JAZYKOVÉ KORPUSY A VÝZKUM AFÁZIE

LANGUAGE CORPORA AND APHASIA RESEARCH

Mgr. Michal Láznička, Ph.D.[1, 2]

Michal Láznička

Abstrakt

Existence jazykových korpusů, velkých souborů textů obohacených o různá metadata, umožňuje studium jazyka tak, jak je v různých situacích užíván jeho mluvčími, a dává představu o tom, jak časté jsou jazykové prostředky v různých kontextech, což má souvislost i s organizací a fungováním jazykové znalosti. Tento text představuje možnosti využití korpusových dat a nástrojů především ve studiu jazyka v afázii, ale i v klinické praxi. Korpusy jsou zde představeny především jako zdroj referenčních dat, pomocí kterých lze predikovat a vysvětlovat zpracování jazyka v afázii (např. lepší výbavnost slov s vyšší frekvencí), a také jako přístup ke sběru a zpracování afaziologických dat. V ukázkové analýze produkce bigramů (dvouslovných spojení) v korpusu češtiny v afázii je pak s pomocí metody klíčových slov identifikována konstrukce používající *tam být* jako strukturní a komunikační opora dvou mluvčích s nonfluentní afázií. Analýza fluence pak ukázala, že vyšší frekvence bigramů podporuje fluentnější produkci v podobné míře jak u mluvčích s afázií, tak u typických mluvčích.

Klíčová slova

afázie, korpusová lingvistika, frekvence jazykových jednotek, fluence, analýza klíčových slov, užívání jazyka

Abstract

The existence of language corpora – large collections of texts enriched with different types of metadata – allows for language to be studied as it is actually used by speakers in different situations, providing a good estimate of how common linguistic items are in different contexts. This is related to the ways in which linguistic knowledge is organised and processed. This paper presents an overview of the possibilities of employing corpus linguistic data and tools in the study of language in aphasia, as well as in clinical practice. Corpus linguistics and corpora are discussed here mainly as sources of reference data, which can be used to predict and explain language processing in aphasia (e.g. better lexical retrieval of high-frequency words), as well as an approach to the collection and processing of aphasic data. An analysis of bigram production using a corpus of Czech aphasia is presented. Using keyness analysis, the construction using *tam být* ('there be') was identified as a structural and communicative support for two speakers with nonfluent aphasia. An analysis of fluency shows that higher bigram frequency predicts more fluent production to a comparable extent in both typical speakers and speakers with aphasia.

Keywords

aphasia, linguistics, language corpora, frequency of linguistic items, fluency, keyness analysis, language usage

Úvod

V komunikaci s logopedkami a logopedy často narážím na pozorování, že mluvčí s afázií používají slova, která jsou obvyklá nebo působí jako ustálená spojení, mají s nimi méně problémů a bývá na ně zaměřena i terapie. To souvisí i s předpokladem, že častá slova a slovní spojení budou i komunikačně užitečnější a jejich trénink může přispět k úspěšnější funkční komunikaci, na niž se v poslední době v afaziologii soustřeďuje pozornost (Doedens a Meteyard, 2020, 2022). Ačkoli se ukazuje, že mluvčí mají poměrně dobrý odhad, jak obvyklá jednotlivá slova v úzu jsou (Brysbaert a Cortese, 2011), existence velkých jazykových korpusů, počítačově

[1] Mgr. Michal Láznička, Ph.D. Ústav lingvistiky, Filozofická fakulta Univerzity Karlovy, nám. Jana Palacha 1/2, 116 38 Praha 1, Česká republika. E-mail: Michal.Laznicka@ff.cuni.cz.

[2] Department of Modern Languages, College of Arts and Law, University of Birmingham, Edgbaston, Birmingham B15 2TT, Spojené království.

https://doi.org/10.36833/lkl.2026.010 Přijato k recenzi: 2. 3. 2026 Přijato k publikaci: 21. 4. 2026

zpracovaných databází textů, umožňuje tyto informace snadno získat pro velké množství slov a bez zkreslujících vlivů individuální jazykové zkušenosti.

Právě korpusová lingvistika, tedy jazykovědná disciplína zabývající se konstrukcí a využitím korpusů, je jednou z oblastí, v nichž může lingvistika přispět afaziologii či šířeji logopedii, aniž by to vyžadovalo hlubší teoretické či technické znalosti. Užší propojení klinické logopedie a lingvistiky je přitom žádoucí a vzájemně přínosné: na jedné straně může lingvistika přispívat k přesnějšímu popisu i explanaci jazykového chování lidí s afázií, z druhé strany potom představuje jazyk v afázii specifický zdroj dat, s jejichž pomocí lze ověřovat funkčnost lingvistických modelů. Realita za očekáváními však poněkud zaostává. Na straně lingvistiky je zájem o výzkum jazyka v afázii v českém i světovém kontextu poměrně omezený.[3] To může být dáno i praktickými hledisky: lingvistický výzkum afázie je náročný především z hlediska dostupnosti, zpracování a vyhodnocení dat. Většímu zapojení lingvistických teorií do výzkumu afázie i klinické praxe může kromě nezájmu z lingvistické strany bránit i častá potřeba širších teoretických znalostí a časové nároky s tím spojené.

Cílem tohoto textu je prezentovat jazykové korpusy a možnosti jejich využití ve výzkumu afázie i v klinické praxi. Po stručném obecném uvedení do problematiky korpusů budou představeny konkrétní nástroje dostupné v rámci Českého národního korpusu (ČNK), které by bylo možné přímo využít v klinické praxi. Následně bude demonstrováno, jak může být speciální korpus jazyka využit v afázii pro zpřesnění diagnostiky a plánování terapie.

Korpusy: základní charakteristika

Jazykové korpusy jsou velké soubory textů, které jsou počítačově zpracované, obohacené o metadata a prohledavatelné. Texty v korpusech mluveného jazyka obsahují přepisy zvukových nahrávek (často i zdrojové nahrávky) a korpusy multimodální komunikace zaměřené na propojení jazykové komunikace s gesty obsahují vedle textů zachycujících řeč i gesta také audiovizuální stopu. Vznik a rozvoj korpusové lingvistiky od 60. let 20. století (McEnery a Hardie, 2013) umožnil, aby popisy jazyka a jeho užívání byly založeny na relativně velkých a reprezentativních souborech dat namísto introspekce či neformálních pozorování.[4] Texty zařazené do korpusů jsou vybírány s ohledem na svůj účel (viz Tabulka 1). Zatímco obecné korpusy lze vnímat jako určitý odraz průměrné jazykové zkušenosti mluvčích, specializované korpusy ukazují užívání jazyka specifickými populacemi či ve specifických tematických, časových nebo situačních kontextech.[5] Velikost korpusů se pohybuje v řádech tisíců až miliard slov v závislosti na účelu a dostupnosti jazykového materiálu. Hlavním zprostředkovatelem jazykových korpusů v Česku je ČNK, který přes webové rozhraní (https://korpus.cz) umožňuje po bezplatné registraci přístup k široké škále korpusů i specializovaných nástrojů pro práci s nimi. Čeština přitom patří ve světovém měřítku mezi jazyky s nejlepší korpusovou infrastrukturou.

