Univerzita Karlova v Praze

Filozofická fakulta

Ústav obecné lingvistiky

Bakalářská práce

Eliška Konývková

Experimentální validace metrik lexikální diverzity

Experimental validation of lexical diversity metrics

Praha 2021

Vedoucí práce: PhDr. Jiří Milička, Ph.D.

Na tomto místě bych chtěla poděkovat vedoucímu práce PhDr. Jiřímu Miličkovi, Ph.D. za jeho podporu, konstruktivní připomínky a časovou flexibilitu při konzultacích. Moc děkuji také hodnotitelkám a hodnotitelům textů pro experiment za jejich čas a energii, kterou tomu věnovali. V neposlední řadě patří velké díky Bětce, Ivě, Vojtěchovi a kolektivu Jako Doma.

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 30.12.2021 ……..

Jméno a příjmení: Eliška Konývková

Abstrakt:

Lexikální diverzita se v lingvistice dotýká mnoha oblastí (např. výuka L1 i L2, diagnostika afázie, forenzní lingvistika, kvantitativní lingvistika obecně). Její hodnoty zjišťujeme pomocí mnoha tradičních metrik, jako je počet typů, interval opakování stejných slov, sémantické propojení slov apod. Tyto metriky byly zavedeny, aniž by se zkoumalo, jak korelují se subjektivním hodnocením lexikální diverzity. První externí subjektivní evaluaci validity těchto metrik tak uskutečnili teprve během posledního desetiletí v několika článcích týmy kolem Scotta Jarvise. V experimentu, který popisují v článku z března roku 2021 američtí lingvisté Kristopher Kyle, Scott A. Crosley a Scott Jarvis, nechali respondenty ohodnotit lexikální diverzitu cca šesti set textů a jejich hodnocení porovnali s indexy vypočítanými již zmiňovanými tradičními metrikami. Téma lexikální diverzity a její validace od vydání článku v lingvistické komunitě velmi rezonuje. Cílem práce je zopakovat experiment Scotta Jarvise a kol. pro češtinu.

Abstract:

Lexical diversity touches many areas in linguistics (e.g. L1 and L2 teaching, aphasia diagnosis, forensic linguistics, quantitative linguistics in general). Its values are ascertained using many traditional metrics such as the number of types, the repetition interval of the same words, the semantic connectivity of words, etc. These metrics have been introduced without examining how they correlate with subjective assessments of lexical diversity. Thus, the first external subjective evaluation of the validity of these metrics was only carried out during the last decade in several papers by the teams around Scott Jarvis. In an experiment described in a March 2021 paper by American linguists Kristopher Kyle, Scott A. Crosley and Scott Jarvis, they had respondents rate the lexical diversity of about six hundred texts and compared their ratings with indices calculated by the traditional metrics mentioned earlier. The topic of lexical diversity and its validation has resonated strongly in the linguistic community since the paper was published. The aim of this paper is to replicate the experiment of Scott Jarvis et al. for English.

Klíčová slova: lexikální diverzita, metriky lexikální diverzity, opakování experimentu, kvantitativní lingvistika

Keywords: lexical diversity, metrics of lexical diversity, repetition of the experiment, quantitative linguistics

Obsah

1. Úvod

Na základě článku od kolektivu autorů Kristophera Kylea, Scotta A. Crossleyho a Scotta Jarvise *Assessing the Validity of Lexical Diversity Indices Using Direct Judgements* reprodukuji jejich experiment. Cílem experimentu bylo vyhodnotit, do jaké míry metriky lexikální diverzity odrážejí holistické subjektivní hodnocení LD zhotovené lidskými respondenty. Kolektiv autorů porovnal indexy vztahující se ke třem dimenzím LD (*abundance, variety a volume*) a holistické hodnocení dvou „trénovaných" respondentů. Data získali ohodnocením textů od dvou skupin mluvčích, pro jednu skupinu byla angličtina prvním jazykem, pro druhou byl anglický jazyk druhým nebo dodatečným jazykem.

Autoři si položili dvě výzkumné otázky, první: „Jaký je vztah mezi lidským hodnocením LD a objektivním měřením lexikální *volume, abundance* a *variety*?" a druhou: „Do jaké míry se vztahy mezi lidským hodnocením LD a objektivním měřením lexikální *volume, abundance* a *variety* doplňují?

Cílem této práce je ve vlastním experimentu zodpovědět za pomoci dostupných prostředků první výzkumnou otázku a výsledky porovnat s těmi, ke kterým došli autoři původního experimentu. Metodologie použitá v této práci se od původního experimentu částečně liší, v práci průběžně popisujeme v čem, a z jakého důvodu ji neopakujeme přesně.

Druhá výzkumná otázka se zabývá vytvořením predikčních modelů lidského hodnocení, které by byly schopny vysvětlit odchylky v subjektivním hodnocení lexikální diverzity. Autoři též zkoumali, jak se hodnocení liší u respondentů vzhledem k tomu, jestli je angličtina jejich prvním či doplňkovým jazykem. S přihlédnutím k omezenému rozsahu práce se druhou výzkumnou otázkou nezabýváme.

2. Teoretická část

2.1 Koncept lexikální diverzity

„Je všeobecně uznávané, že jsou indexy lexikální diverzity užitečné, přestože výzkumníci věnující se jazyku zanedbali otázku, co vlastně měří.“[1]

S rozvojem oblasti kvantitativní lingvistiky se definice pojmu lexikální diverzity od 30.let 20.století měnila a zpřesňovala. Už G. K. Zipf se při svém výzkumu pořadí slov podle frekvence zmiňuje o jevu, který můžeme zahrnout do konceptu LD. Ve své knize Psycho-biologie jazyka popisuje tendenci k vyváženosti při distribuci slov a uvádí, že „vysoký stupeň uspořádanosti rozložení slov v proudu slovní zásoby neomylně ukazuje na tendenci udržovat rovnováhu v řeči mezi frekvencí na jedné straně a tím, co lze předběžně označit jako rozmanitost na straně druhé“ (Zipf, 1965).

Nicméně první zmínku o diverzitě slovní zásoby jako takové najdeme v článku od J.B. Carrolla. V článku z roku 1938 tento koncept popisuje jako: „Vzorky řeči nebo textu jednotlivců různého věku, inteligence a původu se budou lišit způsobem, který můžeme nazvat diverzita, to je, relativní počet opakování nebo naopak relativní různorodost slovníku“ (Carroll, 1938, str.379).

Jarvis (2017) srovnává přístup výše uvedených autorů k problematice LD a popisuje, jak Carrollova „rigidní“ definice ovlivnila další vývoj vnímání konceptu LD. Ve stejném článku srovnává dva různé přístupy k definování konceptu lexikální diverzity. Popisuje, že vytvořené metriky LD vznikly aplikací etických řešení na emickou problematiku (Jarvis, 2017, str. 539). Přístup etický, objektivizující, vysvětluje Jarvis jako založený na pozorování z vnějšku. Naopak přístup emický bere v úvahu kontext, rozvíjí problematiku zevnitř. Klíčem k uchopení konceptu LD z emické perspektivy je podle Jarvise zapojení subjektivních holistických hodnocení LD (Jarvis, 2017,str. 539).

[1] „Thus, it is recognized that indices of lexical diversity are useful, even though language researchers have neglected the question of what it is that they are actually measuring.“(Jarvis, 2013, str.94)

Tento komplexní přístup k chápání a měření LD rozvíjí Jarvis již ve svém článku Capturing the Diversity in Lexical Diversity z roku 2013. Lexikální diverzitu popisuje na porovnání s diverzitou ekosystémů. Badatelům na poli kvantitativní lingvistiky vytýká, že se soustředí pouze na frekvence typů a na rozdíl od ekologů nevnímají LD jako multidimenzionální fenomén (Jarvis, 2013). Identifikované dimenze LD jsou popsány autory opakovaného experimentu v jiné kapitole (viz 2.4.1).

2.2 Využití objektivních metrik LD

Na tomto místě považuji za vhodné zmínit, proč je validace metrik LD důležitá. Přestože se koncept lexikální diverzity rozvinul na poli lingvistiky a objektivní měření LD zde stále nachází uplatnění, zasahuje využití metrik LD také do mnoha dalších oblastí. Jak uvádí McCarthy & Jarvis (2007), „tvůrci a testeři jakékoli metriky LD tedy musí přijmout zodpovědnost za zavádění těchto nástrojů“ (str. 476). Níže uvádím několik oblastí, které s lexikální diverzitou a jejími metrikami pracují.

Metriky LD se používají při zkoumání a diagnostice neurobiologických poruch, např. afázie. Harris Wright et al. (2003) ve svém článku popisují využití metriky D (vyvinutou Malvernem a Richardsem,1997) při analýze promluv dospělých s afázií. V této práci testujeme oproti lidským hodnocením konkretizovanou verzi metriky, HD-D.

Další oblastí, ve které mohou být metriky LD využity je forenzní lingvistika. Psycholog Kevin Colwell (2002) použil metriku LD, konkrétně Type Token Ratio (TTR), aby v experimentu odlišil křivé a pravdivé výpovědi svědků. V této práci testujeme aktualizovanou verzi TTR, a totiž „*moving average*“ TTR (MATTR).

Dále jsou metriky lexikální diverzity široce využívány ve zkoumání jazykové akvizice (Malvern, 2008), v akvizici L2 (Jarvis & Daller, 2013a). Někteří z autorů tohoto článku se zabývali metrikami LD vzhledem k nastavení hodnocení kvality psaní „*writing quality*“ (A. Crossley & S. McNamara, 2016).

2.3 Shrnutí dosavadních výzkumů validace metrik LD

Zkoumáním platnosti metrik lexikální diverzity v porovnání s hodnocením uděleným lidskými respondenty se zabýval Scott Jarvis, jeden z autorů opakovaného experimentu již v předchozích letech. Spolu s dalšími lingvisty provedl dva experimenty, v nichž se blíže zaměřoval zkoumání lidských hodnocení lexikální diverzity.

2.3.1 Experiment 2013

V článku z roku 2013 zkoumá kolektiv autorů, zda běžně užívané lexikální indexy odpovídají konstruktům, které by měli měřit. Autoři porovnávají lidská hodnocení s několika kategoriemi jako je adekvátnost kolokací, specifičnost slov a v neposlední řadě s lexikální diverzitou.(Jarvis & Daller, 2013a).

Korpus textů byl ohodnocen třemi rodilými mluvčími angličtiny, kteří byli v hodnocení „vytrénovaní“ („*trained raters*“). Hodnotitelům byl poskytnut trénink na dvaceti textech z „testovacího korpusu“, který nebyl součástí následného experimentu. Při hodnocení samotných textů měli respondenti za úkol udělit první známku v rozmezí 1-6 pro každou z analytických kategorií (mezi nimi i lexikální diverzitu), a druhou známku v rozmezí 1-5 udělovali jako holistické skóre pro celý text. Skóre pro jednotlivé kategorie i holistická hodnocení byly následně zprůměrovány a použity při validaci s metrikami LD (Jarvis & Daller, 2013b).

Pro kategorii lexikální diverzity bylo s holistickými hodnocením nejsilněji korelováno TTR, druhá byla metrika D, třetí pak MTLD. Nejslabší korelaci měla metrika „*Yuleovo K*“ (Jarvis & Daller, 2013a, str. 121).

2.3.2 Experiment 2017

V druhém článku z roku 2017 *„Grounding lexical diversity in human judgements“* Scott Jarvis velmi detailně popisuje metodu přidělování holistických hodnocení lidskými respondenty. Metodu upravoval několik let a v článku popisuje různé způsoby, jak nakonec zajistil, aby respondenti dosáhli vysoké mezianotátorské shody *(„inter-rater reliability*“).

Jarvis ve svém článku zařazuje mnoho detailů, které mi pomohli při zadávání textů respondentům pro tuto práci. Sběr hodnocení probíhal v průběhu několika let a Jarvis postupně upravoval podmínky. Zpřesňoval instrukce pro hodnotitele, opravoval chyby v textech

k hodnocení a experimentoval s tréninkem hodnotitelů a poskytování motivace. Níže uvádím přehled jeho postupu.

2.3.2.1 Korpus

Z důvodu možného ovlivnění hodnotitelů různými proměnnými se Jarvis rozhodl zařadit do korpusu texty, které si byly obsahově blízko. Zvolil tedy texty, ve kterých autoři psanou formou převypravovali scénku z filmu Charlieho Chaplina („*narrative retelling*"), dohromady obsahoval korpus 276 textů, zařazeni byli autoři, pro které byla angličtina prvním jazykem i autoři, kteří se angličtinu naučili jako druhý nebo doplňkový jazyk.

Jarvis texty zadal nejprve dvěma hodnotitelům – expertům, aby texty ohodnotili v kategorii kvality psaní („*writing quality*"). Výsledky poté seřadil a z korpusu vybral padesát textů tak, aby reprezentovali celou škálu hodnocení udělenou těmito hodnotiteli.

