UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

METÓDY SPRACOVANIA PRIRODZENÉHO JAZYKA

DIPLOMOVÁ PRÁCA

2023 **Bc. Marko Penzeš**

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

METÓDY SPRACOVANIA PRIRODZENÉHO JAZYKA

DIPLOMOVÁ PRÁCA

Študijný odbor:	9.2.9 Aplikovaná informatika
Študijný program:	Aplikovaná informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	prof. RNDr. Michal Munk, PhD.

Nitra 2023 Bc. Marko Penzeš

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied a informatiky

223740

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Marko Penzeš
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: Diplomová práca**Jazyk záverečnej práce:** slovenský
Sekundárny jazyk: anglický

Názov: Metódy spracovania prirodzeného jazyka

Anotácia: Práca spadá do oblasti spracovania prirodzeného jazyka, konkrétne sa zameriava na predspracovanie (tokenizácia a lematizácia) a na základnú deskripciu textu pomocou automatických NLP nástrojov a mier.

Cieľom práce je pomocou NLP metód porovnať kvalitu strojového a humánneho prekladu vzhľadom na lexikálnu rôznorodosť a hustotu. V teoretickej časti práce je žiadúce popísať použité miery, vrátane potrebných knižníc v Pythone. V praktickej časti práce je žiadúce implementovať vybranémiery na základnú deskripciu textu v Pythone a demonštrovať ich použitiena konkrétnych kolekciách dokumentov.

Charakter práce:
Výskumný – stanovenie výskumného problému, metodika výskumu, výsledky výskumu (štatistická interpretácia), interpretácia výsledkov výskumu (vecná interpretácia).

Predmetové prerekvizity:
Web mining (1., Mgr.), Počítačová analýza dát (2., Bc.), Umelá inteligencia (2.,Bc.), Programovanie a údajové štruktúry (1., Bc.)

Najdôležitejšie kompetentnosti získané spracovaním témy:
analyzovať veľké dáta (big data)
postupy analýzy dát z pohľadu ich integrity a
kvalityreferovať o výsledkoch analýzy
vykonať analýzu údajov
metodika vedeckého výskumu
zhodnotenie a navrhnutie algoritmu pre riešenie konkrétnej zadanej úlohy

Školiteľ: prof. RNDr. Michal Munk, PhD.
Oponent: Mgr. Nataliia Časnochová Zozuk
Katedra: KI - Katedra informatiky

Dátum zadania: 08.11.2021
Dátum schválenia: 09.01.2023

RNDr. Ján Skalka, PhD., v. r.
vedúci/a katedry

POĎAKOVANIE

Rád by som využil túto príležitosť a vyjadril vďaku svojmu školiteľovi a rodinným príslušníkom, za ich neoceniteľnú podporu a vedenie počas prípravy tejto práce. Ich výstižné pripomienky a odborná pomoc boli kľúčové pri formovaní tejto práce a som im skutočne vďačný za ich čas a úsilie.

ABSTRAKT

PENZEŠ, Marko: Metódy spracovania prirodzeného jazyka. [Diplomová práca]. Univerzita Konštantína Filozofa v Nitre. Fakulta prírodných vied a informatiky. Školiteľ: prof. RNDr. Michal Munk, PhD. Stupeň odbornej kvalifikácie: Magister odboru Aplikovaná informatika. Nitra: FPVaI, 2023. s. 69.

Cieľom našej práce je pomocou metód spracovania prirodzeného jazyka porovnať kvalitu strojového a humánneho prekladu vzhľadom na lexikálnu rôznorodosť a hustotu a tiež hlbšie porozumenie, predstavenie a analýza problematiky v tejto oblasti. Pri dosahovaní cieľa používame rôzne knižnice jazyka Python, ktoré slúžia na spracovanie prirodzeného jazyka. Výsledky ukazujú, že zatiaľ čo v lexikálnej hustote nie sú medzi ľudskými a strojovými prekladmi významné rozdiely, v lexikálnej diverzite a expanznom pomere významné rozdiely sú. Konkrétne Simpsonov index a miera HD-D MTLD poukazujú na väčšiu lexikálnu rozmanitosť v ľudských prekladoch a pomer rozšírenia poukazuje na väčšiu variabilitu ľudských prekladov. Výskum poukazuje na potrebu ďalšieho skúmania použitých metód. Postupom pri dosahovaní stanoveného cieľa je uskutočnenie analýzy výsledkov, ich kritické zhodnotenie a interpretácia. Práca odhaľuje nedostatky voľne dostupných knižníc jazyka Python na spracovanie prirodzeného jazyka pri spracovaní slovenského jazyka.

Kľúčové slová: Spracovanie prirodzeného jazyka. Lexikálna rôznorodosť. Lexikálna hustota.

ABSTRACT

PENZEŠ, Marko: Natural language processing methods. [Final Thesis]. Constantine the Philosopher University in Nitra. Faculty of Natural Sciences and Informatics. Supervisor prof. RNDr. Michal Munk, PhD. Degree of Qualification: Master of Applied Informatics. Nitra: FNSaI, 2023. p. 69.

The goal of our work is to use natural language processing methods to compare the quality of machine and human translation with regard to lexical diversity and density, as well as a deeper understanding, presentation and analysis of issues in this area. To achieve this goal, we use various Python libraries that are used for natural language processing. The results show that while there are no significant differences in lexical density between human and machine translations, there are significant differences in lexical diversity and expansion ratio. Specifically, Simpson's index and the HD-D MTLD measure indicate greater lexical diversity in human translations, and the expansion ratio indicates greater variability in human translations. The research points to the need for further investigation of the methods used. The procedure for achieving the set goal is the analysis of the results, their critical evaluation and interpretation. The work reveals the shortcomings of freely available Python language libraries for natural language processing when processing the Slovak language.

Keywords: Natural language processing. Lexical diversity. Lexical density.

OBSAH

ÚVOD

Spracovanie prirodzeného jazyka je v informatike náročná téma vzhľadom na rozmanitosť a rôznorodosť ľudského jazyka. Je to preto, že počítače majú problém interpretovať informácie sprostredkované prostredníctvom prirodzených jazykov. Existujú rôzne dôvody, prečo NLP zlyháva (Garbade, 2018).

Garbade (2018) porovnáva problém, ktorý môže nastať pri NLP strojom k výmene informácií medzi ľuďmi. V rámci komunikácie medzi ľuďmi môže nastať situácia, kedy dochádza k nepochopeniu, prípadne k nedorozumeniam medzi komunikujúcimi. Uvedená situácia môže nastať v prípade, ak komunikujúce osoby využívajú pri vzájomnej interakcií odlišný kód, teda iný prirodzený jazykový systém. Ako uvádza Garbade (2018) je možné, že podobná situácia nastane aj pri spracovaní prirodzeného jazyka strojom.

Môžeme konštatovať, že programy na NLP v súčasnosti fungujú, akoby boli "ľudské." Za väčšiny okolností obe strany chápu kontext a podstatu obsahu správy, čo umožňuje jej interpretáciu v kontexte komunikácie medzi človekom a strojom. Napriek tomu sa môže stať, že si jedna zo strán nesprávne vysvetlí nejaký pojem a príjemca z nejakého dôvodu nepochopí kontext komunikácie. K zlyhaniu zo strany počítača môže dôjsť aj pri textovej komunikácii, ktorá nastáva vtedy, keď nástroje na NLP nie sú dostatočne inteligentné (Ishaq, 2019).

Výsledkom ľudského poznania a inteligencie je prirodzený jazyk. Prirodzený jazyk obsahuje mnoho nejednoznačných a neurčitých fráz a výrokov, čo vedie k chybám v základných poznávacích myšlienkach. Napríklad frázy ako "vysoký," "krátky" alebo "horúci" sa pre svoju subjektívnosť ťažko prekladajú, najmä pre počítače. V dôsledku toho sa porovnateľné frázy musia spresniť, aby systémy nemali problémy s ich pochopením. Takéto výrazy sa v ľudskej reči vyskytujú často, preto by sa malo vyvíjať úsilie na ich zahrnutie do NLP systémov (Friedenberd, Silverman, 2006).

Garbade (2018) sa zaoberá podobným problémom v oblasti strojového NLP. Podľa autora sa v ľudskej reči môžu vyskytnúť sarkastické poznámky pri oznamovaní informácií, ktoré môže mať počítač následne problém pochopiť. Podľa Garbadeho (2018), spracovaniu prirodzeného jazyka počítačmi bráni aj nepresnosť a nejednoznačnosť prirodzeného jazyka. Dôkladné pochopenie ľudského jazyka si vyžaduje nielen porozumenie slov, ale aj prepojenie mnohých pojmov tak, aby správa dávala zmysel.

Vzhľadom k vyššie uvedenému je cieľom diplomovej práce je porovnať kvalitu dvoch prekladov, a to humánneho a strojového prostredníctvom metód a vybraných automatických metrík NLP.

1 ANALÝZA SÚČASNÉHO STAVU

1.1 Umelá inteligencia

Umelá inteligencia (AI) je interdisciplinárny odbor, ktorý spája informatiku, matematiku, štatistiku, kognitívnu psychológiu a ďalšie. Zahŕňa vytváranie inteligentných strojov, ktoré sa dokážu učiť z údajov, prijímať rozhodnutia a vykonávať úlohy, ktoré si zvyčajne vyžadujú ľudskú inteligenciu. AI možno rozdeliť na dva hlavné typy: úzku alebo slabú AI a všeobecnú alebo silnú AI (Silver, 2011).

AI je rýchlo sa rozvíjajúca oblasť, ktorej cieľom je vytvoriť inteligentné systémy schopné vykonávať úlohy, ktoré si bežne vyžadujú ľudskú inteligenciu, ako je učenie, riešenie problémov a rozhodovanie. V posledných rokoch zaznamenala umelá inteligencia prudký nárast záujmu a investícií vďaka pokroku v oblasti strojového učenia, spracovania prirodzeného jazyka a ďalších podoblastí. Potenciál využitia umelej inteligencie je obrovský a zahŕňa odvetvia od zdravotníctva a finančníctva až po dopravu a zábavu. Rýchle tempo vývoja však vyvolalo aj otázky týkajúce sa etických, právnych a spoločenských dôsledkov umelej inteligencie, najmä pokiaľ ide o otázky, ako sú súkromie, zaujatosť a zamestnanosť (Gepperth, Hammer 2016).

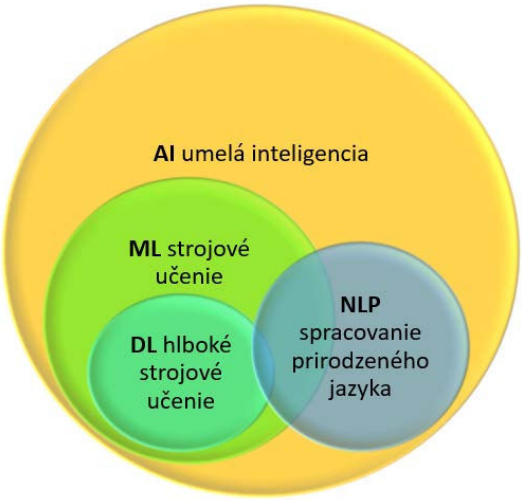
Všeobecne definovaná umelá inteligencia zahŕňa akúkoľvek techniku, ktorá umožňuje počítačom napodobňovať ľudské správanie a reprodukovať alebo prekonávať ľudské rozhodovanie pri riešení zložitých úloh samostatne alebo s minimálnym zásahom človeka (Janiesch a kol., 2021). Ako taká sa zaoberá rôznymi ústrednými problémami, vrátane reprezentácie znalostí, uvažovania, učenia, plánovania, vnímania a komunikácie a vzťahuje sa na rôzne nástroje a metódy (napr. prípadové uvažovanie, systémy založené na pravidlách, genetické algoritmy, fuzzy modely, multiagentové systémy) (Chen a kol., 2008).

Úzka alebo slabá umelá inteligencia je určená na vykonávanie špecifických úloh, ako je napríklad rozpoznávanie obrazu, rozpoznávanie reči alebo NLP. Tieto systémy sú navrhnuté na vykonávanie jednej úlohy a nie sú schopné všeobecnej inteligencie (Silver, 2011).

Na druhej strane, všeobecná alebo silná umelá inteligencia je navrhnutá na vykonávanie akejkoľvek intelektuálnej úlohy, ktorú môže vykonávať človek. Tento typ umelej inteligencie zatiaľ neexistuje, ale je konečným cieľom tejto oblasti (Janiesch a kol., 2021).

AI má široké spektrum aplikácií v rôznych oblastiach, vrátane zdravotníctva, financií, dopravy, vzdelávania a ďalších. Medzi obľúbené aplikácie umelej inteligencie, podľa autora Millera (2019), patria napr.:

- rozpoznávanie obrazu a reči: AI sa používa v systémoch na rozpoznávanie obrazu a reči, ako sú napríklad Siri a Alexa,
- spracovanie prirodzeného jazyka: AI sa používa v systémoch na NLP, ako sú chatboty a virtuálni asistenti,
- robotika: umelá inteligencia sa používa v robotike, aby stroje mohli vykonávať úlohy, ktoré si vyžadujú zručnosť a rozhodovanie podobné ľudskému rozhodovaniu,
- zdravotníctvo: umelá inteligencia sa používa v zdravotníctve na diagnostiku chorôb, vypracovanie liečebných plánov a analýzu lekárskych snímok,
- financie: umelá inteligencia sa používa vo financiách na odhaľovanie podvodov, predpovedanie trhových trendov a automatizáciu obchodovania.



Obrázok 1 Vyjadrenie vzťahu medzi umelou inteligenciou, strojovým učením, hlbokým učením a spracovaním prirodzeného jazyka [1]

Na obrázku 1 sú graficky znázornené vzťahy medzi umelou inteligenciou, strojovým učením, hlbokým učením a spracovaním prirodzeného jazyka s cieľom uľahčiť ich pochopenie ilustráciou.

[1] Zdroj Obrázok 1: https://umelainteligencia.sk/ako-je-mozne-ze-nam-siri-rozumie/

1.2 Strojové učenie

Podľa Jabeen a kol. (2018) je strojové učenie odvetvie umelej inteligencie, ktoré sa zaoberá technikami a algoritmami, ktoré umožňujú programom vhodne sa prispôsobiť rôznym vstupným hodnotám bez toho, aby na to museli byť explicitne naprogramované. Namiesto toho tak robia na základe poznatkov, ktoré v priebehu času nazbierali (Cibula, 2017). Jednou z najpopulárnejších metód umelej inteligencie na spracovanie veľkých objemov údajov je strojové učenie. Ide o algoritmus, ktorý má schopnosť samočinne sa korigovať a časom sa zlepšuje, keď sa učí zo skúseností alebo zapracováva nové údaje (Investopedia, 2020).

1.3 Hlboké učenie

S nedávnym pokrokom v oblasti digitálnych technológií sa zväčšila veľkosť súborov, analýza komplexných, vysokorozmerných a šumom kontaminovaných údajov je však veľkou výzvou a je nevyhnutné vyvinúť nové algoritmy, ktoré dokážu zhrnúť, klasifikovať, extrahovať dôležité informácie a previesť ich do zrozumiteľnej podoby (Wang, Wang, 2018; Pourpanah a kol., 2019).

Hlboké učenie je podmnožina strojového učenia, ktorá zahŕňa trénovanie umelých neurónových sietí s cieľom dosiahnuť vysokú úroveň presnosti v úlohách, ako je rozpoznávanie obrazu, rozpoznávanie reči a NLP. Na rozdiel od tradičných algoritmov strojového učenia, ktoré sú založené na explicitných inštrukciách, algoritmy hlbokého učenia sa učia zo samotných údajov (Young a kol., 2018).

Hlavnými výhodami hlbokého učenia sú schopnosť učiť sa z veľkých súborov údajov, schopnosť zvládať zložité úlohy a schopnosť prispôsobiť sa novým údajom. S hlbokým učením sú však spojené aj výzvy vrátane potreby veľkého množstva označených údajov, nadmerného prispôsobovania a ťažkostí pri interpretácii modelov. Celkovo je hlboké učenie výkonná a rýchlo sa rozvíjajúca oblasť, ktorá mení spôsob, akým uvažujeme o umelej inteligencii (Young a kol., 2018).

1.4 PRIRODZENÝ JAZYK

Nakonečný (1998) opisuje jazyk ako súbor semiotických ukazovateľov, ktoré ľudská civilizácia používa ako nástroj myslenia, aj ako prostriedok jazykovej komunikácie.

Prirodzený jazyk sa vzťahuje na ľudské jazyky, ako je slovenčina. Umelý jazyk, strojový jazyk a jazyk formálnej logiky sú protikladom prirodzeného jazyka (Nordquist, 2020). Čermák (2007) ponúka definíciu prirodzeného jazyka, a to ako „*možnosť (opakovateľnej) komunikácie medzi aspoň dvoma partnermi, založenej na systéme, ktorý je komplexný, dynamický a ktorý umožňuje svoje znaky podľa daných pravidiel kombinovať"* (Čermák, 2007, s. 14).

Podľa Millward a Hayes (2011) sú základné koncepty prirodzeného jazyka:

- *systematickosť*, čo znamená, že sú riadené viacerými vzájomne prepojenými systémami vrátane fonológie, grafiky, morfológie, syntaxe, lexiky a sémantiky,
- *normy*, napríklad priradením konkrétneho pojmu k určitej veci alebo pojmu, a sú konvenčné a arbitrérne,
- *redundantnosť*, čo znamená, že informácie vo vete sú vyjadrené viac ako jedným spôsobom,
- *dynamickosť*.

Ako uvádza Rakesh (2018) vo svojej publikácií, známa esej Alana Turinga *"Výpočtová technika a inteligencia"* bola prvýkrát publikovaná v roku 1950. Ako meradlo intelektu Alan Turing navrhol Turingov test, ktorý sa v súčasnosti široko používa. Turingov test posudzuje, či je počítačový program dostatočne inteligentný, a to do tej miery, že posudzovateľ nie je schopný dôsledne zistiť, či ide o program, alebo skutočnú osobu len na základe obsahu konverzácie.

Podľa Hancoxa (2010), začal revolúciu v roku 1957 Noam Chomsky, a to vývojom svojho systému syntaktických štruktúr, založeného na pravidlách. Jeden z efektívnejších systémov NLP, SHRDLU, bol vytvorený v 60. rokoch 20. storočia, ale s obmedzenými slovníkmi (Hutchins, 2005).

Ďalšou osobnosťou, ktorej meno je spojené s históriou spracovania prirodzeného jazyka, je Roger Schank, ktorý v roku 1969 navrhol konceptuálnu teóriu závislosti pre porozumenie prirodzenému jazyku (Schank, 1969).

Väčšina NLP systémov sa až do 80. rokov 20. storočia spoliehala na zložité súbory, ktoré boli vytvorené pomocou ručne písaných pravidiel. Vývoj algoritmov strojového učenia na spracovanie jazyka od konca 80. rokov však viedol k revolúcii v oblasti NLP. Pokrok v oblasti NLP možno pripísať viacerým faktorom. Vyplynulo to z dôsledného nárastu výpočtového výkonu, ako aj z postupného poklesu dominancie lingvistických teórií od čias Chomského (Hancox, 2010). Proces spracovania prirodzeného jazyka a metódy používané v procese spracovania sú uvedené v nasledujúcich kapitolách.

