, že různé typy textů (diskurzu) mohou odrážet jiné jazykové zacházení mluvčích/pisatelů s klíčovými lexikálně-sémantickými pojmy. Vztáhneme-li to k výroku Lakoffa a Johnsona, že „metaforické jsou do značné míry procesy lidského myšlení" (Lakoff — Johnson, 2002: s. 18), mohla by tato případová studie naznačovat, že metaforické myšlení není univerzální, že může být diskurzivní, tedy reflektující různé komunikační situace a vázané na konkrétní jazykové (kon)texty.

14 Do těchto subanalýz nebyly zahrnuty texty stylu řečnického a hovorového, a to pro nedostatečnou velikost dat — dohromady obsahují jen cca 100 slov.

Podobné závěry jsou ale vzhledem k fázi výzkumu prozatím spíše hypotetické, přestože určité souvislosti tato případová studie naznačuje. Je totiž známo, že výsledné kolokace jsou podmíněny mj. i výběrem asociační míry (srov. Křen, 2006), a tak bude třeba tyto dva možné faktory — vliv typu textu a/nebo asociační míry — v budoucnu prověřit dalšími analýzami.

3. PŘÍPADOVÁ STUDIE 2: KOLOKAČNÍ SÍTĚ V ANALÝZE DISKURZU

Podívejme se nyní na vnímání „východoevropských přistěhovalců" analýzou kolokačních asociací s výrazy *immigrant* (přistěhovalec) a *immigrants* (přistěhovalci) v komentářích čtenářů pod články dvou britských novin: *The Guardian* a *Daily Mail*. Každé z těchto periodik přitahuje jiný typ čtenáře. Zatímco *Guardian*, který lze charakterizovat jako typ seriózních novin politicky nakloněných doleva, přitahuje čtenáře, jejichž hodnoty se s pozicí novin typicky shodují, *Daily Mail* jsou novinami pro masový trh politicky tíhnoucí doprava a mají obecnější čtenářskou obec (McNair, 2009). Komentáře vyjadřují subjektivní názory, perspektivy a ideologie čtenářů v reakci na konkrétní obsah novinových článků. Těžištěm této dílčí analýzy je otázka imigrace, téma, jež bylo již dříve široce zkoumáno nejrůznějšími metodami, včetně metod korpusové lingvistiky (např. Blinder — Allen, 2016; KhosraviNik, 2009; Gabrielatos — Baker, 2008).

Samotnou analýzu je nutno uvést malou historickou poznámkou, jež doplňuje kontext výzkumu: v lednu 2014 otevřela Británie svůj trh práce občanům z Rumunska a Bulharska. V přípravě na tento politický krok probíhaly v britském tisku časté diskuse o možných dopadech tohoto rozhodnutí na britskou ekonomiku a kvalitu života v Británii. Média také srovnávala podobnou událost, jež se stala o deset let dříve (2004), kdy se trh práce otevřel občanům Polska, Maďarska, České republiky a Slovenska. Zajímavé je, že po referendu o brexitu v roce 2016, jež přineslo rozhodnutí Británie opustit Evropskou unii, lze debatu o přistěhovalectví před brexitem považovat za faktor přispívající k výsledku referenda, a tedy otázku s vysokým společenským významem a dopadem.

3.1 METODA A DATA

Data použitá k této analýze jsou, jak jsme již uvedli, vzorkem čtenářských komentářů vyskytujících se pod články v novinách *The Guardian* a *Daily Mail*. K identifikaci příslušných článků těchto dvou novin z let 2010 až 2013 jsme jako vyhledávací výraz použili *East/Eastern European(s)* (Východoevropan(é)). Jde o kolektivní jazykový výraz často používaný britským tiskem při uvádění odkazů na občany z nových zemí Evropské unie (např. Rumunska, Bulharska, České republiky nebo Polska). Celkově bylo z *The Guardian* (dále GU) extrahováno 1 024 495 tokenů, z *Daily Mail* (dále DM) pak 729 042 tokenů. Detailnější informace o těchto subkorpusech viz tabulka 7:

korpus	počet komentářů	počet přispěvatelů (čtenářů)	tokeny	průměrná délka komentáře (tokenu)
GU	10,193	4,072	1,024,495	101
DM	13,265	6,093	729,042	55
celkem	23,458	10,165	1,753,537	75