[3] Například na poslední velké obecnělingvistické konferenci ICLC, která se zaměřuje na studium jazyka v kontextu kognitivních schopností, nebyl mezi více než 220 prezentovanými příspěvky ani jeden věnovaný afázii (program viz https://iclc17.com/wp-content/uploads/2025/07/ICLC17-Programme-2025-07-17.pdf).

[4] Srov. Mluvnici současné češtiny (Cvrček et al., 2010), která je založena na korpusových datech.

[5] Je třeba zdůraznit, že z důvodů omezeného rozsahu zde uvádím pouze velmi stručné základní teze. Pro podrobnější úvod viz Čermák (2017) či Stefanowitsch (2020).

Hledisko	Základní dělení	Příklady (ČNK)
Produkce textu	psaný jazyk	SYN2025: reprezentativní korpus současné češtiny složený stejným dílem z beletristických, odborných a publicistických textů; obsahuje 100 milionů slov
	mluvený jazyk	ORAL: přepisy nahrávek, především spontánní neformální komunikace; 5,3 milionu slov
Časové hledisko	současný jazyk	SYN2025
Časové hledisko	současný jazyk	SYN2025
	historický jazyk	Diakorp: české texty z 14.–20. století; necelých 3,5 milionu slov
Zaměření	obecný jazyk	SYN2025
	specializovaný korpus	CzeSL-SGT: osvojování češtiny jako druhého jazyka, obsahuje písemné práce nerodilých mluvčích češtiny; necelý milion slov
		ONLINE1: texty z internetového zpravodajství a příspěvky ze sociálních sítí z let 2017–2021; více než 7 miliard slov
	paralelní korpus	InterCorp: umožňuje vyhledávat v různojazyčných (překladových) verzích textů; texty v 61 jazycích v celkovém rozsahu téměř 5 miliard slov

Tabulka 1: ***Druhy korpusů s příklady z ČNK***

Texty obsažené v korpusech jsou do různé míry opatřeny metadaty. Tato metadata zahrnují jednak informace o textech jako takových (např. rok vydání či datum pořízení nahrávky v mluveném korpusu), dále bývají v různém rozsahu zahrnuty informace o jednotlivých slovech v textu. Základem bývá tzv. lemmatizace, při níž je každému slovu přiřazen jeho „slovníkový tvar", např. všechny vyskloňované tvary slova *kočka* mají v korpusu přiřazeno společné lemma KOČKA. Obvyklé jsou tzv. morfologické značky, které každému slovu přiřazují slovní druh a případně další gramatické kategorie, např. tvar *kočkami* by měl morfologickou značku *NNFP7-----A----*, která indikuje, že se jedná o podstatné jméno (NN) ženského rodu (F) v množném čísle (P) a instrumentálu, tj. 7. pádě (7). Některé korpusy mají navíc i syntaktické značky zachycující větnou strukturu. Ve větě *Máte kočku?* je tak slovo *kočku* v korpusu SYN2025 označeno jako přímý předmět s řídícím slovesem *mít*. Přítomnost těchto metadat umožňuje vyhledávání na různých úrovních obecnosti. V korpusu tak lze např. vyhledat

- podstatná jména končící na *-o*, která nejsou středního rodu,
- tvary slovesa *hledat* v přítomném čase,
- dokonavá slovesa v minulém čase nebo
- přídavná jména ve funkci přívlastku slova *kočka*.

Výstupem takového dotazu pak je tzv. KWIC (key word in context, klíčové slovo v kontextu), který ukazuje četnost a kontexty výskytu dané jazykové jednotky. Popis a výzkum jazyka s využitím korpusů je založen na distribučních charakteristikách jazykových jednotek, které lze popsat s využitím frekvence výskytu a dalších na ní založených měr, a to disperze a asociační síly (pro základní, ač poněkud technický přehled viz Gries, 2024, kap. 2). Základní charakteristiky těchto tří konceptů shrnuje Tabulka 2.

Míra	Podtyp	Komentář
Frekvence	hrubá	počet výskytů jednotky v daném korpusu
	standardizovaná	frekvence přepočtená na relativní počet výskytů v milionu slov; umožňuje srovnání korpusů různých velikostí
	proporční	procentuální užití jednotky v daném kontextu
	tokenová	počet všech výskytů konkrétního tvaru (např. frekvence slovesa *pít* s přímým předmětem)
	typová	počet unikátních tvarů či variant (např. počet podstatných jmen v roli přímého předmětu slovesa *pít*)
Disperze		rovnoměrnost výskytů jednotky napříč jednotlivými texty v korpusu
Asociační síla	slovo + žánr / konkrétní text	identifikuje slova užívaná ve zkoumaném textu disproporčně vzhledem k referenčnímu korpusu, tzv. klíčovost
	slovo + slovo	identifikuje slova, která se spolu vyskytují disproporčně v porovnání s výskytem s jinými slovy, tzv. kolokace
	slovo + gramatická struktura	identifikuje afinitu konkrétních lexémů k určité gramatické struktuře (např. konkrétní slovesa a minulý čas), tzv. kolostrukce

Tabulka 2: ***Přehled distribučních charakteristik užívaných v korpusové lingvistice***

Pro lepší pochopení potřeby frekvence i disperze lze použít lemmasta *kolonista* a *povodeň* v korpusu ORAL. Ačkoli je jejich frekvence srovnatelná (37 a 39), *povodeň* je intuitivně mnohem obvyklejším slovem. To je potvrzeno distribucí obou slov

napříč jednotlivými texty (v tomto případě konverzacemi) v korpusu, jinak řečeno jejich disperzí: *kolonista* se vyskytuje pouze v jedné z 1 546 konverzací, zatímco *povodeň* ve 33 různých konverzacích. Frekvence 37 slova *kolonista* je tak způsobena jedním rozhovorem na specifické téma, zatímco frekvence 39 slova *povodeň* je relativně spolehlivá, protože v každé konverzaci zazní průměrně o málo více než jednou.

