UNIVERZA V LJUBLJANI
SKUPNI INTERDISCIPLINARNI PROGRAM DRUGE STOPNJE
KOGNITIVNA ZNANOST

V SODELOVANJU Z UNIVERSITÄT WIEN,
UNIVERZITA KOMENSKÉHO V BRATISLAVE
IN EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Sabina Gorenc

NENADZOROVANO UČENJE ZA AVTOMATSKO POENOSTAVLJANJE BESEDIL

Magistrsko delo

Ljubljana, 2022

UNIVERZA V LJUBLJANI
SKUPNI INTERDISCIPLINARNI PROGRAM DRUGE STOPNJE
KOGNITIVNA ZNANOST

V SODELOVANJU Z UNIVERSITÄT WIEN,
UNIVERZITA KOMENSKÉHO V BRATISLAVE
IN EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Sabina Gorenc

NENADZOROVANO UČENJE ZA AVTOMATSKO POENOSTAVLJANJE BESEDIL

Magistrsko delo

Mentor: prof. dr. Marko Robnik-Šikonja
Somentor: prof. dr. Marko Stabej

Ljubljana, 2022

Povzetek

Za povečanje dostopnosti in raznovrstnosti lahkega branja v slovenščini, ki vsebuje jezikovno prilagojena besedila, smo izdelali prototip sistema, ki avtomatsko poenostavlja besedila. To je prvi sistem za samodejno pretvarjanje slovenskih povedi in besedil v enostavnejšo obliko. Pripravili smo podatkovno množico za slovenski jezik s poravnanimi enostavnimi in kompleksnimi stavki, ki bo uporabna za nadaljnje raziskave. Uporabili smo model T5 za slovenski jezik, ki je naučen na drugih nalogah s področja naravne obdelave jezika. Model uporablja strojno učenje s prenosom znanja na globokih nevronskih mrežah z arhitekturo kodirnik-dekodirnik. Za iskanje optimalnih vrednosti hiperparametrov in evalvacijo uspešnosti sistema smo uporabili avtomatske mere ROUGE in BERTScore, ki so dokaj visoke in kažejo na uspešnost sistema. Sistem generira enostavčne ali enostavne večstavčne povedi s preprostimi priredji in podredji in ne uporablja trpnika ali posebnih simbolov. S stališča skladenjske preprostosti je sistem uspešen, bolj podrobno pa smo njegovo uspešnost ocenili še s pomočjo človeške evalvacije z uporabo vprašalnika, ki bi se ga lahko uporabilo za preverjanje razumljivosti in smiselnosti avtomatsko zgeneriranih stavkov tudi v nadaljnjih študijah. Z vprašalnikom smo ugotovili, da model ni preveč uspešen pri tvorjenju smiselnih in razumljivih odstavkov. Večina ocenjevalcev je menila, da so skoraj ali čisto nerazumljivi. Raziskovali smo še kriterije razumljivosti za avtomatsko generirana besedila in ugotovili, da so pomembni kriteriji razumljivosti jedrnatost, jezikovna pravilnost, leksikalna preprostost, skladenjska preprostost, koherenca in povzemalna ustreznost. Določitev kriterijev razumljivosti za avtomatsko generirana besedila je pomemben doprinos k nadaljnjemu razvoju in evalvaciji modelov avtomatskega poenostavljanja besedil, saj omogočajo objektivno oceno razumljivosti takih besedil. Naš sistem se je najboljše odrezal po kriterijih skladenjske in leksikalne preprostosti, najslabše pa v povzemalni ustreznosti, koherenci in jedrnatosti. Sistem je delno uporaben kot pomoč poenostavljalcem, potencialno pa bi se ga dalo izkoristiti v kombinaciji s povzemanjem za zagotavljanje preprostejšega besedišča in preproste skladenjske strukture.

Ključne besede: obdelava naravnega jezika, poenostavljanje besedil v slovenščini, lahko branje, globoke nevronske mreže, model zaporedje v zaporedje, T5 model, razumljivost besedil, kriteriji razumljivosti.

Abstract

In order to increase the accessibility and variety of easy reading in Slovenian, which contains stylistic and language adaptations, we created a prototype of a system that automatically simplifies texts. This is the first system for automatically converting Slovenian sentences and texts into a simpler form. We have prepared a dataset for the Slovenian language that contains aligned simple and complex sentences, which can be used for further development of models for simplifying texts in Slovenian. We used the slovene T5 model, which is pretrained on other tasks. Namely, the model uses machine learning with knowledge transfer using deep neural networks with an encoder-decoder architecture. To find good values of hyperparameters and evaluate the performance of the system, we used automatic measures ROUGE and BERTScore, which are high and indicate a good performance of the system. The system generates single-clause or simple multi-clause sentences and does not use adverbs or special symbols. From the syntactic simplicity point of view, the system is successful, but we assessed its success in more detail with the help of human evaluation using a questionnaire that could be used to check the comprehensibility and meaningfulness of automatically generated sentences in further studies. With the questionnaire, we found that the model was not successful in generating comprehensible paragraphs. Most reviewers found them to be almost or completely unintelligible. We also investigated the comprehensibility criteria for automatically generated texts and found that the important comprehensibility criteria are conciseness, linguistic correctness, lexical simplicity, syntactic simplicity, coherence and summary relevance. Our system performed the best in syntactic simplicity and lexical simplicity, and the worst in summary relevance, coherence and conciseness. The system is partly useful as an aid to simplifiers, and could potentially be used in combination with summarization to provide simpler vocabulary and simple syntactic structure.

Keywords: natural language processing, text simplification in Slovene, easy reading, deep neural networks, sequence-to-sequence model, T5 model, text comprehensibility, comprehensibility criteria.

Kazalo

1 Uvod

Za ljudi, ki zaradi različnih kognitivnih ali socialno-kulturnih razlogov težko razumejo standardna besedila, je potrebno besedila poenostavljati. Cilj poenostavljanja je iz kompleksnih ustvariti enostavnejša besedila, ki so lažje dostopna in razumljiva, pri tem pa ohranijo najpomembnejšo originalno vsebino in pomen. Taka besedila imajo besedoslovne, oblikoslovne, skladenjske in oblikovne prilagoditve (slednjih se v tej nalogi ne bomo dotikali). Z leksikalnega stališča zahtevne besede in tujke zamenjamo s preprostejšimi sinonimi. Zahtevne besede so v primerjavi s preprostejšimi navadno daljše in tvorjene ter se manjkrat pojavijo v korpusu. S sintaktičnega stališča zgradbo stavkov poenostavimo tako, da povedi razbijemo na krajše enostavčne povedi, spustimo manj pomembne informacije in po potrebi pomembne informacije parafraziramo (Haramija in Knapp, 2019). V Sloveniji se s poenostavljanjem besedil aktivno ukvarjata Zavod Risa, kjer so ročno izdelali nekaj prirejenih besedil (Risa, 2020a) in MMC RTV SLO, ki poenostavljena besedila objavlja na novičarskem portalu https://www.rtvslo.si/dostopno/lahko-branje/?. Da bi postopek pospešiliax in povečali število priredb v lahko branje, želimo s pristopi obdelave naravnega jezika izdelati prototip sistema, ki bo avtomatsko poenostavljal besedila. Ker tak sistem za slovenščino še ne obstaja, bi s tem postavili osnovo za izdelavo orodja, ki bi se uporabljalo za pripravo gradiv v lažje berljivih oblikah, didaktičnih gradiv in gradiv za učenje slovenščine kot tujega jezika. To bi tudi olajšalo dostop do informacij v obliki, ki je ljudem, ki imajo težave z branjem in s pisanjem, razumljiva, jih izobraževalo in opismenjevalo ter prispevalo k njihovemu aktivnejšemu vključevanju v družbo.

Ker je ročno poenostavljanje besedil zamudno in drago, se v tehnološko dobro podprtih jezikih v ta namen uporabljajo avtomatizirana orodja. Obdelava naravnega jezika je področje umetne inteligence, ki se ukvarja z obdelavo besedil, napisanih v človeških oz. naravnih jezikih. Poleg avtomatskega poenostavljanja besedil so tipične naloge s tega področja razpoznavanje govora, strojno prevajanje, sistemi za odgovarjanje na vprašanja, klasifikacija dokumentov, tekstovno rudarjenje (ki odkriva znanje v tekstovnih podatkovnih bazah), generiranje in povzemanje besedil ter prilagajanje besedila glede na kontekst. Te naloge so med seboj prepletene: npr. pri poenostavljanju besedil želimo besedila tudi povzemati, kar lahko naredimo z ekstraktivnim povzemanjem, ki kopira pomembne stavke iz originalnega besedila, ali z abstraktivnim povzemanjem, ki v povzetek vključi tudi nove stavke in besede (Kryściński, Paulus, Xiong in Socher, 2008). Avtomatskega poenostavljanja besedil se lahko lotimo tudi s strojnim prevajanjem iz kompleksnega v enostavni jezik. Obstaja tudi pristop, ki zajame več vrst nalog obdelave besedil naenkrat. Imenuje se strojno učenje s prenosom znanja (angl. transfer learning). Ta vrsta učenja uporabi predhodno naučeno znanje pri reševanju enega tipa nalog za učenje naloge drugega tipa.

Danes se v namen poenostavljanja besedil večinoma uporabljajo globoke nevronske mreže, zgodovinsko pa so se uporabljali pristopi PBMT (angl. Phrase-based machine translation), TBMT (angl. Tree-based machine translation) ali SBMT (angl. Syntax-based machine translation). Med modeli globokih nevronskih mrež za poenostavljanje besedil prevladujejo modeli vrste kodirnik – dekodirnik z arhitekturo zaporedje v zaporedje (Aprosio, Tonelli, Turchi, Negri in Di Gangi, 2019). Ti modeli so večinoma preizkušeni le na angleškem jeziku (Štajner in Saggion, 2018) in nekaj jezikih, ki imajo na voljo dovolj virov, kot so kitajščina, francoščina in nemščina. V teh jezikih je zaradi velikih učnih množic omogočena primerjava različnih modelov in posledično je razvoj različnih meto-

dologij avtomatskega poenostavljanja besedil v teh jezikih hiter. Veliki jezikovni modeli so vnaprej naučeni na nenadzorovan oz. samonadzorovan način ter tako pridobijo obsežno jezikovno znanje. Ti modeli so nato prilagojeni specifičnim nalogam, kot je poenostavljanje.

V slovenskem jeziku učna množica za problem poenostavljanja besedil še ne obstaja, niti še ni raziskano, v kakšni meri so modeli kodirnik – dekodirnik z arhitekturo zaporedje v zaporedje za poenostavljanje besedil, ki temeljijo na angleškem jeziku, uporabni za slovenski jezik. Sprva smo preizkusili direktno nenadzorovano učenje po zgledu Surya, Mishra, Laha, Jain in Sankaranarayanan (2019), vendar neuspešno, poleg tega so trenutni pristopi z modeli tipa kodirnik - dekodirnik vrste T5 (Raffel idr., 2021) mnogo uspešnejši pri generativnih nalogah. Z nastankom SloT5 (Ulčar in Robnik-Šikonja, 2022) se je pojavila možnost, da ta pristop preizkusimo.