Garbade (2018) a Nordquist (2020) nám poskytujú podobné vysvetlenia NLP. Zhodujú sa v tom, že NLP, je podoblasťou umelej inteligencie, ktorá sa zaoberá interakciou počítačov a ľudí pomocou prirodzeného jazyka.

Ako vysvetľuje Garbade (2018), hlavným cieľom spracovania prirodzeného jazyka je čítať, dekódovať a pokúsiť sa interpretovať ľudský jazyk spôsobom, ktorý môže počítač spracovať.

Následné NLP závisí, podľa autora Garbadeho (2018), od rôznych interakcií medzi človekom a strojom. Medzi tieto interakcie patrí napr.:

- človek komunikuje so strojom pomocou ľudskej reči,
- zvuk je zachytávaný zariadením,
- konverzia zvuku na text,
- spracovaním údajov v texte,
- dáta sa menia na zvuk,
- zariadenie interaguje s človekom prostredníctvom zvuku.

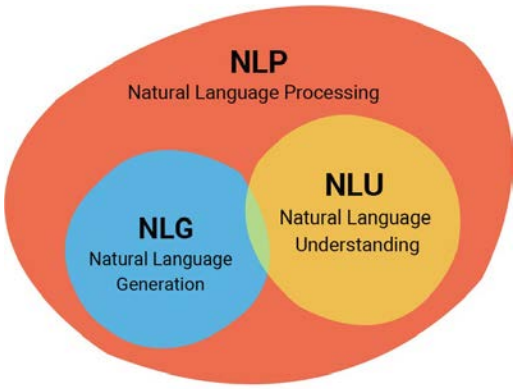
Na transformáciu neštruktúrovaného jazykového vstupu do formátu zrozumiteľného pre počítače sa pri NLP používajú algoritmy na rozpoznávanie a extrahovanie pravidiel prirodzeného jazyka. Počítač využíva algoritmy na extrakciu významu spojeného s každou vetou a na získanie základných údajov z nich po zadaní textu (Garbade, 2018).

1.4.1 Proces spracovania prirodzeného jazyka

V oblasti umelej inteligencie, význam techník spracovania prirodzeného jazyka, rastie. V podoblastiach umelej inteligencie sa skúmajú obrovské množstvá neštruktúrovaných textových údajov. Tieto texty sa analyzujú prostredníctvom rôznych knižníc (Alberola a kol., 2019).

Počiatočné fázy NLP boli zverené do rúk ľudí. Keďže ručné kalkulačky "rozumeli" číslam a vyvinuli sa tak, aby dokázali vykonávať zložitejšie výpočty v matematike, geometrii a štatistike, umožnili výraznú úsporu času. Porozumenie a NLP v súčasnosti sleduje podobný trend (Ishaq, 2019).

NLP zahŕňa dve metódy (Obrázok 2): porozumenie prirodzenému jazyku (NLU) a generovanie prirodzeného jazyka (NLG) (Alberola a kol., 2019).



Obrázok 2 Metódy spracovania prirodzeného jazyka [2]

Generovanie prirodzeného jazyka (Natural Language Generation - NLG) je proces generovania fráz, viet a odsekov, je to transformácia údajov do jazyka čitateľného pre človeka. Pod pojmom porozumenie prirodzenému jazyku rozumieme proces transformácie textu do formálnejších reprezentácií (Khurana a kol., 2017).

Porovnanie procesu spracovania prirodzeného jazyka s procesom spracovania ľudských informácií uľahčí jasnejšie pochopenie jeho rastu. V procese strojového spracovania prirodzeného jazyka existuje korelácia s ľudským získavaním znalostí. Deti sa zaoberajú neštruktúrovaným materiálom, ktorý môže pozostávať z rôznych podnetov, a transformujú ho na informácie. Na základe postupného hromadenia týchto informácií ich začíname skúmať, aby sme pochopili udalosti a ťažkosti v živote človeka. Z neorganizovaných údajov sa nakoniec stávajú vedomosti. Metóda strojového učenia sa riadi podobným prístupom. Systém najprv prevádza neštruktúrovaný textový materiál na zmysluplné pojmy, potom zisťuje vzťahy medzi týmito pojmami a nakoniec získava porozumenie kontextu (Ishaq, 2019).

Zatiaľ čo všetky tieto úlohy sú pre stroj náročné, porozumenie prirodzenému jazyku (NLU), ktoré zahŕňa sémantickú (a pragmatickú) úroveň, je obzvlášť náročné

[2] Zdroj Obrázok 2: https://medium.com/nerd-for-tech/nlp-nlu-and-nlg-7aa3d8d64e2f

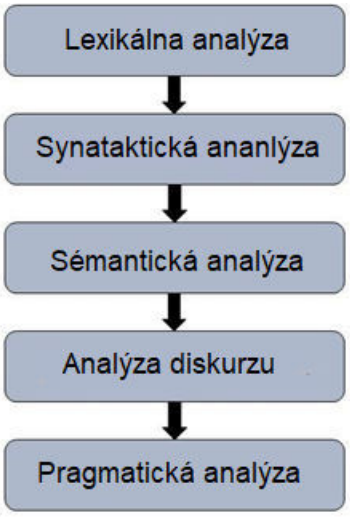
vzhľadom na všadeprítomnú nejednoznačnosť jazyka a jemne odlišné vnímanie významu slov, fráz a viet ľuďmi. Cieľom NLU je dať jazyku zmysel tým, že počítačom umožní čítať a chápať text. Kľúčovou otázkou preto je, ako získať význam z prirodzeného jazyka prekonaním jeho prirodzenej zložitosti (Navigli, 2018).

Zatiaľ čo generovanie textu zahŕňa vytváranie nového textu na základe určitých pravidiel a vstupov, NLU sa zaoberá analýzou a interpretáciou existujúceho textu. Konkrétne sa naša práca zameriava na vývoj automatizovaných meraní lexikálnej rozmanitosti a hustoty, ktoré sú kľúčovými ukazovateľmi bohatosti a zložitosti používania jazyka v danom texte. Uplatňovaním techník NLU sa snažíme hlbšie pochopiť jazykové vlastnosti v ľudskom a strojovom preklade. Je dôležité poznamenať, že náš výskum sa primárne nachádza v oblasti NLU a nie v oblasti tvorby textov.

1.4.2 Prístupy spracovania prirodzeného jazyka

Mnohé prístupy k spracovaniu prirodzeného jazyka sú buď verejne prístupné, alebo ide o laboratórne vyvinuté výskumné nástroje, ktoré nie sú dostupné širokej verejnosti (Chowdhary, 2020).

Prístupy spracovania prirodzeného jazyka sa riadia určitým poradím (Obrázok 2).



Obrázok 3 Proces spracovania prirodzeného jazyka [3]

V rámci procesu prebieha:

[3] Zdroj Obrázok 3: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm

- *Lexikálna analýza* - o vložení údajov do systému spracovania sa vykoná lexikálna analýza textu, ktorá zahŕňa identifikáciu a analýzu slovných štruktúr. Lexikálna analýza rozdeľuje celý text na odseky, vety a slová. Ak systém musí analyzovať zvukové údaje a nie písaný dokument, fonologická analýza textu predchádza lexikálnej analýze. Fonológia je štúdium porozumenia zvukov reči v rámci slov a medzi nimi a uplatňuje sa v dialogických systémoch (Liddy, 2001).
- *Syntaktická analýza* - syntaktická analýza nasleduje po lexikálnej analýze. Usporiadanie slov vo vete, ktoré dodáva gramatický význam, sa označuje ako syntax. Pri spracovaní prirodzeného jazyka sa syntaktická analýza používa na posúdenie toho, ako dobre prirodzený jazyk zvláda gramatické pravidlá (Garabík a kol., 2004). Počítačové algoritmy uplatňujú gramatické princípy na odvodenie významu skupiny slov. V dôsledku toho uvádzame niekoľko syntaktických metód.
- *Lematizácia,* postup redukcie nájdeného slova na jeho základný tvar na jednoduchšie skúmanie, je dobre známy syntaktický prístup. Lematizácia je proces prevodu podstatných mien do tvaru nominatívu jednotného čísla. Príkladom lematizácie je prevod výrazu "stroju" na "stroj" (Garabík a kol., 2004). Cieľom lematizácie, ako ju definujú Laclavk a Ciglan (2006), je objavenie koreňovej formy slova.

Špecializovaným druhom lematizácie je *stemming*, ktorý sa vzťahuje na izoláciu koreňa slova (Paralič, 2010). Stemming je odstránenie prípon a predpôn, čím sa slovo normalizuje. To znamená, že napríklad slovo "naučiť" sa bude písať ako "učiť" (Garbade, 2018).

- *Parsovanie je ďalším syntaktickým prístupom (Garabík a kol., 2004). Parsing zahŕňa rozbor slov z hľadiska štruktúry vety, t. j rozdelenie podmetovej a prísudkovej časti vety a ich následnú klasifikáciu.*

Technika *tokenizácie* pre NLP je tiež dobre známa. Podľa Garabíka a kol. (2004) je tokenizácia kľúčovou zložkou NLP. Podľa Staša a kol. (2015) je tokenizácia najvýznamnejším aspektom predspracovania textu. Cieľom tokenizácie je identifikovať jednotlivé slová a hranice viet, ktoré by mohli slúžiť ako vstupné údaje pre následné spracovanie (Staš a kol., 2015). Tokenizácia zahŕňa rozdelenie vety na jednotlivé slová alebo tokeny, s ktorými potom počítač manipuluje (Garbade, 2018). V praxi sa zo spracovaného textu odstraňujú interpunkčné znamienka a iné netextové značky (Hotho a kol., 2005).

So segmentáciou fráz sa často stretávame v súvislosti s tokenizáciou. Podstatná časť súvislého textu sa segmentuje na elementárne textové komponenty (Garbade, 2018). Na lepšie pochopenie segmentácie využijeme príklad od Paraliča (2010), ktorý uvádza, že príkladom segmentácie je napríklad veta: *Príklad na segmentáciu (textu).* Počas segmentácie sa predchádzajúca veta rozdelí na nasledujúce elementárne textové jednotky: *[Príklad] [na] [segmentáciu] [(] [textu] [)] [.]*

- *Sémantická analýza* je jednou z náročných častí spracovania prirodzeného jazyka, ktorá ešte nie je úplne vyriešená (Garbade, 2018). Na zvýšenie schopnosti počítačových systémov porozumieť prirodzenej reči a jazyku je potrebné implementovať sémantickú analýzu a porovnať extrahovaný význam so znalostnou bázou (Staš a kol., 2015). Jednoducho povedané, cieľom sémantickej analýzy je určiť význam vety (Chowdhary, 2020).

Sémantika je aplikácia počítačových algoritmov na pochopenie významu a interpretácie slov a štruktúry viet (Garbade, 2018). Prístupom sémantickej analýzy je rozpoznávanie pomenovaných entít, všeobecne známe pod skratkou NER, čo znamená Named Entity Recognition. Rozpoznávanie pomenovaných entít je proces rozhodovania, ktoré časti textu možno identifikovať a zaradiť do vopred definovaných skupín (Garbade, 2018).

- *Analýza diskurzu*

Doteraz sme sa venovali štruktúre a významu jednotlivých viet, ale naším cieľom je pochopiť celý text a jeho kontext. Na dosiahnutie tohto cieľa použijeme metódu analýzy diskurzu, ktorá sa snaží pochopiť vzťahy medzi jednotlivými vetami (Chowdhary, 2020).

- *Pragmatická analýza*

Poslednou fázou spracovania prirodzeného jazyka je pragmatická analýza. Pragmatika analyzuje použitie danej vety v rôznych kontextoch, ako aj vplyv na interpretáciu vety (Liddy, 2001).

1.5 Zložitosť jazyka ako problém spracovania prirodzeného jazyka

V spojení s takým komplikovaným procesom, akým je NLP počítačom, je potrebné denne riešiť niekoľko otázok a treba sa pokúsiť tento proces zlepšiť.

V štúdiách týkajúcich sa rôznych jazykov sa tvrdí, že strojové preklady vykazujú niektoré opakujúce sa trendy, ktoré naznačujú prekladateľské univerzálie, ako je

napríklad strata lexikálnej hustoty a lexikálnej variability, pričom iné štúdie ukazujú pravý opak. Zdá sa teda, že tieto "univerzálie (strojového) prekladu" nie sú až také univerzálne, ale do veľkej miery závisia od žánru textu, jazykovej kombinácie a smeru prekladu, ako aj od ďalších faktorov, napríklad od konkrétnych architektúr príslušných systémov strojového prekladu (Brglez, Vintar, 2022).

Druhým problémom náročnosti je samotná flektívnosť slovenského jazyka, ktorá vytvára predpoklad pre mnohonásobne väčší slovník, než je to v prípade jazyka anglického. Autori Staš a kol. (2015) považujú komplexné porozumenie významu za jednu z najnáročnejších úloh v súčasnosti, ktorej sa budeme bližšie venovať v kapitole výsledkov.

Munková a kol. (2020) okrem toho upozorňujú, že slovenčina má voľnejší slovosled ako angličtina, ktorá je charakterizovaná ako analytický jazyk.

Autori Khurana a kol. (2022) vidia ďalší problém práve v kontextových slovách a frázy v jazyku, kde rovnaké slová a frázy môžu mať rôzne významy vo vete, ktoré sú pre človeka ľahko zrozumiteľné, no môžu byť ťažko pochopiteľné pre systémy NLP. Takémuto typu výziev možno čeliť aj pri riešení synoným pretože ľudia používajú mnoho rôznych slov na vyjadrenie tej istej myšlienky. Ďalej sa v jazyku používajú homonymá, slová, ktoré sa vyslovujú rovnako, ale majú rôzne definície, sú tiež problematické pre systémy spracovania prirodzeného jazyka. Vety používajúce sarkazmus a iróniu môžu byť niekedy ľuďmi pochopené opačne, a tak navrhovanie modelov na prácu s takýmito vetami je v NLP skutočne náročnou úlohou. Okrem toho vety v jazyku, ktoré majú akýkoľvek typ nejednoznačnosti v zmysle interpretácie viacerými spôsobmi, sú tiež oblasťou, na ktorej treba pracovať a kde možno dosiahnuť väčšiu presnosť (Khurana a kol., 2022).

Jazyk obsahujúci neformálne frázy, výrazy, idiómy a kultúrne špecifický žargón sťažuje navrhovanie modelov určených na široké použitie, avšak s množstvom údajov, na ktorých sa pravidelne trénuje a aktualizuje, sa modely môžu zlepšiť, ale je naozaj náročnou úlohou zaoberať sa slovami, ktoré majú v rôznych geografických oblastiach rôzny význam. V skutočnosti sa takéto typy problémov vyskytujú aj pri práci s rôznymi oblasťami, napríklad význam slov alebo viet môže byť iný v oblasti vzdelávania, ale majú iný význam v zdravotníctve, práve, obrane atď. Ako uvádzajú autori Khurana a kol., (2022) modely pre NLP teda môžu dobre fungovať pre jednotlivú doménu, geografickú oblasť, ale pre široké použitie je potrebné takéto problémy riešiť.

Ďalším problémom, podľa autorky Romanyshynovej, (2019) je, že NLP má mnoho aplikácií v podnikaní aj pri vývoji softvéru, ale prekážky v ľudskom jazyku spôsobili, že text je náročné analyzovať a replikovať.

Autori Brglez a Vintar (2022) analyzovali výstupy troch strojových prekladov z angličtiny do slovinčiny z hľadiska lexikálnej rozmanitosti v troch rôznych žánroch a zistili, že prekladové systémy, najmä neurónové, vykazujú väčšiu lexikálnu diverzitu ako ich ľudské preklady. Napriek tomu kvalitatívna metóda ukazuje, že tieto kvantitatívne výsledky nie sú vždy spoľahlivým nástrojom na posúdenie skutočnej lexikálnej diverzity a že veľa lexikálnej "kreativity," najmä zo strany neurónových prekladových systémov, je často nespoľahlivá, nedôsledná a chybná.

1.6 Programovací jazyk Python

Python sa stal jedným z najpopulárnejších programovacích jazykov na svete, a to z dobrého dôvodu. Vďaka svojej všestrannosti a jednoduchému používaniu je vhodnou voľbou pre vývojárov všetkých úrovní. Jednou z oblastí, v ktorej Python skutočne zažiaril, je NLP (Gupta, 2023).

Python je populárnym programovacím jazykom, ktorý sa široko používa vo výskume, vrátane spracovania prirodzeného jazyka a analýzy údajov. V tomto kontexte je viacero knižníc, na ktoré sa výskumníci často spoliehajú, a to sú napríklad knižnice Pandas, NLTK, Stanza (Bird a kol., 2009).

Python je vysokoúrovňový programovací jazyk, ktorý v posledných rokoch získava na popularite vďaka svojej flexibilite, čitateľnosti a jednoduchosti používania. V tejto kapitole si bližšie priblížime programovací jazyk Python a jeho využitie v rôznych oblastiach (Gupta, 2023).

Python je univerzálny programovací jazyk, ktorý sa dá použiť v rôznych aplikáciách. Široké uplatnenie má vo vedeckých výpočtoch, analýze údajov, umelej inteligencii a pri tvorbe webových stránok. Jednou z hlavných výhod jazyka Python je jeho jednoduché používanie a čitateľnosť, vďaka čomu je obľúbený pre začiatočníkov aj skúsených programátorov (Pedregosa a kol., 2011).

Python sa bežne používa vo vedeckých výpočtoch vďaka rozsiahlym knižniciam, ktoré poskytujú širokú škálu nástrojov na analýzu údajov, vizualizáciu a vedecké výpočty. Medzi obľúbené knižnice patria NumPy, SciPy, Pandas a Matplotlib. Tieto knižnice poskytujú výkonné nástroje na analýzu údajov, strojové učenie, vizualizáciu údajov a vedecké výpočty (Bird a kol., 2009).

Python sa vďaka svojej jednoduchosti a flexibilite široko používa aj pri vývoji webových stránok. Bežne sa používa na vývoj webových aplikácií, systémov správy obsahu a webových stránok elektronického obchodu (Kozaczko, 2018).

Popularita jazyka Python v oblasti umelej inteligencie je spôsobená jeho jednoduchosťou a flexibilitou. Bežne sa používa pri strojovom učení, spracovaní prirodzeného jazyka a počítačovom videní. Medzi obľúbené knižnice na strojové učenie v jazyku Python patria TensorFlow, Keras a Scikit-Learn (Pedregosa a kol., 2011).

Vo výskume používame programovací jazyk Python na získanie potrebných dát pre náš výskum.

1.6.1 Výhody jazyka Python

Python má oproti iným programovacím jazykom niekoľko výhod. Jednou z hlavných výhod jazyka Python je jeho jednoduchosť a ľahká použiteľnosť. Syntax jazyka Python je jednoduchá a zrozumiteľná, čo uľahčuje jeho učenie a používanie začiatočníkom. Okrem toho rozsiahle knižnice a vstavané funkcie jazyka Python uľahčujú a urýchľujú programovanie (Bird a kol., 2009).