Na spojení *strašně moc* lze ilustrovat asociační sílu. Při jejím výpočtu je uvažována distribuce jednotek společně i každé z nich odděleně. Pracuje se s kontingenčními tabulkami (jako např. Tabulka 3) a pomocí statistických testů lze vypočítat, nakolik se distribuce odlišuje od distribuce předpokládané. Spojení *strašně moc* se tak ukazuje jako slovní spojení, které je v mluvené češtině prominentní a zároveň je součástí obecnější pravidelné struktury rozvitého příslovce.[6]

	druhé slovo je *moc*	druhé slovo není *moc*
první slovo je *strašně*	288	frekvence *strašně* − 288 = 3 315
první slovo není *strašně*	frekvence *moc* − 288 = 5 486	ostatní dvouslovná spojení = 4 422 128

Tabulka 3: ***Kontingenční tabulka pro výpočet asociační síly spojení strašně moc v korpusu ORAL***

Korpusová data ve výzkumu zpracování jazyka

Význam frekvence, disperze a asociace nespočívá pouze v popisu jazyka a jeho užívání, tyto míry mají totiž souvislost i s uspořádáním a fungováním jazykové znalosti. Jak bylo opakovaně prokázáno v různých výzkumných kontextech, pokud se mluvčí s nějakou jazykovou jednotkou setkává velmi často (a napříč různými kontexty), je její paměťová stopa silnější a její aktivace snazší. Silněji asociované jednotky pak můžou být uloženy v paměti jako „prefabrikované" kusy jazyka (*chunks*). Korpusová data (nejčastěji frekvence) proto slouží jako prediktory jazykového zpracování v psycholingvistice (např. rychlost rozpoznání slova v úloze lexikálního rozhodování). K tomu viz např. Diessel (2019) nebo Goldberg (2019) a příslušné odkazy.

Jazykový výzkum používající korpusová data jako kritickou či kontrolní proměnnou se netýká pouze typických mluvčích, ale i různých specifických populací, včetně lidí s afázií. Takové studie ukazují např. to, že korpusová frekvence funguje jako prediktor úspěšného vybavení slov v úloze pojmenování obrázku (např. Kittredge et al., 2008 nebo Nozari et al., 2010).

Novější afaziologické studie se zaměřují také na slova v kontextu a využívají proporční frekvenci a asociační sílu. Např. Rachel Hatchard a Elena Lieven (2019) provedly kvalitativní analýzu gramatických chyb v tvarech podstatných jmen, které se objevily v projevech anglických mluvčích s afázií. Ukazují mimo jiné, že tvary množného čísla podstatných jmen můžou být snazší na vybavení či být zdrojem paragramatických chyb, pokud jsou v korpusech angličtiny používány převážně v plurálu (např. *shoe*), a to přesto, že množné číslo je jako kategorie méně frekventované a komplexnější (obsahuje příponu *-s*). Zimmerer et al. (2018) zjistili, že angličtí mluvčí s nonfluentní afázií obecně produkují slovní spojení, v nichž jsou jednotlivá slova vzájemně silněji asociovaná.

Možnost přímého využití nástrojů ČNK v klinické praxi

Zmíněné afaziologické studie využívající korpusová data zpřesňují popis jazykového chování v afázii a pomáhají vysvětlit jeho specifika. Korpusová data zde přitom slouží jako vstupní data, z nichž jsou odvozeny příslušné kvantitativní ukazatele použité v analýzách. Korpusy přitom nemusí mít využití pouze ve výzkumu; lze si představit jejich využití v klinické praxi jako pomůcky, která podává spolehlivé komplexní informace o využití slov v současném živém jazyce. Tyto informace můžou sloužit při výběru kontextů, do nichž lze slova umístit, nebo přímo slovních tvarů, které jsou pro dané slovo typické. ČNK navíc na svých webových stránkách https://korpus.cz nabízí několik aplikací, které podobné informace zpřístupňují bez nároků na hlubší teoretické či technické znalosti. V této části krátce představím tři takové nástroje, pro něž si lze představit přímé využití v praxi, a to frekvenční seznamy a seznamy kolokací a dále aplikace *Gramatikat* (Kováříková a Kovářík, 2023) a *Slovo v kostce* (Machálek, 2020b).

Frekvenční seznam umožňuje získat v přehledné podobě základní představu o tom, nakolik jsou určité jednotky frekventované. Díky tomu, že korpusy ČNK jsou opatřeny zmíněnými lemmaty a informacemi o gramatice, lze vyhledávat do různé míry zobecněné struktury. Pokud by bylo cílem např. posílit trpný rod, lze ve vyhledávací aplikaci *KonText* (Machálek, 2020a) vybrat vhodný korpus (např. SYN2025) a s pomocí nástroje *Vložit tag* sestavit dotaz [tag="V.....P.*"], který vyhledá všechny tvary trpného rodu, aniž by bylo třeba vyhledávací jazyk aktivně ovládat. Následně lze přes kontextovou nabídku *Frekvence* vytvořit frekvenční seznam podle lemmat, případně konkrétních slovních tvarů. Vygenerovanou tabulku lze uložit ve vhodném formátu a následně jednotlivě procházet a vytipovat vhodná slovesa, která se v češtině často používají v trpném rodě. Pak je možné tato konkrétní slovesa vyhledat v korpusu znovu a získat kontexty, v nichž se vyskytují, či porovnat jejich frekvence v trpném rodě s frekvencemi celkovými, a získat tak frekvence proporční. Existující studie přitom ukazují, že ač je pro mluvčí s afázií trpný rod obecně považovaný za obtížný, může vysoká frekvence tuto obtížnost kompenzovat (Gahl a Menn, 2016; Jap et al., 2016). Zjistíme tak například, že 17. nejfrekventovanějším slovesem v trpném rodě je uložit (frekvence 2 401), u nějž tyto tvary tvoří více než třetinu všech výskytů (6 889), a toto sloveso by tedy mohlo být dobrým kandidátem pro cvičné věty. Podobně by např. při tvorbě cvičných vět s tranzitivními slovesy mohly pomoci syntaktické značky v SYN2025, které umožňují vyhledat pro jednotlivá slovesa časté přímé předměty. Komplikovanější dotaz [tag="NN.S4.*"&afun="Obj"&p_pos="V"] tak vyhledá podstatná jména v jednotném čísle a čtvrtém pádě, jejichž řídícím členem je sloveso a která mají větněčlenskou platnost přímého předmětu. Následně lze přes možnost *Vlastní* v kontextové nabídce

[6] Použití chí kvadrátu pro ověření těsnosti tohoto spojení ukazuje signifikantní asociaci mezi slovy, ovšem se zanedbatelnou velikostí efektu ($\chi 2 = 17\,446{,}85$, $p < 0{,}001$, Cramérovo $V = 0{,}0629$).