Uporabnost modela določimo z analizo zgeneriranih poenostavljenih stavkov oz. besedil, kjer lahko poleg umetne inteligence s pridom uporabljamo znanje jezikoslovja. Prva disciplina omogoča analizo z objektivnimi avtomatskimi merami, druga pa subjektivno presojo z vidika bralca. Bolj kot so ustvarjeni stavki oz. besedila razumljivi in bolj kot ohranjajo bistveno vsebino, za bolj uspešen lahko proglasimo model.

Slovensko jezikoslovje se je vprašanja razumljivosti na podlagi medsebojnega vplivanja med bralcem in besedilom začelo dotikati npr. z delom Ferbežar (2009), vsekakor pa še ni raziskano tako kot v nemškem jezikoslovju. Iz tam izhaja več modelov razumljivosti, ki poudarjajo pomen medsebojnega vplivanja med bralcem in besedilom. Modeli na razumljivost ne gledajo kot na enovito lastnost, temveč vpeljujejo različne dejavnike razumljivosti. Ker se razumljivost lahko od jezika do jezika razlikuje (Osolnik Kunc, 2007), nas bodo zanimale predvsem raziskave razumljivosti za slovenski jezik. Model razumljivosti za slovenski jezik je postavila Ferbežar (2012) s sedmimi kriteriji razumljivosti:

- jedrnatost,
- koherenca oz. logična povezanost besedila v smiselno celoto,
- upoštevanje razumevalca,
- leksikalna preprostost,
- skladenjska preprostost,
- jezikovna pravilnost,
- komunikativnost besedila.

Empirične raziskave razumljivosti so potekale na različnih tipih slovenskih besedil, kot so slovenska poročevalna besedila iz slovenskih dnevnikov (Leskovec, 2009), besedila v javni in uradni komunikaciji (Ferbežar, 2009), navodila za uporabo zdravil (Brinovec in Krečan, 2010), pravna besedila (Ambrožič, 2019) ipd. Na slovenski leposlovni literaturi razumljivost še ni raziskana.

Pri merjenju razumljivosti besedil je potrebno upoštevati specifičnost besedilne zvrstnosti (Ferbežar, 2012). Nas bodo zanimale posebnosti leposlovne literature, avtomatsko generiranih besedil ter posebnosti besedil v lahki obliki. Pričakujemo lahko, da generirani stavki ne bodo nujno napisani slovnično pravilno, da ne bodo nujno vsebovali smiselne ali

pravilne vsebine ter da ne bodo nujno pravilno povzeli prenesenega pomena (v metaforah, frazeologiji in ironiji), kar bo oteževalo razumljivost besedila. Pri besedilih v lahki obliki želimo, da so povedi kratke in pretežno enostavčne, ko pa so povedi večstavčne, morajo biti enostavne s preprostimi priredji in podredji, pazimo, da ne uporabljamo nikalnih povedi ali trpnika, da so isti pomeni vedno enako poimenovani, tako da ne uporabljamo sinonimov, namesto tujk uporabljamo slovenske izraze, namesto prenesenih pomenov in prispodob pa konkretna poimenovanja. Prav tako ne uporabljamo večmestnih števil in posebnih simbolov, kot so odstotki, podpičja, tropičja, narekovaji ipd.

V tem delu bomo testirali hipotezo, da je možno po vzoru T5 modela (Raffel idr., 2020) z manjšo podatkovno množico s skoraj tisoč poravnanimi stavki v kompleksni in poenostavljeni obliki ter z uporabo globokih nevronskih mrež zgraditi sistem, ki bo uporabno poenostavljal besedila v slovenščini. Podvprašanja, ki nas bodo zanimala, so:

1. Ali lahko z modelom, ki je uspešen za angleški jezik, tvorimo smiselne poenostavljene slovenske stavke in besedila?

2. V kakšni meri bodo stavki in besedila, zgenerirani z našim sistemom, razumljivi bralcem? Pri tem smo si na dveh z našim sistemom avtomatsko generiranih odstavkih in njima pripadajočima ročnima poenostavitvama zastavili tri ničelne hipoteze:

   (a) H_0: Prvo ročno in prvo avtomatsko besedilo sta enako razumljivi.

   (b) H_0: Drugo ročno in drugo avtomatsko besedilo sta enako razumljivi.

   (c) H_0: Prvo avtomatsko in drugo avtomatsko besedilo sta enako razumljivi.

3. Kateri so kriteriji za razumljivost avtomatsko generiranih poenostavljenih besedil v slovenščini?

Magistrsko delo je sestavljeno iz šestih poglavij. Drugo poglavje vsebuje opis in delovanje nevronskih, predvsem globokih nevronskih mrež, arhitekturo zaporedje v zaporedje in modele tipa kodirnik – dekodirnik. Vsebuje pregled obstoječih raziskav z globokimi nevronskimi mrežami ter pregled raziskav o razumljivosti besedil. Tretje poglavje predstavi podatkovno množico, opisuje metodologijo, ki jo uporabimo za razvoj modela za poenostavljanje slovenskih besedil ter predstavi postopek evalvacije modela. Razloži, katere hiperparametre smo izbrali za naš model glede na avtomatske mere ter predstavi vprašalnik. V četrtem poglavju so predstavljeni rezultati evalvacije modela z avtomatskima merama ROUGE in BERTScore. Predstavimo tudi človeško evalvacijo s pomočjo analize objektivnih lastnosti stavkov in vprašalnika, ki meri razumljivost poenostavljenih odsekov besedil, zgeneriranih z našim modelom. Peto poglavje je namenjeno zaključkom naše raziskave.

2 Izhodišča in pregled obstoječih raziskav

V devetdesetih letih prejšnjega stoletja so se raziskovalci lotili poenostavljanja, tako da so določili slovnična pravila poenostavljanja, vendar je tak način vezan na posamezni jezik in je časovno potraten (Štajner in Saggion, 2018). Z nastankom vzporednih korpusov, ki vsebujejo originalne in poenostavljene stavke, so se uveljavile podatkovne tehnike poenostavljanja. Uporabljajo se tudi hibridne tehnike, ki so osnovane na pravilih in podatkovnih bazah. V zadnjih letih so na področju poenostavljanja prevladale globoke nevronske mreže (Štajner in Saggion, 2018), ki preko hierarhične arhitekture in numerične predstavitve besed v visokodimenzionalnem vektorskem prostoru učinkoviteje obvladujejo kompleksnost jezika (Raaijmakers, 2019). Da zagotovimo dovolj veliko zmogljivost nevronskih mrež, ki sočasno procesirajo veliko število podatkov, uporabljamo grafične procesorje (imenovane tudi grafične procesne enote – GPU).

V tem poglavju bomo predstavili nadzorovano, nenadzorovano in spodbujevalno učenja, kaj so globoke nevronske mreže in kako delujejo, podrobneje bomo pregledali modele zaporedje v zaporedje na modelu kodirnik – dekodirnik ter različne modele globokih nevronskih mrež na modelih zaporedje v zaporedje. Nadaljevali bomo s pregledom raziskav razumljivosti na slovenskih besedilih ter analizirali, kak pristop je trenutno najprimernejši za raziskovanje razumljivosti avtomatsko poenostavljenih besedil.

2.1 Osnove strojnega učenja

Algoritme strojnega učenja lahko glede na lastnosti učnih podatkov kategoriziramo v več skupin:

- Nadzorovano učenje (angl. supervised learning): Algoritmi med učenjem uporabljajo učne podatke, ki so opremljeni z oznako oz. s ciljnim razredom, kar pomeni, da so vnaprej na voljo pravilni odgovori, naloga algoritma strojnega učenja pa je najti model, ki jih napoveduje na podlagi vhodnih podatkov. Za predvidevanje pravilnih razredov se mora model naučiti klasificirati (razvrščati) podatke v vnaprej določene razrede ali jim pripisovati številske vrednosti.

- Nenadzorovano učenje (angl. unsupervised learning): Algoritmi uporabljajo učne podatke, ki nimajo oznak ali vnaprej določenih pravilnih izhodnih vrednosti, naloga modela je ugotoviti strukturo podatkov in odkriti še neznane povezave med podatki. Takšne naloge so npr. gručenje, ki razdeli podatke v smiselne gruče podobnih primerov, ki se razlikujejo od podatkov v drugih gručah. Drugi primer naloge iz nenadzorovanega učenja je vizualizacija, pri kateri so si podobni elementi blizu, različni elementi pa so med seboj bolj oddaljeni. Modela ne moremo zgraditi tako, da bi se ujemal s pravilnimi odgovori v učnih podatkih, zato kakovosti takih modelov ne moremo optimizirati.

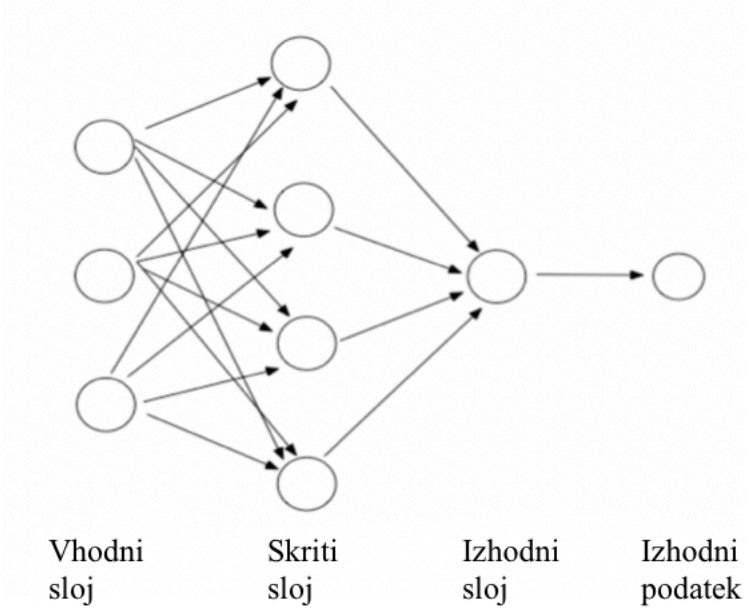
- Spodbujevano učenje (angl. reinforcement learning): uporablja povratno informacijo v obliki številske nagrade ali kazni. Uporablja se npr. v igrah, pri katerih se lahko izid določi šele na koncu.

Obstaja tudi delno nadzorovano učenje, ki je kombinacija obojega.

Pri nalogi poenostavljanja besedil se avtorji odločajo za različne vrste učenja, nekateri uporabijo nadzorovano, nekateri nenadzorovano in drugi delno nadzorovano učenje. Poleg različnih vrst učenja se avtorji odločajo tudi za modele globokih nevronskih mrež z različnimi arhitekturami.