Ďalšou výhodou jazyka Python je jeho čitateľnosť. Kód jazyka Python je ľahko čitateľný a zrozumiteľný, čo uľahčuje jeho údržbu a úpravy. To je dôležité najmä pri veľkých projektoch, kde na tom istom kóde pracuje viacero vývojárov (Kozaczko, 2018).

Veľkou výhodou jazyka Python je aj jeho flexibilita. Python sa dá použiť v rôznych aplikáciách a dá sa ľahko integrovať s inými programovacími jazykmi. Okrem toho objektovo orientované programovacie funkcie jazyka Python uľahčujú písanie opakovane použiteľného kódu (Pedregosa a kol., 2011).

2 NLP METÓDY NA POROZUMENIE PRIRODZENÉHO JAZYKA

Kvantitatívne metódy NLP sme rozdelili do dvoch skupín, a to na lexikálnu rôznorodosť a lexikálnu hustotu.

2.1 Lexikálna rôznorodosť

Na globálnu diverzitu slovnej zásoby sme použili automatické metriky TTR a MTLD (miera textovej lexikálnej diverzity), ktorú prvýkrát navrholi McCarthy a Jarvis (2010) a neskôr ju považovali za jednu z najlepších mier lexikálnej diverzity.

Lexikálna rôznorodosť skúma počet slov použitých v texte alebo v prejave. V texte s vysokou lexikálnou rozmanitosťou sa používa väčšia rozmanitosť slov, zatiaľ čo v texte s nízkou lexikálnou rozmanitosťou sa používa menej slov alebo sa často opakujú tie isté slová (Kurdi, 2016). Lexikálnu rôznorodosť sme počítali nasledujúcimi metrikami:

TTR

Type Token Ration (TTR) sa meria ako jednoduchý celkový pomer medzi všetkými slovami a unikátnymi slovami. Tento pomer je však veľmi citlivý na dĺžku textu, pretože veľmi časté a funkčné slová sa určite opakujú, preto kratšie texty majú tendenciu mať vyšší TTR a dlhšie texty majú tendenciu mať nižšie TTR.

Podľa autora Templin (1957), TTR predstavuje pomer počtu rôznych slov *n* k celkovému počtu slov *N*. TTR meria pomer jedinečných slov k celkovému počtu slov v texte. Vysoké TTR poukazuje na rozmanitejšiu slovnú zásobu.

$$TTR = \frac{n}{N} * 100$$

N=počet všetkých slov; n= počet unikátnych slov

sTTR

Štandardizované TTR (sTTR) bolo navrhnuté na riešenie citlivosti na dĺžku textu tak, že sa vypočíta TTR pre každé n-násobne opakujúce sa slová a tieto pomery sa spriemerujú pre konečný výsledok.

Metóda sTTR rozdeľuje text na úseky určitej dĺžky (zvyčajne 100 slov). Zvyšné slová sa vyradia. Vypočíta sa TTR pre každý celý segment a konečná hodnota TTR je priemerom všetkých úplných segmentov. Takýto prístup funguje dostatočne dobre pri textoch, ktoré sú veľmi dlhé (napr. viac ako 1 000 slov) (Freeman, Cameron, 2008).

GTTR

Hoci sa počet rôznych slov delí veľkosťou textu, ukázalo sa, že TTR je stále ovplyvnené veľkosťou textu, pretože pomer klesá s rastúcim *N* (Hess a kol., 1986; Richards, 1987; Arnaud, 1992). Na kompenzáciu veľkosti textu, a teda na premenu TTR na konštantu v celom texte, sa vyskúšalo niekoľko matematických transformácií TTR. Guiraudovo korigované TTR (GTTR) je jednou z týchto transformácií. Vypočítava pomer jedinečných slov k druhej odmocnine z celkového počtu slov v texte (Guiraud, 1960).

$$GTTR = \frac{n}{\sqrt{N}} * 100$$

N=počet všetkých slov; n= počet unikátnych slov

CTTR

Carroll (1964) navrhuje ďalšiu transformáciu: Carollovo korigované TTR (CTTR). GTTR a CTTR majú rovnako silnú veľkosť účinku, ktorá je približne desaťkrát väčšia ako veľkosť účinku TTR. To potvrdzuje teoretickú výhodu týchto dvoch transformácií oproti pôvodnému TTR (Kurdi, 2020).

$$CTTR = \frac{n}{\sqrt{N * 2}} * 100$$

N=počet všetkých slov; n= počet unikátnych slov

Voc-D

Výpočet voc-D je výsledkom série náhodných výberov textu. Tento prístup začína svoj výpočet tým, že sa z textu vyberie 100 náhodných vzoriek po 35 tokenov. Pre každú z týchto vzoriek sa vypočíta TTR a uloží sa priemerné TTR. Rovnaký postup sa potom opakuje pre vzorky od 36 do 50 tokenov.

McCarthy a Jarvis (2007) teoreticky dokazujú, že to, čo voc-D v konečnom dôsledku predstavuje, je v skutočnosti súčet pravdepodobností výskytu akéhokoľvek slova z textu vo vzorke s daným počtom tokenov, inými slovami, voc-D aproximuje pravdepodobnosť výberu konkrétneho typu v náhodnej vzorke s daným počtom tokenov z cieľového textu. Meranie diverzity slovnej zásoby (voc-D) meria diverzitu slovnej zásoby textu počítaním počtu jedinečných kmeňov (napr. run, running, run) vo vzťahu k celkovému počtu slov.

$$Vocd = (log\ log\ N\ )/(log\ log\ n\ )\text{^}2$$

N=počet všetkých slov; n= počet unikátnych slov

HD-D

Pravdepodobnosti pre všetky lexikálne typy v texte sa sčítajú a súčet sa použije ako index lexikálnej rôznorodosti textu. McCarthy a Jarvis (2007) nazvali túto metriku HD-D alebo hypergeometrické rozdelenie diverzity. HD-D sa určuje tak, že pre každý lexikálny typ v texte sa vypočíta pravdepodobnosť výskytu ktoréhokoľvek z jeho tokenov v náhodnej vzorke 42 slov vybraných z textu (McCarthy, Jarvis, 2010). Výber 42 tokenov je v podstate ľubovoľný: je to stredný bod rozsahu 35 - 50 tokenov, ktorý použili Malvern a Richards (2012). McCarthy a Jarvis (2007) však ukázali, že akýkoľvek počet tokenov vo vzorke medzi 35 a 50 je v podstate zameniteľný. Ako uvádzajú McCarthy a Jarvis (2007), tieto výpočty sú síce ťažkopádne a mimo rámca tohto výskumu, ale ľahko sa vykonávajú počítačom a keďže si nevyžadujú metódu viacnásobného výberu vzoriek ani opakovanie voc-D, sú menej náročné na výpočty ako generovanie voc-D.

Miera diverzity hypergeometrického rozdelenia (HD-D): HD-D je ďalšia miera diverzity slovnej zásoby, ktorá zohľadňuje pravdepodobnosť náhodného výskytu slova. Vypočítava rozmanitosť slovnej zásoby textu porovnaním skutočného počtu jedinečných slov s počtom, ktorý by sa očakával náhodne (McCarthy a Jarvis 2007).

$$HDD = (log\ log\ N\ - log\ log\ n\ )/(log\ log\ N\ )\text{^}2$$

N=počet všetkých slov; n= počet unikátnych slov

MTLD

Ďalším spôsobom, ako zohľadniť lexikálnu rozmanitosť, je miera textovej lexikálnej rozmanitosti (MTLD). Je navrhnutá tak, aby znížila vplyv dĺžky textu. MTLD sa vypočíta ako priemerná dĺžka sekvenčných reťazcov slov v texte, ktorý si zachováva danú hodnotu TTR (McCarthy, Jarvis 2010).

Podľa štúdie na segmentoch hovorených textov, ktoré vytvorilo 20 stredne pokročilých nerodených hovoriacich angličtiny, je MTLD menej ovplyvnený dĺžkou textu ako TTR a GTTR, ak sa používa pri textoch s aspoň 100 tokenmi (Koizumi a In'nami, 2012).

Miera textovej lexikálnej diverzity (MTLD) meria rozmanitosť slovnej zásoby textu meraním bodu, v ktorom sa v texte prestanú objavovať nové slová. Metrika MTLD, sa taktiež snaží zohľadniť a vyhnúť vplyvu dĺžky textu. Výpočet MTLD je postupná

analýza textových častí v oboch smeroch, ktorej výsledok nám hovorí o priemernej dĺžke textu, ktorá zachováva vopred definovaný prah TTR. MTLD bola vypočítaná pomocou knižnice LexicalRichness. Lexikálnu diverzitu počítame len pre slovenské preklady, keďže lexikálna diverzita nie je priamo porovnateľná medzi angličtinou a slovenčinou vzhľadom na extrémne skloňovaný charakter slovenčiny.

Maas index

Z koncepčného hľadiska je založený na predpoklade, že krivku TTR možno relatívne dobre prispôsobiť logaritmickej krivke (McCarthy, Jarvis 2007).

Maasov index meria čitateľnosť textu na základe priemernej dĺžky viet a frekvencie slov (Fergadiotis a kol., 2015).

$$Maas = \frac{log\ log\ (N) - log\ (n)}{log\ (N)\text{^}2}$$

N=počet všetkých slov; n= počet unikátnych slov

Hapax index

Mieru lexikálnej rôznorodosti sme taktiež zisťovali pomocou hapaxov. Kedy predpokladáme, že preklad s väčším počtom hapaxov bude rozmanitejší, teda lexikálne bohatší.

Hapax je slovo, ktoré sa v jednom texte alebo súbore údajov vyskytuje len raz. Podiel hapaxov odráža množstvo rôznych slov použitých v texte alebo bohatosť jeho slovnej zásoby. Čím je Hapax index vyšší, tým je jazyk zložitejší (Popescu a kol., 2008).

Týmto spôsobom zachytávame mieru rozmanitosti, ktorá je v zhode s ľudským prekladom a meriame, do akej miery sa strojové prekladové systémy odchyľujú od navrhovaných (vhodných) riešení.

Simpsonov index

Simpsonov index je index diverzity, ktorý sa vo všeobecnosti používa ako kvantitatívna miera, ktorá vyjadruje, koľko rôznych typov/druhov (tak, ako by sa slovo dostalo do slovníka) sa nachádza v súbore údajov/spoločnosti. Táto miera vypočítava pravdepodobnosť, že dve entity náhodne vybraté zo súboru údajov, ktoré sú predmetom záujmu, predstavujú rovnaký typ. Čím vyššia je pravdepodobnosť, tým menší je počet rôznych prvkov prítomných v súbore údajov (Jarvis, 2013).

Simpsonov index meria koncentráciu slov v texte výpočtom pravdepodobnosti, že dve náhodne vybrané slová budú rovnaké (Simpson, 1949).

$$D = 1 - (\Sigma n(n-1)/N(N-1))$$

N=počet všetkých slov; n= počet unikátnych slov

2.2 Lexikálna hustota

Lexikálnu hustotu sme počítali na základe pomeru kontextových slov voči všetkým slovám. Text s vysokou lexikálnou hustotou má vyšší podiel obsahových slov, zatiaľ čo text s nízkou lexikálnou hustotou má vyšší podiel funkčných slov. Meranie lexikálnej hustoty môže poskytnúť informácie o zložitosti a čitateľnosti textu alebo reči.

Lexikálna hustota, lexikálna rozmanitosť alebo lexikálne bohatstvo (Daller, van Hout, Treffers-Daller, 2003) sú termíny, ktoré sa vzťahujú na štatistické miery, ktoré merajú lexikálne bohatstvo textov a môžu sa použiť aj na hodnotenie celkového pokroku študentov. Lexikálna bohatosť textu zodpovedá tomu, koľko rôznych slov sa v texte používa, zatiaľ čo lexikálna hustota poskytuje mieru podielu lexikálnych položiek (t. j. podstatných mien, slovies, prídavných mien a niektorých prísloviek) (Johansson, 2008). Obe miery sme použili pri počítačových analýzach korpusových údajov. Spravidla texty s nižšou hustotou sú ľahšie zrozumiteľné a hovorené texty majú nižšiu úroveň lexikálnej hustoty ako písané texty (Halliday, 1985). Ako však tvrdí Johansson (2008), text môže mať vysokú lexikálnu diverzitu (t. j. obsahovať veľa rôznych slovných druhov), ale nízku lexikálnu hustotu (t. j. obsahovať skôr veľa zámen a pomocných slovies než podstatných mien a lexikálnych slovies) alebo naopak.

Vzhľadom k vyššie uvedenému je v súvislosti so spracovaním prirodzeného jazyka potrebné venovať pozornosť výzvam, ktoré s danou problematikou súvisia. V nasledujúcej kapitole ozrejmíme cieľ našej práce.

3 CIEĽ PRÁCE A POSTUP DOSAHOVANIA CIEĽOV

Cieľom našej práce je pomocou NLP metód porovnať kvalitu strojového a humánneho prekladu vzhľadom na lexikálnu rôznorodosť a hustotu, hlbšie porozumenie, predstavenie a analýza problematiky v oblasti spracovania prirodzeného jazyka. Pre hlbšie porozumenie prirodzeného jazyka sme upriamili našu pozornosť na metódy, ktoré skúmajú rôznorodosť a hustotu prirodzeného jazyka.

Na základe vyššie uvedených teoretických a empirických zistení, týkajúcich sa porovnania strojového a ľudského prekladu formulujeme nasledujúce hypotézy:

H1: Predpokladáme, že strojový preklad má nižšiu lexikálnu rozmanitosť ako ľudský preklad.

H2: Predpokladáme, že lexikálna hustota je nižšia v strojových prekladoch v porovnaní s ľudskými prekladmi.

H3:Predpokladáme, že strojový preklad je menej lexikálne zložitejší ako humánny preklad na základe absolútnych početností slovných druhov a dĺžky viet.

H4: Predpokladáme, že strojový preklad má nižší pomer rozšírenia objemu textu ako ľudský preklad.

Počas nášho výskumu sme sa stretli s mnohými problémami, ktoré boli väčšinou zapríčinené nekompatibilitou knižníc jazyka Python so slovenským jazykom. Pri použití rovnakých metrík rôznymi knižnicami, sa výsledky často líšili, čo sme považovali práve za nekompatibilnosť, alebo nedokonalú znalosť slovenského jazyka. V mnohých prípadoch sme sa taktiež stretávali s nezmyselnými výsledkami.

4 METODIKA VÝSKUMU

V tejto kapitole sa budeme venovať výskumným metódam, ktoré sú základným nástrojom pri skúmaní a objavovaní nových poznatkov v rôznych oblastiach.

Zameriavame sa na analýzu prekladov anglických textov pomocou rôznych techník spracovania prirodzeného jazyka. Údaje použité v tejto štúdii sú zbierkou anglických textov preložených do slovenčiny. Na základe väčšiny predchádzajúcich štúdií (Castilho, a kol., 2017; Vanmassenhove a kol., 2021; Toral, 2019) predpokladáme, že strojové preklady vykazujú nižšiu lexikálnu diverzitu ako ľudské preklady.

V tejto štúdii sa zameriavame na preklady z angličtiny do slovenčiny a rozhodli sme sa konkrétne preskúmať lexikálnu rozmanitosť v ľudských prekladoch v porovnaní so strojovým prekladom. Nižšia lexikálna diverzita v MT by naznačovala menej pestrý a "kreatívny" výstup s menším súborom prekladových ekvivalentov, než aký navrhuje ľudský prekladateľ.

Na analýzu prekladov sme použili niekoľko techník vrátane tagovania, stemmovania a analýzy kontextových pomerov. Na vykonanie týchto analýz sme používali programovací jazyk Python a rôzne knižnice vrátane knižníc Pandas, NLTK, LexicalRichness a Stanza.

Po dokončení analýz sme výsledky interpretovali a vyvodili závery o prekladoch. Budeme tiež diskutovať o silných stránkach a obmedzeniach použitých metód a predložíme odporúčania pre budúci výskum.

4.1 Knižnice použité pre náš výskum

Obsahom kapitoly je priblíženie knižníc, ktoré patria medzi najpoužívanejšie pri NLP v programovacom jazyku Python a ktoré v našej práci používame.

4.1.1 Pandas

Pandas je výkonná knižnica na analýzu údajov, ktorá poskytuje jednoduchý a efektívny spôsob manipulácie a analýzy veľkých a zložitých súborov údajov. Ponúka rôzne dátové štruktúry a funkcie na čistenie, pretváranie, spájanie a vizualizáciu údajov, vďaka čomu je to knižnica, ktorá je určená pre výskumníkov pracujúcich s tabuľkovými údajmi (McKinney, 2022).

4.1.2 NLTK (Natural Language Toolkit)

NLTK knižnica patrí medzi najznámejšie a najpoužívanejšie knižnice v oblasti NLP. Podľa autora McFarlanda (2022) patrí knižnica NLTK za najlepšiu knižnicu jazyka Python pre NLP. NLTK je základná knižnica, ktorá podporuje úlohy ako klasifikácia, tokenizácia, stemming, parsovanie a sémantické úpravy. Je to jedna z najpopulárnejších dostupných knižníc NLP a výskumníci, študenti a vývojári ju používajú už viac ako desať rokov. Obsahuje aj nástroje ako napríklad Brownov korpus a Penn Treebank, ktoré možno použiť na trénovanie a testovanie modelov NLP.

Podľa autora Chowdhary (2020), NLTK je súbor modulov a programov na báze jazyka Python na symbolické a štatistické NLP angličtiny. NLTK je nástroj na NLP, ktorý je vhodný pre expertov aj začiatočníkov.

Jednou z hlavných výhod NLTK je jeho používateľsky prívetivé rozhranie. Ľahko sa používa a na internete je k dispozícii množstvo dokumentácie a návodov. Vďaka tomu je skvelou voľbou pre začiatočníkov, ktorí s NLP ešte len začínajú. NLTK je tiež veľmi dobre prispôsobiteľný, pričom pre každý z jeho nástrojov je k dispozícii veľa možností konfigurácie. To umožňuje vývojárom doladiť výkon svojich modelov NLP a vytvárať presnejšie a efektívnejšie riešenia (Chowdhary, 2020).

Celkovo je NLTK výkonná a všestranná knižnica, ktorú možno použiť na širokú škálu úloh NLP. Vďaka svojej obľúbenosti, jednoduchosti používania a flexibilite je skvelou voľbou pre každého, kto chce pracovať so spracovaním prirodzeného jazyka v jazyku Python. NLTK je výkonná knižnica v jazyku Python, ktorá poskytuje rôzne nástroje a prostriedky pre NLP. NLTK sa používa v rôznych aplikáciách, ako je klasifikácia textu, analýza sentimentu a strojový preklad (Bird a kol., 2009).