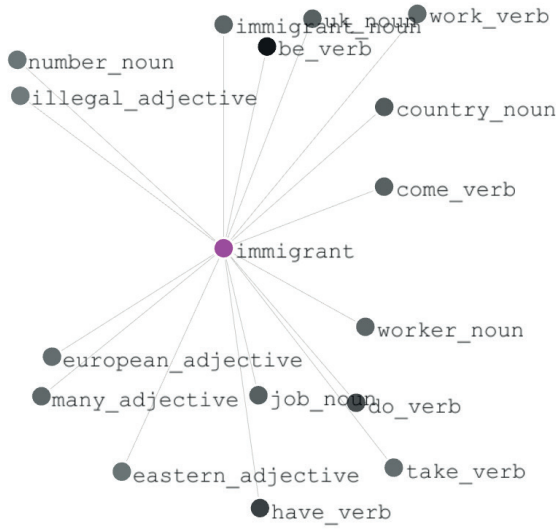
TABULKA 7. Charakteristika subkorpusů *The Guardian* (GU) a *Daily Mail* (DM)

Z tabulky 7 je patrné, že v každém z těchto dvou subkorpusů je více než deset tisíc komentářů, přičemž čtenáři *Guardian* přispívají méně často, ale v průměru delšími komentáři než čtenáři *Daily Mail*. Průměrná délka komentáře činila 101 tokenů v GU a 55 tokenů v DM. Průměrný komentář čtenáře v *Guardian* byl tedy téměř dvakrát delší než průměrný komentář v *Daily Mail*. Hledaným výrazem v kolokační analýze bylo nominální lemma *immigrant* (přistěhovalec). Tentokrát jsme zvolili jako asociační měřítko k identifikaci častých a výlučných asociací skóre logDice (Rychlý, 2008; Gablasova et al., 2017b; Brezina, 2018). Je důležité si uvědomit, že témata, názory i povaha publikovaných novinových článků mají vliv i na typ či charakter čtenářských komentářů, které se pod články objevují. Od tohoto vztahu mezi novinovými články a typy čtenářských komentářů odhlížíme. Důraz je kladen výhradně na diskurz vytvořený čtenáři v reakci na články, jež v daném časovém období obsahovaly hledaný výraz *East/Eastern European(s)*.

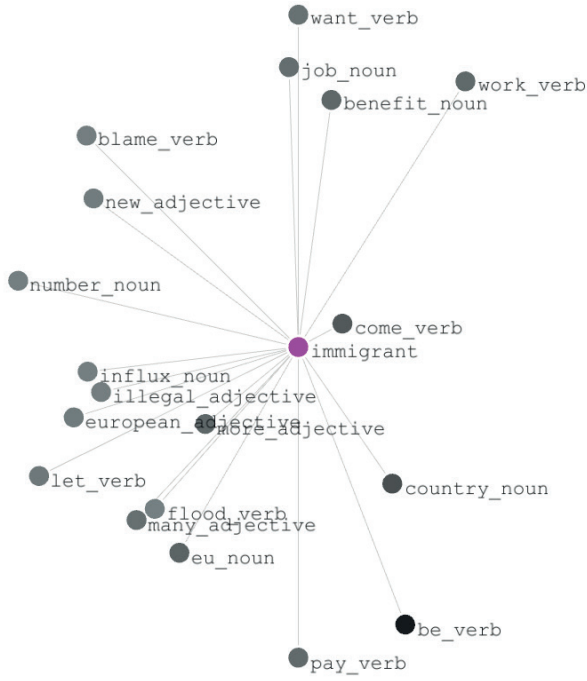
3.2 VÝSLEDKY A DISKUSE

Výsledky analýzy ve formě kolokačních grafů lemmatu *immigrant* (imigrant) v obou subkorpusech (GU a DM), zahrnující tvary singuláru i plurálu, uvádíme na obrázcích 8 a 9. V analýze čtenářských komentářů nás zajímaly především konceptuální souvislosti patrné v diskurzu: zaměřili jsme se proto pouze na substantiva, slovesa a adjektiva jako kolokáty nesoucí sémantický (nikoli gramatický) význam. Poloha kolokátů kolem uzlu je užitečným indikátorem typu syntaktických rámců (či odrazem textových pozic), v nichž se uzel i jeho kolokáty vyskytují. Poziční zobrazení v grafech na obrázcích 8 a 9 tedy vykresluje kolokáty vlevo od uzlu, pokud se před klíčovým slovem *immigrant* objevují i v textu, zatímco kolokáty zobrazené vpravo od uzlu klíčové slovo (uzel) v textu následují.