2.1.1 Pregled avtomatskega poenostavljanja besedil z globokimi nevronskimi mrežami

Nevronska mreža je matematični model, katerega osnovna ideja temelji na delovanju bioloških nevronskih mrež. Sestavljena je iz množice medsebojno povezanih osnovnih gradnikov umetnih nevronov ter uteženih povezav med njimi. Skupek istovrstnih nevronov, ki so medsebojno nepovezani, imenujemo sloj, zaporedje medsebojno povezanih slojev pa model nevronske mreže (prikazan na Sliki 1). Vsaka umetna nevronska mreža ima vhodni in izhodni sloj, ostali sloji, ki niso vhodni ali izhodni, se imenujejo skriti sloji.



Slika 1 – Prikaz modela nevronske mreže z vhodnim slojem, skritim slojem in izhodnim slojem nevronov, ki poda izhodni podatek. Slika povzeta po Raaijmakers (2019).

Uteži pri osnovnem modelu umetnega nevrona vključujejo nabor prilagodljivih parametrov in se uporabljajo kot množitelji vhodnih podatkov nevrona, ki se seštejejo. Vsaka utež je sprva naključno določena, nato pa se v procesu učenja poveča ali zmanjša. Vsota zmnožkov uteži in vhodnih podatkov se imenuje linearna kombinacija vhodnih podatkov. Ko je ta izračunana, nevron linearno kombinacijo pošlje skozi tako imenovano aktivacijsko funkcijo in s tem izračuna izhod. Učenje ali prilagajanje v mreži se zgodi, ko se uteži prilagodijo tako, da mreža daje pravilne izhodne podatke. Aktivacijska funkcija naj bi bila v splošnem nelinearna. Med pogosteje uporabljenimi aktivacijskimi funkcijami so:

- Linearna funkcija (identiteta): $f(x) = x$ se navadno uporablja na vhodnem sloju.
- Sigmoidna funkcija (sigmoid): $f(x) = \frac{1}{1+e^{-x}}$ se navadno uporablja pri binarnih klasifikacijskih problemih.

- Funkcija hiperbolični tangens: $f(x) = \tanh(x)$ predstavlja razširjeno obliko sigmoidne funkcije.
- Popravljena linearna enota (angl. rectified linear units, ReLU): $f(x) = \max(0, x)$ predstavlja preprosto nelinearno preslikavo, ki navzdol omejuje linearno funkcijo. Je ena izmed najpriljubljenejših aktivacijskih funkcij zaradi enostavne implementacije, časovne učinkovitosti ter odpornosti na pojav ničelnega gradienta.
- Softmax funkcija: $f(x)_i = \frac{e^{x_i}}{\sum_{m=1}^{M} e^{x_m}}$ se uporablja predvsem na izhodnih slojih pri reševanju večrazrednih klasifikacijskih problemov.

Globoke nevronske mreže so večnivojske nevronske mreže z vsaj dvema skritima nivojema med vhodnim in izhodnim slojem. Vhodna plast dobi podatke neposredno iz vhoda. Skrite plasti za vhodne podatke uporabljajo izhodne podatke prejšnje plasti. Njihovi izhodi se uporabljajo kot vhodni podatki za naslednje plasti nevronov. Na koncu izhodna plast ustvari izhodne podatke celotne mreže. Uteži se pri globokih nevronskih mrežah spreminjajo vzvratno med dvema nivojem, tako da se izračuna želeni izhod naslednjega sloja. S tem se doseže, da se globoke nevronske mreže iz sloja v sloj učijo bolj abstraktne predstavitve podatkov. Takih zaporednih slojev je lahko več deset ali tudi več sto. Število zaporednih slojev nevronske mreže imenujemo globina modela.

Vhodni podatki za globoke besedilne nevronske mreže so besede ali stavki (včasih tudi črke), predstavljeni kot vektorji, ki ležijo v visokodimenzionalnem prostoru. Metodam, ki spremenijo vhodne podatke v vektorsko predstavitev, pravimo vložitve. Vložitve zakodirajo pomembne informacije o besedah kot razdaljo in smeri med vektorji, semantično podobne besede imajo vrednosti blizu druga drugi. Besede, ki se nahajajo v podobnih kontekstih, imajo podoben pomen in zato bližje reprezentacije. Spodaj je opisano nekaj primerov besednih vložitev, ki so pogosto uporabljene:

- Word2vec (angl. word to vector representation) (Mikolov, Grave, Bojanowski, Puhrsch in Joulin, 2017) besede vektorizira s plitko nevronsko mrežo, ta pa se predstavitve uči iz konteksta besed. Word2vec uporablja enega izmed naslednjih dveh algoritmov: zvezno vrečo besed (angl. Continuous bag of words, CBOW) in preskočni n-gram (angl. n skip-gram). CBOW deluje tako, da ciljno besedo napoveduje iz konteksta, pri čemer upošteva določeno velikost okna. Velikost okna je število besed okoli ciljne besede, ki jih bo sistem vzel za določanje konteksta. Vsaka beseda v oknu je predstavljena kot funkcijski vektor. Sistem ji določi vektor na podlagi ostalih funkcijskih vektorjev. Preskočni n-gram deluje ravno obratno kot CBOW, in sicer ciljno besedo uporabi za določitev konteksta. Za napoved osrednje besede uporablja n predhodnih in naslednjih besed. Algoritem najprej preveri, v kakšnih kontekstih se je ciljna beseda največkrat uporabila in kakšno je bilo zaporedje besed v kontekstu, upoštevajoč velikost okna. Tako kot pri CBOW se tudi pri delovanju preskočnega n-grama modela proces učenja konča šele, ko so na podlagi velikosti okna in korpusa pokrite vse besedne relacije konteksta. Word2vec besedilo spremeni v številčno obliko, ki jo lahko razumejo globoke nevronske mreže.
- Dict2vec (Tissier, Gravier in Habrard, 2017) obstoječi metodi word2vec, ki uporablja algoritem preskočni n-gram, doda negativno vzorčenje (angl. negative sampling) na podlagi obstoječih spletnih slovarjev z definicijami besed. Tissier, Gravier

in Habrard (2017) so za vsako besedo, ki se je v Wikipediji (full Wikipedia dump) pojavila vsaj petkrat, združili vse definicije angleških spletnih slovarjev Cambridge, Oxford, Collins in dictionary.com. Večkrat kot se dve besedi v spletnih slovarjih pojavita skupaj, bolj sta semantično povezani in bližje ležita njuna vektorja. Pri tem ločujejo močen in šibek par besed. Par besed je močen (angl. strong pair), kadar definiciji obeh besed uporabita druga drugo oz. šibek (angl. weak pair), kadar prva beseda nastopa v definiciji druge besede, druga pa ne nastopa v definiciji prve. S tem so besedam dodali kontekst in osnovno semantiko. Učenje je potekalo na različno velikih učnih množicah s približno 50M besed, 200M besed in celotni Wikipediji. Vsaki množici so dodali še pripadajoče definicije. Izkazalo se je, da dajejo množice z dodanimi definicijami boljše rezultate.

- FastText (Bojanowski, Grave, Joulin in Mikolov, 2017) uporablja preskočni n-gram in CBOW, v kombinaciji z negativnim vzorčenjem in softmax aktivacijsko funkcijo. V primerjavi z Word2vec, ki deluje na nivoju besed, ta deluje na nivoju podnizov. Pri vsaki besedi glede na določeno velikost okna n naredi vektor za prvih n znakov besede, nadaljuje z naslednjim n-gramom znakov v besedi in tako naprej do konca besede. Ta metoda dodeli vektorsko vložitev tudi novi, neobstoječi ali narobe zapisani besedi.

- GloVe (angl. Global Vectors for word representation) (Pennington, Socher in Manning, 2014) iz celotnega korpusa konstruira matriko sopojavitve vseh parov besed, predstavitev besed pa dobi z matrično faktorizacijo in učenjem nevronske mreže.

- GN-GloVe (angl. Gender-Neutral Global Vectors) (Zhao, Zhou, Li, Wang in Chang, 2018) uporablja GloVe med učenjem besednih vložitev in jih nadgradi z identifikacijo besed nevtralnega spola (in ohranitvijo besed, ki se uporabljajo specifično za ženski ali moški spol – npr. dojilja), da izloči spolne stereotipe, ki jih spol besede lahko pripiše družbenemu spolu (npr. angleška beseda ”programmerše uporablja tako za moški kot ženski spol in je v svoji definiciji nevtralnega spola, vendar besedna vložitev pri GloVe te besede leži bližje besedi ”moški”kot besedi ”ženska”).

Za slovenski jezik obstajajo vektorske besedne vložitve, dostopne na repozitoriju spletne strani CLARIN.SI (Ljubešić in Erjavec, 2018). Te vložitve so tipa fastText in uporabljajo algoritem preskočni n-gram iz odprtokodne knjižnice fastText. Podatki za vložitve so črpani iz več korpusov, med njimi iz korpusa Gigafida in JANES (Jezikoslovna analiza nestandardne slovenščine). Vseh besed v tej zbirki je slaba dva milijona in pol.

Primeri kontekstnih vložitev so:

- doc2vec (Le in Mikolov, 2014) je nadgradnja word2vec algoritma z dodanim atributom identifikator odstavka, s katerim označi unikaten identifikator dokumenta oz. stavka v korpusu, omogoča učenje dokumentnih vektorjev in s tem podobnosti med dokumenti. Z dodanim identifikatorjem lahko v korpusu izvajamo poizvedbe in računske operacije med dokumenti, njegova vrednost pa je določena kot povprečje besednih vložitev danega dokumenta.

- skip through vectors (Kiros idr., 2015) deluje na trojici neprekinjenih stavkov, kjer na podlagi srednjega stavka poskuša rekonstruirati predhodni in naslednji stavek.

- ELMo (angl. Embeddings from Language Model) (Peters idr, 2018) predstavi vsako besedo glede na kontekst, v katerem je ta uporabljena in se zato imenuje tudi kontekstna vektorska vložitev. V svojem bistvu je to jezikovni model, ki se značilnosti jezika uči tako, da poskuša napovedati naslednjo ali prejšnjo besedo v zaporedju. Ima arhitekturo s tremi komponentami: nevronsko mrežo z dolgim kratkoročnim spominom LSTM za napovedovanje besed, ki sledijo dani besedi, LSTM mrežo za napovedovanje predhodnih besed in konvolucijsko nevronsko mrežo na znakih za kontekstno neodvisno predstavitev besed. Ta vložitev omogoča tudi analizo večpomenskih besed, saj za kontekst uporablja stavek, v katerem beseda nastopa. Zaradi tega ima vsaka beseda mnogo različnih predstavitev.

- BERT (angl. Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee in Toutanova, 2018) je prav tako kontekstna vektorska vložitev, ki je še naprednejša od ELMo vložitve, saj se uči napovedati skrito besedo kjerkoli v stavku.

- T5 (angl. Text-to-Text Transfer Transformation) (Raffel idr., 2020) napove besedo glede na tarčno oznako (primer tarčne oznake: TLDR). Uporablja T5 model, ki ga bomo predstavili v nadaljevanju.