4.1.3 Knižnica Stanza

Stanza je open-source knižnica jazyka Python, ktorú vyvinula Stanford Natural Language Processing Group. Je navrhnutá tak, aby vývojárom uľahčila vykonávanie širokej škály úloh NLP, od základnej tokenizácie až po komplexné rozpoznávanie pomenovaných entít. V rámci kapitoly približujeme knižnicu Stanza a niektoré zo spôsobov, ako ju možno použiť na zlepšenie projektov NLP.

Stanza je ďalšia populárna knižnica pre NLP, ktorá ponúka širokú škálu funkcií a vlastností na spracovanie textu v rôznych jazykoch. Podporuje niekoľko úloh vrátane tokenizácie, označovania častí reči, rozpoznávania pomenovaných entít, rozboru

závislostí a analýzy sentimentu. Jednou z jeho jedinečných vlastností je podpora neurónových modelov, ktorá umožňuje presnejšie a efektívnejšie spracovanie textu (Qi a kol., 2020).

4.1.4 LexicalRichness

V tejto kapitole predstavíme knižnicu LexicalRichness a priblížime niektoré jej kľúčové vlastnosti a funkcie.

Ďalšou z knižníc použitých pre náš výskum je knižnica LexicalRichness, ktorá je určená na pomoc vývojárom pri analýze lexikálnej rozmanitosti textových údajov. Knižnica je užitočná najmä na analýzu zložitosti a rozmanitosti slovnej zásoby v texte, čo môže byť dôležité v oblastiach, ako je lingvistika, vzdelávanie a literatúra.

LexicalRichness sa vzťahuje na rozsah a rozmanitosť slovnej zásoby, ktorú hovoriaci/spisovateľ používa v texte (McCarthy, Jarvis 2007). Lexikálna bohatosť sa používa zameniteľne s lexikálnou rozmanitosťou, lexikálnou variabilitou, lexikálnou hustotou a bohatosťou slovnej zásoby a meria sa pomocou rôznych indexov. Využíva sa, okrem iného, na meranie kvality písania a znalosti slovnej zásoby (Šišková, 2012).

Knižnica LexicalRichness je knižnica jazyka Python, ktorá poskytuje celý rad funkcií na analýzu lexikálneho bohatstva textových údajov. Vyvinuli ju Tim Fawcett a Piotr Pezik a je k dispozícii na bezplatné stiahnutie na indexe balíkov jazyka Python.

Knižnica poskytuje množstvo užitočných funkcií vrátane merania lexikálnej diverzity, bohatosti slovnej zásoby a distribúcie frekvencie slov. Je navrhnutá tak, aby sa ľahko používala a bola flexibilná, čo umožňuje používateľom prispôsobiť si analýzy svojim špecifickým potrebám (Shen, 2021).

Skôr ako začneme knižnicu LexicalRichness používať, musíme ju nainštalovať do nášho systému. Našťastie ide o jednoduchý proces, ktorý možno vykonať pomocou pip, inštalátora balíkov pre Python. Ak chceme nainštalovať knižnicu LexicalRichness, otvoríme príkazový riadok alebo terminálové okno a zadáme nasledujúci príkaz:
„pip install lexicalrichness“ tým sa do nášho systému stiahne a nainštaluje najnovšia verzia knižnice.

Po nainštalovaní knižnice ju môžeme začať používať na analýzu našich textových údajov. Prvým krokom je import knižnice do nášho skriptu Python, napríklad takto:
„from lexicalrichness import LexicalRichness,“ čo nám umožní prístup k funkciám a triedam, ktoré knižnica poskytuje.

Ďalej musíme vytvoriť inštanciu triedy LexicalRichness, ktorú budeme používať na analýzu nášho textu. Môžeme to urobiť zavolaním konštruktora triedy:

„lex = LexicalRichness(text)“

kde „text“ je reťazcová premenná obsahujúca text, ktorý chceme analyzovať. Po vytvorení nášho objektu, „lex“ môžeme začať používať jeho funkcie na analýzu našich textových údajov. Môžeme napríklad použiť funkciu „TTR“ na výpočet pomeru typov a znakov nášho textu:

„print("Type-token ratio:", lex.ttr)“

Vypíše sa pomer type-token nášho textu, ktorý je mierou rozmanitosti slovnej zásoby v texte.

Na výpočet miery lexikálnej rozmanitosti nášho textu môžeme použiť aj funkciu „MTLD“:

„print("MTLD:", lex.mtld())“

Vypíšeme skóre MTLD nášho textu, ktoré je mierou celkovej lexikálnej rozmanitosti textu.

Knižnica LexicalRichness je výkonný nástroj na analýzu lexikálnej bohatosti textových údajov v jazyku Python. Poskytuje celý rad funkcií a mier, ktoré možno použiť na analýzu zložitosti a rozmanitosti slovnej zásoby v texte. Pomocou knižnice môžu vývojári získať prehľad o lingvistických vlastnostiach svojich textových údajov a využiť tieto poznatky na zlepšenie svojich modelov. Či už pracujeme v oblasti lingvistiky, vzdelávania alebo literatúry, knižnica LexicalRichness je cenným zdrojom informácií na analýzu textových údajov v jazyku Python (Shen, 2021).

Celkovo uvedené knižnice poskytujú výskumníkom základné nástroje na prácu s veľkými súbormi údajov a textovými dátami a ich analýzu, vďaka čomu sú cenným prínosom v oblasti dátovej vedy a výskumu NLP.

4.2 Opis dát/dátového setu

Opisu dát predchádzal výber textu v angličtine s existujúcim ľudským prekladom v slovenčine. Aby sme mohli uskutočniť pomerne spoľahlivú kvantitatívnu štúdiu lexikálnej diverzity, vyberali sme rôzne žánre a dlhšie texty.

Náš výskum sme vykonávali na datasete, ktorý obsahoval zdrojový text, dva rôzne preklady a ID dokumentu z ktorého bol zdrojový text získaný. Dataset sme si upravili tak,

že sme doň pridali pomocnú premennú, ktorá reprezentovala, či daný preklad je ľudský, alebo strojový. Takto upravený dataset sme v Pythone načítali pomocou knižnice Pandas.

Našou úlohou bolo v rámci výskumu využiť už spomínané metódy spracovania prirodzeného jazyka a následne ich použiť pri spracovaní prekladov. Vety sme združili do textov a metriky sme vyrátavali na textoch.

Tabuľka 1 Kompozícia dát

Početnosti	**HT**	**MT**
Abstraktné podst. m	**772**	**757**
Konkrétne podst. m	**156497**	**153824**
Podstatné mená	**157269**	**154581**
Prídavné mená	**25358**	**24230**
Slovesá	**20177**	**20577**
Príslovky	**1626**	**1605**
Slová	**256092**	**253894**
Znaky	**15555624**	**1527805**
Slovesá v prítomnom čase	**12052**	**12214**
Slovesá v minulom čase	**15778**	**14713**
Slovesá v budúcom čase	**25**	**31**
Jednoduché vety	**103**	**91**
Zložené vety	**16431**	**14644**
Vety	**16534**	**14735**

Korpusová štatistika (Tabuľka 1) zobrazuje početnosti rôznych textových vlastností (slovných druhov a i,), ktoré sme v našej práci analyzovali a určovali rozdiely medzi ľudským a strojovým prekladom.

4.2 Predspracovanie dát

Pracovali sme s dátovým setom, ktorý už bol zarovnaný na vety, tým pádom nám niektoré kroky predspracovania textu odpadli. Pomocou knižnice Pandas sme náš dataset načítali do prostredia Python a následne sme upravovali štruktúru textu do tvaru, kedy jeden segment obsahoval text. Ďalším krokom bolo pridanie pomocnej premennej s názvom *„MT_T,“* ktorá rozdeľuje preklady do dvoch skupín a to, či ide o ľudský, alebo strojový preklad.

		Translations
DOC_ID	MT_T	
1.0	0.0	Seamus Heaney sa narodil v katolíckej vlastene...
	1.0	Seamus Heaney sa narodil v katolíckej nacional...
2.0	0.0	NASLEDUJÚCI deň som cestoval vlakom pozdĺž ved...
	1.0	Na druhý deň som CESTOVAL vlakom po vedľajšej ...
3.0	0.0	Počas mnohých expedícií začiatkom 19. storočia...
...	...	...
84.0	1.0	Prečo vaše internetové návyky nie sú také čist...
86.0	0.0	Predstavme si planétu Zem bez vírusov.Mávneme ...
	1.0	Predstavme si planétu Zem bez vírusov.Mávneme ...
87.0	0.0	Nemôžeme plne pochopiť súčasné nepokoje alebo ...
	1.0	Nemôžeme úplne pochopiť súčasné prevraty alebo...

Obrázok 4 Štruktúra dát

4.3 Použité automatické metriky

V tejto podkapitole si predstavíme využité metriky, ktoré sme použili na spracovanie oboch prekladov, ktoré sme rozdelili do troch rozličných kategórii a to na: lexikálnu rôznorodosť, lexikálnu hustotu a početnosti.

4.3.1 Lexikálna rôznorodosť

Vypočítali sme lexikálnu rozmanitosť pre každý preklad, aby sme potvrdili alebo vyvrátili naše predpoklady, a to, že lexikálna diverzita je nižšia v strojových prekladoch v porovnaní s ľudskými prekladmi. Lexikálnu diverzitu alebo rozmanitosť možno vypočítať rôznymi metódami, napríklad voc-D, HD-D, MTLD, TTR, Maas a inými.

Ako prvé sme počítali metriku TTR, pomocou knižnice NLTK sme načítali tokenizér, ktorý sme následne použili na vypočítanie počtu všetkých a unikátnych slov v prekladoch pomocou ktorých sme následne počítali TTR. Podobne sme počítali aj metriky sTTR, GTTR, CTTR, ktoré sú transformácie metriky TTR. Pri týchto metrikách sa taktiež používa počet všetkých a unikátnych slov v prekladoch. Pomocou počtu všetkých slov a unikátnych slov sme taktiež počítali Simpsonov index.

Následne sme si do nášho prostredia Python nainštalovali knižnicu LexicalRichness, pomocou ktorej sme počítali nasledujúce metriky lexikálne rôznorodosti: voc-D, HD-D, MTLD, Maas index a Hapax index.

Do premennej *lex* sme importovali knižnicu s naším textom a použitím rôznych metód ktorými knižnica už disponuje ako napríklad „*lex.vocd()*“, „*lex.hdd()*“ a ďalšie, sme následne mohli vypočítať spomínané metriky lexikálnej rôznorodosti.

4.3.2 Lexikálna hustota

Vyššia lexikálna hustota môže naznačovať zložitejší a náročnejší text, pretože obsahové slová nesú väčší význam a vyžadujú si väčší stupeň spracovania.

Na druhej strane nižšia lexikálna hustota môže naznačovať jednoduchší a prístupnejší text, keďže funkčné slová poskytujú rámec na pochopenie obsahu. Lexikálnu hustotu sme počítali na základe týchto pomerov:

- pomer podstatných mien ku všetkým slovám,
- pomer prídavných mien ku všetkým slovám,
- pomer prísloviek ku všetkým slovám,
- pomer slovies ku všetkým slovám,
- pomer kontextových slov ku všetkým slovám.

Celkovo je lexikálna hustota jedným z viacerých faktorov, ktoré prispievajú ku komplexnosti a čitateľnosti textu alebo prejavu a môže byť užitočným nástrojom na analýzu a zlepšenie písanej a hovorenej komunikácie (Laufer, Nation, 1995).

Pomery podstatných mien, prídavných mien, prísloviek a slovies sme najprv počítali v Pythone pomocou knižnice Stanza pre slovenský jazyk. Po vizualizácií výsledkov sme spozorovali prípady, kedy nastala situácia, že knižnica Stanza nebola schopná detegovať žiadne podstatné mená, prídavné mená a ani príslovky v určitých prípadoch, kedy sme si boli istí, že sa v konkrétnom prípade nachádzajú. Preto sme sa pre náš výskum rozhodli použiť knižnicu NLTK za pomoci ktorej sa nám podarilo vypočítať pomery podstatných a prídavných mien, slovies a prísloviek.

4.3.3 Početnosti

S cieľom zvýšiť vizualizáciu výskumu sme realizovali zisťovanie viacerých početností v ľudskom a strojovom preklade, na základe ktorých sme dva rozličné preklady zanalyzovali. Zamerali sme sa na zisťovanie nasledujúcich početností:

- počet podstatných mien v prekladoch,
- počet prídavných mien v prekladoch,
- počet prísloviek v prekladoch,
- počet slovies v prekladoch,
- počet všetkých slov v prekladoch,
- počet všetkých znakov,
- počet slovies v prítomnou, budúcom a minulom čase v prekladoch,

- počet jednoduchých a zložených viet v prekladoch.

Po predchádzajúcich skúsenostiach s nesprávnymi výsledkami knižnice Stanza, sme sa rozhodli znovu použiť knižnicu NLTK na zistenie vyššie uvedených početností.

Na základe zisťovania uvedených početností sme určili rozdiely medzi ľudským a strojovým prekladom. Výsledkom týchto analýz sa budeme venovať v nasledujúcej kapitole.

4.3.4 Miera rozšírenia (expanding ratio)

Rozširujúci pomer je metrika používaná v spracovaní prirodzeného jazyka a počítačovej lingvistike na meranie stupňa rozšírenia alebo zúženia medzi zdrojovým textom a jeho prekladom. Vypočíta sa porovnaním dĺžky cieľového textu (t. j. preloženého textu) s dĺžkou zdrojového textu a vyjadrením rozdielu v percentách dĺžky zdrojového textu.

Vzorec na výpočet rozširujúceho pomeru (ER) je:

$$ER = (\frac{Dĺžka\ cieľa - Dĺžka\ zdroja}{Dĺžka\ zdroja}) * 100$$

Kladný rozširujúci pomer znamená, že cieľový text je dlhší ako zdrojový text, zatiaľ čo záporný rozširujúci pomer znamená, že cieľový text je kratší ako zdrojový text. Rozširujúci pomer je jednou z mnohých metrík používaných na hodnotenie kvality strojových prekladov a môže byť užitočný pri identifikácii oblastí, v ktorých sa v preklade mohli pridať alebo vynechať informácie v porovnaní s pôvodným textom. Na určenie celkovej kvality prekladu by sa však mala používať v spojení s inými metrikami a ľudským hodnotením.

Mieru rozšírenia sme vypočítali vyššie uvedeným vzorcom po tom, ako sme zistili počet slov v prekladanom texte a prekladoch.

5 VÝSLEDKY

Obsahom nasledujúcej kapitoly sú výsledky, ku ktorým sme dospeli v rámci nášho výskumu.

5.1 Výsledky riešenia

Na vykonanie analýzy sme použili štatistický softvér Statistica, ktorý nám vygeneroval krabicové grafy na základe ktorých sme schopní interpretovať získané dáta. Variabilitu výsledkov sme overovali F-testom, kde porovnávame rozdiely v smerodajných odchýlkach a pomocou tohto testu sme schopní zistiť, či je štatisticky významný rozdiel vo variabilite prekladov. Následne sme použili dva t-testy, jeden sme použili v prípade, keď variabilita humánneho aj strojového prekladu je približne rovnaká a druhý t-test sa používa, keď je variabilita prekladov rozdielna. Pomocou daných t-testov sme schopní určiť, či sú rozdiely vo výsledkoch metrík štatisticky významné.

V kapitole cieľov sme si stanovili hypotézy, ktoré sme následne overovali, a to:

H1: Predpokladáme, že strojový preklad má nižšiu lexikálnu rozmanitosť ako ľudský preklad.

Použitým vyššie uvedených metrík TTR, sTTR, GTTR, CTTR, voc-D, HD-D, MTLD, Maas index, hapax index a Simpsonov index sme dosiahli výsledky zobrazené v Tabuľke 2, ktoré predchádzali výpočtu štatistických významností.

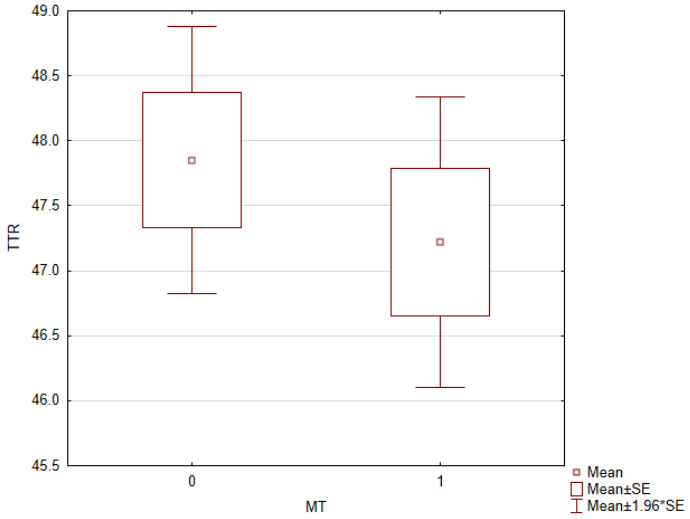
Na overenie prvej hypotézy, že lexikálna diverzita je v strojových prekladoch nižšia v porovnaní s ľudskými prekladmi, sme použili niekoľko metrík na meranie lexikálnej diverzity v oboch typoch prekladov.

Tabuľka 2 Lexikálna rôznorodosť pre dokument s ID 1

MT_T	TTR	sTTR	GTTR	CTTR	voc-D	HD-D	MTLD	Maas	Hapax	Simpsonov index
0	40,705	31,875	3187,451	2253,868	252,505	0,928	293,877	0,01184	1937	0,99007
1	40,593	29,625	2962,478	2094,788	191,835	0,909	168,496	0,01199	1703	0,98563

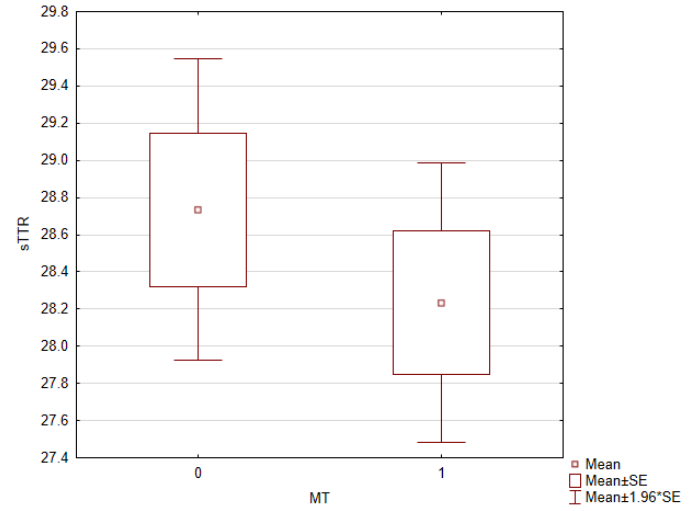
Na všetkých grafoch vidíme vypočítaný bodový odhad priemeru, intervalový odhad priemeru, taktiež známy ako 95% interval spoľahlivosti pre priemer, smerodajnú chybu odhadu priemeru pre všetky metriky pre oba preklady. Zo smerodajnej chyby odhadu priemeru a smerodajnej odchýlky vyplýva variabilita, či sú hodnoty homogénnejšie alebo heterogénnejšie.