Frekvence vygenerovat seznam sloves (výběr atributu p_lemma) a přímých předmětů (atribut lemma). Výslednou tabulku lze opět uložit a dále procházet. Ukáže se tak například, že vůbec nejčastějším takovým spojením je *mít pocit* a že nejčastějším doplněním slovesa *zvednout* je *hlava*. Podobně lze využít funkci *Kolokace*, která pro dané slovo vygeneruje silně asociovaná spojení. Při vyhledání lemmatu OTEVŘÍT a následném využití funkce *Kolokace* (s nastavením atributu na lemma) se jako nejsilněji asociovaná slova ukáží *dveře*, *oko*, *ústa*, *okno* atd. Tyto informace lze opět využít při sestavování konkrétních cvičných položek, ať už bychom se chtěli zaměřit na jevy časté, a tedy zřejmě snazší, nebo naopak jevy vzácné, a tedy obtížnější.

Podobným způsobem funguje aplikace *Slovo v kostce*, která pro jedno konkrétní slovo vygeneruje komplexní profil (viz např. profil slovesa *auto* https://www.korpus.cz/slovo-v-kostce/search/cs/auto), který obsahuje informaci o frekvenci lemmatu, včetně srovnání frekvence v psaném a mluveném jazyce, a proporční frekvence jednotlivých vyskloňovaných či vyčasovaných tvarů. Dále jsou zobrazeny nejčastější kolokace a také slova, která se v korpusu vyskytují v podobných kontextech. Aplikace rovněž umožňuje porovnat z těchto hledisek dvě různá lemmata (např. dvojice auto – vlak, viz https://www.korpus.cz/slovo-v-kostce/compare/cs/auto--vlak). S využitím aplikace tak lze ověřit, jaké jsou obecné parametry užití daných slov v současné češtině a zda jim odpovídá způsob, kterým je slovo v určité úloze použito, či způsob užití daným mluvčím. S pomocí kolokací můžeme identifikovat typické přívlastky podstatných jmen nebo předmětová doplnění sloves či vytvářet skupiny významově spřízněných slov.

Aplikace *GramatiKat* poskytuje ve zpracované podobě proporční frekvence jednotlivých gramatických kategorií u podstatných a přídavných jmen a sloves. Tímto způsobem lze získat informaci např. o použití konkrétního podstatného jména v jednotném a množném čísle či v jednotlivých pádech včetně jeho srovnání s ostatními podstatnými jmény. Nástroj pomůže identifikovat slova, která jsou v daných gramatických kategoriích užívána nejčastěji, a proto můžou být vhodná pro jejich nácvik, nebo naopak najít kategorie, v nichž se dané slovo vyskytuje nejčastěji. To může pomoct určit, ve kterých tvarech dává smysl dané slovo posilovat v souladu s požadavkem na rozvoj funkční komunikace. Lze totiž předpokládat, že gramatické tvary pro dané slovo obvyklé budou zároveň komunikačně užitečnější, protože jsou motivované významově (Janda, 2019), srov. výrazně vyšší podíl sedmého pádu jednotného čísla u dopravních prostředků (jet vlakem/tramvají/autem). Podíváme-li se tímto způsobem na užití podstatných jmen v korpusu mluvené češtiny (apelativa s minimální frekvencí 50), ukazuje se, že se ve všech teoreticky možných 12 kombinacích čísla a pádu (při vynechání vokativu) vyskytuje pouze menší množství jmen (126 z 1 724). Uvažujeme-li proporční frekvenci tvarů, ukazuje se, že u nezanedbatelného množství lemmat (přes 14 %) odpovídá za 75 % a více veškerých užití šest či méně kombinací čísla a pádu. Tyto informace pak lze opět využít při tvorbě cvičných položek, ale také například při analýze chyb. Jak již bylo zmíněno výše, frekventované tvary můžou být zdrojem paragramatických substitucí (Hatchard a Lieven, 2019). Z opačné perspektivy si pak lze představit situaci, kdy by vysoká frekvence konkrétních slovních tvarů mohla maskovat reálnou výkonnost u dané gramatické kategorie, a při tvorbě testových položek zaměřených na konkrétní jev je tedy žádoucí tyto proporční frekvence sledovat.

Analýza slovních spojení ve výpovědích českých mluvčích s afázií

Text uzavírá ukázka toho, jak lze ve výzkumu i praxi využít specializované korpusy, v tomto případě jazyka v afázii. Jejich existence umožňuje pracovat v afaziologickém výzkumu s relativně většími objemy dat, což přináší spolehlivější a lépe zobecnitelné výsledky a umožňuje využití širšího spektra analytických technik (např. strojového učení). Zároveň tyto korpusy umožňují přístup k datům, jejichž sběr a zpracování by jinak byly pro výzkumníky velmi nákladné, a mají také významné pedagogické využití.

Největším takovým dostupným korpusem je AphasiaBank (MacWhinney et al., 2011), obsahující v současné době přepisy nahrávek více než 500 anglických mluvčích s afázií, které jsou opatřeny lingvistickými i klinickými metadaty. Tato analýza je založena na menším korpusu češtiny v afázii, který je zatím v předpublikační fázi a obsahuje data 11 osob s afázií a tří typických mluvčích (Láznička, 2022).

Analýza je zaměřena na bigramy, dvojice po sobě následujících slov vymezených tak, že např. spojení *analýza bigramů v korpusu české afázie* obsahuje bigramy *analýza bigramů – bigramů v – v korpusu – korpusu české – české afázie*. Ve dvou dílčích krocích ukazuju měření fluence a její vztah s frekvencí, které můžou být součástí diagnostiky, dále pak identifikaci konkrétního slovního spojení a jeho využitelnost v terapii.[7]

Data

Pro analýzu byla využita část korpusu češtiny v afázii obsahující jazykovou produkci elicitovanou na základě tří úloh: převyprávění úryvku z filmu, popis obrázku a tvorba krátkého příběhu na základě obrázků. Korpus je opatřen lemmatizací a slovnědruhovými značkami. Transkribovány jsou i veškeré dysfluence (Láznička, 2022). Tabulka 4 obsahuje základní charakteristiku použitých dat.