Poslední úpravou textu bylo odstranění chyb. Jarvis předvídal, že chyby v textu mohou budoucí hodnotitele lexikální diverzity ovlivnit, proto v textech opravil všechny zásadní gramatické chyby.

Jarvis testoval korelace metrik LD a hodnocení respondentů v letech 2011, 2012, 2014 a 2015. Respektive se v prvních třech experimentech zaměřoval na nastavení hodnocení tak, aby hodnotitelé dosáhli vysoké mezianotátorské shody.

2.3.2.2 Hodnocení 2011

V prvním experimentu z roku 2011 texty ohodnotilo 11 respondentů, kteří byli buď učitelé angličtiny jako cizího jazyka nebo studenti lingvistiky. Texty hodnotili odděleně podle náhodných setů a Jarvis uvádí, že neprošli žádným školením na hodnocení LD (Jarvis, 2017, str. 543). Jisté vodítko hodnotitelům však poskytl, a to v podobě jednoho textu hodnocením lexikální diverzity 5. Respondentům byl koncept lexikální diverzity popsán jako „rozmanitost slov v textu" (str. 543) a byli požádáni, aby texty ohodnotili na škále od jedné do desíti (1 pro nejnižší LD, 10 pro nejvyšší LD).

Jarvis naměřil velmi nízkou mezianotátorskou shodu (měřil vždy mezi dvěma hodnotiteli), průměr Pearsonovy korelace byl r = 0,30.

2.3.2.3 Hodnocení 2012

O rok později experiment zopakoval, tentokrát s dvaceti hodnotiteli – studenty lingvistiky. Hodnotitelům také zadal přesnější instrukce k samotnému hodnocení.

Instrukce zněly:

„Přečtěte si text a ohodnoťte jeho lexikální diverzitu na škále 1-10 (1 je nejnižší možné hodnocení, 10 nejvyšší). LD není ekvivalentní jazykové odbornosti nebo úrovni psaní, ujistěte se tedy, že vaše hodnocení odráží opravdu LD – rozmanitost slov, nikoli jak dobře je text napsaný. Vaše vnímání a intuice je při hodnocení LD nejdůležitější. Nepřemýšlejte o hodnocení příliš, snažte se však být konzistentní. Pro snazší nastavení vašeho hodnocení předkládám esej, který na škále LD reprezentuje level 5. Ohodnoťte lexikální diverzitu ostatních textů v souladu s tímto příkladem (Jarvis, 2017, str. 543).“

Při zadávání textů respondentům jsem vycházela právě z těchto instrukcí (kromě poskytnutí příkladu hodnocení LD na levelu 5), viz kapitolu 3.3.

Přestože měli hodnotitelé k dispozici vzorově ohodnocený text, korelace průměru hodnocení respondentů byla r = 0,32, tedy jen nepatrně silnější než v roce 2011.

Přestože byl koeficient spolehlivosti Cronbachova alfa (α = 0.90) vysoký, pojal Jarvis podezření, že respondenti neberou svoji úlohu příliš vážně (chodili pozdě, odcházeli dřív apod.) Na základě této zkušenosti se rozhodl respondenty v dalším experimentu motivovat, aby „udělali, co je v jejich silách“(Jarvis, 2017, str. 544).

2.3.2.4 Hodnocení 2014

V roce 2014 provedl Jarvis další experiment, 21 hodnotitelů oznámkovalo 50 textů z let 2011 a 2011 a navíc ještě deset textů ze stejného korpusu. Navíc se rozhodl poskytnout hodnotitelům čtrnáct dní před samotným hodnocením LD výcvik v hodnocení ve Společném Evropském referenčním rámci pro jazyky (CEFR). Respondenti poté texty ohodnotili podle tohoto rámce a o týden později hodnotili LD textů. Navíc oproti minulým experimentům poskytl za účast na úloze hodnotitelům motivaci v podobě získání plusových bodů v kurzu lingvistiky.

Průměr korelace mezi dvojicemi hodnotitelů byl znatelně vyšší, r = 0,57. U Jarvise však vyvstala obava, zda není skóre lexikální diverzity ovlivněno právě tréninkem hodnocení textů

podle CEFRu. Když porovnal hodnocení podle rámce CEFR a hodnocení lexikální diverzity, zjistil mezi nimi silnou korelaci r=0,85.

2.3.2.5 Hodnocení 2015

V roce 2015 zopakoval Jarvis experiment znovu, opět jako respondenty oslovil studenty lingvistiky z Ohio University, dohromady texty hodnotilo dvacet respondentů. Aby Jarvis potvrdil, že hodnocení z minulého roku není ovlivněno možným zkreslujícím efektem tréninku v hodnocení podle CEFR rámce, zadal hodnotitelům nejdříve úlohu hodnocení lexikální diverzity. Pro tuto úlohu byly poskytnuty stejné instrukce a motivace jako pro hodnotitele v roce 2014 a stejně jako v roce 2012 a 2014 jim bylo k ohodnocení zadáno šedesát textů. Hodnotitelé mezi sebou dosahovali korelací od r=0,10 do r=0,84, Cronbachova alfa dosáhla hodnoty α =0.95, což Jarvis považuje za dostatečný důkaz, že výsledky hodnocení z roku 2014 jsou replikovatelné a odrážejí lexikální diverzitu a nikoli jiný koncept.

Až o týden později byl respondentům poskytnut výcvik v hodnocení podle CEFR rámce, o dva dny později ohodnotili respondenti podle tohoto rámce stejných šedesát textů jako při hodnocení LD, avšak v jiném pořadí. Párová Pearsonova korelace mezi hodnotitely variovala mezi r=0,3 až r=0,74. Pearsonova korelace mezi hodnocením textů na lexikální diverzitu a hodnocením podle CEFR rámce byla velmi silná, r=0,89. (Jarvis, 2017, str. 546).

2.3.3 Shrnutí

Jarvis z výsledků vyvozuje, že mluvčí jednoho jazyka mají velmi podobné intuice o lexikální diverzitě (Jarvis, 2017, str. 548). Tvrdí, že lidská hodnocení mohou sloužit jako standard vůči kterému mohou být validovány metriky LD.

2.4 Popis experimentu Kylea et al.

Před praktickou částí této bakalářské práce zařazuji detailnější popis dat a postupů, které využívali Kyle et al. ve svém experimentu.

2.4.1 Lexikální diverzita má mnoho dimenzí

Lexikální diverzitu autoři experimentu vnímají jako mnohodimenzionální fenomén, vycházejí z Jarvisovy (2013a, 2017) studie, ve které porovnává měření a popis lexikální

diverzity s diverzitou biologickou. V článku z roku 2013 (Jarvis & Daller, 2013b) uvádí Jarvis pouze šest dimenzí (*properties*) LD, v článku autoři Kyle et al. uvádí jednu navíc, a to abundanci (*abundance*). Při definování *konstruktu* LD v témže článku popisují TTR jako užitečnou metriku, která je však s měřením LD nekompatibilní, a to hlavně z důvodu závislosti na délce textu. Domnívám se, že z tohoto důvodu neuvádí ani počet typů jako vlastnost lexikální diverzity. U jedné dimenze Jarvis dvakrát změnil terminologické pojmenování, původně ustavenou jako *variegation*, o rok později přejmenoval na *variability* (Jarvis & Daller, 2013b) a nyní o ní referuje jako o *variety*. Mění se pouze pojmenování, popis dimenze zůstává beze změny.

Podle autorů zahrnuje koncept LD tyto dimenze:

Rozsah *volume* (počet tokenů)

Abundance *abundance* (počet typů)

Variace *variety* (relativní poměr jedinečných slov)

Vyváženost *evenness* (míra vyjadřující rovnost počtu opakování typů)

Disparita *disparity* (sémantická příbuznost slov)

Neobyčejnost *specialness* (přítomnost konkrétních slov vnímaných jako obohacující diverzitu)

Rozptyl *dispersion* (velikost intervalů opakování stejného slova)

V experimentu se objevují z výše uvedených dimenzí LD dimenze tři. Jsou to *abundance, volume a variety.*

2.4.2 Popis indexů a jejich spolehlivosti

K výpočtu indexů LD používali autoři software TAALED (Tool for the Automatic Analysis of Lexical Diversity) vyvinutý Kristopherem Kylem ve spolupráci se Scottem Jarvisem a Scottem Crossleym. Níže popsané metriky autoři článku identifikovali jako relativně nezávislé na délce textu. Čtyři indexy níže byly použity k vypočítání hodnot *variety*, jedné z dimenzí LD.

HD-D: *hypergeometric distribution diversity index* udává pravděpodobnost, s jakou se určitý počet tokenů jednoho typu vyskytne v náhodně vybraném vzorku textu. Výsledná hodnota je

složená z pravděpodobností vypočítaných pro každé slovo v textu, vzorek textu obsahuje 42 tokenů. Čím vyšší hodnota HD-D, tím vyšší lexikální diverzita (McCarthy & Jarvis, 2010).

HD-D, tak jak ho uvádí kolektiv autorů v experimentu, je založené na indexu voc-D, uvedeném v roce 1997 Davidem Malvernem a Brianem Richardsem. Podle McCarthyho a Jarvise (2010) je tento index závislý na délce textu jen velmi málo. Nezávislost indexu na délce textu ověřovali i Zenker a Kyle (2021), a to na poměrně objemném korpusu 4 542 textů.

MATTR: *moving average type-token ratio* snižuje efekt délky textu na *type-token ratio* (TTR) tím, že počítá průměrné TTR v překrývajících se výsecích (oknech) textu. Z TTR vypočítaných na jednotlivých výsecích poté software spočte průměr. Počet tokenů v jednom výseku lze nastavit, Jarvis et al. uvádí, že pro svůj experiment počítali MATTR na výsecích textů po 50-ti tokenech. Na kolik tokenů je výsek nastavený ovlivňuje výslednou hodnotu TTR, viz kapitola 4.1. Stabilitu vzhledem k různým délkám textů ověřili Zenker a Kyle (2021).

MTLD: *the measure of textual lexical diversity* analyzuje řetězec tokenů, sekvenci zahajuje s TTR = 1 a podle opakování typů v sekvenci se zmenšuje. Když dosáhne předem stanovené hodnoty TTR (podle autorů článku .720, také McCarthy 2010) připočítá hodnotu – faktor 1. Při sčítání faktorů zůstávají zbytkové sekvence, které nedosáhnou hodnoty 1. MTLD tyto sekvence nevyřadí, ale vypočte hodnotu částečného faktoru. Aby nedošlo k chybě (zbytkové sekvence mají různou délku) zpracuje software data i v opačném směru. Konečná hodnota MTLD je vypočítána z průměrné hodnoty po směru i proti směru analyzovaných sekvencí tokenů (McCarthy & Jarvis, 2010).

MTLD – W: *the measure of textual lexical diversity – moving window* je varianta MTLD popsaného výše, normovaná obdobně jako MATTR. U této metriky bylo též ověřeno, že je nezávislá na délce textu (Zenker & Kyle, 2021).

Použité statistické metody

Pro posouzení vztahu mezi indexy lexikální diverzity a holistickým hodnocením lidskými respondenty byla v článku Kylea et al. použita Pearsonova korelace se dvěma proměnnými.

2.4.3 Korpus

Co se týče žánru textů v experimentu Kylea et al., všechny texty jsou zastřešeny žánrem polemiky (argumentative essay). Texty byly vybrány ze dvou korpusů. První část korpusu obsahovala texty mluvčích, jejichž prvním jazykem byla angličtina (315 textů), druhá část

korpusu sestávala z textů, které napsali mluvčí, pro které byla angličtina druhým nebo doplňkovým jazykem (300 textů). Dohromady shromáždili autoři experimentu 615 textů k ohodnocení. Téma bylo předem určené a studentům byl zadán časový limit. Tyto texty poté roztřídili podle známek („*writing score*“) a z každé úrovně ohodnocení vybrali poměrný počet textů (Kyle et al., 2021).

2.4.4 Hodnotitelé

Kolektiv autorů vybrané texty následně nechal ohodnotit dvěma „trénovanými“ hodnotiteli. Způsob nácviku hodnocení LD v tomto experimentu se lišil od předchozích experimentů zabývajícími se validací metrik LD.

Výcvik hodnotitelů a kalibrace hodnocení probíhala v několika fázích. V první fázi byly hodnotitelům poskytnuty tři texty s již přiděleným hodnocením LD. Ohodnocené texty s hodnotami 2,89, 5,00 a 8,04 byly převzaty z Jarvisova experimentu z roku 2017 a hodnotitelé byli požádáni, aby v souladu s nimi nastavili svoji škálu hodnocení. Ve druhé fázi ohodnotili respondenti na lexikální diverzitu sto textů, které nijak nesouvisely s výzkumem. Poté co bylo dosaženo vysoké „mezianotátorské shody“ (Kappa >0,70), ohodnotili respondenti samotné texty do experimentu.