Za slovenski jezik obstaja več vrst vektorskih stavčnih vložitev tipa ELMo (Ulčar in Robnik-Šikonja, 2020), BERT (Ulčar in Robnik-Šikonja, 2021) in T5 (Ulčar in Robnik-Šikonja, 2022). Učenje modelov vložitev je samonadzorovano, saj uporablja nadzorovan način, za tvorjenje učne množice pa posebej ne označuje podatkov.

Med modeli globokih nevronskih mrež za poenostavljanje besedil prevladuje arhitektura zaporedje v zaporedje (angl. Sequence-to-sequence), ki jo lahko implementiramo v modelih vrste kodirnik-dekodirnik. Ti modeli so večinoma preizkušeni le na angleškem jeziku (Štajner in Saggion, 2018) in nekaj jezikih, ki imajo na voljo dovolj virov, kot so kitajščina, francoščina in nemščina.

2.2 Modeli zaporedje v zaporedje

Osnovni komponenti modelov zaporedje v zaporedje sta kodirnik in dekodirnik. Kodirnik spremeni vsak vhodni element v skriti vektor, ki vsebuje prestavitev vhodnega elementa in njegovega konteksta, dekodirnik pa spremeni skriti vektor v izhodni element in uporabi prejšnji izhod kot kontekst. Takšne mreže so sestavljene iz celic GRU (angl. gated recurrent unit), LSTM (angl. long short term memory) ali transformer (Vaswani idr., 2017). Mreža GRU vsebuje vrata posodobitve, ki odločijo, ali prejšnji izhodni podatek posredujejo naslednji celici. Mreža LSTM je sestavljena iz vhodnih vrat, ki ščitijo trenutni korak pred nepomembnimi vhodi, izhodnih vrat, ki preprečujejo trenutnemu koraku posredovanje nepomembnih informacij naslednjim korakom in iz vrat pozabe, ki omejujejo informacijo, ki se posreduje med celicami (Raaijmakers, 2019). Transformer (Vaswani idr., 2017) ne uporablja niti rekurentne niti konvolucijske mreže, temveč deluje na podlagi mehanizma samopozornosti. Ta med procesiranjem niza dopolni vsak element z uteženim povprečjem preostalega niza. Vhodni niz preslika v niz vložitev, ki je potem vhod kodirnika. Kodirnik je sestavljen iz več blokov, ki vsak vsebuje dve podkomponenti: plast samopozornosti, ki ji sledi majhna usmerjena nevronska mreža. Dekodirnik ima podobno

strukturo kot kodirnik, le da za vsako plastjo mehanizma samopozornosti, ki pride iz kodirnika, uporabi standardni mehanizem pozornosti. Izhod zadnjega sloja dekodirnika gre do polno povezanega sloja, ki na izhodu uporabi funkcijo softmax, njihove uteži pa so deljene z vhodno vložitveno matriko. Prednost transformerja pred GRU in LSTM je, da se lahko mreža zaradi mehanizma pozornosti, ki zmore večjo paralelizacijo, hitreje uči.

Nisioi, Štajner, Ponzetto in Dinu (2017) so prvi uporabili model zaporedje v zaporedje za poenostavljanje besedil. Model istočasno izvaja leksikalno poenostavljanje, reducira vsebino in ohranja pomen ter slovnično pravilnost. Za vhod so avtorji uporabili vložitve word2vec in originalno angleško Wikipedijo, za izhod pa poenostavljeno angleško Wikipedijo. Mrežo so zgradili iz dveh plasti LSTM celic s 500 skritimi nevroni (Štajner in Saggion, 2018). Model se je izkazal za boljšega kot klasični pristopi PBMT, SBMT in nenadzorovano leksikalno poenostavljanje na osnovi besednih vložitev.

Čeprav so bili osnovni modeli zaporedje v zaporedje uspešni v mnogih nalogah NLP, za poenostavljanje stavkov niso bili idealni, ker so mnogokrat namesto zamenjave, brisanja besed ali zamenjave vrstnega reda besed, stavek pustili enak originalu (Zhang in Lapata, 2017). Zhang in Lapata (2017) sta model zaporedje v zaporedje izboljšala z mehanizmom pozornosti. Ta mehanizem skuša oponašati človeško pozornost, ki nam omogoča, da iz celotne vsebine izluščimo le vsebinsko pomembne dele oz. smo na njih bolj pozorni. Mehanizem poizkuša pozornost implementirati z vpeljavo posebnih blokov, ki odločajo, katero stanje je trenutno relevantno. Mreža, ki sta jo avtorja predstavila, je zgrajena iz GRU celic, hrani več skritih stanj, mehanizem pozornosti pa odloča, katero je trenutno relevantno. Uporabila sta spodbujevano učenje, ki raziskuje prostor možnih poenostavitev tako, da maksimizira pričakovano funkcijo nagrade, ki jo sestavljajo trije kriteriji: preprostost, relevantnost in tekočost besedila. Algoritem spodbujevanega učenja je soočen tako z učnimi podatki kot tudi s povratno informacijo, ki učni sistem in podatke povezuje s povratno zanko. Povratna informacija je podana kot pozitiven ali negativen dražljaj sistemu, odvisno od izida učenja.

Vu, Hu, Munkhdalai in Yu (2018) so v modelu zaporedje v zaporedje namesto LSTM ali GRU celic uporabili arhitekturo RNN z dodatnim pomnilnikom (angl. Memory-augmented RNN), ki se je izkazala za boljšo pri dolgih in zahtevnih stavkih. Njihov model uporablja za kodirnik nevronski semantični kodirnik NSE (angl. Neural semantic encoders) in LSTM za dekodirnik.

Surya, Mishra, Laha, Jain in Sankaranarayanan idr. (2019) so za poenostavljanje besedil pri modelu kodirnik - dekodirnik uporabili nenadzorovano učenje, kar pomeni, da za vhodni kompleksni stavek niso vnaprej določili izhodnega poenostavljenega stavka. Uporabili so ločeni učni množici preprostih in kompleksnih besedil, ki med sabo nista povezani in poravnani ter imata različno vsebino. Takšni množici sta cenejši in lažje dosegljivi kot vzporedni in poravnani množici ter tako omogočata poenostavljanje tudi v jezikih, ki poravnanih množic nimajo. Avtorji so uporabili en kodirnik in dva različna dekodirnika, kjer je eden generiral preproste stavke, drugi pa kompleksnejše. Za kodirnik so uporabili dva nivoja dvosmerne GRU mreže, za vsakega od dekodirnikov pa dva nivoja enosmerne GRU mreže. V vsakem koraku dekodirnika se je ustvaril kontekstni vektor, ki so mu pripeli uteži pozornosti. Tak vektor (oz. matriki vseh vektorjev za oba dekodirnika za vsak korak) je bil podatek za naslednji dekodirnik. Surya, Mishra, Laha, Jain in Sankaranarayanan idr. (2019) so pokazali, da njihov model z nenadzorovanim učenjem dosega primerljive rezultate kot modeli z nadzorovanim učenjem. Ob dodajanju nekaj poravna-

nih stavkov se je delovanje še izboljšalo (Surya, Mishra, Laha, Jain in Sankaranarayanan idr., 2019).

Raffel idr. (2020) so se nalog generiranja besedil, kot so poenostavljanje besedil, povzemanje, odgovarjanje na vprašanja ali strojno prevajanje, lotili s strojnim učenjem s prenosom znanja. Pri tem so vse naloge spremenili v nalogo tipa besedilo v besedilo. To pomeni, da njihov model zaporedje v zaporedje, imenovan T5, na vhodu prejme besedilo in na izhodu zgenerira novo besedilo. Model T5 uporablja transformer oblike kodirnik - dekodirnik in deluje z mehanizmom samopozornosti. Vhodni niz preslika v niz vložitev, ki je potem vhod kodirnika. Le-ta je sestavljen iz več blokov, kjer vsak vsebuje dve podkomponenti: plast samopozornosti, ki ji sledi majhna usmerjena nevronska mreža. Dekodirnik v modelu T5 ima podobno strukturo kot kodirnik, le da za vsako plastjo mehanizma samopozornosti, ki pride iz kodirnika, uporabi standardni mehanizem pozornosti. Izhod zadnjega sloja dekodirnika gre do polno povezanega sloja, ki na izhodu uporabi funkcijo softmax, njihove uteži pa so deljene z vhodno vložitveno matriko.

Po zgledu Raffel idr. (2021) sta Ulčar in Robnik-Šikonja (2022) izdelala T5 model za slovenski jezik, poimenovan SloT5 [1]. Izdelala sta dva različna modela, t5-sl-small, ki ima 8 plasti kodirnikov in 8 plasti dekodirnikov, s skupno okoli 60 milijoni parametrov, ter t5-sl-large s 24 kodirniki in 24 dekodirniki, s skupno okoli 750 milijoni parametrov. Pri obeh modelih sta uporabila globoko nevronsko mrežo arhitekture transformer, učenje s prenosom znanja ter model kodirnik-dekodirnik z mehanizmom pozornosti. Oba modela sta bila učena na velikem slovenskem korpusu. Učna množica je tako vsebovala okoli 4 milijarde besed. Pred začetkom učenja sta avtorja uporabila slovar za stavčne vložitve tipa SloBERTa (Ulčar in Robnik-Šikonja, 2021) z uporabo modela sentencepiece. Iz nabora nalog, na katerih so Raffel idr (2021) učili svoje modele, sta bila oba slovenska modela trenirana na dveh izmed njih: razšumljanje (angl. denoising; tj. izločanje šuma iz podatkov) in maskiranje delov besedila (angl. span corruption; pri tem gre za maskiranje besed, tako da jih zamenjamo z drugimi in napovemo, kakšen je bil originalen stavek). Avtorja sta uporabila 1 milijon korakov, velikost niza 4096 žetonov (oz. besed) in (malo manj kot) eno epoho učenja (tj. število ponovitev učenja preden se algoritem ustavi). Novo naučena slovenska modela sta evalvirala z nalogo povzemanja, kjer sta uporabila dve podatkovni množici, novice Slovenske tiskovne agencije (STA) in novice AutoSentiNews (ASN). Pred tem sta modela prilagodila s knjižnico HuggingFace. Vhodni podatki za naloge povzemanja so bili besedila člankov, izhodni pa njihovi povzetki. Uspešnost modelov sta primerjala z obstoječimi večjezičnimi T5 modeli: mT5-large in mT5-small. Ugotovila sta, da je mT5-large uspešnejši od obeh slovenskih modelov, medtem ko sta slovenska modela uspešnejša kot mT5-small model.