[7] K textu je připraven doprovodný „protokol", který obsahuje veškerý kód použitý pro zpracování a analýzu dat i podrobnější informace k metodologii. Protokol je společně s veškerými zdrojovými daty k dispozici na repozitáři Zenodo, viz https://doi.org/10.5281/zenodo.19695424.

Participant	Skupina	Počet výpovědí	Počet pozic (včetně dysfluencí)	Počet slov	Průměrná délka výpovědi (medián)	Počet bigramů
AA1	anomická	103	996	790	7,67 (5)	686
AA2	motorická transkortikální	159	1 034	508	3,19 (3)	354
AA3	kondukční	168	1 445	957	5,70 (5)	793
AA4	motorická transkortikální	77	430	225	2,92 (2)	152
BA1	Brocova	41	313	156	3,80 (3)	120
BA2	anomická	122	1 177	663	5,43 (5)	548
BA3	anomická	114	1 161	759	6,66 (5)	647
BA4	Brocova	89	352	178	2 (1)	63
PA1	anomická	87	624	488	5,61 (4)	405
PA2	motorická transkortikální	51	287	221	4,33 (4)	170
PA3	kondukční	106	1 157	806	7,60 (6)	698
AC1	bez afázie	108	1 122	915	8,47 (7,5)	807
AC2	bez afázie	108	841	621	5,75 (4)	516
AC3	bez afázie	63	426	360	5,71 (5)	297

Tabulka 4: ***Základní parametry použitého subkorpusu***

Z transkriptů rozčleněných na takzvané konverzační jednotky (viz Loban, 1966, přibližně odpovídají výpovědi) byly extrahovány bigramy v hranicích výpovědí v celkovém počtu 6 256. Dysfluence oddělující dvě slova v rámci výpovědi byly ignorovány a bigram označen jako dysfluentní (srov. Příklad 1).

Výpověď	*zistila že je na <pauza> s- sloupu <pauza> cirgusovym*
Bigramy	*zistila že – že je – je na – na sloupu – sloupu cirgusovym*

Příklad 1: ***Segmentace výpovědi na bigramy***

Stejným způsobem byly extrahovány všechny bigramy z korpusu mluvené češtiny ORAL (Kopřivová et al., 2017), který byl pro tuto analýzu zvolen jako referenční. Celkem bylo získáno 4 411 828 bigramů vyprodukovaných 2 787 mluvčími v 1 546 konverzacích.

Výpočet klíčovosti

Součástí analýzy byl výpočet tzv. klíčovosti získaných bigramů. Jde o typ asociační míry, která vyhodnocuje, do jaké míry jsou ve srovnání s nějakým referenčním korpusem v určitém textu či souboru textů jazykové jednotky pod- či nadreprezentované. Přístup je často využíván v analýze diskurzu, kde slouží k identifikaci slovní zásoby a gramatických struktur, pomocí nichž je určité téma prezentováno (pro typický příklad srov. Hořejší, 2017). Hodnota klíčovosti byla počítána na bigramech lemmat.[8] Na úrovni lemmat by tedy bigramy z Příkladu 1 vypadaly: *zjistit že – že být – být na – na sloup – sloup cirkusový*.

Tímto způsobem bylo analyzováno 2 748 unikátních bigramů vyprodukovaných mluvčími s afázií. Pro výpočet klíčovosti byla použita Kullbackova-Leiblerova divergence (Gries, 2024), která vyjadřuje, nakolik se pozorovaná distribuce určitého jevu (zastoupení bigramu v korpusu afázie) odchyluje od určité teoretické distribuce (zastoupení bigramu v obou korpusech dohromady).

Výsledky

Při analýze textů s využitím klíčovosti jsou zkoumaná slova typicky seřazena dle své hodnoty a následně je blíže analyzováno prvních *n* výskytů. Pro potřeby této analýzy je v Tabulce 5 uvedeno 30 bigramů s nejvyšší hodnotou, které mají frekvenci pět a vyšší. Bigramy s pěti a více výskyty lze v tomto případě považovat za informativní a analyticky zajímavé vzhledem k počtu nahrávek (11) a počtu bigramů (2 748) a také vzhledem k jejich frekvenční distribuci (frekvence pět a více je dosaženo pouze u 4,26 % všech unikátních bigramů).

[8] To bylo vedeno jednak velikostí afatických dat, a především mnoha idiosynkratickými tvary slov, které vyplývají z přepisu jazykové produkce v obou korpusech (např. slovo *jsou* se v korpusu ORAL objevuje jako *jsou – sou – sú – só – sóu*).

Lemma 1	Lemma 2	Frekvence (afázie)	Frekvence (ORAL)	Klíčovost
ten	lev	18	2	0,0322
myš	myš	10	1	0,0179
do	do	20	604	0,0153
ten	prase	11	61	0,0137
do	trouba	9	30	0,0123
ten	Chaplin	6	0	0,0109
ten	ten	54	12 756	0,0108
ten	klec	7	11	0,0107
tam	běhat	7	33	0,00898
z	z	11	315	0,00857
pán	tam	5	5	0,00804
horký	voda	6	24	0,00795
se	probudit	8	112	0,00789
ano	ano	11	488	0,00717
ten	kočka	7	90	0,00707
no	a	44	12 821	0,00671
tam	být	51	17 046	0,00633
ten	myš	5	32	0,00602
k	k	5	62	0,0051
dávat	do	6	136	0,00508
být	sníh	5	64	0,00506
na	země	7	270	0,00484
být	známý	5	78	0,00477
ten	dveře	6	170	0,00469
a	tam	16	2 666	0,00455
na	na	13	1 771	0,00438
a	a	17	3 258	0,00424
a	jít	10	1 137	0,00385
na	stůl	5	196	0,00344
no	ano	5	224	0,00324

Tabulka 5: ***Seznam 30 bigramů s nejvyšší hodnotou klíčovosti v korpusu afázie***

Při pohledu na bigramy je vidět, že většina z nich buď ukazuje na obsah elicitačních úloh, nebo na fakt, že se jedná o projevy mluvčích s afázií. V použitém videu je Ch. Chaplin v cirkusu uvězněn v kleci se lvem, tvorba příběhu je založena na třípanelovém komiksu, ve kterém kočka honí myš a způsobí, že se jídlo dané do trouby spálí, a na použitém Ladově obrázku je zabijačka na dvoře zasněženého venkovského stavení. Osm bigramů s celkovou frekvencí 141 má první a druhé lemma shodné. Pohled na konkrétní výskyty ukazuje, že většina z těchto bigramů souvisí s problémy způsobenými zřejmě vybavováním slov. To se týká i dvojice *ten – ten*, která je uvedena specificky proto, že lemmata ohebných slov zastupují širokou škálu slovních tvarů (zde všechny kombinace rodu, čísla a pádu), takže dvě stejná lemmata vedle sebe nutně nemusí znamenat opakování, srov. např. *tim to skončilo* (PA2). Bližší pohled však ukazuje, že i zde jde (kromě tří výskytů z celkových 54) o dysfluentní produkci, kdy je dané zájmeno zopakováno, opraveno jiným tvarem nebo dochází k reformulaci. V nahrávkách se tato opakování často vyskytují s dalšími dysfluentními projevy (hezitace, pauzy, slovní fragmenty) a 62 těchto bigramů (ze 141) je přímo dysfluentních.