Ve dvou výše zobrazených grafech s uzlem *immigrant* lze pozorovat srovnatelný počet kolokátů: 16 v GU a 19 v DM. Z toho je 9 kolokátů sdíleno oběma subkorpusy: *be_verb* (být_verbum), *come_verb* (přijít_verbum), *country_noun* (země_substantivum), *european_adjective* (evropský_adjektivum), *illegal_adjective* (ilegální_adjektivum), *job_noun* (zaměstnání_substantivum), *many_adjective* (mnoho_adjektivum), *number_noun* (počet_substantivum), *work_verb* (pracovat_verbum). To ale neznamená, že se tyto kolokáty používají v obou subkorpusech ve shodných kontextech a s týmiž konotacemi. Např. adjektivum *illegal* (nelegální) je v komentářích databáze DM užito převážně v negativně laděných kontextech (viz př. 13 a 14 níže), zatímco subkorpus GU obsahuje řadu případů, kdy je deskriptor *illegal* zpochybněn (př. 15 a 16).



OBRÁZEK 8. Kolokace lemmatu *immigrant* v subkorpusu GU (9a-MI (9), R5–L5, C:10-NC:10; zobrazena pouze substantiva, slovesa a adjektiva)



OBRÁZEK 9. Kolokace lemmatu *immigrant* v subkorpusu DM (9a-MI (9), R5–L5, C:10-NC:10; zobrazena pouze substantiva, slovesa a adjektiva)

(13) *… must come OUT of the EU NOW and send home all **illegal immigrants** NOW I am sick to death of reading articles like this.* (DM)
(14) *I won't even get into the **ILLEGAL immigrants** and how easy we have made it for them…* (DM)
(15) *Rarely is the distinction made between asylum seekers, immigrants and **illegal immigrants**. Personally, I have no time for people who easily take a swipe at hard working low-paid legal migrants who often take jobs that unemployed UK citizens sometimes find unpalatable.* (GU)
(16) *… also irritating when the Daily Mail/BNP crowd posting here repeatedly confuse "**illegal**" **immigrants** with EU citizens, who have every right to be in Britain.* (GU)

Zaměřme se nyní na jedinečné kolokace v každém z obou subkorpusů. GU obsahuje řadu neutrálních kolokátů jako *do_verb* (dělat_verbum), *eastern_adjective* (východní_adjektivum), *have_verb* (mít_verbum), *immigrant_noun* (imigrant_substantivum) a *uk_noun* (UK_substantivum). Tyto kolokáty naznačují, že se diskuse vede kolem přistěhovalců z východní Evropy do Velké Británie, což je mj. odrazem toho, jak byl subkorpus vytvořen: vyhledáváním článků s touto problematikou. Verba *do* (dělat) a *have* (mít) se vyskytují jednak jako pomocná slovesa, jednak jako plnovýznamová slovesa ve spojení s výrazem *immigrant(s)* (imigrant(i)) v syntaktické roli podmětu a předmětu; proto neodkazují na žádné konkrétní téma diskuse o přistěhovalcích.

V subkorpusu GU byla nalezena pouze dvě jedinečná spojení, jež zdůrazňují konkrétní téma nebo část diskuse: *take_verb* (vzít/získat_verbum) a *worker_noun* (pracovník_substantivum). To souvisí s otázkou, zda přistěhovalečtí pracovníci berou práci britským pracovníkům. Příklady této diskuse jsou uvedeny níže:

(17) *What some "working class" Brits fail to understand is that non-British workers (both **immigrants** and Eastern European **workers**, again, different categories) put a lot more into the British welfare state than get out of.* (GU)
(18) *His wife may not need to work because of the high value of the cash that the **immigrant worker** can send home, so relatively fixed costs like childcare and mortgages become irrelevant.* (GU)
(19) *Brown says British jobs for British people, then we get the results in. **Immigrants take** 81% of new jobs.* (GU)