V poslední fázi dostali respondenti možnost upravit ta hodnocení, která se od ostatních lišila o více než jeden bod. Touto úpravou se zvýšila shoda mezi hodnotiteli z Kappa = 0,667 na Kappa = 0,748. Průměr takto kalibrovaných hodnocení byl zařazen do analýzy.

Jarvis ve svém článku z roku 2013 uvádí, že „by byla chyba trénovat respondenty v hodnocení lexikální diverzity“(Jarvis, 2013, str. 101), dále však doplňuje, že nevylučuje poskytovat hodnotitelům trénink v pozdějších fázích výzkumu, až bude jisté, že lidé vnímají koncept lexikální diverzity a také bude popsáno jakým způsobem. V experimentu, který opakujeme, kolektiv autorů hodnotitelům poskytuje nejen trénink v hodnocení, ale i možnost upravit svoje odlišné hodnocení podle druhého respondenta (Kyle et al., 2021, str.159). Můžeme tedy předpokládat, že jsou si autoři v souladu s Jarvisovým tvrzením jisti, že lidé mají společný rámec pro vnímání lexikální diverzity a je i dostatečně popsáno jaký tento rámec je.

V závěru článku autoři shrnují, že dva hodnotitelé mohou být nedostateční a je větší šance, že se v jejich hodnocení vyskytne chyba, a tudíž nebudou nic vypovídat o hodnotách LD.

Pro další výzkum tedy vyzývají k zapojení více hodnotitelů. Tento požadavek naplňuje ve svých dřívějších výzkumech Scott Jarvis a kolektiv (Jarvis & Daller, 2013b),(Jarvis, 2017), kde do hodnocení zapojuje až dvě desítky hodnotitelů.

2.4.5 Výsledky experimentu Kylea et al.

Po porovnání holistických hodnocení a výsledků vypočítaných pro *volume, abundance* a *variety* došli Kyle et al. k závěru, že „všechny indexy LD vykazují střední až vysokou korelaci s lidskými hodnocením LD“(Kyle et al., 2021, str. 162).

Nejsilněji byla s holistickým hodnocením korelována dimenze abundance (počet typů), následně volume (počet tokenů), HD-D a další indexy LD (Kyle et al., 2021, str.163). Viz přiloženou tabulku níže.

Index	r	p
Volume (Tokeny)	0.687	<.001
Abundance (Typy)	0.847	<.001
MATTR	0.492	<.001
HD-D	0.602	<.001
MTLD	0.505	<.001
MTLD-W	0.524	<.001

2.4.6 Shrnutí

Kyle et al. z experimentu vyvozují, že z různých dimenzí lexikální diverzity nejlépe koreluje holistické lidské hodnocení s dimenzí abundance (počet typů v textu) a s indexy „variety“ nezávislé na délce textu (Kyle et al., 2021).

Ze silné korelace subjektivních skóre LD a dimenze abundance (počet typů) autoři článku vyvozují, že představování nových myšlenek v textu („*idea generating*“) je důležitý faktor při vnímání lexikální diverzity textu (Kyle et al., 2021, str. 168). Vyzývají k dalšímu prozkoumání této souvislosti. Toto téma bakalářská práce více nerozvíjí, výsledky follow-up interview ovšem naznačují, že i když na to hodnotitelé nebyli předem upozornění, při analýze svých hodnocení

uváděli tuto kategorii (různě konceptualizovanou – „idea, myšlenka, pointa“) jako důležitý faktor, který jejich hodnocení ovlivnil (viz kapitolu 4.9).

3. Praktická část

3.1 Popis dat

Při výběru vhodných texů k analýze jsme přihlíželi k praxi předchozích experimentů validace metrik LD. Zvolili jsme tedy žákovský korpus. V článku z roku 2017 popisuje Jarvis parametry vzniku takových textů. Ty by měly být „vytvořeny volně (i když obvykle jako odpověď na výzvu, a obvykle v předem stanoveném časovém limitu) a způsobem, který se velmi podobá přirozené jazykové produkci“(Jarvis, 2017, str. 538).

Žákovský korpus Script 2015 (*Skript 2015*, b.r.) obsahoval minimum textů, u kterých by si studenti a studentky mohli určit žánr a téma písemné práce. Vybírala jsem tedy texty podle určeného žánru, a v tomto žánru poté texty se stejným tématem (pokud bylo na výběr z více témat). Další údaje ke vzniku samotných textů nebyli v korpusu k dispozici, předpokládám, že žákům mohly být zadány ještě další podmínky, např. minimální počet slov.

Textů je celkem čtyřicet, z každé žánrové kategorie deset. Zapojit víc žánrů bylo nutné, protože v rámci jednoho žánru nebylo v korpusu k dispozici dostatek textů. Případně bylo nevyvážené zastoupení autorek a autorů textů. Výběr více žánrů s sebou přináší několik možných komplikací – některé texty podle určitého žánru jsou často delší (vypravování), ať už je to tím, že na vystavění příběhu je potřeba víc prostoru, nebo již zmiňované normy ohledně minimálního počtu slov.

Přehled textů a počtu jejich tokenů:

Identifikace	Popis prostředí	Osobní dopis	Vypravování	Charakteristika
1	448	311	384	323
2	361	554	427	398
3	428	345	468	397
4	428	357	476	300

5	379	541	400	328
6	514	351	470	346
7	297	380	396	355
8	269	304	465	360
9	282	381	459	354
10	551	343	511	297

Další dimenzí metadat, které lze z korpusu Script zjistit, je ohodnocení konkrétního textu známkou udělenou vyučujícím, zde na škále A – D. Jedná se tedy o hodnocení pravopisu, stylistiky textu. V rámci možností jsem texty vybírala tak, aby byla zastoupená co nejširší škála známek. Stejně tak Kyle et al. vybírali do svého výzkumu texty tak, aby byla reprezentována široká škála ohodnocení. „Vybrané texty byly rozvrstveny podle hodnocení, aby byla zastoupena škála skóre *[...]"* (Kyle et al., 2021, str. 159). Přestože známka udělená vyučujícím odráží spíše pravopis a stylistiku textu (i když lexikální diverzita se do celkového hodnocení, známky, také promítá), v analýze ji zařazuji ke srovnání s holistickými hodnocením LD, viz kapitolu 4.4.1.

Texty v korpusu byly ohodnoceny známkami A – D. Poměr známek ve vybraných textech je následující:

Známka	Výskyt celkem
A	10x
B	14x
C	11x
D	5x

K autorům a autorkám textů – všichni v době psaní textu studovali první ročník gymnázia, věk 15 let. Vybrané texty byly napsané dvaceti ženami a dvaceti muži.

3.2 Výběr hodnotitelů a hodnotitelek

V experimentu, který opakujeme byly texty ohodnoceny dvěma respondenty. Oba prošli nácvikem hodnocení lexikální diverzity (viz kapitolu 2.4.4). Respondenti, kteří hodnotili texty

pro tuto práci, však udělovali známku bez možnosti hodnocení jakkoli „natrénovat“. Respondentka E a respondent K mají pedagogické vzdělání a věnovali se ve svých povoláních výuce českého jazyka. Neprošli však žádným speciálním tréninkem a bylo na nich, jak si LD konceptualizují. Respondenti B, J a S dosáhli také vysokoškolského vzdělání, nikdy se však systematicky nevěnovali lingvistice nebo českému jazyku.

Tím, že měli všichni respondenti alespoň minimální vysokoškolské vzdělání byla alespoň minimálně splněna podmínka, aby texty hodnotili „jednotlivci, kteří sdílejí stejný původ („*background*“)“(Jarvis, 2017, str. 542).

Celkem texty ohodnotilo pět respondentů, tři ženy a dva muži.

Informace o respondentech uvádím v tabulce:

Respondent/ka	Muž/Žena	Věk	Nejvyšší dosažené vzdělání
E	Ž	59	Mgr.
K	M	66	PaedDr.
B	Ž	28	Bc.
J	M	29	Bc.
S	Ž	33	Ing.

3.3 Instrukce pro hodnotitelky a hodnotitele

Vzhledem k navazujícímu rozhovoru (follow-up interview) s respondentkami a respondenty bylo před hodnocením textů sděleno minimum instrukcí. Hodnotitelé byli obecně seznámeni s tématem bakalářské práce a poté instruování po vzoru Scotta Jarvise v článku z roku 2017.

„Přečtěte si text a ohodnoťte jeho lexikální diverzitu na škále 1-10 (1 je nejnižší možné hodnocení, 10 nejvyšší). LD není ekvivalentní jazykové odbornosti nebo úrovni psaní, ujistěte se tedy, že vaše hodnocení odráží opravdu LD – rozmanitost slov, nikoli jak dobře je text napsaný. Vaše vnímání a intuice je při hodnocení LD nejdůležitější. Nepřemýšlejte o hodnocení příliš, snažte se však být konzistentní. Pro snazší nastavení vašeho hodnocení předkládám esej,

který na škále LD reprezentuje level 5. Ohodnoťte lexikální diverzitu ostatních textů v souladu s tímto příkladem (Jarvis, 2017, str. 543).“[2]

Cílem navazujícího rozhovoru po odevzdání hodnocení bylo zjistit, na co se hodnotitelé zaměřovali, jaké jevy pro ně spadaly pod termín „lexikální diverzita“. To odpovídá Jarvisově experimentu z roku 2013, kde uvádí, že podání informací o dimenzích LD by bylo zkreslující, protože „účelem studie bylo v první řadě zjistit, zda těchto šest faktorů ovlivňuje lidské vnímání lexikální rozmanitosti, aniž by posuzovatelům bylo řečeno, co mají hledat“(Jarvis & Daller, 2013, str. 34).

Jediným, avšak podstatným rozdílem v instrukcích zadaných respondentům pro tuto práci a těm z roku 2013 byla nemožnost nastavit svá hodnocení podle poskytnutého příkladu, v případě Jarvise textu ohodnoceného objektivně jako textu s úrovní LD 5.

[2] Quickly read each essay. As soon as you finish reading an essay, rate its level of lexical diversity on a scale of 1 to 10, where 1 is the lowest possible and 10 is the highest possible. Lexical diversity is not the same thing as language proficiency or writing quality, so make sure that your rating reflects only the lexical diversity – or variety of words – in the essay, not how well the essay is written. Your perception or intuition of lexical diversity is what matters the most. Don’t think too much about your ratings, but do try to be consistent. To help calibrate your ratings, assume that the essay below represents a level 5 on the lexical diversity scale. Rate the other essays in relation to how lexically diverse they are in comparison to this example.(Jarvis, 2017, str. 544)

4. Výsledky

4.1 Program pro analýzu

Analýza textů byla původně plánovaná ve stejném softwaru, který použili autoři článku (TAALED) (Kyle et al., 2021). Program však z textů nevygeneroval přesvědčivý výstup (počty typů a tokenů neodpovídaly, pro různé indexy byly generovány stejné výsledky). Program jsem otestovala zadáním textů nejen v češtině, ale také v angličtině, výsledky neodpovídaly ani v jednom případě.

Vzhledem k výše uvedeným komplikacím, bylo textům přiřazeno objektivní hodnocení LD vypočítané v softwaru koRpus (Michalke, 2020). K propočtu indexů jsem použila verzi aplikace veřejně přístupnou na webových stránkách (koRpus text analysis, b.r.).

Bylo nutné zjistit, jestli program dokáže správně vyhodnotit i texty zadané v českém jazyce. Empiricky bylo zjištěno, že program pracuje s UTF-8. Přičemž nic dalšího než porovnávat řetězce v UTF-8 nebylo pro analýzu potřeba, neboť slova již byla předsegmentována (každé slovo na novém řádku) a taktéž lemmatizace textů byla zpracována předem.

Parametry nastavené při generaci hodnocení jsou uvedené v tabulce níže.

MSTTR segment size:	100
MTLD/MTLD-MA factor size:	0.72
MTLD-MA min. tokens/factor:	50
MTLD-MA step size:	50
HD-D sample size:	42
MATTR moving window:	50 – 100 – 150 – 200 - 250
Base for logarithm:	10

4.2 Lemmata a slovní tvary

Z analýzy textů vzniky dva sety dat, první pro lemmata a druhý pro slovní tvary. Kyle et al. neuvádějí, zda texty ve svém experimentu analyzovali ve formě lemmat nebo slovních tvarů. Motivací pro toto rozdělení byl Jarvisův a Hashimotův článek o vlivu operacionalizace slovních typů na metriky LD (2021). V něm porovnávali výsledky operacionalizací jednotlivých „slovoforem“. Autoři článku však neuvádějí, zda k analýze použili lemmata nebo slovní tvary, nebylo by tedy možné tyto dva datasety porovnat. Pro analýzu v této bakalářské práci jsme tedy zvolili lemmata a slovní tvary neanalyzovali.