Z tematicky nespecifických „plnovýznamových" bigramů je potenciálně zajímavé spojení *tam být*. Jedná se o komponenty konstrukce, která se označuje jako existenciální a prezentační (srov. anglické *there is*). Pokud zahrneme i varianty se slovesem *být* na první pozici nebo se zájmenným příslovcem tady, je celková frekvence konstrukce 72 výskytů v celkem 64 výpovědích, alespoň jednou byla přitom použita všemi participanty s výjimkou BA4. Zvýšená frekvence konstrukce je částečně dána elicitačními úlohami, které nutně obsahují určité prvky popisu či umisťování participantů do daných scén. Přesto ale zasluhuje další komentář.

Konstrukce sama o sobě není v češtině atypická, jedná se o gramatickou a obvyklou vazbu se specifickou funkcí. Co lze ovšem považovat za potenciálně atypické, je její nadužívání a výskyt v neočekávaných či nezcela patřičných kontextech (např. spíše deskriptivní oproti očekávanému narativnímu v úloze převyprávění filmové scény). To lze sledovat v Příkladu 2, který je reprezentativní ukázkou produkce participanta AA4.

20	AA4	osel <D> ktere terej <D> pána <D> poslal <D> ke <D> k lvoli
26	AA4	a tam <D> pan <D> tam p(r)án byl
27	AA4	a: <D> s- <D> lav <D> byl <D> tam
28	EXP	mh
30	AA4	pán <D> tam byl ale <D> taky
31	AA4	pán tam byl <D> hm

Příklad 2: ***Ukázka z participanta AA4; EXP = examinátor, pauzy a hezitace jsou nahrazeny značkou <D>***

Ve 42 výpovědích v této úloze participant kromě sledované vazby vyprodukoval samostatně pět výpovědí s jinými slovesy (u tří z nich se jedná o parafázie) a sedm výpovědí s konstrukcí tam být. Podobné komunikační chování je patrné u participanta AA2, který navíc produkoval i některé strukturní paralelismy, jak ukazuje Příklad *3*. Ve výpovědi 91 můžeme vidět nedokončenou výpověď, která je ve výpovědi 92 následována spojením *tam byl*, zřejmě s funkcí jisté opravy a dokončení předchozích dvou výpovědí (89 a 91). Zdá se, že u obou participantů s transkortikální motorickou afázií a s prvky agramatické produkce (jak ji popisuje Faroqi-Shah, 2023) slouží tato konstrukce jako jakási komunikační a strukturní opora, která umožňuje produkci plných větných rámců a zajišťuje určitou plynulost diskurzu.

16	AA2	šlo <D> osel nebo <D> osel
17	AA2	<D> a <D> tam <D> běal Chaplin <D> no
24	AA2	tam <D> dal <D> ten <D> ten ten <D> tam <D> běhal <D> běhal
89	AA2	tam <D> byl <D> (závt) <D> s- <D> vy- ne <D> pe- ne <D> lev
91	AA2	lev <D> dojde <D> dojde <D> n: do pos- ne
92	AA2	<D> tam <D> byl k tomu

Příklad 3: ***Ukázka z produkce participanta AA2***

Právě identifikace takového specifického větného rámce by mohla být následně využita v terapii. Lze se zde opřít o studii kolektivu autorů (Bruns et al., 2021), která se věnuje nácviku a je zaměřena na možnost využití podobných konkrétních slovních spojení s vysokou frekvencí pro ukotvení a nácvik dalších podobných, méně ustálených spojení (např. využití *I don't know* pro nácvik záporu obecně). Výsledky nebyly zcela jednoznačné, ovšem u části participantů ke zlepšení skutečně došlo. Pokud lze např. u participanta AA2 identifikovat spojení *tam být* jako jeden z „ostrůvků" fluence a zároveň jsou zohledněny zmíněné strukturní paralelismy, mohlo by být výhodné explicitně na danou strukturu upozornit a zařadit ji mezi používané komunikační strategie. Zároveň by mohlo být do terapie zařazeno vědomé rozšiřování tohoto rámce dalšími slovesy (např. *tam stál*, *tam ležel*, *tam četl* apod.).

Fluence bigramů a vztah k frekvenci

Na pozadí zmíněných dysfluencí, které mezi klíčovými bigramy zaujímají významné místo, lze v poslední části analýzy ukázat, jak by právě fluence mohla být využita pro porovnání řečové produkce mluvčích s afázií.

Nižší fluence, resp. vyšší výskyt dysfluencí, je jednou z nejnápadnějších charakteristik jazyka v afázii. To dobře ukazuje srovnání proporce dysfluentních bigramů mezi oběma korpusy. Zatímco v korpusu ORAL tvoří dysfluentní bigramy 5,14 % všech výskytů, v korpusu afázie je to 29,12 %. To může být do určité míry dáno i specifickým kontextem nahrávek, ovšem srovnání s třemi typickými mluvčími v korpusu afázie to nepotvrzuje. Při porovnání fluence jednotlivých participantů s afázií se sloučenými daty tří typických mluvčích je fluence statisticky významně nižší u všech participantů s výjimkou AA1, mluvčího s lehkou anomickou až reziduální afázií (viz Tabulka 6). Výrazně nižší podíl fluentních bigramů lze pozorovat především u participantů s nonfluentními afáziemi AA2, AA4, BA1 a BA4. Lze přitom usuzovat i na rozdíly mezi jednotlivými participanty. Např. afázie participanta PA2 byla charakterizována jako motorická transkortikální stejně jako u participantů AA2 a AA4, ovšem jejich výrazně nižší fluence ukazuje na závažnější zasažení jazykových schopností. Ukazuje se tak vhodnost návrhů pro využití produkce bigramů či trigramů jako dobře dostupné a informativní míry celkové fluence (srov. Bruns et al., 2021). V případě rozšíření korpusu češtiny v afázii by bylo možné srovnávat výkony konkrétních mluvčích s celým korpusem, a získat tak ucelenější představu o jejich afázii i jejím případném vývoji. Podobné nástroje na výpočty fluence i dalších parametrů můžou být za předpokladu dostatečného množství dat zabudovány přímo do korpusu (srov. Zimmerer et al., 2018).