Nyní se podívejme na jedinečné kolokáty v subkorpusu DM. Jedná se o *benefit_noun* (dávky_substantivum), *blame_verb* (obviňovat_verbum), *eu_noun* (EU_substantivum), *flood_verb* (zaplavit_verbum), *influx_noun* (příliv/nával_substantivum), *let_verb* (nechat/dovolit_verbum), *more_adjective* (více_adjektivum),[15] *new_adjective* (nový_adjektivum), *pay_verb* (platit_verbum), *want_verb* (chtít_verbum). Lze je rozdělit do tří hlavních skupin: na i) kontextové (*eu_noun*); ii) popisné/hodnoticí (*flood_verb, influx_noun, more_adjective, new_adjective, benefit_noun*) a iii) orientované na činnost (*blame_verb, pay_verb, want_verb, let_verb*). Jediným kontextovým koloká-

15 V angličtině je výraz *more* používán jako determinant, adverbium nebo zájmeno. Automatická analýza zde v rozporu s konvencí označuje prenominální výskyty *more* jako adjektiva.

tem je *EU*, má však v kontextu diskuse o přistěhovalectví často negativní konotace. Popisné/hodnoticí kolokace ukazují hlavní příběh diskuse: existuje velké množství nových přistěhovalců, kteří již ve Velké Británii jsou nebo se do Británie chystají; tito přistěhovalci berou podporu v nezaměstnanosti atd. Silně negativní výrazy jako *flood* (zaplavit) či *influx* (příliv) slouží k tomu, poukázat na souvislosti mezi přistěhovalectvím a přírodní katastrofou (viz př. 20–21).

(20) *Jobs, **benefits** and housing will all be given to a massive **influx** of **immigrants**.* (DM)
(21) *Its no surprise that we are in this state, Labour desperately wanted a **flood** of **immigrants** into the country.* (DM)

Poslední skupina kolokátů je zaměřena v širokém smyslu na činnost — všechny tyto kolokáty jsou verbální. Vyskytují se v různých kontextech, a jak je patrno z pozic zobrazených v grafu na obrázku 10, přistěhovalci jsou v těchto spojeních syntaktickými podměty nebo předměty. Viz příklady tohoto aspektu diskuse níže:

(22) *Dont **blame** the **immigrants**, they are just after a better quality of life.* (DM)
(23) *... wait only when the country is completely over run will they learn, because the **immigrants** wont be **paying** any tax to fund the madness.* (DM)
(24) *should I **pay** council tax to fund **immigrants** to ruin our Country?* (DM)
(25) *we don't **want** or need anymore **immigrants**.* (DM)
(26) *don't **let immigrants** in to the country.* (DM)

4. POTENCIÁL A LIMITACE TECHNIKY KOLOKAČNÍCH GRAFŮ A SÍTÍ

V závěru studie je třeba zhodnotit, nebo přinejmenším tematizovat potenciál i praktická omezení grafické vizualizace kolokací ve formě kolokačních grafů a sítí. Obecně platí, že neexistuje žádný nástroj nebo metoda vyhovující všem výzkumným účelům. Ani autoři tohoto textu si nenárokují univerzálně platné a většinové použití nástroje #LancsBox v kolokačních analýzách. Nicméně jsme přesvědčeni, že vizualizace kolokací má nesporný potenciál nového analytického vhledu, který je dán schopností výzkumníka nahlížet na kolokace komplexněji. Ačkoli všechny informace o kolokacích lze prezentovat prostřednictvím tradičních kolokačních seznamů, ba dokonce i velmi složitou kolokační síť lze nahradit řadou kolokačních tabulek se statistickými údaji — ty jsou ostatně k dispozici v nástroji #LancsBox vedle kolokačních grafů —, efektivní analýze propojení mezi různými lexikálními jednotkami v jazyce a diskursu velmi napomáhá právě grafický formát (tj. vizualizace). Mimoto mohou grafy zobrazit více dimenzí kolokace současně, a to včetně pozice kolokace v textu a diskursu, slovnědruhové příslušnosti a frekvence kolokátu — to vše navíc k síle kolokace vyjádřené konkrétní asociační mírou. Kolokační sítě jsou tedy informačně bohaté a mají vysoké *data-ink ratio* (Tufte, 2001), což je jedním z kritérií efektivní vizualizace. Užitečnost této techniky, prezentované i v tomto textu, byla prokázána mj. i popularitou nástroje #LancsBox s více než 50 tisíci uživateli ve více než 150 zemích po celém světě.