Priemerná bodová hodnota TTR pre ľudský preklad je vyššia ako pre strojový preklad (Graf 1), ale rozdiel medzi týmito dvoma výsledkami nie je natoľko veľký, aby sme mohli predpokladať, že je prítomný štatisticky významný rozdiel. Na druhej strane, oba preklady majú pomerne vysoké priemerné hodnoty TTR (nad 47), čo poukazuje na pomerne vysoký stupeň lexikálnej diverzity. To by mohlo naznačovať, že oba preklady sú kvalitné a efektívne vyjadrujú význam východiskového textu. Variabilita je v tomto prípade podobná.

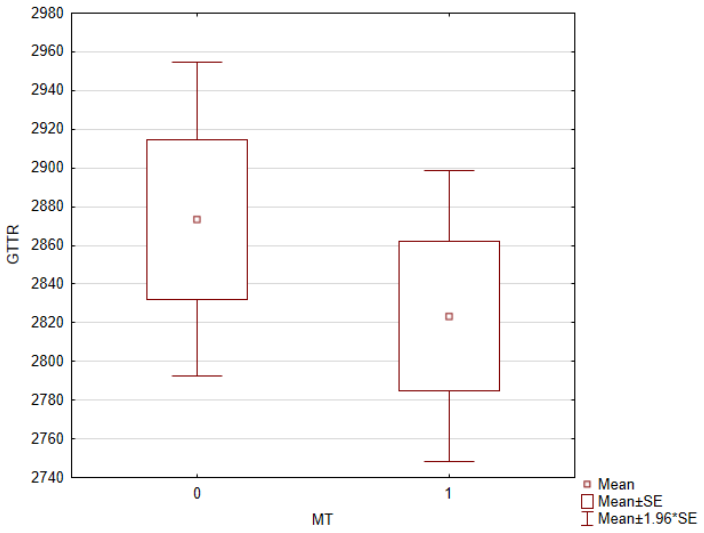


Graf 1 TTR ľudského a strojového prekladu

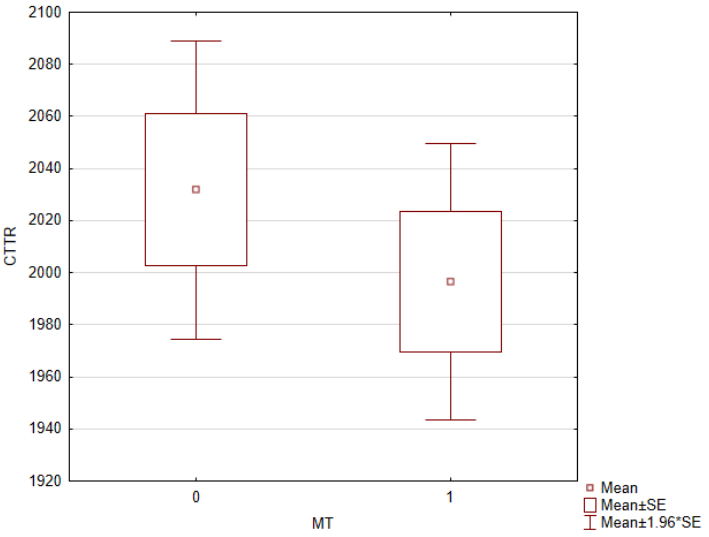
Na Grafe 2, ktorý znázorňuje výsledky metriky sTTR, ktorá nie je ovplyvnená dĺžkou textu do takej miery ako metrika TTR, vidíme podobný výsledok ako pri metrike TTR, kedy priemer hodnôt metriky sTTR je vyšší pre ľudský preklad, ale tento rozdiel nie je dostatočne veľký, aby sme mohli predpokladať, že sa tu nachádza štatisticky významný rozdiel. Pre Grafy 3 a 4, ktoré znázorňujú výsledky metrík GTTR a CTTR, môžeme konštatovať rovnaké výsledky, keďže sa jedná o transformácie pôvodnej metriky TTR, ktoré tiež nie sú natoľko ovplyvnené dĺžkou textu. Variabilita je tiež v týchto prípadoch veľmi podobná.



Graf 2 sTTR ľudského a strojového prekladu



Graf 3 GTTR ľudského a strojového prekladu

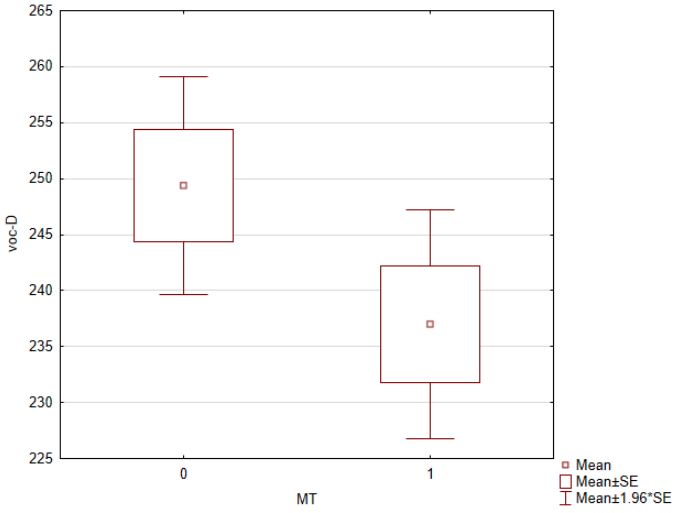


Graf 4 CTTR ľudského a strojového prekladu

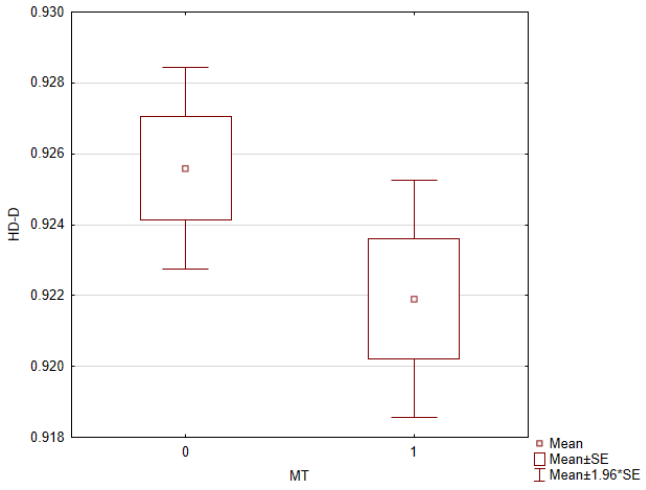
F-test nepotvrdil prítomnosť štatisticky významných rozdielov vo variabilite pre Grafy 1,2,3,4 (*viď prílohu A1*) kedy hodnota *p* bola približne *0,5*.

T-testy pre tieto metriky nepotvrdili štatisticky významný rozdiel vo výsledkoch, *p* hodnota dosahovala hodnotu približne *0,38* (*viď prílohu A2*).

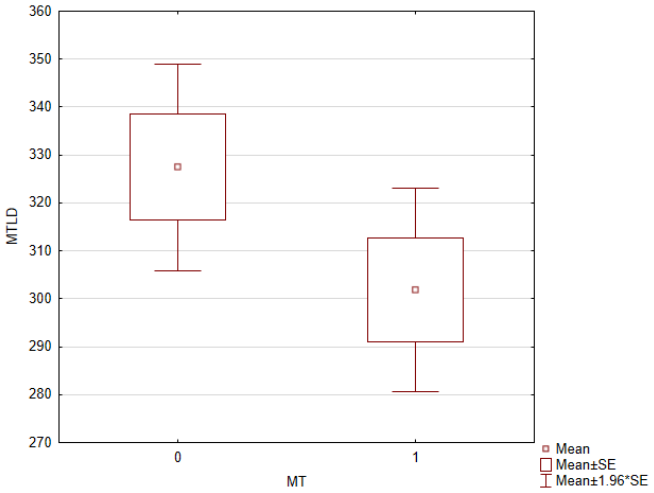
Na Grafoch 5,6 a 7, ktoré zobrazujú výsledky metrík voc-D, HD-D a MTLD, môžeme pozorovať, že priemerný bod výsledkov pre ľudský preklad je stále vyšší ako pre strojový preklad a priemerný bod výsledkov ľudského prekladu už nezasahuje do intervalového odhadu priemerov pre strojový preklad. Na základe vyššie uvedeného môžeme predpokladať, že uvedený rozdiel je štatisticky významný.



Graf 5 voc-D ľudského a strojového prekladu



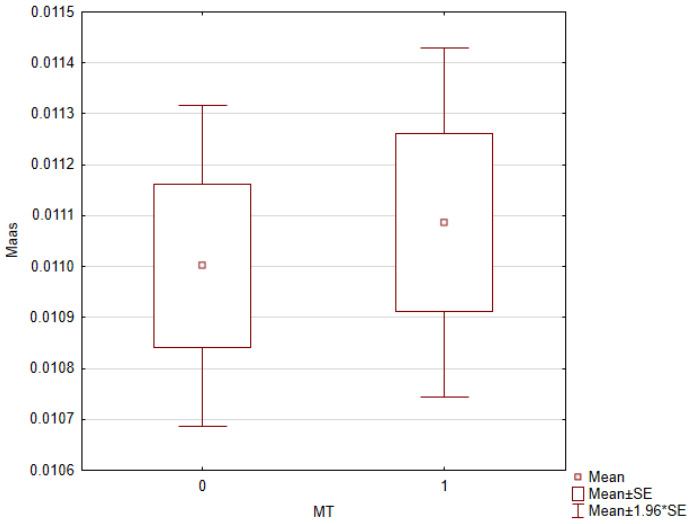
Graf 6 HD-D ľudského a strojového prekladu



Graf 7 MTLD ľudského a strojového prekladu

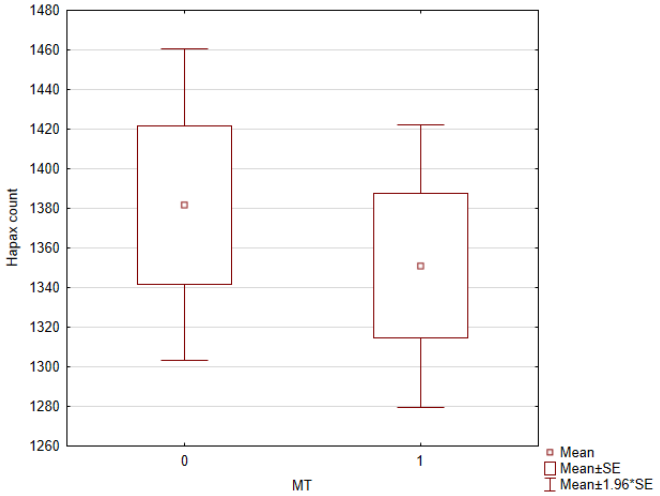
Výsledky testov pre metriky voc-D, HD-D a MTLD nepotvrdili prítomnosť štatisticky významných rozdielov vo variabilite a výsledkoch, ale hodnoty *p* pre t-test sa približujú k hodnote *0,05*, čo by mohlo naznačovať, že pri použití väčšieho množstva dát by sa pravdepodobne potvrdila prítomnosť štatisticky významného rozdielu vo výsledkoch, čo by potvrdzovalo našu hypotézu, že strojový preklad má nižšiu lexikálnu rozmanitosť ako ľudský preklad.

Výsledky Grafu 8 ukazujú, že priemerná hodnota Maas indexu je pre ľudský preklad mierne nižšia ako pre strojový preklad, rozdiel je však v rámci štandardnej chyby merania, takže nie je jasné, či je štatisticky významný alebo je spôsobený len náhodnými odchýlkami.



Graf 8 Maas index výsledky

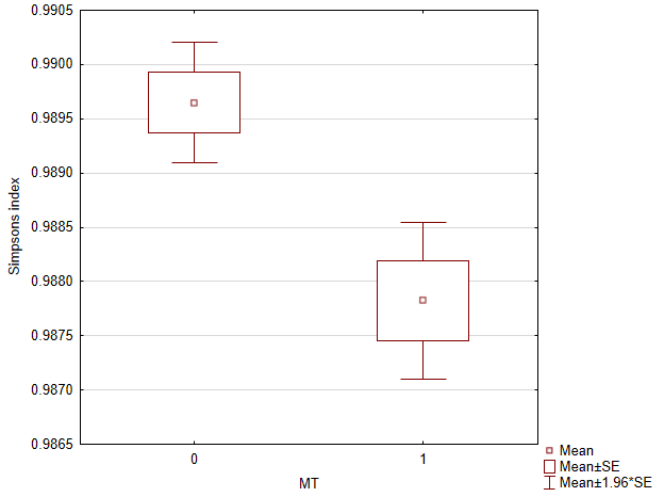
Na Grafe 9 môžeme pozorovať podobný trend ako v prípade Grafov 1, 2, 3 a 4 kedy bodový priemer pre ľudský preklad je vyšší ako bodový priemer strojového prekladu, ale predpokladáme, že tento rozdiel nie je štatisticky významný.



Graf 9 Hapax index

Testy nepotvrdili štatisticky významný rozdiel pre metriky Maasov index a hapax index (viď prílohu *A1* a *A2*) .

Z Grafu 10 môžeme vidieť, že rozdiel medzi priemernými hodnotami ľudského prekladu je doteraz najväčší spomedzi všetkých ostatných metrík. Na základe toho môžeme predpokladať, že rozdiel bude štatisticky významný. Tento predpoklad sme overili testom, ktorý uvádzame nižšie. Variabilita je väčšia pre strojový preklad čo naznačuje, že hodnoty pre strojový preklad v tomto prípade sú heterogénnejšie ako hodnoty ľudského prekladu.



Graf 10 Simpsonov index ľudského a strojového prekladu

F-test nepotvrdil rozdiely vo variabilite metriky Simpsonov index, ale t-test potvrdil prítomnosť štatisticky významného rozdielu pre túto metriku, čo potvrdzuje našu hypotézu, že strojový preklad má nižšiu lexikálnu rozmanitosť ako ľudský preklad.

Záverom možno konštatovať, že naše zistenia potvrdzujú hypotézu H1, že lexikálna rôznorodosť je v strojových prekladoch nižšia v porovnaní s ľudskými prekladmi, iba čiastočne a to pre metriku Simpsonov index.

H2: Predpokladáme, že lexikálna hustota je nižšia v strojových prekladoch v porovnaní s ľudskými prekladmi.

Na overenie hypotézy H2, že lexikálna hustota je v strojových prekladoch nižšia v porovnaní s ľudskými prekladmi, sme použili pomer kontextových slov (podstatné meno, sloveso, prídavné meno a príslovka) ku všetkým slovám.

Tabuľka 3 výsledky lexikálnej hustoty pre dokument s ID 1

MT_T	**NOUN/all**	**VERB/all**	**ADJ/all**	**ADV/all**	**contextual_ratio**
0	0,6037	0,0789	0,115134	0,005382	0,694553164
1	0,5789	0,0914	0,096132	0,004318	0,669921142

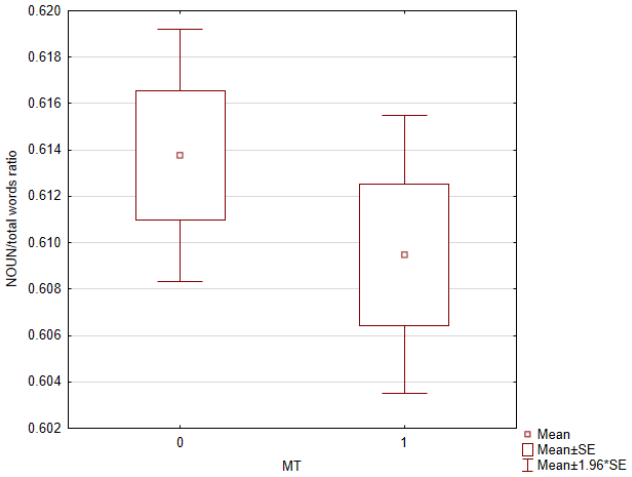
Uvedené metriky slúžia na meranie podielu slov v texte, ktoré nesú kontextový význam, čo môže poskytnúť informácie o celkovej hustote textu. Na analýzu lexikálnej hustoty každého súboru prekladov sme pre každý text v našej vzorke vypočítali pomer kontextových slov ku všetkým slovám. Potom sme porovnali priemerný pomer pre strojové preklady a ľudské preklady a vykonali sme štatistickú analýzu na overenie významných rozdielov medzi oboma skupinami.

Ako prvý sme vypočítali bodový odhad priemeru a intervalový odhad priemeru pre každú z daných metrík lexikálnej hustoty prekladov, aby sme získali grafy, na základe ktorých sme schopní interpretovať výsledky nášho výskumu. Následne sme použili F-test, ktorý nám hovorí o variabilite výsledkov (viď prílohu *B1*).

Následne sme použili t-testy pre určenie štatistických rozdielov metrík lexikálnej hustoty.

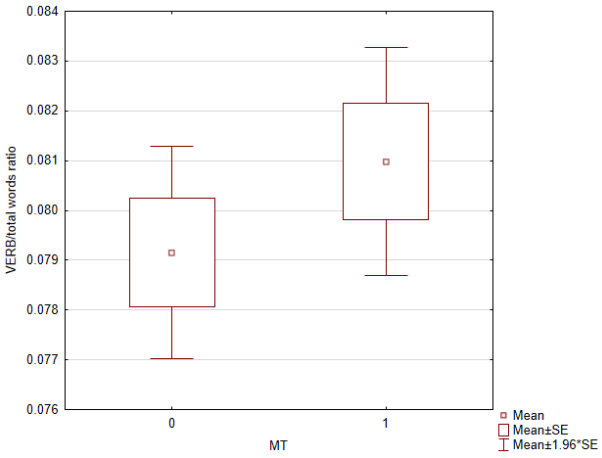
Výsledky indikujú (Grafu 11) že pomer použitých podstatných mien k všetkým slovám je vyšší v ľudskom preklade ako pomer podstatných mien ku všetkým slovám v

strojovom preklade, ale nepredpokladáme štatisticky významné rozdiely vo variabilite a výsledkoch.



Graf 11 Pomer podstatných mien ku všetkým slovám

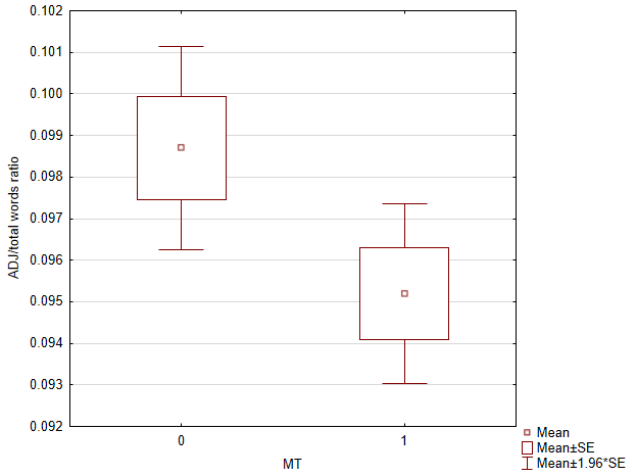
Z Grafu číslo 12 vyplýva, že v strojovom preklade bolo použitých viac slovies ako v ľudskom preklade, ale nepredpokladáme, že tento rozdiel bude štatisticky významný.