Strojno učenje s prenosom znanja se je v zadnjem času uveljavilo kot eden izmed učinkovitejših pristopov naslavljanja problema pomanjkanja kakovostnih in dovolj obsežnih podatkovnih zbirk (Vrbančič, 2021). Pri tej vrsti učenja se zaradi določenega predhodnega znanja, ki ga je globoka nevronska mreža pridobila na sorodnih problemih, zmanjša potreba po veliki količini podatkov. Prednaučeno znanje je shranjeno v obliki uteži slojev globoke nevronske mreže, ki se jih lahko prenese iz prednaučenega modela v model, ki naslavlja podobno nalogo. S takšnim prenosom znanja se zmanjša tudi časovna zahtevnost učenja modela na novem problemu, saj v splošnem učenje zahteva manj ponovitev kot pri klasičnem učenju globoke mreže (Vrbančič, 2021). Tak pristop je obetaven tudi

[1] `https://huggingface.co/cjvt/legacy-t5-sl-small`

za avtomatsko poenostavljanje besedil v slovenskem jeziku, kjer lahko uporabimo naučen model za slovenski jezik na nalogah razšumljanja in maskiranja delov besedila na nalogi poenostavljanja.

2.3 Razumljivost besedil

Eden od namenov poenostavljanja besedil je, da so informacije v poenostavljenih besedilih razumljive tudi ljudem, ki imajo težave z branjem originalnih besedil. V širšem družbenem kontekstu je razumljivost vsakdanja kategorija, ki jo ljudje dojemamo kot enovito lastnost, v jezikoslovju pa nanjo gledamo kot na formalno lastnost besedila, sestavljeno iz več dejavnikov (Ferbežar, 2009).

2.3.1 Kaj je razumljivost besedil

V našem delu bomo skušali ocenjevati razumljivost generiranih besedil. SSKJ razumljivost razlaga kot ”lastnost, značilnost razumljivega”, slednje pa kot takšno, ”ki se da razumeti”oz. ”ki se mu da ugotoviti pomen.”(Leskovec, 2009) ”Razumljivost kot lastnost besedil lahko razumevanje olajša ali pa ga oteži.”(Leskovec, 2009) Vidimo, da je izrazu razumljivost besedila soroden izraz razumevanje besedila, pri čemer prvi poudarja vidik besedila, drugi pa bralčev pogled na besedilo. Na razumljivost lahko gledamo kot na stopnjo razumevanja besedila, na katero bralec različno vpliva s svojimi predispozicijami in predznanjem (Osolnik Kunc, 2007). Zavedati se moramo, da razumljivost vseeno ni besedilna značilnost sama na sebi, temveč obstaja le v interakciji med besedilom in bralcem (Leskovec, 2009). Razumljivost besedil z jezikoslovnega stališča je zapletena, njihovo razumljivost oz. nerazumljivost je težko demonstrirati, saj je relativna (Ferbežar, 2009), dodatne zaplete pa predstavlja tudi dejstvo, da ne moremo z zagotovostjo reči, ali gre pri razumljivosti za enotno ali sestavljeno lastnost (Ferbežar, 2012). S teorijo razumljivosti besedil se je v 20. stoletju ukvarjalo nemško in ameriško jezikoslovje, a je zaradi kompleksnosti preučevanega pojava vsaki izmed njih mogoče očitati pomanjkljivosti (Leskovec, 2009).

2.3.2 Pregled raziskovanja razumljivosti besedil

Z raziskovanjem razumljivosti besedil se ukvarja več različnih disciplin, med drugimi psihologija, kognitivna znanost in jezikoslovje. Vsaka disciplina uvaja različne pristope raziskovanja tega fenomena, od računskih berljivostnih formul, zelo abstraktnih kognitivnih modelov, do jezikoslovnih modelov razumljivosti. V nadaljevanju bomo predstavili izsledke teh pristopov, njihove prednosti in morebitne pomanjkljivosti.

2.3.2.1 Raziskovanje razumljivosti besedil v psihologiji

Raziskovanje razumljivosti se je začelo v sedemdesetih letih dvajsetega stoletja v psihologiji z določanjem zahtevnosti besedil na podlagi besediščne obremenjenosti ter z razvojem formul berljivosti. Te izračunajo berljivostni indeks in napovejo stopnjo razumljivosti poljubnega besedila (Leskovec, 2009). Formule upoštevajo lastnosti besedil, kot so dolžina besed, dolžina stavkov, relativna pogostost besed, uporaba splošnega besedišča, kratkih povedi in nezapletenih skladenjskih struktur ipd. Takšna metodologija objektivno meri kompleksnost besedila, zaradi česar se formule berljivosti še vedno pogosto uporabljajo

kot metoda za merjenje razumljivosti. Merijo le precej površinske lastnosti besedila, ne upoštevajo pa povezave z bralčevim razumevanjem in tako ne dajejo nobenih specifičnih navodil za boljši zapis besedil.

2.3.2.2 Raziskovanje razumevanja in razumljivosti besedil v kognitivni znanosti

Raziskovanje razumljivosti besedil se je začelo v kognitivni znanosti z razvojem precej abstraktnih modelov razumljivosti oz. razumevanja besedil. Tak model je Kintschev in Van Dijkov model razumevanja (1983), ki temelji na teoriji mentalnih modelov. Ta izhaja iz predpostavke, da si ljudje za vse, kar nas obdaja, kar doživljamo, počnemo, sprejemamo ipd., ustvarjamo mentalne modele oz. mentalne predstave, ki so osnova našega celotnega razumevanja sveta, samih sebe in naše interakcije s svetom (Verdonik, 2011). Dve osebi nikoli ne vidita sveta v povsem enaki luči, ker si vsaka ustvarja svoje lastne mentalne modele, obstajajo pa nekakšni splošno veljavni vzorci in skupni elementi, zaradi česar so do določene mere mentalni modeli različnih oseb vseeno skupni (Verdonik, 2011). Teorija mentalnih modelov tako zavrže pojmovanje, da mentalna reprezentacija vsebine besedila ustreza njeni lingvistični reprezentaciji in skuša razložiti, kako se informacija iz različnih delov besedila integrira v koherentno mentalno reprezentacijo (Vizjak Pavšič, 2006). Van Dijk in Kintsch v svojem modelu izpostavita, da pri razumevanju ustvarjamo dva tipa mentalnih modelov: predstavo besedila, ki pomeni semantično predstavo besedila, in predstavo situacije, ki pomeni kognitivno predstavo dogodkov, dejanj, oseb in splošne situacije, o kateri govori besedilo (Verdonik, 2011). V tem modelu ima v procesu razumevanja ključno vlogo besedišče, razumevanje se začne z analizo besed oz. trditev iz besedila. Razumevanje se dogaja v ciklih, mentalni modeli pa se gradijo postopoma, vsak nov stavek dograjuje model in gradi kontekst za interpretacijo naslednjega stavka (Vizjak Pavšič, 2006). Če je besedilo daljše, naj bi najprej v prvem ciklu v delovnem spominu analizirali le določeno število trditev, ki naj bi se shranjevale v kratkoročnem spominu, tem uskladiščenim trditvam naj bi se v novem ciklu pridružile nove trditve, dokler se naj ne bi vse medsebojno povezane trditve prenesle iz kratkoročnega v dolgoročni spomin (Leskovec, 2009). V dolgoročnem spominu naj bi bila shranjena posameznikova znanja in izkušnje oz. njegova kognitivna shema. Po vsakem ciklu naj bi se na novo konstruirana koherenčna slika prenesla v dolgoročni spomin, težave z razumevanjem besedila pa naj bi nastale, ko med propozicijami v kratkoročnem spominu in delom predstave besedila, shranjenim v kratkoročnem spominu, ni nobene povezave. V tem primeru je nujno sklepanje. Za razumevanje besedila mora bralec dopolniti informacije v besedilu s svojim lastnim znanjem, védenjem in izkušnjami; posameznikovo razumevanje je torej povezano z informacijami, shranjenimi v njegovem dolgoročnem spominu (Ferbežar, 2012). Modele naj bi lažje konstruirali, če je besedilo koherentno in ima referenčno kontinuiteto, kar pomeni, da se trditve v besedilu nanašajo na nekaj, kar je bilo predhodno v besedilu že omenjeno (Vizjak Pavšič, 2006). Referenčno kontinuirana besedila so lažje razumljiva, kar Dijk in Kintschev model razlaga z manjšo obremenjenostjo delovnega spomina (Vizjak Pavšič, 2006). Prednost modela je, da razumevanje besedila dojema ne le v povezavi z lastnostmi besedila, temveč tudi v povezavi z interakcijo med bralcem in besedilom (Leskovec, 2009). Pomanjkljivosti modela so, da razmejitev med tipoma mentalnih modelov ”predstava besedila”in ”predstava situacije”ni enostavna (Verdonik, 2011), da se vsebinsko razčlenjevanje začenja šele s pomensko analizo (Leskovec, 2009) ter da ne vključuje dveh pomembnih segmentov, ki prav tako vplivata na razumevanje: socialne kognicije

(mnenja, odnosa, ideologij, vrednot, emocij itd.) ter kontekstnih modelov.

2.3.2.3 Raziskovanje razumljivosti besedil v jezikoslovju

Z raziskavami razumljivosti besedil v jezikoslovju so nastali modeli razumljivosti, ki so poudarili pomen medsebojnega vplivanja med bralcem in besedilom. Na razumljivost besedil niso gledali kot na objektivno lastnost besedila, temveč da jo vsaj do neke mere določa bralec s svojo motivacijo, z znanjem in védenjem, nameni branja in poslušanja itd. (Ferbežar, 2009). Primeri takih modelov so hamburški model razumljivosti, Groebenov model razumljivosti, Sauerjev minimalni model za analizo in optimizacijo razumljivosti, Baumannov model hierarhičnih dimenzij razumljivosti in karlsruhejski model. Za slovenski jezik je Ferbežar (2012) predstavila model, ki temelji na sedmih kriterijih razumljivosti, prilagojenih iz karlsruhejskega modela.

2.3.2.3.1 Hamburški model razumljivosti

Hamburški model razumljivosti (Langer, Schulz von Thun, Tausch, 1993: v Ferbežar, 2012) temelji na strokovni presoji razumljivosti besedila in vsebuje štiri dejavnike razumljivosti:

1. preprostost ubeseditve, ki je nasprotje zapletenosti,
2. preglednost, ki je nasprotje nepreglednosti oz. nepovezanosti
3. kratkost in jedrnatost, ki je nasprotje dolgoveznosti,
4. dodatno stimulacijo, ki je v nasprotju z odsotnostjo stimulacije.

Očitki tega modela so, da uporablja dihotomije, ki vedno ne ločujejo dveh nasprotujočih se polov, ter da se pri anketirancih dopušča subjektivno tolmačenje dihotomij (Osolnik Kunc, 2007).

2.3.2.3.2 Groebenov model razumljivosti

Groebenov model razumljivosti (Groeben, 1982: v Ferbežar, 2012) razumljivost besedil meri na podlagi:

1. stilistično-slovnične preprostosti,
2. kognitivne členjenosti (uvrščanje novih vsebin med že obstoječe kognitivne koncepte in implikacije pri učenju),
3. pomenske zgoščenosti,
4. konceptualnega konflikta (spodbujanje radovednosti in motivacije za branje).