4.3 Mezianotátorská shoda

Mezianotátorskou shodu jsme vyhodnotili několika způsoby, jednak propočtem korelací mezi jednotlivými hodnotiteli (viz kapitolu 4.4) a také metrikami pro výpočet shody mezi více hodnotiteli (zde mezi pěti hodnotiteli).

První z nich je Krippendorffova alfa, vypočítaná online v aplikaci ReCalc (*ReCal for Ordinal, Interval, and Ratio Data (OIR) – Deen Freelon, Ph.D.*, b.r.). Při hodnotě $\alpha = 0.109$ se její signifikance ukázala jako velmi nízká.

Druhou použitou metrikou je Cronbachova alfa, též vypočítaná online v aplikaci (Wessa, 2021). Její hodnota při započítání všech respondentů je $\alpha = 0.5723$, což se nachází na hraně signifikance.

	Hodnotitelé	Cronbachova alfa
a.	Všichni	0.5723
b.	Bez respondentky E	0.482
c.	Bez respondenta K	0.5935
d.	Bez respondentky B	0.4748
e.	Bez respondenta J	0.3813
f.	Bez respondentky S	0.6208

Z tabulky výše můžeme vyčíst, jak se koeficient mění, když vynecháme jednoho z hodnotitelů. Z přehledu vychází, že hodnocení respondentky S jsou od ostatních respondentů odlehlá. Při vynechání jejího hodnocení se Cronbachova alfa zvýší na $\alpha =$

0.6208. Jakým způsobem se hodnocení liší můžeme pozorovat v grafu subjektivních hodnocení v kapitole 4.4.

Při posuzování shody ve svém experimentu měřili Kyle et al. shodu pouze mezi dvěma respondenty, zvolili proto výpočet koeficientu Kappa. Hodnota koeficientu nejdříve dosáhla Kappa = 0,667, poté co měli hodnotitelé možnost svá hodnocení navzájem porovnat a upravit ta z nich, která se lišila o více než jeden bod, se hodnota zvýšila na Kappa = 0,748.

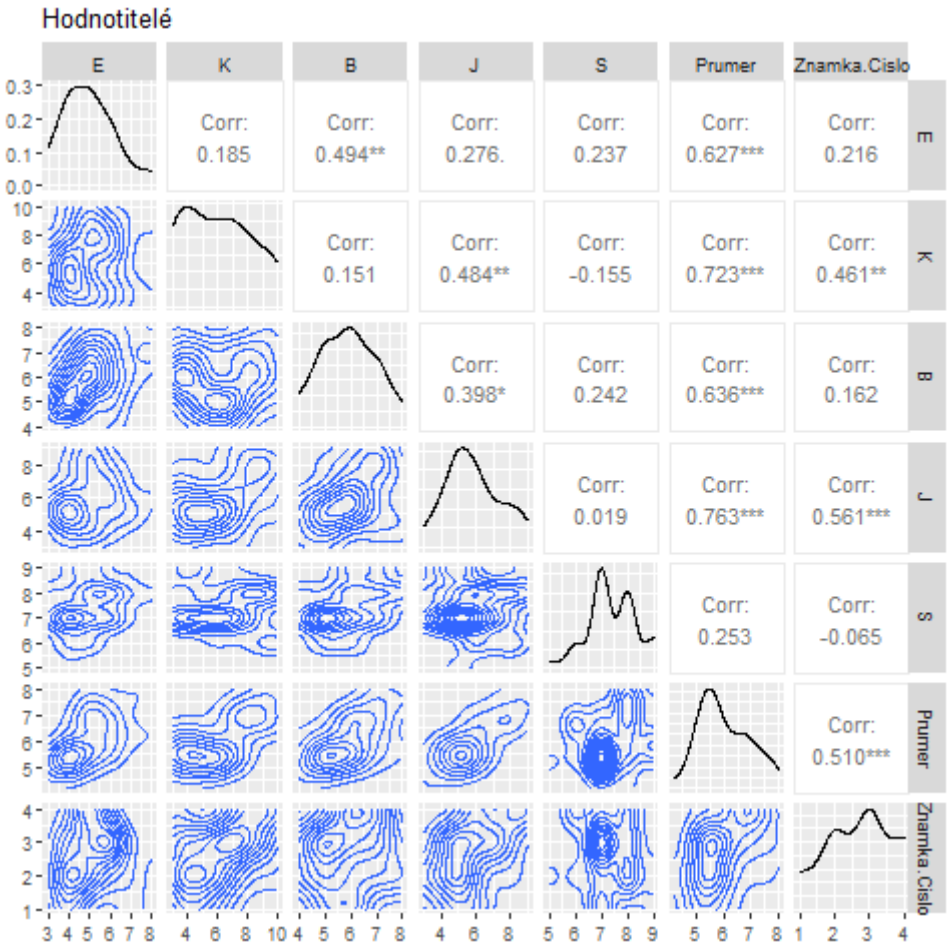
4.4 Pairwise korelace

4.4.1 Korelace mezi hodnotiteli a mezi hodnotiteli a známkou

Při porovnání korelací jednotlivých respondentů mezi sebou najdeme dvě silnější korelace mezi dvěma páry hodnotitelů. První mezi hodnotitelkami B a E (r = 0,494), druhou mezi hodnotiteli J a K (r = 0,484), hodnotitelka S s nikým výrazněji nekoreluje a potvrzuje se tak i výsledek Cronbachovy alfy – její hodnocení jsou odlehlá.

Korelace mezi dvěma páry hodnotitelů (E+B, K+J) naznačuje, že mají podobné vnímání konceptu lexikální diverzity. Když srovnáme jejich odpovědi ve *follow-up* interview, tak zjistíme, že se v mnohém shodují i ve svých introspekcích.

Při srovnání subjektivních hodnocení a známky, kterou jsme zjistili z korpusu (známky udělené vyučujícím) najdeme nejsilnější korelaci u hodnotitelů J a K (r = 0,561; r = 0,461). Hodnotitel K má pedagogické vzdělání a věnoval se systematicky výuce českého jazyka, hodnotitel J má humanitní vzdělání a věnoval se výuce anglického jazyka. Hodnotitelka S je se známkou korelována slabě negativně (r = -0,065).



Graf 1: Korelace mezi hodnotiteli, Korelace subjektivních hodnocení a známky udělené vyučujícím

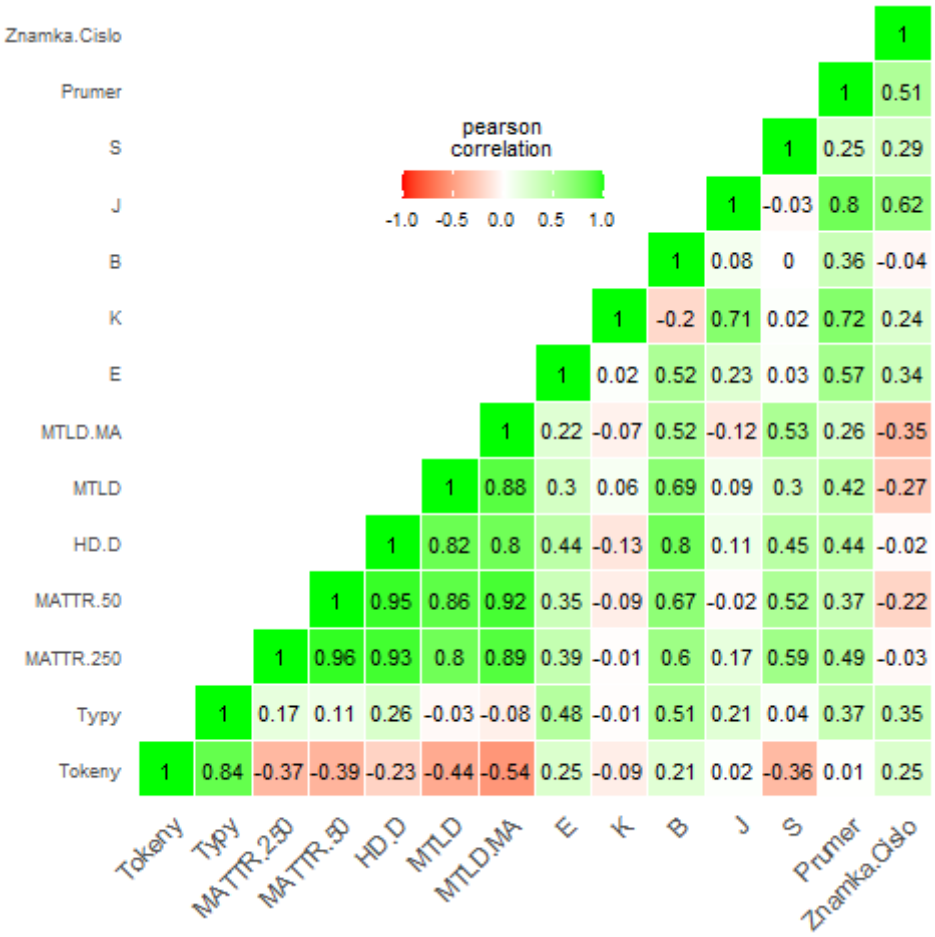
4.5 Korelace hodnocení a žánrů

V této kapitole porovnáváme korelace subjektivních i objektivních hodnocení v rámci jednoho žánru. Zařazení textů, které spadají do více žánrů je odlišné od postupu Kylea et al., zároveň však můžeme pozorovat, jak se subjektivní i objektivní hodnocení u jednotlivých žánrů proměňovalo.

4.5.1 Popis místa

Silnou korelaci vykazují metriky *variety* mezi sebou (viz také kapitolu 4.7), u všech čtyř žánrů se naměřené hodnoty pohybují mezi r=0,7 až 0,98. *Volume* (tokeny) a metriky *variety* nabývají hodnot středně silné inverzní korelace. Korelace tokenů a metrik je u každého žánru jiná, zápornou korelaci vykazuje také žánr „dopisu“. Pro přehled uvádím tabulku průměrného počtu typů a tokenů a jejich korelací.

Žánr	Počty typů (průměr)	Počet tokenů (průměr)	Korelace typy (abundance)	Korelace tokeny (volume)
Popis místa	229,9	409,6	r= 0,26 až – 0,08	r= - 0,54 až – 0,23
Vyprávění	229,6	443,3	r= 0,87 až 0,58	r= 0,67 až 0,33
Charakteristika	207	344,5	r=0,57 až 0,4	r= 0,09 až – 0,07
Dopis	203,2	381,9	r= - 0,5 až – 0,11	r= - 0,66 až – 0,28

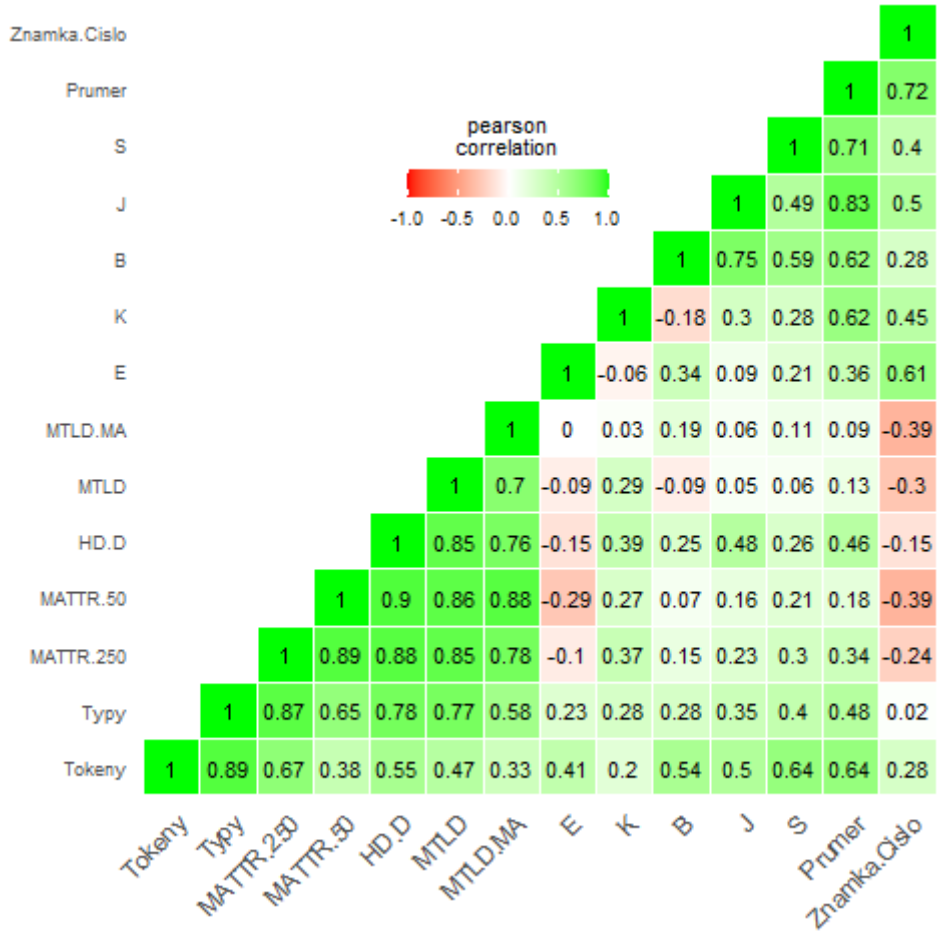


Graf 2: Pearsonova korelace hodnocení a metrik u žánru „popis místa"

4.5.2 Vyprávění

Korelace typů a tokenů s metrikami v tomto žánru je v porovnání s ostatními nejsilnější.