Graf 12 Pomer slovies ku všetkým slovám

Testy potvrdili náš predpoklad a pomer medzi počtom slovies a všetkých slov nie je štatisticky významný pre ľudský a strojový preklad s hodnotou *p=0,5777*.

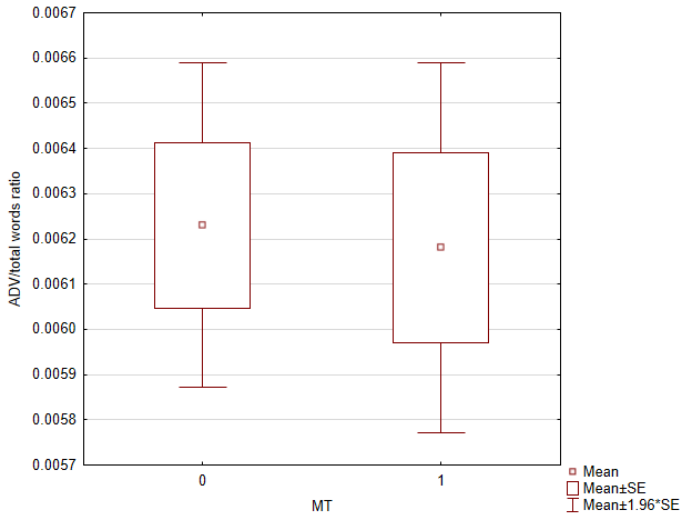
Graf číslo 13 zobrazuje výsledky pomeru prídavných mien pre oba preklady. Predpokladáme, že v tomto prípade môže byť prítomný štatisticky významný rozdiel pre ľudský a strojový preklad v prospech ľudského prekladu.



Graf 13 Pomer prídavných mien ku všetkým slovám

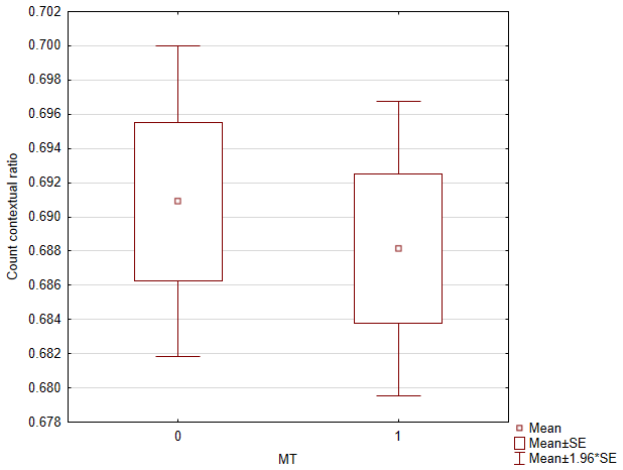
Výsledky F-testu nenaznačujú rozdiely vo variabilite (viď prílohu *B1*). T-testom sme potvrdili prítomnosť štatisticky významného rozdielu pre danú metriku lexikálnej hustoty medzi ľudským a strojovým prekladom, kedy hodnota *p* bola *0,036* (viď prílohu *B2*).

Z Grafu 14 vyplýva, že pomer počtu použitých prísloviek je veľmi podobný pre oba preklady, variabilita je heterogénnejšia pre strojový preklad, ale nepredpokladáme žiadne významné rozdiely pre metriku pomerov prídavných mien ku všetkým slovám.



Graf 14 Pomer prísloviek ku všetkým slovám

Na základe F-testov a t-testov pomer prísloviek ku všetkým slovám, ktoré znázorňuje *Graf* 14 môžeme konštatovať, že významné rozdiely pre metriku pomeru prísloviek ku všetkým slov nie sú prítomné. *P* hodnota pre F-test bola *0,272* (viď prílohu B1) a pre t-test bola *0,859 (*viď prílohu B2*).* Podobné výsledky naznačuje aj *Grafu 15*.



Graf 15 Pomer kontextových slov ku všetkým slovám

Na porovnanie bodových priemerov pre oba typy prekladov sme použili t-test pre nezávislé vzorky a F-test s hladinou významnosti $p < 0,05$. F-test nepreukázal štatisticky významný rozdiel ani pri jednej z použitých metrík vo variabilite a predpokladáme, že rozptyl premenných je v oboch prípadoch podobný. Naše výsledky ukázali (p*ríloha B2*), že v priemere bola lexikálna hustota strojových prekladov rovnaká ako hustota ľudských prekladov. Uvedený rozdiel bol štatisticky významný iba pre metriku pomeru prídavných mien ku všetkým slovám pre ľudský preklad s hodnotou $p=0,0363$. *Tým sme našu hypotézu, že lexikálna hustota je nižšia v strojových prekladoch v porovnaní s ľudskými prekladmi, potvrdili iba čiastočne.*

H3:Predpokladáme, že strojový preklad je menej lexikálne zložitejší ako humánny preklad na základe absolútnych početností slovných druhov a dĺžky viet.

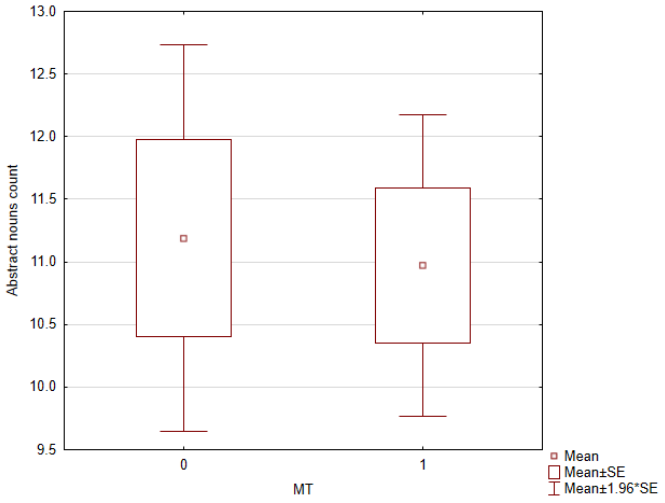
V nasledujúcej časti práce sú uvedené opisné štatistiky pre rôzne metriky početností pre analýzu lexikálnej rôznorodosti ľudského a strojového prekladu, ako je počet podstatných mien, počet prídavných mien, počet slovies, počet prísloviek, počet slov, počet znakov, počet jednoduchých viet, počet zložených viet a počet viet. Výsledky poskytujú prehľad o rozdelení týchto metrík v rámci súboru údajov vrátane priemeru, štandardnej odchýlky a štandardnej chyby.

Tabuľka 4 zistené dáta pre dokument ID 1

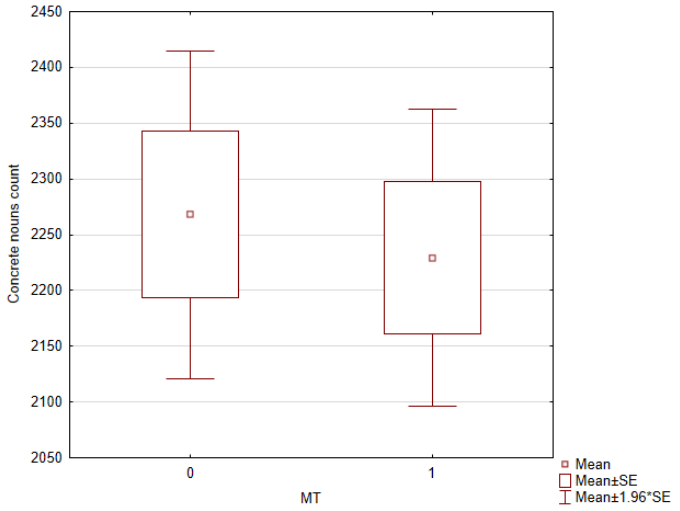
MT_T	Podst.m.	Príd. m.	Slovesá	Príslovky	Slová	Znaky	Prítomný č.	Minulý č.	Budúci č.	Počet slov zdroj
0	3702	706	484	33	6132	34724	248	370	0	6227
1	3083	512	487	23	5326	29402	206	368	1	6227

Na overenie hypotézy *H3*, že absolútna hodnota pomeru dĺžok v strojových prekladoch je bližšia východiskovému textu ako v ľudských prekladoch, sme použili niekoľko metrík vrátane počtu podstatných mien, prídavných mien, prísloviek a slovies v prekladoch, počtu všetkých slov a všetkých znakov v prekladoch, počtu slovies v prítomnom, budúcom a minulom čase a počtu jednoduchých a zložených viet v prekladoch. Podobne ako pri predchádzajúcich hypotézach, sme taktiež testovali variabilitu týchto metrík F-testom a následne overovali štatistické významnosti t-testom.

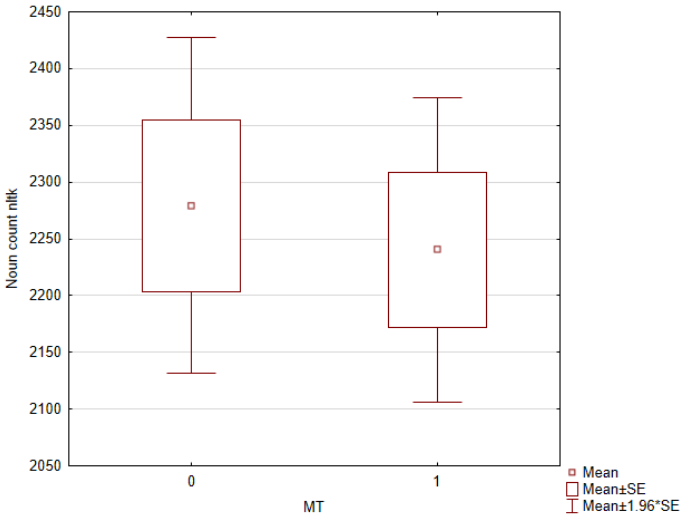
Z Grafov 16, 17, 18, 19, 20, 21, 22, 23, 24 vyplýva, že počet abstraktných podstatných mien v prekladoch sa príliš nelíši, ale Grafy 16 a 24 ukazujú rozdiel vo variabilite metrík, kedy ľudský preklad bol hetorogénnejší a obsahoval vyšší počet unikátnych abstraktných podstatných mien a vyšší počet unikátnych slov v jednoduchých vetách ako strojový preklad. Z bodového priemeru výsledkov nepredpokladáme štatisticky významný rozdiel medzi prekladmi pre tieto metriky.



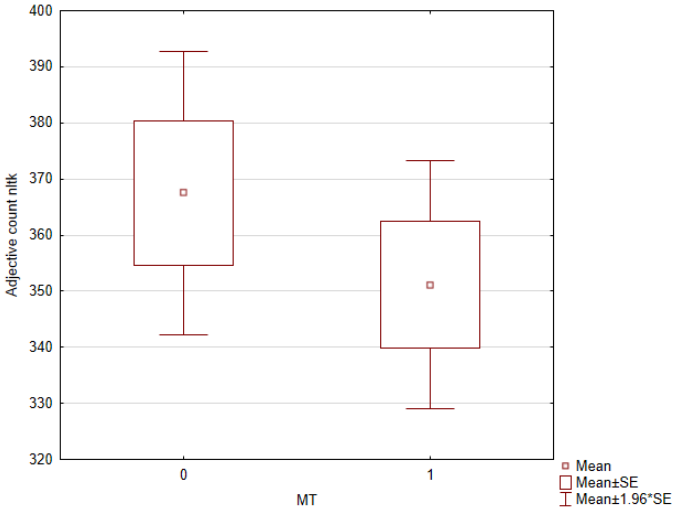
Graf 16 Počet abstraktných podstatných mien v prekladoch



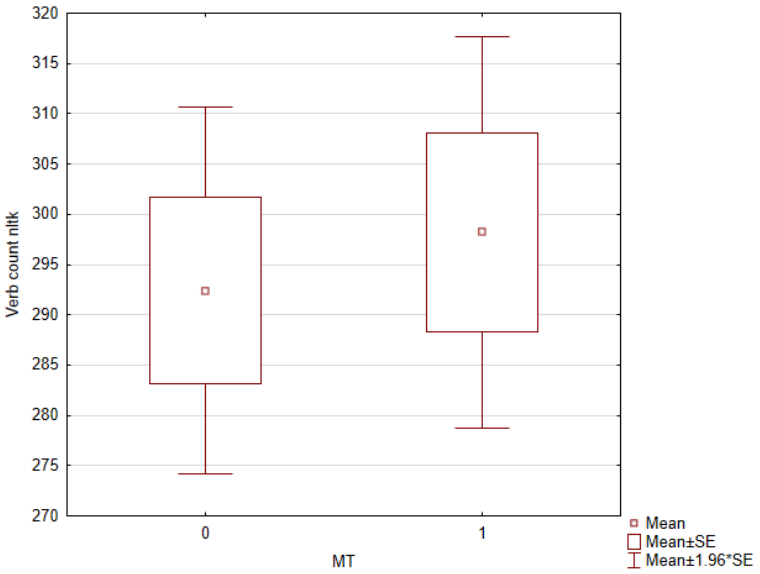
Graf 17 Počet konkrétnych podstatných mien v prekladoch



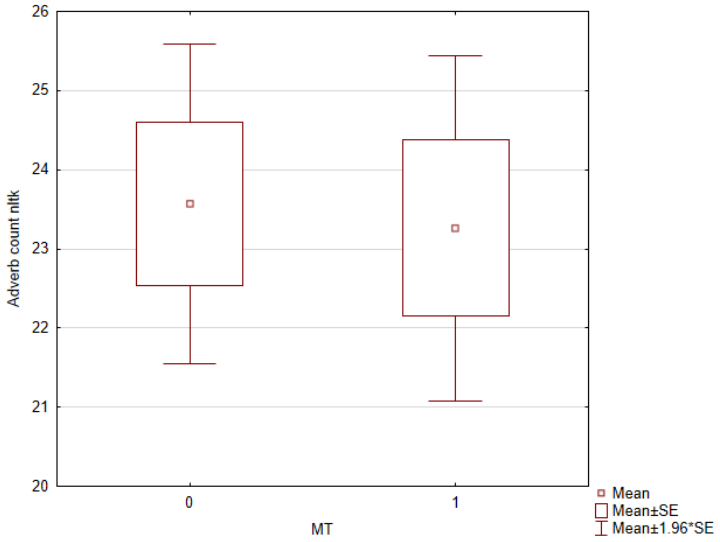
Graf 18 Počet podstatných mien v prekladoch



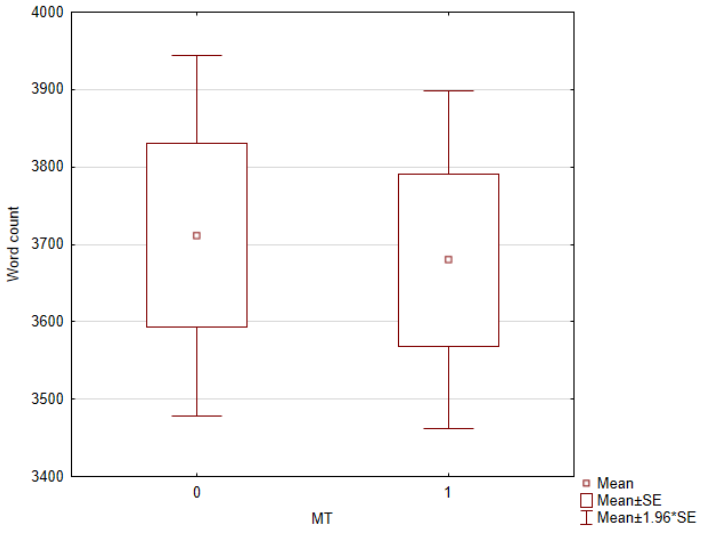
Graf 19 Počet prídavných mien v prekladoch



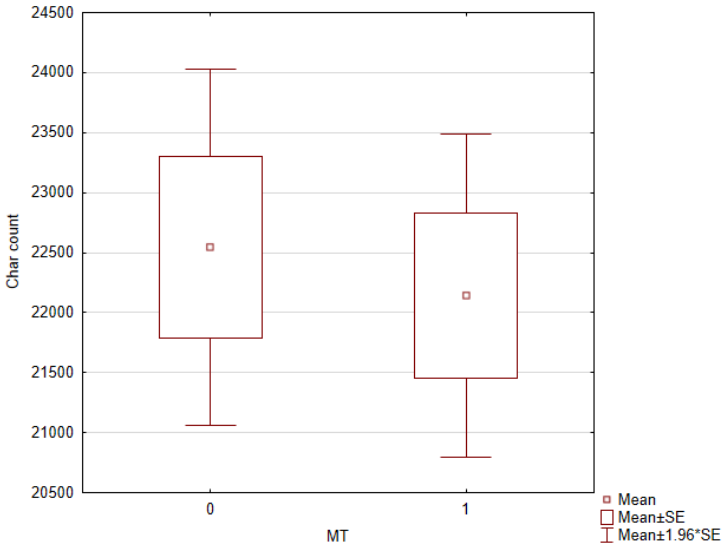
Graf 20 Počet slovies v prekladoch



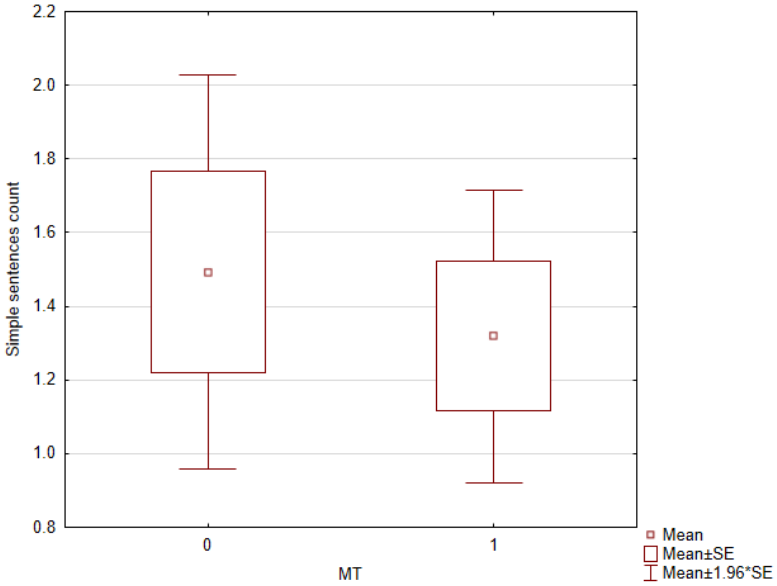
Graf 21 Počet prísloviek v prekladoch



Graf 22 Počet slov v prekladoch



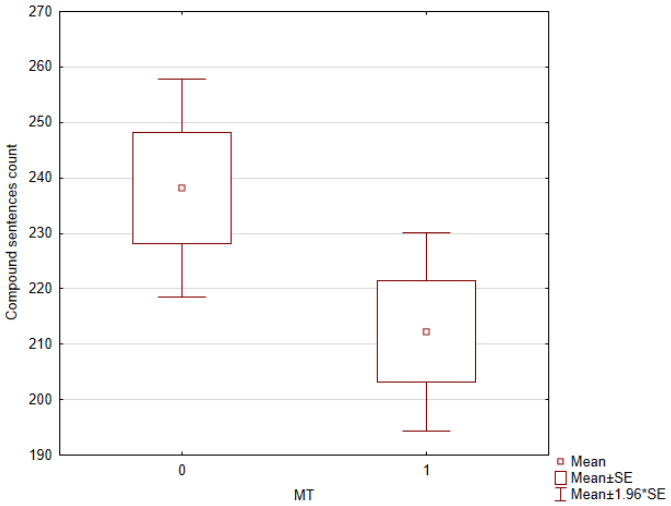
Graf 23 Počet znakov v prekladoch



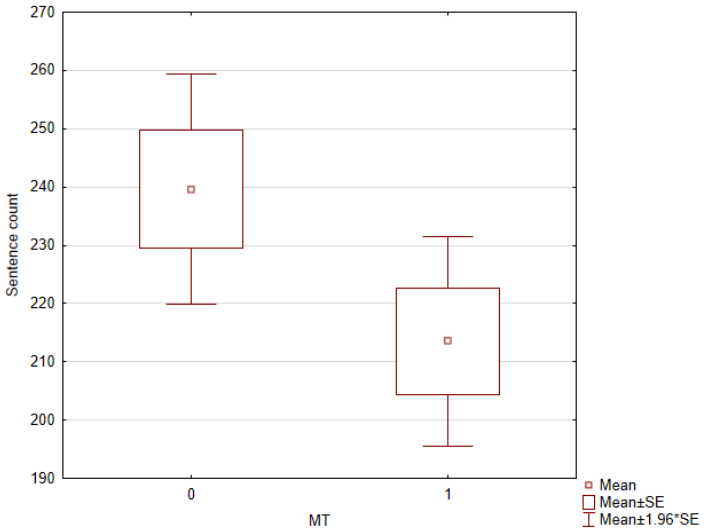
Graf 24 Počet jednoduchých viet v prekladoch

F-test preukázal štatisticky významný rozdiel vo variabilite iba pre metriku počtu jednoduchých viet v prekladoch, Graf 24, (viď prílohu *C1*), kedy *p* hodnota bola *0,01555*. Vo výsledkoch, t-testy pre tieto metriky, nepreukázali rozdiel (viď príloha *C2*).