Groeben kot najpomembnejši dejavnik v tem modelu izpostavlja kognitivno členjenost, kjer bralcu pripisuje kognitivno vlogo pri obdelavi besedila. Predpostavlja, da absolutne razumljivosti besedila ni, ampak je ta usmerjena k bralcu. Očitki tega modela so, da mu primanjkuje prepoznavnih teoretičnih izhodišč, iz katerih bi se dalo izpeljati nastavke, po katerih posega, da je nedorečen in da ne upošteva zunajbesedilnih dejavnikov (Leskovec, 2019). Kljub temu pa sta tako hamburški kot Groebnov model prinesla

praktične predloge, ki pomagajo izboljšati kakovost besedil, saj je iz posameznih ocen v štirih dimenzijah mogoče natančneje ugotoviti, v katerem pogledu je besedilo potrebno izboljšati.

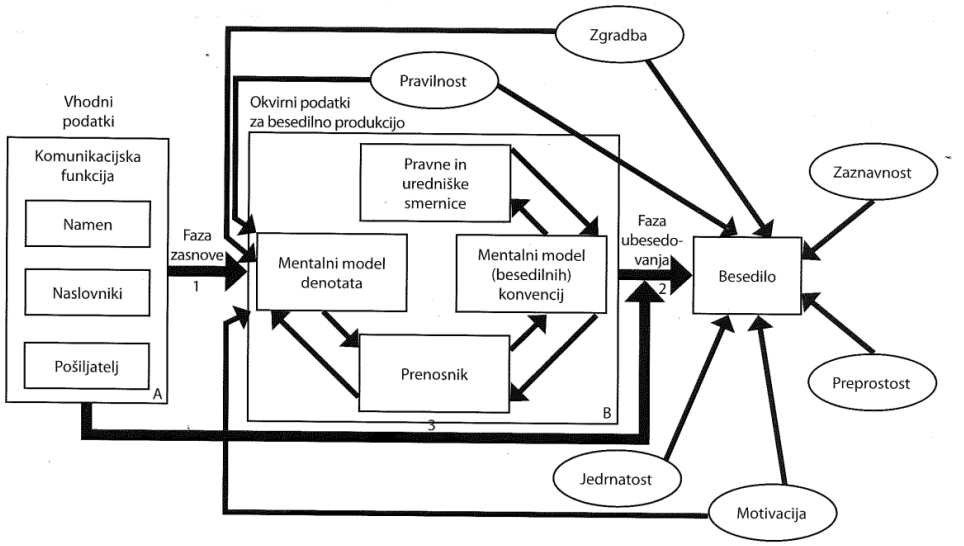
2.3.2.3.3 Sauerjev minimalni model za analizo in optimizacijo razumljivosti

Sauerjev minimalni model za analizo in optimizacijo razumljivosti (Sauer, 1995: v Ferbežar, 2012) se je razvil ob vprašanju razumljivosti besedil pri prevajanju. Temelji na štirih minimalnih kriterijih razumljivosti besedila:

1. berljivosti, ki se nanaša na površinske lastnosti besedila,
2. razumljivosti v smislu enostavnosti besedila za razumevanje,
3. pripravnosti, ki se nanaša na bralčevo zaznavanje besedila s pregledno strukturiranostjo besedila (naslovi in podnaslovi, primeri, opombe itd.),
4. uporabnosti, ki se nanaša na učinek oz. posledice, ki jih ima besedilo na bralčevo mentalno predstavo pomena.

Razumljivost besedila umesti v štiridelno shemo: zunanji, materialni del besedila predstavljata berljivost in pripravnost, notranji, globinski del, ki je povezan s kognitivnimi procesi, pa razumljivost in uporabnost (Ferbežar, 2012).

2.3.2.3.4 Karlsruhejski model razumljivosti



Slika 2 – Karlsruhejski model razumljivosti (povzeto iz Ferbežar, 2012, str. 113).

Karlsruhejski model (Göpferich, 2001: v Ferberžar, 2012) je nadgradnja hamburškega in Groebnovega modela, kamor je avtorica vpeljala nanašanjski okvir (na sliki 2) in štirim obstoječim dimenzijam razumljivosti, ki jih je nekoliko modificirala, dodala še dve.

Pri tem je s primeri prikazala, katere značilnosti pripomorejo k izpolnitvi zahtev, ki jih pripisuje šestim dimenzijam (Leskovec, 2019). Pri nanašanjskemu okvirju se avtorica opira na psihološke razumljivostne mentalne modele. Ogrodje modela predstavljajo vhodni podatki s komunikacijsko funkcijo besedila (v okvirju A) in okvirni podatki za njegovo produkcijo (v okvirju B). Zunaj tega referenčnega okvirja je šest dimenzij razumljivosti. Komunikacijska vloga besedila določa kakovost besedila in povezuje namen besedila z naslovnikom in pošiljateljem. Okvirni podatki, ki jih določa komunikacijska funkcija besedila, se nanašajo na mentalni model denotata, mentalni model besedilnih konvencij, prenosnik ter pravne in uredniške smernice. V mentalnem modelu denotata so bolj ali manj kompleksne kognitivne sheme, ki pošiljatelju omogočajo ubeseditev določene predmetne vsebine, naslovniku pa v procesu razumevanja tvorjenje besedilnega pomena oz. razumetja besedila. V mentalnem model besedilnih konvencij so dogovori oz. konvencije, ki se nanašajo na izbiro ustrezne besedilne vrste in ustreznega ubeseditvenega sloga znotraj nje, pravne in uredniške smernice pa so določila, ki jih podjetja in institucije postavljajo za oblikovanje ali za potrebe pravno-upravnega poslovanja (Ferbežar, 2012).

Dimenzije razumljivosti v karlsruhejskem modelu so:

1. Jedrnatost: besedilo naj vsebuje ravno toliko podatkov, da omogočajo razumljivost. Zahteve po jedrnatosti, ki jih avtorica navaja, so:
    - manjkajoče ali
    - odvečne podrobnosti v mentalnem denotativnem modelu,
    - raba dolgih izrazov namesto primernejših krajših z enako pomembnostjo za besedilo,
    - tavtologije oz. redundanca podatkov.
2. Neoporečnost oz. pravilnost: besedilo naj nima notranje protislovnosti, zajame naj pravilno oceno predznanja prejemnikov, ustrezen mentalni denotativni ali konvencijski model, ustrezen prenosnik in naj nima jezikovnih napak v besedilu.
3. Motivacija: tu gre za stopnjevanje bralčevega interesa za branje besedila. Na ravni mentalnega denotativnega modela lahko motivacijo stopnjujemo s ponazoritvijo z zgledi, ki izvirajo iz sveta prejemnikov, na ravni kodiranja pa na primer z rabo slenga, s pisanjem v prvi osebi množine ipd., odvisno od besedilne vrste. Motivacija je namreč po mnenju avtorice tesno povezana z besedilnimi vrstami.
4. Zgradba: ta dimenzija se za razliko od hamburškega in karlsruhejskega modela nanaša le na vsebinsko zgradbo, in sicer tako na zgradbo mentalnega denotativnega modela kot na zgradbo kodiranja. Avtorica razlikuje zgradbo na makro- in mikroravni, pri čemer makroraven opredeljuje kot raven, ki presega dve sosednji povedi, mikroraven pa kot tisto, ki ju ne. Na makroravni, ki je tesneje povezana z mentalnim denotativnim modelom, gre za členjenje besedila, ki naslovnika smiselno in logično vodi, na mikroravni, ki je tesneje povezana s kodiranjem v besedilu, pa gre za uporabo ustreznih konektorjev, ki informacije posredujejo v logičnem zaporedju.
5. Preprostost: nanaša se zgolj na kodiranje v besedilu, na izbiro besedišča in skladenjskih struktur, neposrednost izražanja, nedvoumnost ter usklajenost izbranega besedišča in skladenjskih struktur z izbrano besedilno vrsto.

6. Zaznavnost: tako kot prejšnja dimenzija se tudi ta nanaša le na kodiranje v besedilu, na oblikovne lastnosti besedil in na nejezikovne nosilce informacij, kot je opremljenost besedila z naslovi, podnaslov, alinejami, slike, ilustracije ipd.

Pomanjkljivost tega kompleksnega modela, ki združuje stališča različnih disciplin, je, da ne vzpostavlja jasne povezave med referenčnim okvirom za produkcijo in vrednotenje besedil in dimenzijami razumljivosti, ki so zunaj tega okvirja ter da ni povsem jasno, kako uporaben je v praksi (Ferbežar, 2012).

2.3.2.4 Pregled raziskav razumljivosti za slovenski jezik

Zavedati se moramo, da vprašanje razumljivosti ni nujno univerzalni jezikoslovni problem (Osolnik Kunc, 2007) ter da je različno med besedilnimi vrstami (Ferbežar, 2012). Empirično se je razumljivost raziskovala na različnih slovenskih besedilih, kot so slovenska poročevalna besedila iz slovenskih dnevnikov (Leskovec, 2009), besedila v javni in uradni komunikaciji (Ferbežar, 2009), navodila za uporabo zdravil (Brinovec in Krečan, 2010), pravna besedila (Ambrožič, 2019) in svetopisemska govorica (Tomažič, 2020). Vloga razumljivosti v slovenskem leposlovju še ni raziskana.

Ferbežar (2009) je ugotovila, da ljudi na splošno pri razumevanju besedil v slovenščini najbolj ovirajo nepoznane besede in nepoznavanje tematike, nekoliko manj jih ovirajo dolgi stavki, velikih ovir pa jim ne predstavljajo gostobesednost, zapleteni stavki in jezikovne napake. Jezikovne napake so za bralce moteče, besedila z veliko jezikovnimi napakami pa manj sprejemljiva, dejansko pa ne ovirajo razumevanja. Leksikalne ovire, kot so uporaba neustreznih besed ali besednih zvez, slovnične neustreznosti, besede z napačnim pomenom, neustrezna tvorba besed, uporaba besede neustrezne besedne vrste ipd., pri domačih govorcih vplivajo na razumevanje nekaterih podrobnosti besedila, med tem ko pri tujih govorcih povzročijo hude motnje v procesu razumevanja na globalni ravni. Tudi druge raziskave na različnih besedilnih vrstah potrjujejo zgornje ugotovitve. Pri navodilih za uporabo zdravil so ljudem najpogosteje nerazumljivi strokovni izrazi in imena spojin, ki so v latinščini (Brinovec in Krečan, 2010). Pri svetopisemskih besedilih pride do težav z razumevanjem in razumljivostjo zaradi posebnosti jezika, ki je uporabljen v teh besedilih (Tomažič, 2020). Besedila so polna retoričnih in slogovnih figur, personifikacij, prilik, alegorij, govornih figur in besednih iger, ki lahko otežujejo razumevanje besedil. Poleg tega je bilo sveto pismo napisano v določenem okolju in določeni kulturi, medtem pa se je sodobni jezik razvijal in spremenil. Jezik v teh delih je povsem drugačen od današnjega. Razumevanje svetopisemskih besedil je dodatno pogojeno s predrazumevanjem, brez določenih predpostavk, ki vodijo razumevanje, bibiličnega besedila namreč ne moremo razumeti. V diplomskem delu je Ambrožič (2019) raziskovala razumljivosti pravnih besedil v povezavi s pleonastičnim zanikanjem. O pleonastičnem zanikanju govorimo, ko je skladenjsko prisotna nikalnica ne, vendar ne določa nikalne pomenske podstave povedi. Primer: »Bojim se, da Janez nima prav.« Ugotovi, da v nekaterih primerih pleonastično zanikanje otežuje razumljivost pravnih besedil. Tako bi zgornji primer raje parafrazirali kot: »Menim, da se Janez moti.«

Razumljivost lahko raziskujemo z različnimi metodologijami. V večini raziskavah je razumljivost merjena kot enovita lastnost na lestvici od 1 do 5, brez dodanih kriterijev. Nekateri jo raziskujejo v okviru jezikoslovnih modelov z vprašalniki, drugi z jezikoslovnimi testi, nekateri avtorji pa določanje stopnje uresničevanja kriterijev razumljivosti preverjajo kar sami.