#LancsBox se také stále častěji používá v publikovaných výzkumech a konferenčních příspěvcích (v květnu 2021 uvádí *Google Scholar* přes 400 odkazů na tento nástroj). Uveďme pro ilustraci témata několika nedávných studií k dalším příkladům použití analytické techniky kolokačních grafů a sítí:

— analýza vulgárních slov na twitteru (Gauthier, 2021);
— analýza jazyka, kterým se hovořilo o globální pandemii během tříměsíčního období kontroly pohybu (lockdown) v Malajsii (Joharry — Turiman, 2020);
— analýza jazyka emocí v britských novinách v kontextu úsporných opatření (austerity) (Ha, 2020);
— analýzy jazyka a sexuality (Baker, 2018).

Dále se #LancsBox a vizualizace kolokačních sítí úspěšně používají v jazykové pedagogice. Například Liu (2020) ukázala, že metodický přístup DDL (data driven learning) s nástrojem #LancsBox byl efektivní při zlepšování receptivních i produktivních znalostí anglických kolokací u čínských studentů. Postoje studentů navíc naznačovaly, že většina z nich byla s aplikací tohoto přístupu k učení kolokací spokojena.

Podívejme se nyní na určitá omezení této techniky. Hlavním omezením je zobrazení kolokací v dvourozměrném prostoru (2D). Vzhledem k velkému počtu uzlů a kolokací může být graf snadno přeplněn, a pak ztrácí schopnost jasně či přehledně zobrazit informace. Existují však techniky pro redukci tohoto problému (srov. Brezina et al., 2015). Tabulková forma prezentace dat nesporně nadále zůstává užitečným a svým způsobem nutným prostředkem pro interpretaci kolokací. Je důležité si ale uvědomit, že kolokační grafy a sítě nejsou v konkurenčním, ale doplňkovém vztahu k tradičním kolokačním seznamům. Lze rovněž zmínit, že jednotliví lingvisté budou pro analýzu kolokací dávat přednost rozličným analytickým technikám a využívat různé nástroje i databáze různých velikostí. Tuto rozmanitost v aplikaci korpusových metod a nástrojů bychom měli přijmout jako pozitivní aspekt, který lze produktivně využít prostřednictvím triangulace (Baker — Egbert, 2016).

5. ZÁVĚR

Hlavním cílem tohoto textu bylo představit korpusový nástroj #LancsBox a techniku dynamické vizuální kolokační analýzy. Zvláštní pozornost byla věnována teoretickému konceptu kolokačních grafů a sítí i metodickým otázkám identifikace kolokace v korpusech, zejména nastavení klíčových parametrů kolokační analýzy. Tyto parametry je možno systematicky zaznamenat pomocí notace CPN umožňující replikaci výsledků. Vzhledem k omezenému rozsahu této studie však nebylo možno provést komplexní (hloubkové) analýzy, ale spíše prezentovat a ilustrovat možnosti kolokačních grafů a sítí při analýze lexikálních (či sémantických) vztahů mezi výrazy ve dvou případových studiích z oblasti lexikografie a analýzy diskurzu na vybraných příkladech z angličtiny a češtiny.

Budoucí výzkum by měl zahrnovat komplexnější kontrastivní (komparativní) analýzu kolokací v češtině a angličtině, detailněji charakterizující jazykové užití výrazů v textu a kontextu či diference chování výrazů v těchto typologicky odlišných jazykových systémech.

LITERATURA

Baker, P. (2018): Language, sexuality and corpus linguistics: Concerns and future directions. *Journal of Language and Sexuality*, 7, 2, s. 263–279.

Baker, P. (2016): The shapes of collocation. *International Journal of Corpus Linguistics*, 21, 2, s. 139–164.

Baker, P. (2009): The BE06 Corpus of British English and recent language change. *International journal of corpus linguistics*, 14, 3, s. 312–337.

Baker, P. — Egbert, J. (eds.) (2016): *Triangulating Methodological Approaches in Corpus Linguistic Research*. New York: Routledge.

Béjoint, H. (2016): Dictionaries for general users: history and development; current issues. In: P. Durkin (ed.), *The Oxford handbook of lexicography*. Oxford: Oxford University Press, s. 7–24.

Blinder, S. — Allen, W. L. (2016): Constructing immigrants: Portrayals of migrant groups in British national newspapers, 2010–2012. *International Migration Review*, 50, 1, s. 3–40.

Brezina, V. (2016): Collocation networks. In: P. Baker — J. Egbert (eds.), *Triangulating Methodological Approaches in Corpus Linguistic Research*. New York: Routledge, s. 90–107.