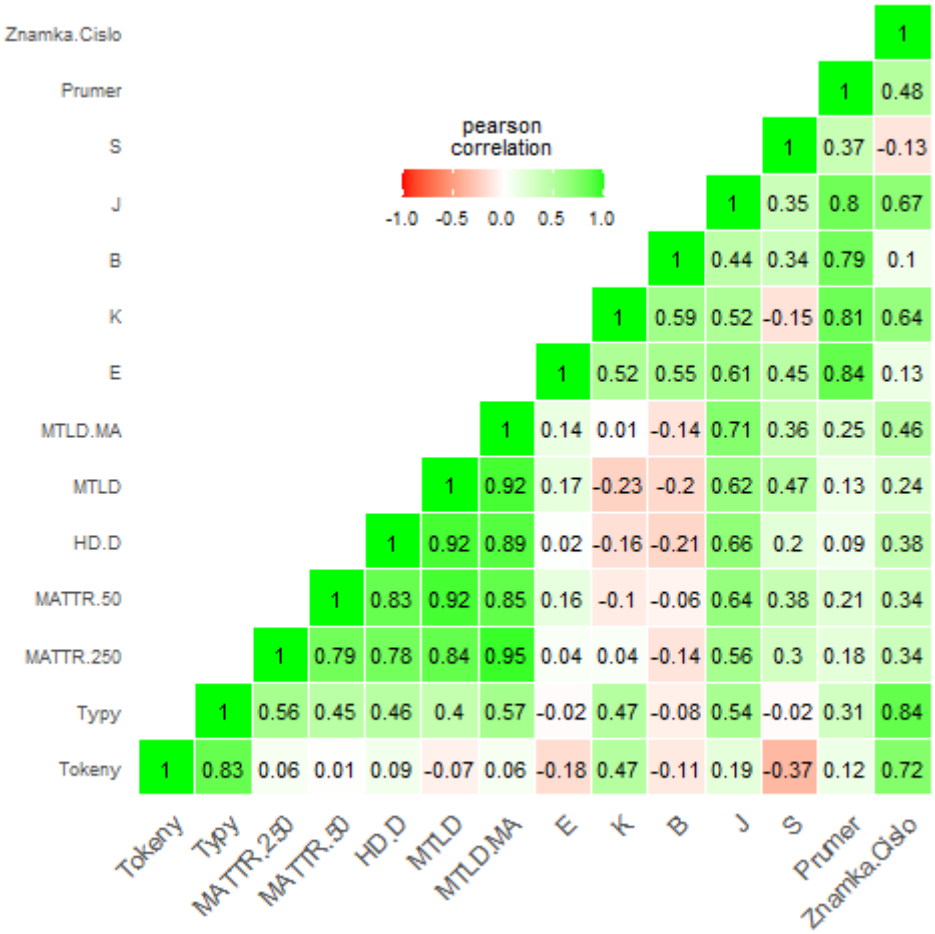
Z grafu můžeme vyčíst, že je zde ze všech žánrů nejsilnější záporná korelace metrik variety se známkou udělenou vyučujícím (r= -0,39 až – 0,15).



Graf 3: Pearsonova korelace hodnocení a metrik u žánru „vyprávění"

4.5.3 Charakteristika

Tento žánr vykazuje velmi nízkou korelaci tokenů i typů vůči metrikám variety. Ze všech žánrů má však nejsilněji korelaci mezi tokeny, typy a známkou udělenou vyučujícím, typy dosahují korelace r= 0,84, tokeny r= 0,72. Oproti ostatním žánrům je tato korelace velmi výrazná (ostatní žánry se pohybují mezi r= - 0,31 až r= 0,02).

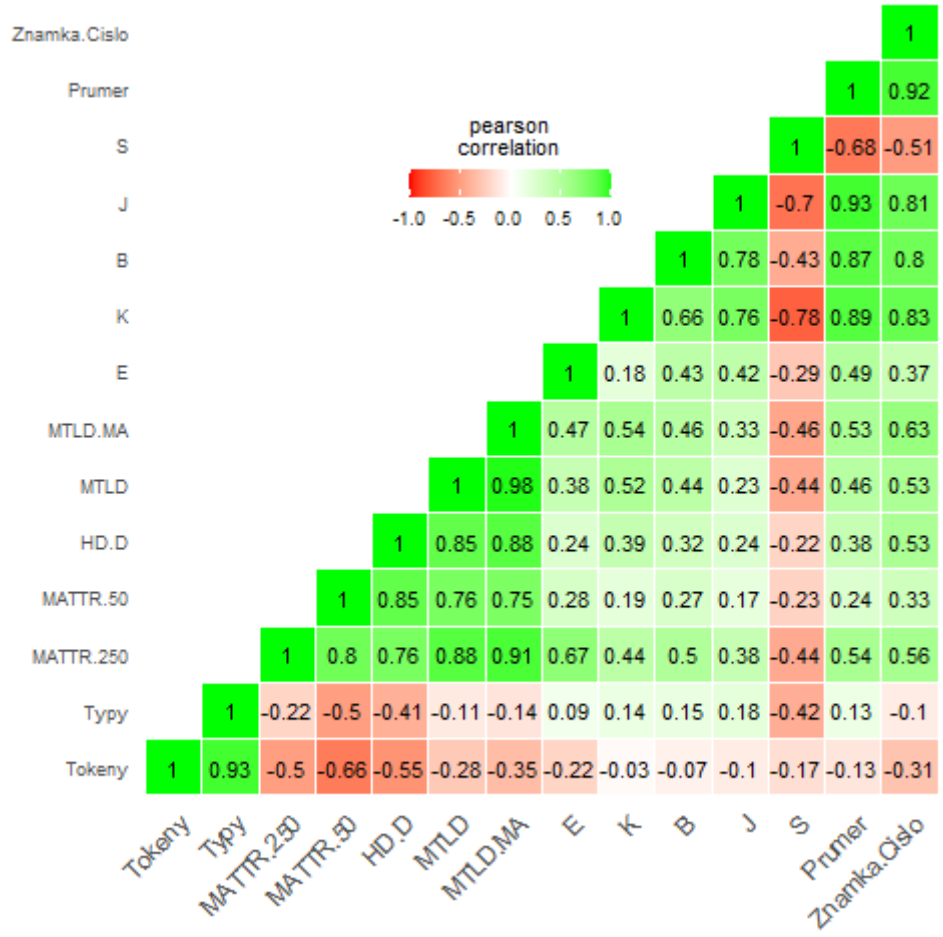


Graf 4: Pearsonova korelace hodnocení a metrik u žánru "charakteristika"

4.5.4 Dopis

Žánr dopisu vykazuje středně silnou negativní korelaci mezi typy, tokeny a metrikami variety. Při porovnání s ostatními žánry z grafu vyčteme, že má nejsilnější korelaci mezi metrikami variety a známkou udělenou vyučujícím.

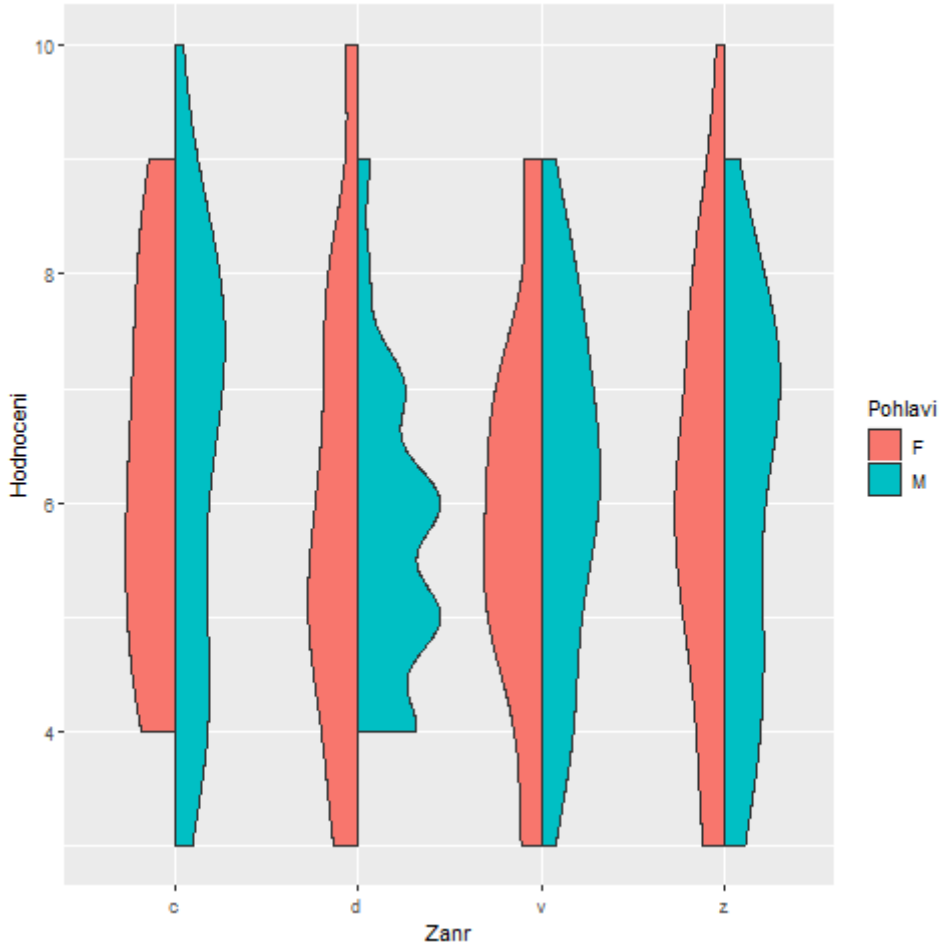
Díky tomu, že jsme vydělili korelace po jednotlivých žánrech, můžeme pozorovat místy velmi silnou zápornou korelaci u žánru dopisu u hodnotitelky S. Když se potom podíváme na celkovou korelaci subjektivních i objektivních hodnocení u všech žánrů dohromady, zjistíme, že hodnotitelka S má i zde nižší korelaci. Důvodem je právě hodnocení tohoto žánru.



Graf 5: Pearsonova korelace hodnocení a metrik u žánru "dopis"¨

4.5.5 Porovnání hodnocení jednotlivých žánrů

V houslovém grafu níže můžeme pozorovat, že byly hodnocení všech respondentů poměrně vyvážené vzhledem k žánru textů i vzhledem k pohlaví autorů textů. Nepotvrdila se tedy analýza vlastního hodnocení respondentek B a S, které tvrdily, že některé žánry (dopis a popis místa) hodnotily nižším skóre.