Na Grafoch 25 a 26, môžeme vidieť rozdiel v bodových priemeroch výsledkov, kedy bodové priemery ľudských prekladov už nezasahujú do intervalového priemeru strojových prekladov, na základe čoho môžeme predpokladať, že bude prítomný štatisticky významný rozdiel medzi týmito metrikami.



Graf 25 počet zložených viet v prekladoch



Graf 26 počet viet v prekladoch

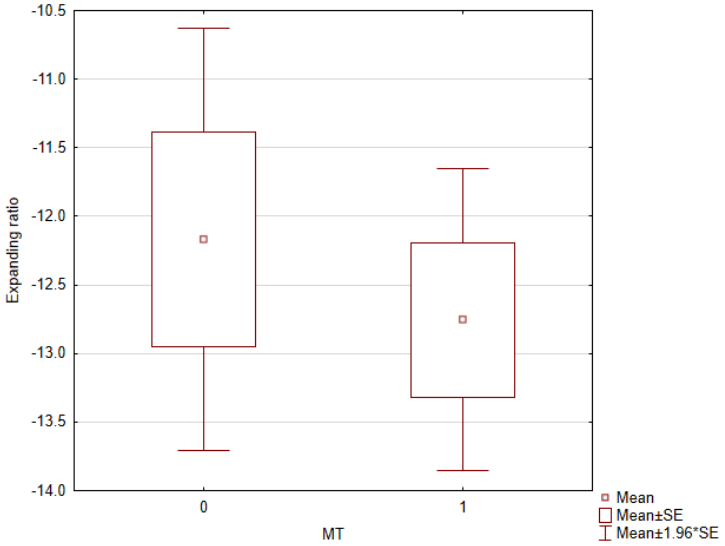
T-testy nepotvrdili náš predpoklad o prítomnosti štatisticky významného rozdielu medzi metrikami, ktoré znázorňujú Grafy 25 a 26, ale hodnota *p* sa blíži k hodnote *0,05* viď prílohu *C2*. *Hypotéza H3 sa nepotvrdila.*

H4: Predpokladáme, že strojový preklad má nižší pomer rozšírenia ako ľudský preklad.

Na overenie hypotézy *H4* sme vypočítali pomer rozšírenia (expanding ratio) pre strojový aj ľudský preklad pomocou vzorca:

$$ER = (\frac{Dĺžka\ cieľa - Dĺžka\ zdroja}{Dĺžka\ zdroja}) * 100$$

Dĺžku zdrojového a preloženého textu sme určili spočítaním počtu slov v jednotlivých častiach textu. Následne sme informácie použili na výpočet pomeru rozšírenia pre každý preklad a porovnali sme ho so zdrojovým textom.



Graf 27 Pomer rozšírenia ľudského a strojového prekladu

Na základe výsledkov je pomer rozšírenia pre ľudský aj strojový preklad záporný, čo znamená, že preklady sú stručnejšie ako pôvodný text. Priemerná miera rozšírenia strojových prekladov je o niečo nižšia ako bodový priemer pre ľudské preklady, čo naznačuje, že strojový preklad je stručnejší ako ľudský preklad. Rozdiel v priemeroch však nemusí byť štatisticky významný, keďže sa intervaly priemerov pre obe skupiny prekrývajú.

Na základe výsledkov t-testov môžeme konštatovať, že pri porovnaní oboch prekladov, neexistuje významný rozdiel medzi mierou rozšírenia strojového a ľudského prekladu. Hodnota $p = 0,546$ naznačuje, že rozdiel v priemeroch medzi oboma skupinami nie je štatisticky významný. Treba však poznamenať, že F-test s p hodnotou *0,007* preukázal, štatisticky významný rozdiel vo variabilite prekladov.

Záverom naše zistenia poukazujú na to, že strojové preklady nemusia mať nižší pomer rozšírenia ako ľudské preklady. *Hypotéza H4 sa nepotvrdila.*

5.2 Interpretácia výsledkov výskumu

Na získanie lexikálnej rôznorodosti sme použili niekoľko metrík spracovania prirodzeného jazyka, ako sú napríklad metriky TTR, HD-D, voc-D, MLTD, Simpsonov index a ďalšie. Za najlepšie metriky pri zisťovaní lexikálnej rôznorodosti považujeme metriky voc-D, HD-D a MTLD, ktorých výsledky t-testov sa na základe *p* hodnoty výrazne nelíšia. McCarthy a Jarvis (2010) metriku MTLD taktiež považovali za jednu z najlepších mier lexikálnej diverzity. Metrika Maasovho indexu sa ukázala v našom výskume ako najmenej efektívnu pri rozpoznávaní rozdielov prekladov. S cieľom zlepšiť prehľadnosť výsledkov t-testov lexikálnej rôznorodosti sme sa rozhodli rozdeliť výsledky výskumu do troch skupín a to na:

- *výsledky, kedy bol malý rozdiel medzi prekladmi,*
- *výsledky, kedy bol rozdiel veľký, ale nebol štatisticky významný,*
- *výsledky, kedy bol rozdiel medzi prekladmi štatisticky významný.*

Na základe vyššie uvedeného rozdelenia sme do prvej skupiny zaradili metriky: TTR, sTTR, GTTR, CTTR, Maasov index a hapax index, kedy hodnota *p*, ktorá vyjadruje mieru rozdielnosti medzi prekladmi, bola v rozmedzí od *1* do *0,102460*. Metriky sTTR, GTTR a CTTR majú identické hodnoty *p* pri t-testoch, čo môže byť spôsobené tým, že všetky patria do skupiny transformovaných metrík TTR. Hodnota *p* je pre tieto metriky nižšia oproti metrike TTR, čo môže potvrdzovať teóriu, že sú menej ovplyvnené dĺžkou textu ako metrika TTR.

Na základe vyššie uvedeného rozdelenia sme do druhej skupiny zaradili výsledky metrík: HD-D, MTLD, a voc-D, kedy hodnota *p* bola v rozmedzí od *0,102460* do *0,05*, čo naznačuje veľký rozdiel vo výsledkoch, ale stále nie je štatisticky významný. Treba sa však zamyslieť nad tým, že je veľká pravdepodobnosť, že by sa pri týchto metrikách rozdiel ukázal ako štatisticky významný, ak by sme mali k dispozícii väčší súbor dát.

Do tretej skupiny sme na základe vyššie uvedeného zaradili výsledky metriky Simpsonov index, kedy rozdiel vo výsledkoch bol štatisticky významný, čo svedčí o väčšej lexikálnej rozmanitosti ľudského prekladu. Štatisticky významný rozdiel sme taktiež zaznamenali vo variabilite výsledkov pre Simpsonov index. Výsledky lexikálnej rôznorodosti naznačujú, že hoci sa niektoré ukazovatele lexikálnej rozmanitosti nemusia medzi oboma prekladmi výrazne líšiť, iné môžu vykazovať významné rozdiely. Tieto zistenia môžu mať dôležité dôsledky pre štúdium lexikálnej rozmanitosti a jej vzťahu k jazykovým znalostiam.

Lexikálnu hustotu prekladov sme zisťovali pomocou pomerov kontextových slov ako sú podstatné mená, prídavné mená, príslovky a slovesá ku všetkým slovám. Výsledky lexikálnej hustoty rozdeľujeme do dvoch skupín kedy sa:

- *štatistické rozdiely potvrdili,*
- *štatistické rozdiely nepotvrdili.*

Štatisticky významné rozdiely sa nepotvrdili pri metrikách: pomer podstatných mien ku všetkým slovám, pomer slovies ku všetkým slovám, pomer prísloviek ku všetkým slovám a pomer kontextových slov ku všetkým slovám. Hodnota *p* dosahovala hodnoty do *0,25* pre uvedené metriky.

Štatisticky významné rozdiely sme pomocou t-testu zaznamenali pre metriku pomeru prídavných mien ku všetkým slovám, kedy hodnota *p* bola *0,036*. Pomocou F-testu sme rozdiely vo variabilite pre metriky spomínané vyššie nezistili. Celkové výsledky pre lexikálnu hustotu preukázali, že v ľudskom preklade bol použitý vyšší počet prídavných mien ako v strojovom preklade, čo môže naznačovať, že ľudský preklad sa snaží poskytnúť podrobnejší a opisnejší text v porovnaní so strojovým prekladom. Naše výsledky naznačujú, že môžu existovať určité rozdiely v lexikálnej hustote ľudských a strojových prekladov z hľadiska podielu prídavných mien, ale rozdiely v podstatných menách, slovesách a príslovkách nie sú štatisticky významné. Je dôležité poznamenať, že vplyv na zistené výsledky majú taktiež knižnice použité v jazyku Python, ktoré sme používali práve pri úlohách NLP a stretli sme sa s diametrálne odlišnými výsledkami pre rôzne knižnice. Tento fakt je potrebné zobrať do úvahy a bolo by žiaduce zopakovať náš výskum, ale s knižnicami, ktoré sú kompatibilnejšie so slovenským jazykom. Pri nahliadaní na výsledky sme si vedomí aj obmedzení nášho výskumu za ktoré môžeme považovať veľkosť vzorku a reprezentatívnosť použitých textov. Pre potvrdenie týchto zistení by bolo vhodné realizovať ďalší výskum.

Lexikálnu zložitosť sme zisťovali pomocou metrík početností ako sú počty: podstatných mien, prídavných mien, slovies, prísloviek, slov, znakov, jednoduchých viet, zložených viet a celkového počtu viet pre ľudský a strojový preklad. Na základe odlišností v spomínaných početnostiach sme sa snažili zistiť, či ľudský preklad je lexikálne zložitejší ako strojový preklad. F-test preukázal štatisticky významný rozdiel vo variabilite výsledkov pre metriku počtu jednoduchých viet. T-testy preukázali, že pri vyššie uvedených metrikách lexikálnej zložitosti sme nezistili štatisticky významný rozdiel pre ani jednu z nich. Treba však poznamenať, že pri metrikách počtu zložených

viet a celkového počtu viet sme zistili veľké rozdiely na základe t-testov. *P* hodnota pre tieto dve metriky bola približne *0,0578*, čo je už veľmi blízko ku štatisticky významnému rozdielu. Na základe toho predpokladáme, že by sme boli schopní zistiť štatisticky významný rozdiel v týchto metrikách, ak by sme pre náš výskum použili väčšie množstvo dát. Na druhú stranu môžeme na základe našich výsledkov taktiež konštatovať, že systém strojového prekladu použitý v našej práci dosiahol úroveň porovnateľnú s presnosťou ľudských prekladateľov, pokiaľ ide o lexikálnu hustotu. Je dôležité poznamenať, že hoci sa lexikálna hustota medzi ľudským a strojovým prekladom výrazne nelíšila, ostatné faktory, ako napríklad gramatická presnosť, môžu stále zohrávať rozhodujúcu úlohu pri určovaní kvality prekladov.

Pomer rozšírenia ľudského a strojového prekladu sme overovali pomocou F-testu a t-testov. F-test preukázal štatisticky významný rozdiel vo variabilite prekladov a ľudský preklad sa preukázal ako heterogénnejší oproti homogénnejšiemu strojovému prekladu. Domnievame sa, že dôvodom získaných výsledkov je fakt, že ľudský preklad bol opisnejší na základe predošlých testov, kedy sa potvrdil štatisticky významný rozdiel pomeru prídavný mien ku všetkým slovám a preklad celkovo obsahoval viac slov čo zvýšilo variabilitu ľudského prekladu. T-testy nepreukázali štatisticky významné rozdiely vo výsledkoch. Tým sa hypotéza, *že strojový preklad má nižší pomer rozšírenia objemu textu ako ľudský preklad, nepotvrdila*.

Na podrobnejšie preskúmanie vzťahu medzi mierou rozšírenia a kvalitou prekladu je však potrebný ďalší výskum, v ktorom by bolo použité väčšie množstvo dát a sofistikovanejšie knižnice kompatibilné so slovenským jazykom.

Na posúdenie kvality metrík je potrebné zohľadniť niekoľko aspektov, ako napríklad výkonnosť metrík, fakt, že všetky tieto metriky majú svoje silné a slabé stránky, interpretovateľnosť a kompatibilnosť metrík s inými jazykmi, ako je anglický jazyk.

Uznávame, že naša štúdia má určité obmedzenia, ako napríklad skutočnosť, že sme testovali len preklady jedného konkrétneho ľudského prekladateľa a použili sme obmedzený počet dát.

ZÁVER

Zavedenie strojového prekladu znamenalo výrazný pokrok v oblasti spracovania prirodzeného jazyka. Kvalita strojového prekladu v porovnaní s ľudským prekladom je však stále predmetom rôznych diskusií.

Hlavným cieľom našej práce bolo pomocou NLP metód porovnať kvalitu strojového a ľudského prekladu vzhľadom na lexikálnu rôznorodosť a hustotu. Výsledky našej práce preukázali, že existuje rozdiel medzi ľudským a strojovým prekladom pri použití metriky Simpsonov index pre lexikálnu diverzitu a metrika pomeru prídavných mien ku všetkým slovám pre lexikálnu hustotu vykazovala významné rozdiely medzi ľudským a strojovým prekladom. Predpokladáme však, že v prípade väčšieho súboru údajov by, metriky ako voc-D, HD-D a MTLD by tiež vykazovali významné rozdiely v prospech ľudského prekladu. Zatiaľ čo vo väčšine metrík početnosti neboli žiadne významné rozdiely, F-test ukázal rozdiely vo variabilite, čo naznačuje, že môžu existovať rozdiely medzi ľudským a strojovým prekladom, ktoré nie sú okamžite zrejmé prostredníctvom tradičných metód štatistickej analýzy. Výsledky však vo väčšine prípadov umiestňujú ľudský preklad nad strojový, aj keď štatisticky významné rozdiely sme pozorovali len pri určitých metrikách. Naše výsledky naznačujú, že ľudské preklady sú efektívnejšie pri interpretácii podstaty a komplexnosti zdrojového textu. Celkovo náš výskum zdôrazňuje dôležitosť zvažovania viacerých metrík pri hodnotení kvality prekladu. Hoci strojový preklad zaznamenal v posledných rokoch významný pokrok, stále nedokáže plne replikovať úplnú podstatu a komplexnú povahu ľudského jazyka. Ľudský preklad ako taký zostáva nevyhnutný na vytváranie vysokokvalitných prekladov, ktoré presne vyjadrujú zamýšľaný význam. Záverom náš výskum ukázal, že hoci strojové preklady majú svoje výhody, nemôžu sa vyrovnať lexikálnej rozmanitosti a hustote ľudských prekladov. Zistenia nášho výskumu majú dôležité dôsledky pre oblasť spracovania prirodzeného jazyka.

Uvedomujeme si určité limitácie našej práce, za ktoré považujeme zameranie sa iba na jednu konkrétnu kombináciu jazykov a smeru prekladu, zameranie sa na konkrétne knižnice, ktoré nemusia byť kompatibilné so slovenským jazykom do vhodnej miery a nedostatok rozsiahlych kvalitatívnych a ľudských experimentov.

Do budúcnosti je dôležité, aby výskumníci pokračovali v hľadaní spôsobov, ako zlepšiť strojové preklady, a zároveň si uvedomovali dôležitosť účasti človeka na procese

prekladu. Budúci výskum s väčšími súbormi údajov by mohol vrhnúť viac svetla na rozdiely medzi ľudským a strojovým prekladom a potenciálne pomôcť zlepšiť kvalitu strojového prekladu.

V konečnom dôsledku naše zistenia podčiarkujú dôležitosť pokračujúceho výskumu a vývoja v oblasti jazykového prekladu s cieľom dosiahnuť presnejšiu a efektívnejšiu komunikáciu medzi kultúrami a jazykmi.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

ALBEROLA, M.A., GALLEGO M.G., MAESTRE, U.G. 2019. Fundamentals of Natural Language Processing. V: *Artificial Vision and Language Processing for Robotics*, 2019. s. 356. ISBN 978-1838552268.

ARNAUD P.L. 1992. *Objective Lexical and Grammatical Characteristics of L2 Written Compositions and the Validity of SeparateComponent Tests.* [online] [cit. 2023-01-6]. Dostupné na: https://doi.org/10.1007/978-1-349-12396-4_13

BIRD, S., KLEIN, E., LOPER, E. 2009. *Natural Language Processing with Python.* [online] [cit. 2023-02-13]. Dostupné na: https://tjzhifei.github.io/resources/NLTK.pdf

BRGLEZ, M., VINTAR, Š. 2022. *Lexical Diversity in Statistical and Neural Machine Translation.* [online] [cit. 2023-02-20]. Dostupné na: https://doi.org/10.3390/info13020093

CARROLL , J. B. 1964. *Language and Thought.* [online] [cit. 2023-01-18]. Dostupné na: https://www.proquest.com/docview/1994303329?pq-origsite=gscholar&fromopenview=true

CASTILHO, S., MOORKENS, J., GASPARI, F., CALIXTO, I., TINSLEY, J., WAY, A. 2017. *Is Neural Machine Translation the New State of the Art?* [online] [cit. 2023-02-1]. Dostupné na: https://doi.org/10.1515/pralin-2017-0013

CIBULA, M. 2017. *Strojové učenie.* [online] [cit. 2023-02-13]. Dostupné na: https://smnd.sk/mcibula/zakl_info/definicia.html

ČERMÁK, F. 2007. Jazyk a jazykověda. 3. vyd. Praha: Univerzita Karlova, 2007. 382 s.