Brezina, V. (2018a): *Statistics for corpus Linguistics: A practical guide*. Cambridge: Cambridge University Press.

Brezina, V. (2018b): Collocation Graphs and Networks: Selected Applications. In: *Lexical Collocation Analysis*. Cham: Springer, s. 59–83.

Brezina, V. — Gablasova, D. (2018): The corpus method. In: J. Culpeper — P. Kerswill — R. Wodak — T. McEnery — F. Katamba (eds.), *English Language: Description, Variation and Context*. Second edition. Basingstoke: Palgrave Macmillan.

Brezina, V. — McEnery, T. — Wattam, S. (2015): Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20, 2, s. 139–173.

Cope, B. — Kalantzis, M. — Magee, L. (2011): *Towards a semantic web: Connecting knowledge in academic research*. Oxford: Chandos.

Council of Europe (2001): *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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

Čermák, F. (2006): Kolokace v lingvistice. In: F. Čermák — M. Šulc (eds.), *Kolokace. Studie z korpusové lingvistiky*. Sv. 2. Praha: Nakladatelství Lidové noviny, s. 9–16.

Deignan, A. (2015): *Figurative language and lexicography*. *International Handbook of Lexicography*. Berlin, Germany: Springer.

Evert, S. (2008): Corpora and collocations. *Corpus linguistics. An international handbook*, 2, s. 1212–1248.

Firth, J. (1957): *Papers in Linguistics*. Oxford: Oxford University Press.

Gablasova, D. — Brezina, V. — McEnery, T. (2017a): Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67, S1, s. 155–179.

Gablasova, D. — Brezina, V. — McEnery, T. (2017b): Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Language Learning*, 67, S1, s. 130–154.

Gablasova, D. — Brezina, V. — McEnery, T. — Boyd, E. (2017): Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, 38, 5, s. 613–637.

Gabrielatos, C. — Baker, P. (2008): Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English linguistics*, 36, 1, s.5–38.

Gauthier, M. (2021): 'Eww wtf, what a dumb bitch': a case study of similitudes inside gender-specific swearing patterns on Twitter. *Corpora*, 16, 1, s. 31–61.

Granger, S. — Paquot, M. (eds.) (2012): *Electronic Lexicography*. Oxford: OUP.

Gries, S. T. (2013): 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18, 1, s. 137–166.

Gullick, D. — Rayson, P. — Mariani, J. — Piao, S. — Taiani, F. (2010): *CONE: COllocational Network Explorer* [software]. https://code.google.com/archive/p/collocation-network-explorer/ (cit. září 2020).

Ha, F. W. (2020). Collocation and Emotions in the Context of Austerity in British Newspapers. In: T. Griebel — S. Evert — P. Heinrich (eds.), *Multimodal Approaches to Media Discourses: Reconstructing the Age of Austerity in the United Kingdom*. Abingdon: Routledge, s. 88–109.

Hanks, P. (2016): Definition. In: P. Durkin (ed.), *The Oxford handbook of lexicography*. Oxford: Oxford University Press, s. 94–122.

Hoffmannová, J. — Homoláč, J. — Chvalovská, E. — Jílková, L. — Kaderka, P. — Mareš, P., Mrázková, K. (2016): *Stylistika mluvené a psané češtiny*. Praha: Academia.

Howarth, P. (1998): Phraseology and second language proficiency. *Applied linguistics*, 19, 1, s. 24–44.

Hunston, S. (2001): Colligation, lexis, pattern, and text. In: M. Scott — G. Thompson (eds.), *Patterns of text: In honour of Michael Hoey*. Amsterdam: John Benjamins, s. 13–33.

Church, K. W. — Hanks, P. (1990): Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, s. 22–29

Joharry, S. A. — Turiman, S. (2020): Collocation Networks and Covid-19 in Letters to the Editor: A Malaysian Case Study. *Asia Pacific Journal of Corpus Research*, 1, 1, s. 1–30.

Karlík, P. — Nekula, M. — Rusínová, Z. (eds.) (1997): *Příruční mluvnice češtiny*. Praha — Brno: Nakladatelství Lidové noviny.

Karlík, P. — Nekula, M. — Pleskalová, J. (eds.) (CzechEncy): *Nový encyklopedický slovník češtiny online*. Dostupné z www: https://www.czechency.org/

Kilgarriff, A. — Rychlý, P. — Smrž, P. — Tugwell, D (2004): *The Sketch Engine*. Information Technology.