Graf 6: Porovnání hodnocení jednotlivých žánrů; „c" označuje charakteristiku, „d" označuje dopis, „v" označuje vyprávění, „z" označuje popis místa

4.6 Korelace subjektivních hodnocení LD a indexů LD

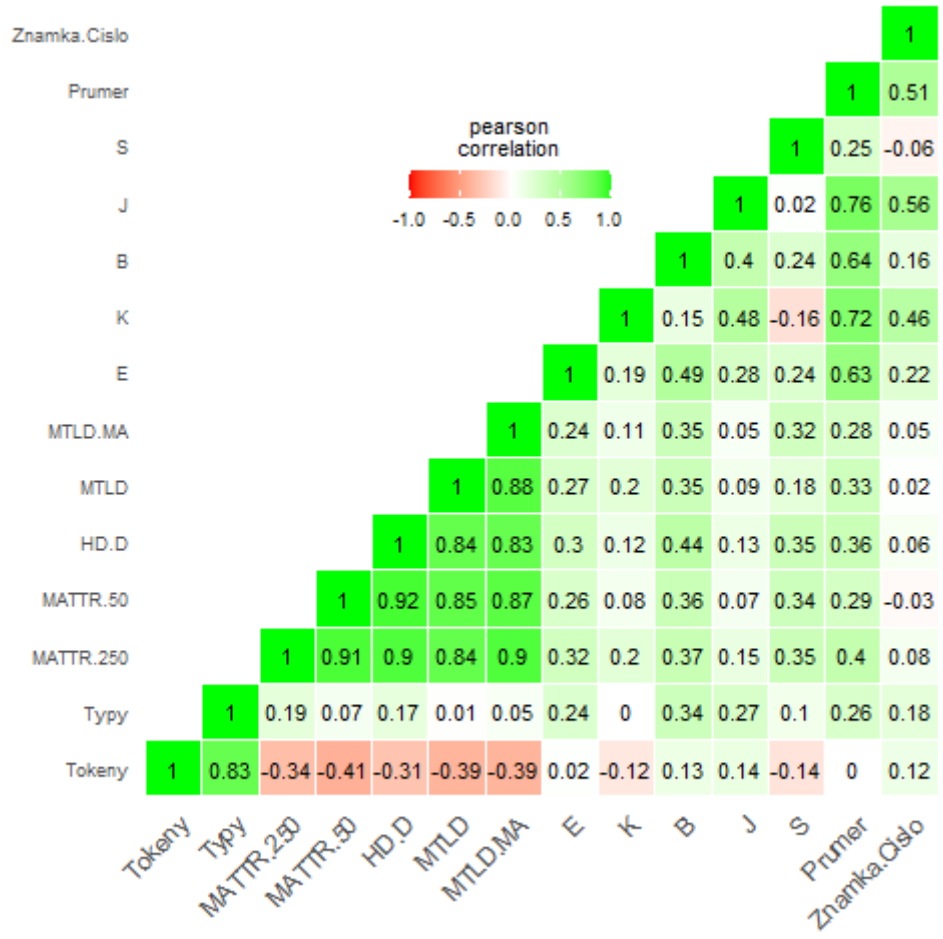
V grafu č. 7 odpovídáme na výzkumnou otázku, kterou si položili Kyle et al. a kterou v této bakalářské práci replikujeme. Otázka zní: „Jaký je vztah mezi lidským hodnocením LD a objektivním měřením lexikální *volume, abundance* a *variety*?"

Výsledky naměřené autory Kylem et al. vykazovaly středně silné až silné korelace mezi všemi indexy *volume, abundance* a *variety* a lidskými hodnoceními. Níže postupně porovnáme výsledky týkající se všech třech dimenzí a holistického hodnocení.

Autoři článku zjistili, že vůbec nejsilněji je s lidskými hodnocením korelována dimenze *abundance*, tedy počet typů. Když se podíváme na graf níže, tak zjistíme, že z našich měření je dimenze *abundance* s lidskými hodnoceními korelovaná jen velice slabě. Personův koeficient korelace se pohybuje mezi r = 0,34 až r = 0,0. Nejsilněji je s počtem typů korelována hodnotitelka B, r = 0,34, tato korelace je však pořád jen středně silná. Korelační koeficient pro počet typů a hodnotitele K je r= 0,002304.

Druhou nejsilněji korelovanou dimenzí v článku Kylea et al. bylo *volume* neboli počet tokenů. V našem měření byla korelace *volume* se subjektivním hodnoceními ještě slabší než u abundance. Naměřené hodnoty se pohybují u všech hodnotitelů kolem nuly, ať už do plusu nebo do minusu, což je přesně to, co bychom očekávali.

Třetí a poslední dimenze, kterou autoři článku vyčlenili pro validaci se subjektivními hodnoceními byla *variety* (variace), v grafu propočítána čtyřmi indexy variety MTLD-MA (jiné označení pro MTLD-W), MTLD, HD-D, MATTR (s výsekem textu 50 a 250). Když porovnáme korelace těchto indexů a subjektivních hodnocení, zjistíme, že se pohybují od slabých po středně silné. Ze všech indexů *variety* a všech hodnotitelů a hodnotitelek je nejsilněji korelovaný index HD-D s hodnocením respondentky B. Korelace zde dosahuje r = 0,44. Nejslabší korelaci pak vidíme u metriky MTLD-MA a hodnotitele J, r = 0,05.



Graf 7: Pearsonova korelace subjektivních a objektivních hodnocení LD ve všech žánrech

4.6.1 Průměr hodnocení LD a metriky LD

Autoři článku k propočítání korelací používají průměr hodnocení respondentů. Když porovnáme průměrné lidské hodnocení s metrikami LD, kromě indexu *volume* (počtu typů), tak se korelace pohybuje mezi r = 0,26 až r= 0,4. V případě indexu *volume* a průměrného subjektivního hodnocení je r = 0. Z grafu můžeme pozorovat, že „zprůměrováním“ hodnocení se korelace zvýšila.

4.6.2 Jednotliví hodnotitelé a metriky LD

Zaměříme-li se v grafu č. 7 na konkrétní hodnotitele, tak zjistíme, že silnější korelace hodnocení a metrik nacházíme u hodnotitelek B a S (obzvlášť index *variety*). Obě hodnotitelky prošly přírodovědeckým vzděláním, naopak hodnotitelé E a K dosáhli pedagogického vzdělání, hodnotitel J dosáhl vzdělání v humanitně zaměřeném oboru. Je otázkou, zda jejich hodnocení neovlivňovala (ne)zkušenost s hodnocením školních prací.

4.6.3 Známka udělená vyučujícím a metriky LD

Z grafu č.7 můžeme vyčíst, že známka udělená vyučujícím koreluje s metrikami mnohem méně než holistická hodnocení, pravděpodobně to znamená, že známka neodráží koncept lexikální diverzity.

4.7 Korelace metrik LD mezi sebou

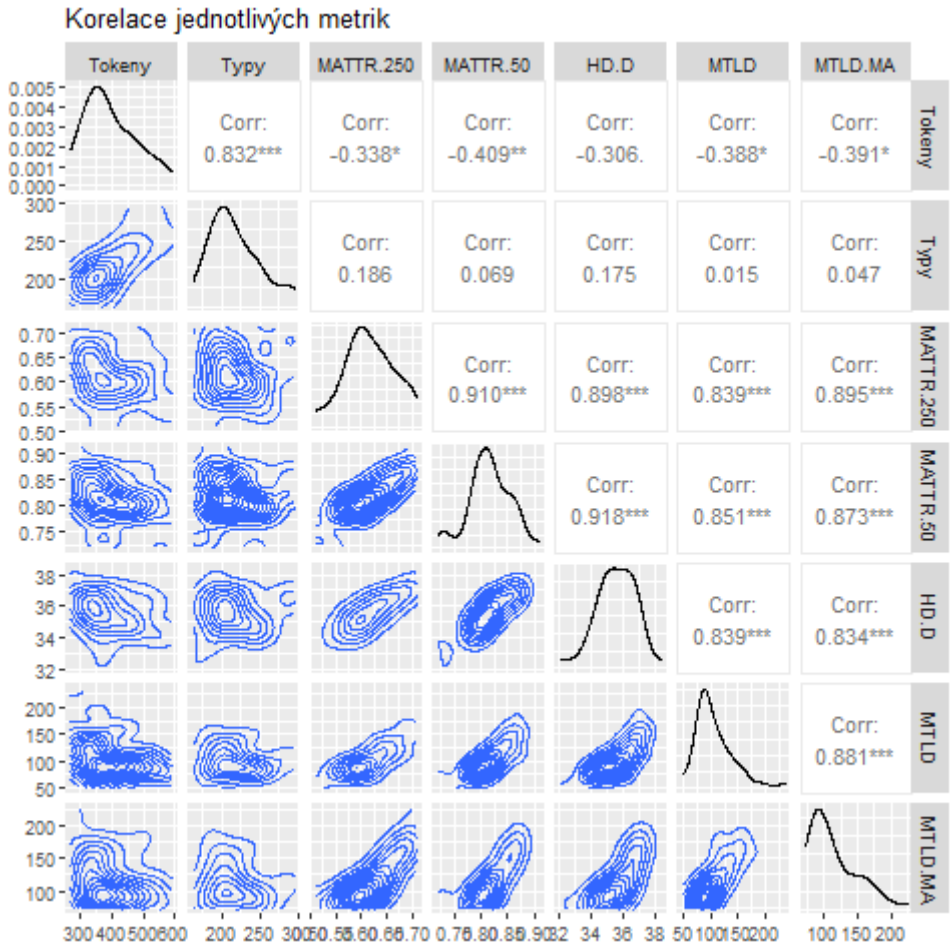
V grafu korelací jednotlivých metrik (graf č.8) vidíme, že indexy variety spolu korelují velmi silně. To odpovídá výsledkům naměřeným Kylem et al. Ti ve svém článku uvádějí, že „indexy variety byly silně korelované (r=0,837 až r=0,954), což naznačuje, že všechny měří velmi podobný aspekt lexikální diverzity (Kyle et al., 2021, str.165).

Nejsilněji byla korelované metriky HD-D a MATTR 50 (r=0,918). To je překvapivé, protože metriky MATTR 250 a MATTR 50 se liší pouze ve velikosti výseku textu, který analyzují. Domněnka, že právě tyto metriky budou korelovány nejsilněji se nepotvrdila. Stejně tak MTLD a MTLD-MA jsou stejné metriky, druhá zmíněná však text analyzuje s „*moving window*“ (viz kapitolu 2.4.2). Z toho vyplývá, že je u každé metriky (MATTR,

MTLD) při analýze textů nutné promyslet nastavení parametrů, rozdílné parametry generují znatelně rozdílné výsledky.

V čem se výsledky analýzy této práce a Kylea et al. rozcházejí jsou korelace *abundance* a indexů *variety*. Autoři ve svém článku naměřili střední až silnou korelaci (r=0,493 až 0,619) abundance (typů) a indexů variety (MATTR, HD-D, MTLD, MTLD-MA) (Kyle et al., 2021, str. 165). Hodnoty naměřené pro tuto práci se však pohybovaly v rozmezí r=0,015 až r= 0,186. Nebyly tedy vůbec korelované.

Pro dimenzi *volume* (tokeny) a indexů *variety* uvádí autoři slabou korelaci v rozmezí r=0,163 až r=0,295. Ani u těchto metrik se výsledky neschází. Tokeny a indexy variety textů ohodnocených objektivními metrikami pro tuto práci dosahují středně silných negativních korelací od r= - 0,306 do r= - 0,409.



Graf 8: Pearsonova korelace jednotlivých metrik mezi sebou

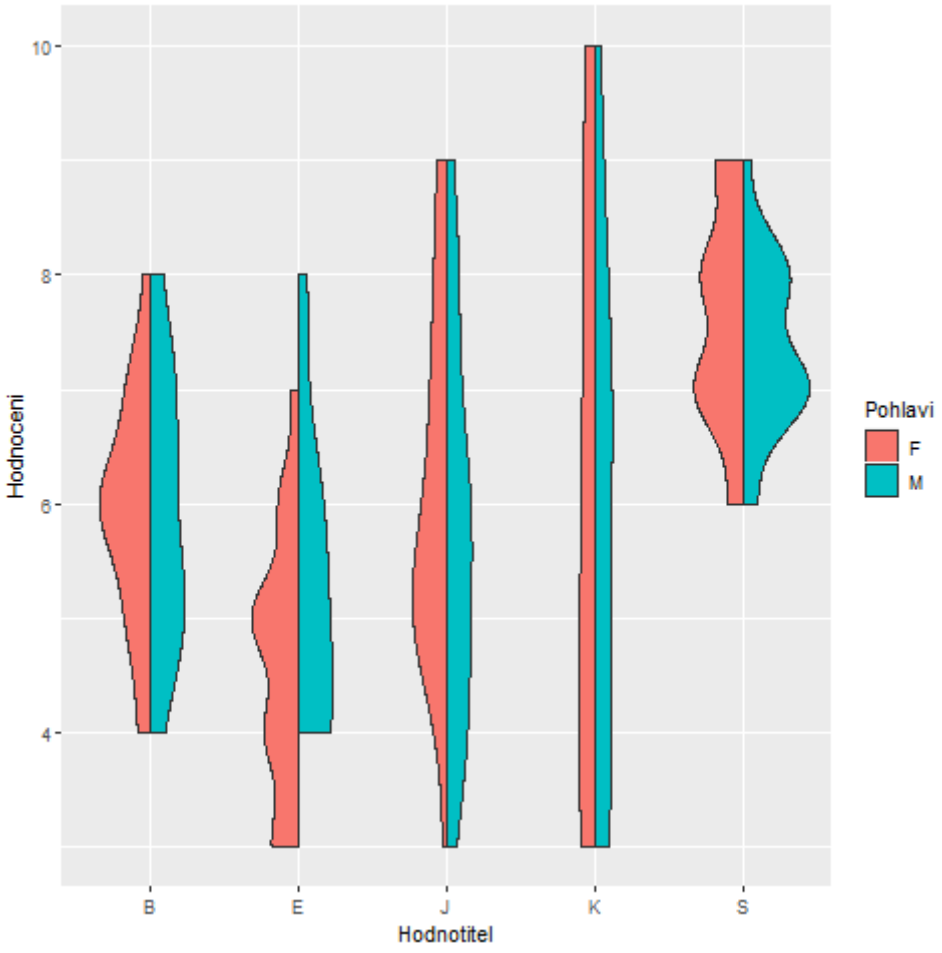
4.8 Hodnocení a pohlaví autorů textů

4.8.1 Subjektivní hodnocení

Z houslového grafu č.9 níže můžeme vyčíst, že hodnotitelé udíleli skóre rovnoměrně a nebyli zaujati ani proti ženám ani mužům. Jediná hodnotitelka E měla škálu hodnocení nastavenou vůči mužům a ženám rozdílně, avšak vzorek textů byl poměrně malý, takže nemůžeme vyvrátit, že to byla náhoda.