Participant	Počet bigramů	Proporce fluentních bigramů	Rozdíl oproti typickým mluvčím	Korigovaná hodnota p
AA1	686	0,8455	−0,0206	n. s.
AA2	354	0,4435	−0,4225	< 0,001
AA3	793	0,7188	−0,1473	< 0,001
AA4	152	0,5526	−0,3134	< 0,001
BA1	120	0,4500	−0,4160	< 0,001
BA2	548	0,5967	−0,2693	< 0,001
BA3	647	0,7372	−0,1288	< 0,001
BA4	63	0,4127	−0,4534	< 0,001
PA1	405	0,8099	−0,0562	< 0,01
PA2	170	0,7882	−0,0778	< 0,01
PA3	698	0,7865	−0,0795	< 0,001

*Tabulka 6: **Výsledky porovnání fluence participantů s afázií se třemi typickými mluvčími (sloučená data); binomický test s Holmovou korekcí pro vícečetné srovnání***

Dostupnost dat z obecného korpusu dále umožňuje zaměřit se na vztah mezi fluencí bigramů a jejich frekvencí. Pro tyto potřeby byly použity všechny bigramy vyprodukované participanty s afázií a doložené i v korpusu ORAL, celkem 3 919 pozorování. Analýza pomocí binomické logistické regrese se smíšenými efekty ukázala signifikantní efekt frekvence ($\beta = -0{,}261$, $p < 0{,}001$): nárůst logaritmované frekvence bigramu o jednu jednotku vede k poklesu šance dysfluentní produkce přibližně o 23 % (OR = 0,77, 95% CI [0,71; 0,83]). Zajímavé je, že model se stejnou strukturou vytvořený pro vzorek 500 000 výskytů z korpusu ORAL ukazuje velmi podobný výsledek ($\beta = -0{,}252$, $p < 0{,}001$). To naznačuje, že efekt frekvence je mezi mluvčími s afázií a typickými mluvčími srovnatelný, což odpovídá obecným předpokladům tzv. usage-based přístupu k afázii. Ten předpokládá, že jazykové zpracování mluvčích s afázií a bez ní je kvalitativně stejné, avšak lidé s afázií mají větší množství problémů z důvodů omezených kognitivních zdrojů (Gahl a Menn, 2016).

Závěr

V tomto textu byly představeny základní principy korpusového přístupu ke studiu jazyka a možnosti jejich využití v lingvistické afaziologii. Příkladová analýza ukázala, že klíčovost identifikuje jak slovní zásobu charakteristickou pro použité elicitační úlohy, tak jevy charakteristické pro jazyk v afázii. Dále byl demonstrován vliv frekvence na fluenci souvislé řeči, který byl v tomto případě kvalitativně srovnatelný u typických mluvčích i mluvčích s afázií. Motivací pro vznik textu bylo především předvést potenciál jazykových korpusů pro afaziologii v klinické lingvistice a logopedii, a to jak pro výzkum jazyka v afázii, jehož výsledky je následně možno převést do klinické praxe (např. efekty frekvence), tak přímo pro klinickou praxi. V ní můžou najít využití nástroje ČNK (tvorba testových a terapeutických materiálů, analýza chyb) i vznikající korpus češtiny v afázii (srovnání klinických profilů, identifikace možných komunikačních strategií).

Zmíněný korpus češtiny v afázii bude tím přínosnější a využitelnější, čím rozsáhlejší soubor dat bude obsahovat, protože tak lépe zachytí široké spektrum specifik češtiny v afázii, a to jak pro pedagogické, tak popularizační účely. Zároveň poslouží jako cenný a snadno dostupný zdroj pro formulaci výzkumných otázek i celých studií.

Dovolím si tedy tento text zakončit nabídkou spolupráce všem logopedkám a logopedům, kteří by byli ochotni zapojit se společně se svými klientkami a klienty do budování korpusu sdílením nahrávek či jejich přepisů.

Dedikace

Tento výstup byl podpořen projektem Evropského fondu pro regionální rozvoj „MSCA Fellowships CZ – UK2“ (reg. č. CZ.02.01.01/00/22_010/0013392).

Acknowledgement

This output was supported by the European Regional Development Fund project “MSCA Fellowships CZ – UK2” (Reg. No. CZ.02.01.01/00/22_010/0013392).

Literatura

LÁZNIČKA, M., 2022. *Discourse production of Czech speakers with aphasia: A Usage-based exploration*. Dizertační práce. Praha: Filozofická fakulta Univerzity Karlovy. Vedoucí práce Eva Lehečková. Dostupné z: Discourse Production of Czech Speakers with Aphasia: A Usage-based Exploration | Digitální repozitář UK.

BRUNS, C., BEEKE, S., ZIMMERER, V. C., BRUCE, C. a VARLEY, R. A., 2021. *Training flexibility in fixed expressions in non-fluent aphasia: a case series report*. Online. International Journal of Language & Communication Disorders, vol. 56, no. 5, s. 1009-1025. DOI: 10.1111/1460-6984.12652. Dostupné z: Training flexibility in fixed expressions in non-fluent aphasia: A case series report - Bruns - 2021 - International Journal of Language & Communication Disorders - Wiley Online Library.

BRYSBAERT, M. a J. CORTESE, M. J., 2011. *Do the effects of subjective frequency and age of acquisition survive better word frequency norms?* Online. Quarterly Journal of Experimental Psychology, vol. 64, no. 3, s. 545-559. DOI: 10.1080/17470218.2010.503374. Dostupné z: Do the effects of subjective frequency and age of acquisition survive better word frequency norms?: The Quarterly Journal of Experimental Psychology: Vol 64 , No 3 - Get Access.

CVRČEK, V., KODÝTEK, V., KOPŘIVOVÁ, M., KOVÁŘÍKOVÁ, D., SGALL, P., ŠULC, M., TÁBORSKÝ, J., VOLÍN, J. a WACLAWIČOVÁ, M., 2010. *Mluvnice současné češtiny.* Praha: Karolinum. ISBN 978-80-246-1743-5.