Kilgarriff, A. — Kovář, V. — Krek, S. — Srdanović, I. — Tiberius, C. (2010): A quantitative evaluation of word sketches. In: *Proceedings of the 14th EURALEX International Congress*. Afûk, Ljouwert: Fryske Akademy, s. 372–79.

Kilgarriff, A. — Baisa, V. — Bušta, J. — Jakubíček, M. — Kovář, V. — Michelfeit, J. — Rychlý, P. — Suchomel, V. (2014): The Sketch Engine: ten years on. *Lexicography*, 1, 1, s. 7–36.

KhosraviNik, M. (2009): The representation of refugees, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005). *Discourse & Society*, 20, 4, s. 477–498.

Křen, M. (2006): Kolokační míry a čeština: srovnání na datech ČNK. In: F. Čermák — M. Šulc (eds.): *Kolokace*. Praha: Nakladatelství Lidové noviny, s. 223–248.

Lakoff, G. — Johnson, M. (1980): *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G. — Johnson, M. (2002): *Metafory, kterými žijeme*. Brno: Host.

Liu, T. (2020): *Evaluating the effect of data-driven learning (DDL) on the acquisition of academic collocations by Chinese learners of English* (Doctoral dissertation, Lancaster University).

McEnery, T. (2006): *Swearing in English: Bad language, purity and power from 1586 to the present*. London: Routledge.

McEnery, T. — Hardie, A. (2011): *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
McNair, B. (2009): *News and Journalism in the UK*. London: Routledge.
Mluvnice češtiny [2]. Tvarosloví (1986). Praha: Academia.
Murphy, L. M. (2016): Meaning Relations in Dictionaries: Hyponymy, meronymy, synonymy, antonymy, and contrast. In: P. Durkin (ed.), *The Oxford handbook of lexicography*. Oxford: Oxford University Press, s. 94–122.
Paquot, M. (2017): The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35, 1, s. 121–145.
Pecina, P. (2010): Lexical association measures and collocation extraction. *Language resources and evaluation*, 44, 1, s. 137–158.
Phillips, M. (1985): *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam, Netherlands: North-Holland.
Pořízka, P. (2014): *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. Olomouc: Vydavatelství FF UP.
Rychlý, P. (2008): A lexicographer-friendly association score. In: P. Sojka – A. Horák (eds.), RASLAN 2008: Recent Advances in Slavonic Natural Language Processing, Brno: Masarykova Univerzita, s. 6–9.
Sinclair, J. M. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Tufte, E. (2001): *Visual display of quantitative information*. Cheshire, CT: Graphics Press.
Williams, G. (1998): Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3, 1, s. 151–171.

ZDROJE A NÁSTROJE

Baker, P. (2006): *British English 2006 (BE06)*. Dostupný z: https://cqpweb.lancs.ac.uk/
Brezina, V. — Weill-Tessier, P. — McEnery, A. (2020): *#LancsBox* (v. 5.1.2) [software]. Dostupný z: http://corpora.lancs.ac.uk/lancsbox
Brezina, V. (2019): *BNC2014 Baby +*. Dostupný z #LancsBox: http://corpora.lancs.ac.uk/lancsbox
Pořízka, P. (2021): *KFS — korpus funkčních stylů (v1.0)*. Olomouc. Dostupný z #LancsBox: http://corpora.lancs.ac.uk/lancsbox
The Sketch Engine [software]. Dostupný z: https://www.sketchengine.eu/
Křen, M. — Cvrček, V. — Čapka, T. — Čermáková, A. — Hnátková, M. — Chlumská, L. — Jelínek, T. — Kováříková, D. — Petkevič, V. — Procházka, P. — Skoumalová, H. — Škrabal, M. — Truneček, P. — Vondřička, P. — Zasina, A. (2015): *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF. Dostupný z: https://www.korpus.cz/kontext/query?corpname=syn2015

Václav Březina | Department of Linguistics and English Language, ESRC Centre for Corpus Approaches to Social Science, Faculty of Arts and Social Sciences, Lancaster University | Lancaster, United Kingdom LA1 4YL
ORCID ID: 0000-0002-1613-6100
v.brezina@lancaster.ac.uk

Petr Pořízka | Katedra bohemistiky, Filozofická fakulta Univerzita Palackého | Křížkovského 10, 779 00 Olomouc
ORCID ID: 0000-0001-6980-9148
petr.porizka@upol.cz