Další informací, kterou z grafu můžeme vyčíst je variabilita hodnocení jednotlivých respondentů. Celou škálu skóre (1-10) nevyužil ani jeden z respondent a respondentek. Největší variabilitu v hodnocení vidíme u respondenta K. Ten jako jediný udělil nejvyšší známku (10).

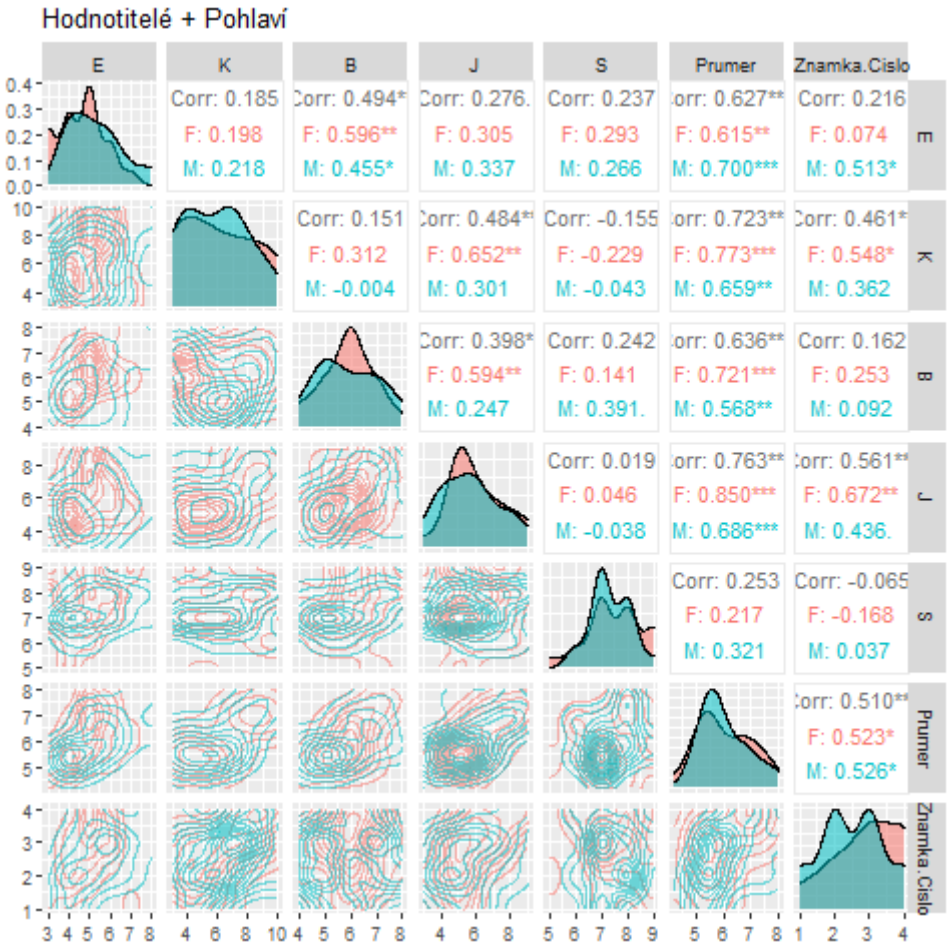
Naopak hodnotitelka S využívala jen velmi malý rozsah z nabídnuté škály, její hodnocení se pohybovala mezi 6 a 9.



Graf 9: Hodnotitelé a pohlaví autorů

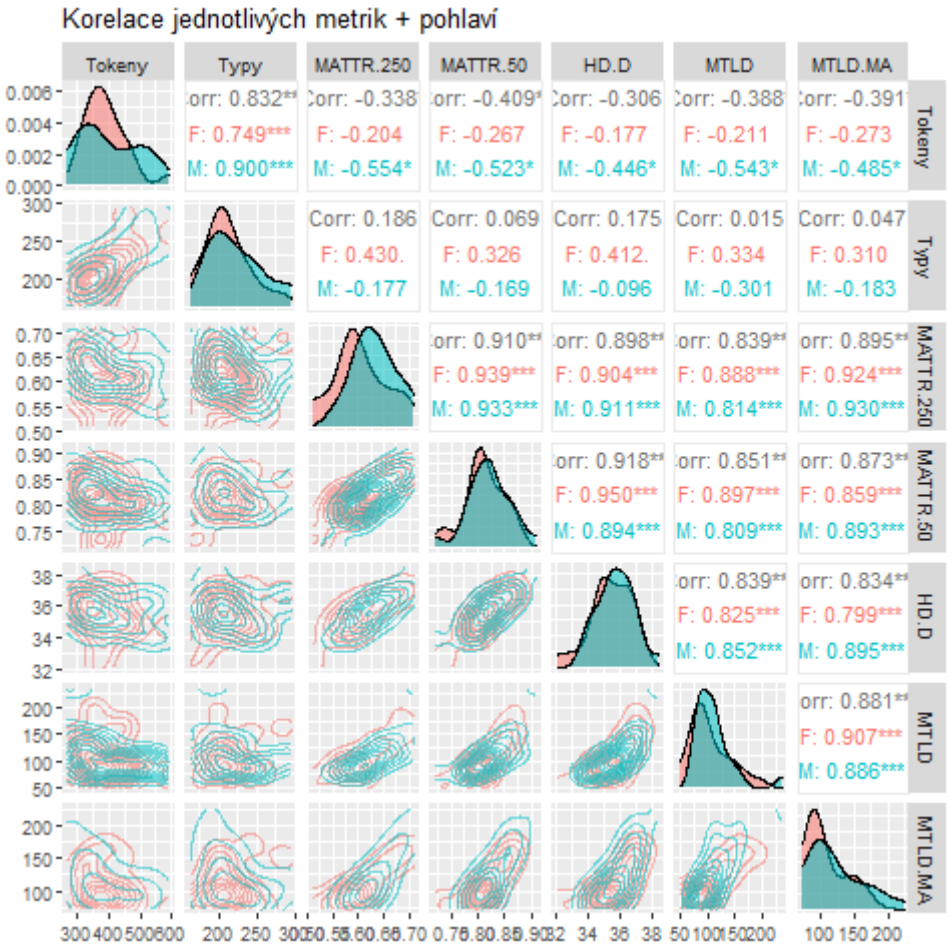
4.8.2 Subjektivní a objektivní hodnocení

Srovnejme dva grafy níže, první (graf č. 10) porovnává korelace subjektivních hodnocení a druhý korelace objektivních hodnocení v závislosti na pohlaví. Z grafů vyčteme, že respondenti byli v hodnocení textů podle pohlaví autorů mnohem vyrovnanější než metriky LD, respektive měli v korelacích vypočítaných pro ženy a pro muže menší variabilitu.



Graf 10: Korelace subjektivních hodnocení v závislosti na pohlaví

U grafu objektivního hodnocení níže (graf č.11), pozorujeme velkou variabilitu korelací mezi muži a ženami u dimenze abundance (typy). Korelace typů a metriky MATTR 250 se pohybuje mezi r= - 0,177 (muži) a r=0,430 (ženy).



Graf 11: Korelace objektivního hodnocení v závislosti na pohlaví

4.9 *Follow-up* interview

Z možných způsobů získání přesnějšího vhledu do procesu holistického hodnocení textů navrhuje Jarvis a kol. požádat hodnotitele o *follow-up interview* (Kyle et al., 2021, str. 168). Do rozhovoru o motivacích pro hodnocení se byli ochotni zapojit všechny respondentky a respondenti. *Follow-up* interview proběhlo do dvou dnů od finalizace hodnocení textů.

Všem respondentům byly položeny tři otevřené otázky:

1. *Čeho jste si všímali při udělování hodnocení?*
2. *Jak by podle vás měl vypadat text se silnou lexikální diverzitou?*
3. *Jak by podle vás měl vypadat text se slabou lexikální diverzitou?*

Přestože každá z respondentek a respondentů přistupovali k hodnocení specificky a při *follow-up* interview používali různé termíny, při bližším prozkoumání jejich rozborů se v mnohém shodovali. Tyto introspekce můžeme roztřídit do několika „subjektivních dimenzí" LD (tabulky níže). Některé z nich se shodují s dimenzemi LD zmiňovanými autory článku.

1. Idea

a.	+	téma, nosná myšlenka	Respondentka E
b.	+	myšlenka, pointa	Respondent J
c.	+	když se neopakovaly stejné motivy	Respondent K

Z odpovědí získaných v následném rozhovoru můžeme vypozorovat, že hodnotitelé odvozovali svoje hodnocení na několika úrovních. Jako zásadní faktor ovlivňující jejich hodnocení uváděli kategorii „ideje". Tento výsledek navazuje na závěr v článku Kylea et al., ve kterém popisují, že „lidské vnímání lexikální diverzity je ovlivněno celkovým počtem rozdílných idejí v textu" (str.167). Další kategorie, které respondenti zmiňovali byly „fantazie" a „projev imaginace".

Autoři článku odvozují tuto hypotézu z výsledků měření, nejsilněji byla totiž se subjektivními hodnoceními korelována dimenze abundance (počet typů). Introspekce u hodnotitelů v této bakalářské práci však neodpovídá naměřeným hodnotám, jejich subjektivní hodnocení bylo s dimenzí abundance korelováno velmi slabě.

2. Srozumitelnost

a.	-	když jsem tomu nerozuměla	Respondentka E
b.	-	když byl text kostrbatý, tak míň bodů	Respondentka B
c.	-	nutnost číst část textu víckrát, abych to pochopil – navazování vět nedávalo smysl	Respondent J

d.	+	srozumitelnost textu	Respondent K
e.	-	kostrbatý dojem způsobený použitím nevhodných slovních spojení	Respondent K
f.	-	nedokončené věty, rozpory v textu	Respondent K

Požadavek na srozumitelnost textu byl zastoupen téměř u všech respondentů. K této kategorii můžeme přistupovat ze dvou stran. Text může být nesrozumitelný kvůli nesprávně zvoleným slovům nebo slovním spojením, příp. rozporům v textu. Většina respondentů vytýkala textům jejich „kostrbatost“, respondentka E (2.a) a respondent K (2.d) však poukazovali na srozumitelnost textu ve smyslu, aby byl „tak akorát“.

Tendenci mluvčích k hledání „vyváženosti“ v jazyce popisoval již Zipf (1965). “Rozpory v textu“, „nelogické navazování vět“ apod., všechny tyto introspekce u jednotlivých hodnotitelů nám v této kategorii potvrzují, že se hodnotitelé zaměřovali i na vyšší celky než slova.

3. Množství informace

a.	-	když je text zdlouhavý, ale nic neřekne – nedokážu si nic představit	Respondentka B
b.	-	hodně „prázdných“, „měkkých“ slov – nemá to obsah, nic se nedovím	Respondent J
c.	+	nemusí to být dlouhý text, když je výstižný	Respondentka B
d.	+	výskyt „ostrých slov“ – když málo slov přesně vystihne popisované = vysoká information load	Respondent J

Výpovědi hodnotitelů a hodnotitelek roztřízených do kategorie „množství informace“ nám naznačují, že respondenti oceňovali, případně kritizovali nedostatek, „information load“. Ať už na úrovni slov; u respondenta J se vyskytovaly metafory „ostrá“, tj. slova s vysokým množstvím informace a „měkká“ slova, tj. slova s nízkým množstvím informace nebo na úrovni

celého textu, např. u respondentky B: „když je text zdlouhavý, ale nic neřekne“ (tak jeho LD hodnotím nižší známkou).