DOEDENS, W. J. a METEYARD, L., 2020. *Measures of functional, real-world communication for aphasia: a critical review.* Online. Aphasiology, vol. 34, no. 4, s. 492-514. DOI: 10.1080/02687038.2019.1702848 Dostupné z: Full article: Measures of functional, real-world communication for aphasia: a critical review.

DOEDENS, W. J. a METEYARD, L., 2022. *What is Functional Communication? A Theoretical Framework for Real-World Communication Applied to Aphasia Rehabilitation.* Online. Neuropsychology Review, vol. 32, s. 937-973. DOI: 10.1007/s11065-021-09531-2. Dostupné z: What is Functional Communication? A Theoretical Framework for Real-World Communication Applied to Aphasia Rehabilitation | Neuropsychology Review | Springer Nature Link.

GAHL, S. a MENN, L., 2016. *Usage-based approaches to aphasia.* Online. Aphasiology, vol. 30, no. 11, s. 1361-1377. DOI: 10.1080/02687038.2016.1140120. Dostupné z: Usage-based approaches to aphasia: Aphasiology: Vol 30 , No 11 - Get Access.

GRIES, S. T., 2024. *Frequency, Dispersion, Association, and Keyness.* Amsterdam, Philadelphia: John Benjamins Publishing Company. ISBN 978-90-272-1492-8.

HATCHARD, R. a LIEVEN, E., 2019. *Inflection of nouns for grammatical number in spoken narratives by people with aphasia: how glass slippers challenge the rule-based approach.* Online. Language and Cognition, vol. 11. no. 3, s. 341-372. DOI: 10.1017/langcog.2019.21 Dostupné z: Inflection of nouns for grammatical number in spoken narratives by people with aphasia: how glass slippers challenge the rule-based approach | Language and Cognition | Cambridge Core.

JANDA, L. A., 2019. *Businessmen and Ballerinas Take Different Forms: A Strategic Resource for Acquiring Russian Vocabulary and Morphology.* Online. Russian Language Journal, vol. 69, s. 31-49. DOI: 10.70163/0036-0252.1045. Dostupné z: „Businessmen and Ballerinas Take Different Forms: A Strategic Resource „ by Laura A. Janda.

JAP, B. A., MARTINEZ-FERREIRO, S. a BASTIAANSE, R., 2016. *The effect of syntactic frequency on sentence comprehension in standard Indonesian Broca's aphasia.* Online. Aphasiology, vol. 30, no. 11, s. 1325-1340. DOI: 10.1080/02687038.2016.1148902. Dostupné z: Full article: The effect of syntactic frequency on sentence comprehension in standard Indonesian Broca's aphasia.

KITTREDGE, A. K., DELL, G. S., VERKUILEN, J. a SCHWARTZ, M. F., 2008. *Where is the effect of frequency in word production? Insights from aphasic picture-naming errors.* Online. Cognitive Neuropsychology, vol. 25, no. 4, s. 463-492. DOI: 10.1080/02643290701674851. Dostupné z: Where is the effect of frequency in word production? Insights from aphasic picture-naming errors: Cognitive Neuropsychology: Vol 25 , No 4 - Get Access.

KOPŘIVOVÁ, M., LUKEŠ, D., KOMRSKOVÁ Z., POUKAROVÁ, P., WACLAWIČOVÁ, M., BENEŠOVÁ, L. a KŘEN, M., 2017. *ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017.* Online. Praha: Ústav Českého národního korpusu. Dostupné z: https://www.korpus.cz.

KOVÁŘÍKOVÁ, D. a KOVÁŘÍK, O., 2023. *GramatiKat (verze 2). Nástroj pro výzkum gramatických kategorií a gramatických profilů.* Praha: Filozofická fakulta Univerzity Karlovy. Dostupné z: GramatiKat.

MACWHINNEY, B., FROMM, D., FORBES, M. a HOLLAND, A., 2011. *AphasiaBank: Methods for studying discourse.* Online. Aphasiology, vol. 25, no. 11, s. 1286-1307. DOI: 10.1080/02687038.2011.589893. Dostupné z: AphasiaBank: Methods for Studying Discourse - PubMed.

MACHÁLEK, T., 2020a. KonText: Advanced and Flexible Corpus Query Interface. In: CALZOLARI, N., BÉCHET, F., BLACHE, P., CHOUKRI, K., CIERI, C., DECLERCK, T., GOGGI, S., ISAHARA, H., MAEGAARD, B., MARIANI. J., MAZO, H., MORENO, A., ODIJK, J. a PIPERIDIS, S. (ed.). *Proceedings of the Twelfth Language Resources and Evaluation Conference. LREC Conferences.* Paříž: European Language Resources Association, s. 7003-7008. ISBN 979-10-95546-34-4. Dostupné z: lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

MACHÁLEK, T., 2020b. Word at a Glance: Modular Word Profile Aggregator. In: CALZOLARI, N., BÉCHET, F., BLACHE, P., CHOUKRI, K., CIERI, C., DECLERCK, T., GOGGI, S., ISAHARA, H., MAEGAARD, B., MARIANI. J., MAZO, H., MORENO, A., ODIJK, J. a PIPERIDIS, S. (ed.). *Proceedings of the Twelfth Language Resources and Evaluation Conference. LREC Conferences.* Paříž: European Language Resources Association, s. 7009-7014. ISBN: 979-10-95546-34-4. Dostupné z: lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

MCENERY, T. a HARDIE, A., 2013. The History of Corpus Linguistics. In: ALLAN, K. (ed.). *The Oxford Handbook of the History of Linguistics.* Oxford: Oxford University Press. ISBN 978-0-19-958584-7.

NOZARI, N., KITTREDGE, A. K., DELL, G. S. a SCHWARTZ, M. F., 2010. *Naming and repetition in aphasia: Steps, routes, and frequency effects.* Online. Journal of Memory and Language, vol. 63, no. 4, s. 541-559. DOI: 10.1016/j.jml.2010.08.001. Dostupné z: Naming and repetition in aphasia: Steps, routes, and frequency effects - ScienceDirect.

ZIMMERER, V. C., NEWMAN, L., THOMSON, R., COLEMAN, M. a VARLEY, R. A., 2018. *Automated analysis of language production in aphasia and right-hemisphere damage: frequency and collocation strength.* Online. Aphasiology, vol. 32, no. 11, s. 1267-1283. DOI: 10.1080/02687038.2018.1497138. Dostupné z: Automated analysis of language production in aphasia and right-hemisphere damage: frequency and collocation strength: Aphasiology: Vol 32, No 11.