Leskovec (2009) je v diplomski nalogi raziskovala razumevanje sodobnih slovenskih poročevalnih besedil s karlsruhejskim razumljivostnim modelom. V vsakem poročilu posebej je preverila stopnjo uresničevanja zahtev, ki jih določa omenjenih šest dimenzij razumljivosti in podala vse primere kršitev. Izkazalo se je, da so bile vse dimenzije razumljivosti dobro upoštevane, s poudarkom na jedrnatosti, neoporečnosti in zaznavnosti. Karlsruhški razumljivostni model se je izkazal kot primeren za uporabo na področju novinarskega sporočanja, toda kot avtorica poudari, upoštevanje vseh dimenzij razumljivosti, ki jih model predvideva, še ne pomeni nujno tudi visoke stopnje razumljivosti.

Brinovec in Krečan (2010) sta raziskovali razumevanje navodil za uporabo zdravil s kvantitativno in kvalitativno raziskavo. V kvantitativni raziskavi so ljudje izražali svoja stališča o tem, koliko se jim zdi, da razumejo navodila za zdravila na podlagi enajstih vprašanj. Primeri takih vprašanj so bili: Ali menite, da informacijo, ki ste jo razbrali iz navodila tudi vedno razumete? Lahko zase rečete, da vsako navodilo za zdravilo, ki ga preberete, res dobro razumete? V kvalitativni raziskavi pa so vprašani reševali test, ki je meril razumevanje navodil za uporabo zdravil Rutacid in Nalgesin ter opravili intervju. Test razumevanja za vsako od teh navodil je bil sestavljen iz dveh delov, prvi je zajemal konkretna vprašanja iz navodila za zdravilo, kjer so morali testiranci najprej najti podatek v navodilu in nato pokazati, ali jo razumejo ali ne, drugi del pa je zajemal vprašanja o razporeditvi rubrik, pisavi, o tem kako so informacije podane in ali so razumljive.

Učinkovit poskus merjenja razumljivosti in razumevanja besedil je izvedla Ferbežar (2009 in 2012). Razumljivost in razumevanje sedmih besedil, vsakega v dveh variantah, v avtentični in prilagojeni obliki, je z vprašalnikom merila pri dveh skupinah. Prva skupina je razumevanje in sprejemljivost besedila za bralce ocenjevala kot enovito lastnost besedila na štiristopenjski lestvici od lahko do zelo težko za razumevanje. Druga skupina, ki je bila sestavljena iz strokovnjakov s področja jezikoslovja, je razumljivosti in sprejemljivosti besedil ocenjevala na štiristopenjski lestvici po sedmih kriterijih: jedrnatost, upoštevanje naslovnika, jezikovna pravilnost, besedišče, skladenjske strukture, komunikativnost in koherenca. Tu se je oprla na domnevo, da so prilagojena besedila lažje razumljiva. Prilagoditve besedil so bile minimalne, npr. v izvirniku je bil odstavek ”Pujs ni bil zadnji. Eno mesto pred njim je bil pes in takoj za njim maček.”, v prilagojeni obliki pa ”Pujs ni bil zadnji. Eno mesto pred njim je bil pes in takoj za pujsom maček.”Obe skupini ocenjevalcev sta bili sestavljeni iz govorcev slovenščine in tujih govorcev. Tu je predpostavila, da domači govorci slovenščine v slovenskih besedilih razumejo več, ker v besedilih naletijo na manj ovir. Opazovala je, ali se ocene strokovnjakov, izračunane kot povprečje vseh sedmih parametrov razumljivosti, ujemajo z oceno razumljivosti pri običajnih bralcih. Jezikovni test sta sestavljali dve uradovalni besedili, vprašanje tujega govorca slovenščine poslanega na institucijo, navodilo računske naloge za tretji razred osnovne šole, dva kratka časopisna komentarja in strokovno besedilo s področja vinogradništva. Na podlagi prebranega besedila so ocenjevalci iz nabora odgovorov na vprašanja obkrožili tistega, ki se jim je zdel pravilen, ter ocenili, kako lahko oz. težko razumljivo se jim je zdelo besedilo. Raziskava zaradi premajhnega števila podatkov ni uspela pokazati, ali je kateri od sedmih kriterijev razumljivosti bolj ključen od drugih

Ferš (2019) je v magistrskem delu pripravila turistični vodnik za mesto Maribor v lahko berljivi obliki v sodelovanju z osebami in za osebe z motnjami v duševnem razvoju. Razumljivost svojega predlaganega turističnega vodnika, ki je upošteval pravila za pisanje besedil v lahkem branju, je preverjala neposredno pri ciljni skupini, to je med člani iz

bralne skupine Črna na Koroškem z motnjami v duševnem razvoju. Ti so podali svoje komentarje o stavkih in besedah, ki jih niso razumeli, predlagali drugačne, povedali, katere povedi niso potrebne ter sodelovali pri izbiri slikovnega materiala. Preverjanja razumljivosti zapisanega pri ciljni publiki je zagotovo dober indikator, da bo besedilo zapisano razumljivo, toda ni primeren za naše delo, kjer bomo generirali večje količine besedil.

Zidarn (2020) je v svoji magistrski nalogi uspešnost povzemanja za avtomatsko povzeta slovenska besedila z globokimi nevronskimi mrežami ocenil z ROUGE mero ter s človeško evalvacijo. Ugotovil je, da visoka ocena ROUGE ne zagotavlja nujno kakovostnega in pravilnega povzetka. Tako je dal 22 osebam v reševanje anketo, sestavljeno iz 8 člankov in povzetkov. Nekateri povzetki so bili referenčni, nekateri pa zgenerirani z njegovim modelom. Anketiranci so povzetke ocenjevali glede na dve lastnosti, točnost in tekočnost, na lestvici od 1 do 5, torej od manj do bolj točnega in od manj do bolj tekočnega besedila. Pri tem je točnost pomenila ali so v povzetku zajeti točni podatki in dejstva, ki so zapisana v članku, tekočnost pa smiselnost in berljivost stavkov oz. pravilen vrstni red besed v stavku. Tudi nas bo zanimala točnost poenostavljenih besedil kot del razumljivosti. Tekočnost nas v naši raziskavi ne bo zanimala, saj ne vpliva neposredno na razumljivost. Po Sauerjevi štiridelni shemi bi jo enačili z berljivostjo.

Za našo raziskavo jezikovni test ni primerno sredstvo, ker je na nivoju (predvsem kratkih) odstavkov težko zastaviti vprašanja, ki bi zajela bistvo odstavka. Kaj je bistvo, je namreč zelo subjektivno, na nivoju odstavkov, ki jim manjka predhodnji kontekst, pa je še težje. Primernejša metoda se zdi ročno pregledovanje zgeneriranih odstavkov in vprašalnik. Odločili smo se, da kršitve objektivnih dimenzij razumljivosti stavkov oz. povedi (npr. slovnične napake in skladenjske strukture) sami pregledamo, preostale pa analiziramo s pomočjo vprašalnika. Vprašalnik je nasploh najbolj uporabljena metoda za človeško evalvacijo avtomatsko poenostavljenih besedil, kjer raziskovalci uporabljajo različne Likertove lestvice (Stodden, 2021). Mi se bomo oprli na model razumljivosti besedil za slovenski jezik (Ferbežar, 2012) in njegovo empirično implementacijo s štiristopenjsko Likertovo lestvico. Zavedamo se, da je splošne kriterije razumljivosti težko postaviti, saj razumljivost obstajala le v interakciji med besedilom in bralci (Erjavec, 1998: v Leskovec 2009), bralci v množičnih medijih, na katerih so narejene empirične študije razumljivosti, pa so zelo heterogena skupina. Pri besedilih, napisanih v lahki obliki, se ciljna skupina zelo zoža. Zato kriterij upoštevanje naslovnika izpustimo, saj vemo, komu so besedila namenjena in katere jezikovne prilagoditve taka besedila potrebujejo. Tudi komunikativnost besedila nas ne bo zanimala, velikost pisave, opremljenost s primeri, slikovnim gradivom ipd. je mogoče dopolniti naknadno, ločeno od našega sistema. Od preostalih kriterijev razumljivosti bomo pregledali, kateri so najpomembnejši za razumljivost poenostavljenih besedil. Zanimale nas bodo jezikovne napake, ki jih zgenerira naš sistem. Jezikovne napake, ki jih pri pisanju delajo materni govorci slovenščine, načeloma ne pomenijo ovire pri razumevanju (Ferbežar, 2012), zanimalo pa nas bo, če velja enako pri avtomatskem poenostavljanju, ali pa morda tu prihaja do takih jezikovnih napak, ki jih materni govorec ne bi nikoli naredil in s tem usodno vpliva na razumevanje besedila. Zanimala nas bo tudi jedrnatost, leksikalna preprostost, skladenjska struktura in koherenca. Ker je pri poenostavljanju besedil pomemben odnos generiranega besedila do izvirnika, bo za nas pomembno, da generirano besedilo ohranja bistveno vsebino izvirnika in da je s tem povzemalno ustrezno. Zato bomo tako kot Zidarn (2020) spraševali po točnosti oz. po obstoju napačnih podatkov.

3 Metodologija

V tem poglavju predstavimo podatkovno množico za poenostavljanje besedil v slovenskem jeziku, ki je predhodno še ni bilo in postopek izdelave prototipa sistema za poenostavljanje besedil, kjer smo se naslonili na arhitekturo T5 modela [2] (Raffel idr., 2020). Za izbiro hiperparametrov in s tem določitev optimalnega modela za poenostavljanje slovenskih besedil uporabimo skupino metrik ROUGE (Lin, 2004) in BERTScore (Zhang, Kishore, Wu, Weinberger in Artzi, 2019). Poleg avtomatske evalvacije sistema uporabimo še človeško evalvacijo, predstavimo objektivne lastnosti stavkov in vprašalnik, ki smo ga sestavili.