DALLER, H., van HOUT, R., TREFFERS-DALLER, J. 2003. *Lexical richness in the spontaneous speech of bilinguals.* [online] [cit. 2023-01-16]. Dostupné na: https://doi.org/10.1093/applin/24.2.197

FERGADIOTIS, G., WRIGHT, H. H., GREEN, S. 2015. *Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects*. [online] [cit. 2023-02-22]. Dostupné na: https://www.researchgate.net/publication/273464919_Psychometric_Evaluation_of_Lexical_Diversity_Indices_Assessing_Length_Effects

FREEMAN, L. D., CAMERON, L. 2008. *Complex systems and applied linguistics.* Oxford University Press: Oxford. ISBN 978-0-19-442244-4.

FRIEDENBERG, J., SILVERMAN, G. 2006. *Cognitive science: an introduction to the study of mind. Thousand Oaks, California: Sage Publications*, 2006. s. 531. ISBN 9781412925686.

GARABÍK, R., GIANITSOVÁ, L., HORÁK, ŠIMKOVÁ, M. 2004. *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu.* [online] [cit. 2023-02-10]. Dostupné na: https://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf

GARBADE, M. 2018. *A Simple Introduction to Natural Language Processing.* [online] [cit. 2023-01-15]. Dostupné na: https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

GEPPERTH, A., HAMMER, B. 2016. *Incremental learning algorithms and applications*. [online] [cit. 2023-02-12]. Dostupné na: https://www.esann.org/sites/default/files/proceedings/legacy/es2016-19.pdf

GUIRAUD, P. L. 1960. *Problèmes et Méthodes de la Statistique Linguistique.* ISBN 978-90-277-0025-4.

GUPTA, A. 2023. *Top 10 Reason Why You Should Learn Python.* [online] [cit. 2023-02-21]. Dostupné na: https://www.simplilearn.com/tutorials/python-tutorial/why-learn-python

HALLIDAY, M. A. K. 1985. *Spoken and written language.* [online] [cit. 2023-01-8]. Dostupné na: https://www.academia.edu/34763868/241798800_M_A_K_Halliday_Spoken_and_Written_Language_BookFi_org_pdf_pdf

HANCOX, J. P. 2010. *A brief history of Natural Language Processing.* [online] [cit. 2023-02-01]. Dostupné na: https://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html>.

HESS C.W., SEFTON K.M., LANDRY R.G. 1986. *Sample size and type-token ratios for oral language of preschool children.* [online] [cit. 2023-01-18]. Dostupné na: https://doi.org/10.1044/jshr.2901.129

HOTHO, A., NEURNBERGER, A., PAASS, G. 2005. *A Brief Survey of Text Mining*. [online] [cit. 2023-02-20]. Dostupné na: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.447.4161&rep=rep1&type=pdf

HUTCHINS, J. 2005. *The history of machine translation in a nutshel.* [online] [cit. 2023-02-16]. Dostupné na: http://www.hutchinsweb.me.uk/Nutshell-2014.pdf

CHEN, S. H., ANTHONY J. J., NORTON, J. P. 2008. *Artificial Intelligence techniques: An introduction to their use for modelling environmental systems*. [online] [cit. 2023-01-14]. Dostupné na: https://doi.org/10.1016/j.matcom.2008.01.028

CHOWDHARY, R.K. 2020. V : *Fundamentals of Artificial Intelligence*, 2020. s. 716. ISBN 978-8-132-23970-3.

INVESTOPEDIA. 2020. *Hlboké učenie*. [online] [cit. 2023-03-1]. Dostupné na: https://investopedia.sk/2020/10/14/hlboke-ucenie/

ISHAQ, M. 2019. *What are some of the challenges we face in NLP today?* [online] [cit. 2023-01-14]. Dostupné na: https://medium.com/datadriveninvestor/what-are-some-of-the-challenges-we-face-in-nlp-today-2e9d94da1f63

JABEEN, A., AHMAD, N. AND RAZA, K. 2018. Machine learning-based state-of-the-art methods for the classification of RNA-Seq Data. V : *Classification in BioApps*. India: Springer, 2018. s. 133-172. ISBN: 978-3-319-65981-7.

JANIESCH C., ZSCHECH, P., HEINRICH, K. 2021. *Machine learning and deep learning.* [online] [cit. 2023-01-20]. Dostupné na: https://doi.org/10.1007/s12525-021-00475-2

JARVIS, S. 2013. *Capturing the Diversity in Lexical Diversity.* [online] [cit. 2023-01-6]. Dostupné na: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2012.00739.x

JOHANSSON, V. 2008. *Lexical diversity and lexical density in speech and writing: a developmental perspective*. [online] [cit. 2023-01-11]. Dostupné na: https://journals.lub.lu.se/LWPL/article/view/2273/1848

KHURANA, D., KOLI A., KHATTER K., SINGH, S. 2017. *Natural Language Processing: State of The Art, Current Trends and Challenges*. [online] [cit. 2023-01-18]. Dostupné na: https://arxiv.org/ftp/arxiv/papers/1708/1708.05148.pdf

KHURANA, D., KOLI A., KHATTER K., SINGH, S. 2022. *Natural language processing: state of the art, current trends and challenges*. [online] [cit. 2023-02-22]. Dostupné na: https://doi.org/10.1007/s11042-022-13428-4

KOIZUMI, R., IN'NAMI, Y. 2012. *Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens.* [online] [cit. 2022-12-14]. Dostupné na: https://doi.org/10.1016/j.system.2012.10.012

KOZACZKO, D. 2018. *8 best Python Natural Language Processing (NLP) libraries*. [online] [cit. 2023-02-11]. Dostupné na: https://sunscrapers.com/blog/8-best-python-natural-language-processing-nlp/

KURDI, M. Z. 2016. Natural language processing and computational linguistics: speech, morphology and syntax. [online] [cit. 2023-02-17]. Dostupné na: https://doi.org/10.1002/9781119145554

KURDI, M. Z. 2020. *Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL*. [online] [cit. 2023-01-11]. Dostupné na: https://jdmdh.episciences.org/6674/pdf

LACLAVÍK, M., CIGLAN, M. 2006. *Dostupné zdroje a výzvy pre počítačové spracovanie informačných zdrojov v slovenskom jazyku.* [online] [cit. 2023-01-2]. Dostupné na: https://pdfslide.net/documents/dostupne-zdroje-avyzvy-pre-pocitacove-spracovanie-informacnych-zdrojov.html

LAUFER B., NATION P. *Vocabulary size and use: lexical richness in L2 written production.* [online] [cit. 2023-02-24]. Dostupné na: https://doi.org/10.1093/applin/16.3.307

LIDDY, E. D. 2001. Natural language processing. V : *Encyclopedia of Library and Information Science*. [online] [cit. 2023-03-12]. Dostupné na: https://surface.syr.edu/cgi/viewcontent.cgi?article=1019&context=cnlp>.

MALVERN D., RICHARDS, B. 2012. *Measures of Lexical Richness.* [online] [cit. 2023-02-24]. Dostupné na: https://doi.org/10.1002/9781405198431.wbeal0755

MCCARTHY, P. M., JARVIS, S. 2007. *vocd: A theoretical and empirical evaluation*. [online] [cit. 2023-02-18]. Dostupné na: 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.* [online] [cit. 2023-02-2]. Dostupné na: https://link.springer.com/article/10.3758/BRM.42.2.381

MCFARLANDA, A. 2022. *10 Best Python Libraries for Natural Language Processing*. [online] [cit. 2023-02-10]. Dostupné na: https://www.unite.ai/10-best-python-libraries-for-natural-language-processing/

MCKINNEY, W. 2022. *powerful Python data analysis toolkit.* [online] [cit. 2023-02-11]. Dostupné na: https://pandas.pydata.org/pandas-docs/version/1.4.4/pandas.pdf

MILLER, T. 2019. *Explanation in artificial intelligence: Insights from the social sciences.* [online] [cit. 2023-02-17]. Dostupné na: https://doi.org/10.1016/j.artint.2018.07.007

MILLWARD, M.C., HAYES, M. 2011. *A Biography of the English Language*. 3th ed. Wadsworth: Cengage learning, 2011. s. 496. ISBN 978-0495906414.

MUNKOVÁ, D., HAJEK, P., MUNK, M., SKALKA, J. 2020. Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy. V : Procedia Computer Science. [online] [cit. 2023-03-24]. Dostupné na: https://www.sciencedirect.com/science/article/pii/S1877050920311212

NAKONEČNÝ, M. 1998. *Základy psychologie*. Praha: Academia, 1998. s. 590. ISBN 8020006893.

NAVIGLI, R. 2018. *Natural Language Understanding: Instructions for (Present and Future) Use.* [online] [cit. 2023-02-22]. Dostupné na: https://iris.uniroma1.it/bitstream/11573/1172143/1/Navigli_Natural.pdf

NORDQUIST, R. 2020. *What is Natural Language?* [online] [cit. 2023-02-16]. Dostupné na: https://www.thoughtco.com/what-is-a-natural-language-1691422

PARALIČ, J. a kol. 2010. *Dolovanie znalostí z textov.* Košice: Equilibria, s.r.o., 2010. s. 182. ISBN 9788089284627.

PEDREGOSA, F. a kol. 2011. *Scikit-learn: Machine learning in Python*. [online] [cit. 2023-03-24]. Dostupné na: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

POPESCU, I., ALTMANN, G. 2008. *Hapax Legomena and Language Typology.* [online] [cit. 2023-02-3]. Dostupné na: *https*://doi.org/10.1080/09296170802326699

POURPANAH, F., WANG, R. LIM, C. P., WANG, X., SEERA, M., TAN, C. J. 2019. *An improved fuzzy ARTMAP and Q-learning agent model for pattern classifcation.* [online] [cit. 2022-12-14]. Dostupné na: https://doi.org/10.1016/j.neucom.2019.06.002

QI, P. a kol. 2020. *Stanza : A Python Natural Language Processing Toolkit for Many Human Languages.* [online] [cit. 2023-01-11]. Dostupné na: https://arxiv.org/pdf/2003.07082.pdf

RAKESH, H. 2018. *Natural Language Processing.* [online] [cit. 2023-03-24]. Dostupné na: https://becominghuman.ai/natural-language-processing-in-a-nutshell-a784b9fea849

RICHARDS, B. 1987. *Type/Token Ratios: what do they really tell us?* [online] [cit. 2023-01-6]. Dostupné na: https://doi.org/10.1017/S0305000900012885

ROMANYSHYN, M. 2019. *Linguistics In NLP: Why So Complex?* [online] [cit. 2023-02-20]. Dostupné na: https://odsc.medium.com/linguistics-in-nlp-why-so-complex-4fd1fb9873a3

SHEN, L. 2021. *Measuring Political Media Slant Using Text Data.* [online] [cit. 2023-02-25]. Dostupné na: https://www.lucasshen.com/research/media.pdf

SCHANK, R. 1969. *A conceptual dependency parser for natural language*. [online] [cit. 2020-12-19]. Dostupné na: https://www.aclweb.org/anthology/C69-0201.pdf

SILVER D. L. 2011. Machine lifelong learning: challenges and benefits for artifcial general intelligence. V: *International conference on artifcial general intelligence*, s. 370–375. ISBN 978-3-642-22886-5

SIMPSON, E. H. 1949. *Measurement of Diversity.* [online] [cit. 2023-01-18]. Dostupné na: https://www.nature.com/articles/163688a0

STAŠ, J., HLÁDEK, D., JUHÁR, J. 2015. Language model speaker adaptation for transcription of Slovak parliament proceedings. V : *Speech and Computer*, 2015. [online] [cit. 2023-03-25]. Dostupné na: https://www.researchgate.net/publication/281774833_Language_Model_Speaker_Adaptation_for_Transcription_of_Slovak_Parliament_Proceedings

ŠIŠKOVÁ, Z. 2012. *Lexical Richness in EFL Students' Narratives.* [online] [cit. 2023-01-18]. Dostupné na: *https://www.researchgate.net/publication/305999633_Lexical_Richness_in_EFL_Students'_Narratives*

TEMPLIN, M. C. 1957. *Certain Language Skills in Children: Their Development and Interrelationships (NED-New edition, Vol. 26). University of Minnesota Press*. [online] [cit. 2023-02-12]. Dostupné na: http://www.jstor.org/stable/10.5749/j.ctttv2st

TORAL, A. 2019. *Post-editese: An Exacerbated Translationese.* [online] [cit. 2023-02-16]. Dostupné na: researchgate.net/publication/334161205_Post-editese_an_Exacerbated_Translationese

VANMASSENHOVE, E., SHTERIONOV, D.S., GWILLIAM, M. 2021. *Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation*. [online] [cit. 2023-02-15]. Dostupné na: https://aclanthology.org/2021.eacl-main.188.pdf

WANG, Z. WANG, X. 2018. *A deep stochastic weight assignment network and its application to chess playing*. [online] [cit. 2023-02-13]. Dostupné na: https://doi.org/10.1016/j.jpdc.2017.08.013

YOUNG, T., HAZARIKA, D., PORIA, S., CAMBRIA, E. 2018. *Recent trends in deep learning based natural language processing*. [online] [cit. 2023-02-13]. Dostupné na: https://arxiv.org/pdf/1708.02709.pdf

ZOZNAM PRÍLOH

Príloha A

- *A1*

Variable	F-test rôznorodosť Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
TTR	69	69	4.34291	4.73127	1.186847	0.481887
sTTR	69	69	3.43570	3.18638	1.162614	0.536140
GTTR	69	69	343.56973	318.63769	1.162614	0.536140
CTTR	69	69	242.94048	225.31087	1.162614	0.536140
voc-D	69	69	41.33378	43.34685	1.099777	0.696073
HD-D	69	69	0.01206	0.01416	1.379829	0.186911
MTLD	69	69	91.74115	89.83098	1.042980	0.862756
Maas	69	69	0.00134	0.00145	1.174439	0.509146
Hapax count	69	69	333.58283	302.62344	1.215073	0.423931
Simpsons index	69	69	0.00235	0.00305	1.682210	0.033605

- *A2*

Variable	T-tests-rôznorodosť Mean 0	Mean 1	t-value	df	p	t separ. var.est.	df	p 2-sided
TTR	47.85206	47.22187	0.815088	136	0.416447	0.815088	135.0144	0.416458
sTTR	28.73451	28.23453	0.886322	136	0.377009	0.886322	135.2354	0.377018
GTTR	2873.45098	2823.45289	0.886322	136	0.377009	0.886322	135.2354	0.377018
CTTR	2031.83667	1996.48269	0.886322	136	0.377009	0.886322	135.2354	0.377018
voc-D	249.37604	237.01639	1.714110	136	0.088787	1.714110	135.6936	0.088792
HD-D	0.92559	0.92191	1.644413	136	0.102401	1.644413	132.6217	0.102459
MTLD	327.41406	301.90605	1.650224	136	0.101205	1.650224	135.9398	0.101206
Maas	0.01100	0.01109	-0.356658	136	0.721901	-0.356658	135.1304	0.721904
Hapax count	1381.71014	1350.89855	0.568254	136	0.570800	0.568254	134.7298	0.570809
Simpsons index	0.98965	0.98782	3.943522	136	0.000128	3.943522	127.7365	0.000132

Príloha B

- *B1*

Variable	F-test-ER Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
NOUN/total words ratio	69	69	0.023114	0.025374	1.205118	0.443733
VERB/total words ratio	69	69	0.009048	0.009683	1.145193	0.577720
ADJ/total words ratio	69	69	0.010337	0.009152	1.275542	0.318018
ADV/total words ratio	69	69	0.001520	0.001738	1.306368	0.272974

- *B2*

Variable	T-tests-hustota Mean 0	Mean 1	t-value	df	p	t separ. var.est.	df	p 2-sided
NOUN/total words ratio	0.613767	0.609490	1.03494	136	0.302535	1.03494	134.8333	0.302551
VERB/total words ratio	0.079149	0.080981	-1.14788	136	0.253032	-1.14788	135.3798	0.253041
ADJ/total words ratio	0.098702	0.095188	2.11441	136	0.036305	2.11441	134.0347	0.036332
ADV/total words ratio	0.006230	0.006181	0.17721	136	0.859608	0.17721	133.6418	0.859612

Príloha C

- *C1*

Variable	F-Test-početnosti Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
Noun count nltk	69	69	627.201	568.606	1.216719	0.420723
Adjective count nltk	69	69	106.768	94.062	1.288419	0.298506
Verb count nltk	69	69	77.370	82.461	1.135936	0.600688
Adverb count nltk	69	69	8.576	9.238	1.160439	0.541213
Word count	69	69	987.404	924.202	1.141447	0.586940
Char count	69	69	6290.998	5702.185	1.217185	0.419817
Simple sentences count	69	69	2.266	1.685	1.809703	0.015553
Compound sentences count	69	69	83.213	75.846	1.203689	0.446633
Sentence count	69	69	83.663	76.083	1.209191	0.435549

- *C2*

Variable	T-tests-početnosti Mean 0	Mean 1	t-value	df	p	t separ. var.est.	df	p 2-sided
Noun count nltk	2279.26	2240.30	0.382241	136	0.702879	0.382241	134.7124	0.702885
Adjective count nltk	367.51	351.16	0.954340	136	0.341605	0.954340	133.8735	0.341632
Verb count nltk	292.42	298.22	-0.425864	136	0.670880	-0.425864	135.4514	0.670883
Adverb count nltk	23.57	23.26	0.200563	136	0.841340	0.200563	135.2541	0.841341
Word count	3711.48	3679.62	0.195651	136	0.845175	0.195651	135.4092	0.845176
Char count	22545.28	22142.10	0.394434	136	0.693878	0.394434	134.7075	0.693884
Simple sentences count	1.49	1.32	0.511586	136	0.609770	0.511586	125.5715	0.609839
Compound sentences count	238.13	212.23	1.910688	136	0.058149	1.910688	134.8479	0.058167
Sentence count	239.62	213.55	1.915146	136	0.057573	1.915146	134.7914	0.057592

Príloha D

- *D1*

Variable	F-test-ER Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
Expanding ratio	69	69	6.521850	4.668863	1.951279	0.006503

- *D2*

Variable	T-tests-ER Mean 0	Mean 1	t-value	df	p	t separ. var.est.	df	p 2-sided
Expanding ratio	-12.1697	-12.7545	0.605642	136	0.545762	0.605642	123.2001	0.545867