4. Neobyčejnost („*specialness“)*

a.	+	neobvyklá slova – „barevná“ – čím víc „barevných“ slov, tím lepší hodnocení	Respondentka B
b.	-	neutrální slova, která člověk úplně nesnáší, např. velký a malý, dobrý a špatný	Respondentka B
c.	+	výskyt „ostrých slov“ – když málo slov přesně vystihne popisované = vysoká information load	Respondent J
d.	+	zajímavá slovní spojení, přirovnání	Respondent J Respondentka S
e.	+	slova, která mi přišla bohatší jsem si zatrhávala; potřeštěné, citově zabarvené; mimo základní rámec, vystihující citlivost, náladovost – něco duševního, nějaká niternost	Respondentka S
f.	-	Když byly použity tři čtyři balíčky základních slov, kterými se dá mluvit o čemkoli	Respondentka S

Zdaleka nejvíce introspekcí se týkalo (a dalo vzniknout) kategorii „neobyčejnosti“. Je to jedna z dimenzí, kterou ve svém článku popisují i Kyle et al. Nazývají ji „specialness“ a popsali ji jako „přítomnost konkrétních slov vnímaných jako obohacující diverzitu“. Hodnotitelé taková slova popisovali jako „barevná“, „bohatá“, „ostrá“, „mimo základní rámec“ apod.

Respondentka B, která tato speciální slova popsala jako „barevná“ není synestetička, tento popis použila jako metaforu. Uvádí i příklady slov na opačném konci škály hodnocení LD, konkrétně zmiňuje slova „dobrý“ a „špatný“, „velký“ a „malý“. Respondentka S hodnotila texty nízkou známkou, pokud byla použita slova „kterými se dá mluvit o čemkoli“.

5. Opakování („*dispersion*“)

a.	-	- když se slova opakují – když se opakují výplňová slova, např. „jakoby“ – (ještě horší), když si člověk všimne, že se ta stejná slova kumulují na dalším řádku	Respondentka B
b.	-	Když se tam pořád opakovalo sloveso být, horší hodnocení	Respondent J
c.	-	Opakování ukazovacího zájmena „to“	Respondent J
d.	-	Opakování slov	Respondent K
e.	-	Nedostatek slovní zásoby nahrazováno tvary slov „být“ a „mít“	Respondent K

Tato kategorie odpovídá dimenzi „*dispersion*“, kterou popsali Kyle et. al. jako „velikost intervalů mezi opakováním jednotlivých slov“ (str. 156). Hodnotitelka B konkrétně popsala, že udělovala nižší známku, když byl interval opakování příliš krátký (viz 5.a). Jako další důvodem k nízkému hodnocení uváděli respondenti opakování konkrétních slov. Zmiňovali zejména opakování sloves „být“, „mít“ a zájmena „to“. Podle respondentů opakování těchto slov poukazovalo na nízkou slovní zásobu autorů textů.

6. Ovlivnění žánrem

a.	Matoucí, že bylo zařazeno více žánrů – chápu, že v určitých žánrech se neutrální slova opakují – u dopisů jsem udělovala nižší hodnocení, věcný popis, je to v pořádku být věcný, když si jenom sdělujeme nějakou informaci	Respondentka B
b.	popisy domova věcné; naopak vzhled, osobnost, popis milované osoby – poetičtější, svádělo to ohodnotit výš	Respondentka S

Někteří z respondentů tématizovali žánr textů. Všímali si, jakým způsobem ovlivňoval jejich hodnocení. Respondentka B uvádí, že texty v žánru „dopis“ hodnotila nižším skóre. Respondentka S při analýze svých hodnocení uvedla, že nižší hodnocení udělovala žánrům, ve kterých se vyskytovaly „věcné“ popisy, např. popis domova.

Když porovnáme introspekce těchto dvou respondentek s výsledky, tak zjistíme, že naměřené hodnocení se s jejich analýzou neshoduje. Respondentka B ani respondentka S nehodnotily nějaký žánr konzistentně nižší známkou, viz kapitolu 4.5.5.

7. Chyby v textu

a.	-	když jsem viděla, že je v textu fakt pravopisná chyba, třeba prvních pět řádků, podvědomě mě to ovlivnilo	Respondentka S

Texty získané z korpusu neprošly korekcí chyb. Při zadávání textů respondentům bylo zmíněno, že chyby v textu nejsou kritériem pro posuzování LD. Z *follow-up* interview však vyplynulo, že tento faktor respondenty, ať už vědomě či podvědomě, ovlivňoval. Respondentka S uvádí, že když na začátku textu identifikovala gramatickou chybu, její hodnocení to ovlivnilo (negativním směrem). Hodnotitel K dodal spolu s hodnocením LD také hodnocení pravopisu.

Scott Jarvis ve svých předchozích experimentech (Jarvis, 2017; Jarvis & Daller, 2013a) týkajících se validace LD chyby v textech opravuje. Kyle et al. ve svém článku neuvádějí, zda byly texty zadané k hodnocení opraveny.

4.9.1 Shrnutí

Hodnotiteli sdělované myšlenky shromážděné v kategoriích výše odrážejí pohled autorů článku na lexikální diverzitu, ti tvrdí, že je k LD nutné přistupovat jako k mnohodimenzionálnímu fenoménu. Kategorie 4. a 5. („neobyčejnost“ a „opakování“) odpovídají dvěma dimenzím LD navrženým autory článku („*specialness“ a „dispersion*“). Další kategorií, která vzešla z *follow up* interview a odpovídá výzkumům autorů je kategorie „ideje“. Souvislost vnímání LD a představování nových myšlenek („*generating new ideas*“) zmiňují autoři v závěru svého článku (Kyle et al., 2021, str. 168).

Pozorování získané z *follow-up* interview odpovídají tezi, že LD je více než jen rozsah slovní zásoby a týká se i způsobu jejího využití (Duran, 2004).

Zajímavou metodou použitou při hodnocení byl způsob, jakým hodnotitelé a hodnotitelky nastavovali svoji škálu hodnocení. Hodnotitelka S zmiňovala, že si při hodnocení vybavovala texty svých oblíbených literárních autorů, jejichž dílům přisuzovala vysokou

lexikální diverzitu. Oproti nim pak porovnávala texty k hodnocení. Metaforu použila i při odpovědi na otázku „jak by podle vás měl vypadat text se slabou lexikální diverzitou?“, její odpověď zněla „Ostrava pičo“, čímž metaforicky vyjádřila, že svoje měřítko vztahuje k určitému modelovému autorovi nikoli nějaké abstraktní entitě.

Po porovnání odpovědí respondentů ve *follow up* interview a jejich subjektivních hodnocení jsme došli k závěru, že introspekce ne vždy odpovídají naměřeným výsledkům. V prvním případě se neshoda projevila u kategorie „ideje“. Tato kategorie se váže k dimenzi abundance (počtu typů), předpoklad tedy byl, že hodnotitelé budou udělovat textům s vyšším počtem typů vyšší hodnocení. Přestože několik hodnotitelů uvedlo tuto kategorii jako důležitý faktor při udělování hodnocení, dimenze abundance s jejich hodnoceními vůbec nekorelovala (viz kapitolu 4.6).

Další nesoulad introspekce a výsledků můžeme pozorovat u hodnocení jednotlivých žánrů. Hodnotitelka B uvedla, že texty v žánrové kategorii „dopis“ hodnotila nízkým skóre. Hodnotitelka S uvedla, že žánry s „věcným popisem“ (konkrétně popis místa) hodnotila nižším skóre. Při porovnání výsledků mezi žánry však nebyl žádný z žánrů hodnocen systematicky nižším skóre.

Dva hodnotitelé (J a K) se v rozboru svého hodnocení LD shodují. Nižší skóre podle jejich výpovědi udělovali, když v textu bylo příliš opakování (5.b,c,d,e), shodli se také na kategorii „srozumitelnosti“ (2.c,d,e,f). Tito dva hodnotitelé dosahují silnější korelace i při porovnání jimi udělených hodnocení LD.

5. Závěr

Za pomoci dostupných prostředků jsme zopakovali experiment Kylea et al. Výzkumnou otázku, tj. „Jaký je vztah mezi lidským hodnocením LD a objektivním měřením lexikální *volume, abundance* a *variety*?“, kterou si pokládají autoři článku a která se tedy prolíná i touto prací je z analýzy v této práci možné zodpovědět následovně.

Hodnocení udělená respondentkami a respondenty pro tuto analýzu nijak významně nekorelují s hodnotami vypočítanými objektivními metrikami.

Důvodů proč spolu hodnoty nekorelují může být více, jako první možný bych však zmínila nízkou shodu hodnocení mezi respondenty. Ti, na rozdíl od respondentů v experimentu, který opakujeme, neprošli žádným výcvikem ani neměli možnost kalibrovat svá hodnocení podle předloženého příkladového ohodnocení textu na lexikální diverzitu. Také neměli možnost svá, od ostatních hodnotitelů odlišná, hodnocení upravit. Tím jsme ovšem narozdíl od od autorů článku měli příležitost pozorovat přirozené subjektivní hodnocení LD bez umělého ovlivnění objektivními metrikami.

Kyle et al. ve svém článku propočítávali korelace z průměru subjektivních hodnocení, v této práci máme k dispozici korelace zprůměrovaného hodnocení a metrik i korelace jednotlivých hodnotitelů a metrik. I když se však podíváme na korelace hodnotitelů odděleně, najdeme pouze slabé korelace s objektivním měřením *volume, abundance a variety*.

Analyzovali jsme i jiné aspekty, které byly hodnocení inherentní. Zkoumali jsme vztah subjektivního i objektivního hodnocení a žánru hodnoceného textu, dále jsme zkoumali, zda se subjektivní i objektivní hodnocení měnilo vzhledem k tomu, jestli texty k ohodnocení napsaly ženy nebo muži. Vyhodnotili jsem i korelace metrik lexikální diverzity mezi sebou.

Kapitolou, která nereplikuje metodologii Kylea et al., avšak vznikla z jejich doporučení pro další badatele je *follow-up interview*. Domnívám se, že tato kapitola přinesla velmi zajímavé výsledky. V první řadě potvrdila bádání Kylea et al. a hlavně také Scotta Jarvise ohledně lexikální diverzity jako mnohodimenzionálního fenoménu. Bez toho, aniž by respondenti byli informováni o dimenzích lexikální diverzity, a vlastně hlouběji o konceptu

lexikální diverzity obecně, tak ve svých odpovědích ve *follow-up* interview mezi sebou došli k podobným závěrům o podstatě lexikální diverzity. Také specifikovali několik dimenzí LD, které odpovídají dimenzím identifikovaným autory článku, jejichž experiment tato práce opakuje.

Seznam použité literatury

A. Crossley, S., & S. McNamara, D. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7*(3 (February 2016)), 351–370. https://doi.org/10.17239/jowr-2016.07.3.02

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *The Psychological Record*, *2*(16), 379–386. https://doi.org/10.1007/BF03393224

Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology*, *16*(3), 287–300. https://doi.org/10.1002/acp.788

Cronbach Alpha—Free Statistics and Forecasting Software (Calculators) v.1.2.1. (b.r.). Získáno 28. prosinec 2021, z https://www.wessa.net/rwasp_cronbach.wasp#output

Duran, P. (2004). Developmental Trends in Lexical Diversity. *Applied Linguistics*, *25*(2), 220–242. https://doi.org/10.1093/applin/25.2.220

Harris Wright, H., Silverman, S., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology*, *17*(5), 443–452. https://doi.org/10.1080/02687030344000166

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity: Lexical Diversity. *Language Learning*, *63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Daller, H. (Ed.). (2013a). *Vocabulary knowledge: Human ratings and automated measures*. John Benjamins Publishing Company.

Jarvis, S., & Daller, H. (Ed.). (2013b). *Vocabulary knowledge: Human ratings and automated measures*. John Benjamins Publishing Company.

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

KoRpus text analysis. (b.r.). Získáno 19. prosinec 2021, z https://ripley.psycho.hhu.de/R/koRpus/

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Malvern, D. (Ed.). (2008). *Lexical diversity and language development: Quantification and assessment* (1. publ. in paperback). Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Michalke, M. (2020). *KoRpus: Text Analysis with Emphasis on POS Tagging, Readability and Lexical Diversity (Version 0.13-4). Availible from https://reaktanz.de/?c=hacking&s=koRpus*.

ReCal for Ordinal, Interval, and Ratio Data (OIR) – Deen Freelon, Ph.D. (b.r.). Získáno 28. prosinec 2021, z http://dfreelon.org/utils/recalfront/recal-oir/

Skript 2015. (b.r.). Získáno 29. listopad 2021, z https://lindat.mff.cuni.cz/services/teitok/skript2015/index.php?action=cqp&cql=%3Ctext%3E%20[]

Wessa, P. (2021). *Free Statistics Software, Office for Research Development and Education, version 1.2.1, URL https://www.wessa.net/*.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zipf, G. K. (1965). *The psycho-biology of language; an introduction to dynamic philology.* M.I.T. Press.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. System, 40(4), 554–564. doi:10.1016/j.system.2012.10.012

Kristopher Kyle, Scott A. Crossley & Scott Jarvis (2021) Assessing the Validity of Lexical Diversity Indices Using Direct Judgements, Language Assessment Quarterly, 18:2, 154-170, DOI: 10.1080/15434303.2020.1844205

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392. doi:10.3758/BRM.42.2.381

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32(5), 323–352. doi:10.1023/A:1001749303137