3.1 Postopek priprave podatkovne množice

Vzporedno podatkovno množico smo naredili ročno z uporabo besedil, ki so zapisana tako v originalni kot v poenostavljeni verziji. Vzeli smo deset že obstoječih poenostavljenih slovenskih besedil, ki so jih pripravili v Zavodu Risa (Risa, 2020b) in njihove originale: Stari Grad (Voranc, 1981), Černjakova terba (Voranc, 1981), Tantadruj (Kosmač, 1980), Tržačan (Tavčar, 2020), Dostopna Ljubljana (2020), Breda (Magajna, 1940), Deček in blaznik (Grum, 1976), Jerinova Lida (Magajna, 1940), Romeo in Julija (Shakespeare, 1974) ter del Visoške kronike (Tavčar, 1987). Vsa besedila, razen Dostopna Ljubljana, so leposlovna dela. Zavod Risa je v lahko branje priredil še delo Pod svobodnim soncem (Saleški Finžgar, 1978), ki ga zaradi predolgega besedila in s tem prevelikega časovnega vložka, ki bi bil potreben za poravnavo z originalom, nismo vključili v množico. Originalna in njim pripadajoča poenostavljena slovenska besedila smo razdelili na posamezne stavke [3] in jih poravnali. Ustvarili smo okoli tisoč takih stavkov. Večinoma smo enemu kompleksnemu stavku priredili en preprost stavek. Včasih smo skupaj združili več kompleksnih stavkov, predvsem kadar je poenostavljeni stavek vseboval informacije iz več kompleksnih stavkov. Dovolili smo tudi, da je en kompleksni stavek razbit na več preprostih stavkov. Podajamo nekaj primerov vzporednih stavkov iz naše učne množice:

Civilisti so začeli takoj kopati in razmetavati razvaline, ki so pokrivale dno žrela.	Moški so začeli takoj kopati.

Tu gre za menjavo manj splošne besede čivilistiš preprostejšo besedo ”moški”ter za izpustitev manj pomembne informacije.

Salva je zagrmela, civilisti so se zrušili in od njih se je pocedila kri.	Možje so umrli.

Tu gre za abstraktivno povzemanje, ker so v povzetek vključene nove besede, pri čemer se pomen ohrani.

Napočila je jesen in dozorel je sad po vejevju.	Prišla je jesen. Sadje je dozorelo.

[2] `https://github.com/sabina-skubic/text-simplification-slovene.git`

[3] Ko govorimo o stavkih pri strojnem učenju, imamo v mislih, kar se v jezikoslovju pojmuje kot poved, strukturo od velike začetnice do končnega ločila, ki lahko vsebuje več stavkov.

Tu gre za razbitje povedi na dva preprosta stavka, za odpravo metonimije (tj. besedna figura, v kateri je ime za neko stvar zamenjano z drugim predmetno, količinsko povezan pojmom) ”vejevjež besedo šadno drevje”ter za zamenjavo manj znane besede šadš pogosteje uporabljeno besedo šadje”.

Oprtil si je tisti košek, s katerim je zahajal pomladi prašičke kupovat po pogorju, in je odšel v Trst, kjer so tedaj še otroke prodajali.	Na rame si je dal koš za prašičke. Šel je v mesto Trst. Tam so včasih prodajali otroke.

Tu gre za razbitje dolge povedi v tri preproste stavke.

Odpasal sem mu kiras ter pograbil mošnjo, ki jo je res nosil na vratu. Truplo sem zavlekel še bolj v goščo ter ga vrgel v jarek.	Oče je vzel tudi njegov denar, Joštovo truplo pa je vrgel v jarek.

Tu gre za zgostitev dveh povedi v eno preprostejšo, za izpustitev manj pomembnih informacij, za zamenjavo besede ’mošnja’ z besedo ’denar’ ter za za zamenjavo pripovedovalne perspektive: iz prvoosebne pripovedi v tretjeosebno.

Nenadno je zagrabil Lido za ramena in jo zvrnil na hrbet na travo, v trenutku je zasadil svoje ustnice v njene.	Prijel jo je za rame, jo položil na travo in jo poljubil.

Tu gre za preneseni pomen: zasadil svoje ustnice v njene zamenjamo z jo poljubil.

Celotno podatkovno množico vzporednih stavkov smo razbili na tri dele. Prvi del je sestavljen iz približno 80 % vzporednih parov stavkov in smo ga uporabili za učno množico, drugi del je sestavljen iz približno 10 % vzporednih stavkov, ki so služili kot evalvacijska množica, preostalih 10 % smo uporabili kot testno množico. Podatke smo razbili na nivoju odstavkov in s tem zagotovili ohranjanje konteksta znotraj odstavkov. Z uvedbo evalvacijske množice smo zagotovili, da se testna množica ne prilagaja pretirano podatkom. Testno množico smo uporabili za določanje uspešnosti delovanja modela nad podatki, ki jih v fazi učenja še ni procesiral. Na teh podatkih smo računali kvantitativne avtomatske mere uspešnosti modela z različno nastavljenimi parametri in jih uporabili še za končno človeško evalvacijo.

3.2 Postopek izdelave sistema za poenostavljanje besedil

Za poenostavljanje besedil smo prilagodili slovenski T5-sl-small model, ki sta ga izdelala Ulčar in Robnik-Šikonja (2022). Zaradi pomnilniške omejitve našega GPU računalnika se nismo odločili za uporabo slovenskega T5-sl-large ali večjezičnega mT5-large modela. Ker model uporablja učenje s prenosom znanja, ga nismo spreminjali, podali smo mu le našo podatkovno množico. Vhodni podatki so bili odstavki besedil, izhodni pa njihove poenostavitve. Ugotoviti smo morali le primerne hiperparametre, ki bodo vračali najboljše poenostavitve odstavkov. Hiperparametri so parametri modela, ki določajo lastnosti modela in kontrolirajo proces učenja, zato jih nastavimo pred učenjem modela. Analizirali smo hiperparametre: število epoh, velikost učnega paketa in gradient akumulacije (angl.

gradient accumulation). Število epoh nam pove, koliko ponovitev učenja uporabimo, preden se algoritem ustavi, pri čemer se kot ena ponovitev učenja šteje, ko učeni model soočimo z vsemi primeri v učni množici preden posodobimo uteži mreže. Vhode podajamo mreži v paketih fiksne velikosti. Gradient akumulacije se uporablja, ko paket razdelimo na več manjših (angl. mini-batch), ki jih zaporedno poženemo, njihove rezultate pa kopičimo oz. akumuliramo. S temi rezultati posodobimo parametre modela na koncu učenja vsakega manjšega paketa. S tem porabimo manj GPU pomnilnika. Vrednosti hiperparametrov smo ročno nastavljali, tako da smo jih eksponentno večali, kar se je izkazalo za časovno potratno.

3.3 Postopek evalvacije sistema

Poenostavitev stavkov iz učne množice, ki smo jih z modeli z različnimi hiperparametri zgenerirali, smo ocenili z avtomatskimi merami za evalvacijo ustvarjenih stavkov ROUGE in BERTScore. Ker ta metodologija ne upošteva bralčevega razumevanja, smo nekatere zgenerirane odstavke evalvirali tudi s pomočjo človeške presoje. Izhajajoč iz jezikoslovnih modelov razumljivosti smo izdelali vprašalnik s kriteriji razumljivosti, ki smo ga dali v reševanje naravnim govorcem slovenščine.

3.3.1 Predstavitev evalvacije sistema z avtomatskimi merami

Za vrednotenje, kako dobro delujejo modeli z različnim hiperparametri, smo uporabili evalvacijsko in testno množico. S slednjo smo ugotavljali, ali se model pretirano prilagaja učnim podatkom.

Za evalviranje povzetkov in poenostavitev se uporabljajo različne avtomatske mere. Te lahko razdelimo na:

- metrike, ki uporabljajo n-grame, od katerih se najpogosteje uporabljajo BLEU (angl. BiLingual Evaluation Understudy) (Papineni, Roukos, Ward in Zhu, 2002), METEOR (angl. Metric for Evaluation of Translation with Explicit ORdering) (Denkowski in Lavie, 2011), ROUGE (angl. Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) in SARI (Xu, Napoles, Pavlick, Chen in Callison-Burch, 2016),
- metrike, ki uporabljajo razdaljo (angl. edit-distance-based metrics),
- metrike, ki uporabljajo vložitve, npr. besedne ali stavčne vložitve,
- naučene metrike, npr BERTScore (Zhang, Kishore, Wu, Weinberger in Artzi, 2019).

Za evalvacijo uspešnosti naših modelov z različnimi parametri smo uporabili meri ROUGE (Lin, 2004) in BERTScore (Zhang, Kishore, Wu, Weinberger in Artzi, 2019). ROUGE smo izbrali zato, ker je med najbolj uporabljenimi merami za uspešnost avtomatskega povzemanja besedil. Pri poenostavljanju besedil se namreč v veliki meri poslužujemo povzemanja, ko spuščamo manj pomembne informacije. BERTScore smo uporabili, ker ne upošteva le pojavitve istih besed med kompleksnim in generiranim poenostavljenim stavkom, temveč upošteva tudi sinonime v stavkih. Pri tej meri dobimo visoko oceno

za generiran stavek tudi, če pri poenostavljanju uporabimo sinonime ali če stavek parafraziramo, kar pa ne velja pri meri ROUGE. Izračunali smo tudi izgubo. Le-ta pove napako napovedi modela za naše podatke. Njeno vrednost izračunamo tako, da primerjamo vrnjene stavke modela s poenostavljenimi stavki iz evalvacijske množice.

3.3.1.1 Avtomatska mera ROUGE

ROUGE ni enotna metrika, ampak vsebuje naslednje metrike: ROUGE-1, ROUGE-2, ROUGE-L in ROUGE-LSum. Število n v imenu prvih dveh metrik predstavlja prekrivanje n-gramov med referenčnim in generiranim stavkom. Tretja metrika, ROUGE-L, deluje na nivoju celotnega stavka, tako da prepozna najdaljši skupni niz besed LCS (angl. Longest common sequence) med referenčnim in generiranim stavkom. Pri najdaljšem skupnem nizu ni potrebno, da besede stojijo skupaj, morajo pa se pojaviti v enakem vrstnem redu. S tem daje globlji vpogled v strukturo generiranih stavkov. Četrta metrika, ROUGE-LSum, deluje na nivoju celotnega povzetka, pri čemer se računa najdaljši skupni niz med vsakim parom referenčnega in generiranega stavka.

Ocena ROUGE potrebuje vrednost priklica R (angl. recall) in točnosti P (angl. precision). Priklic se izračuna kot razmerje med številom prekrivajočih se n-gramov, ki so tako v referenčnem kot v generiranem stavku in številom n-gramov v referenčnem stavku. Točnost se izračuna kot razmerje med številom prekrivajočih se n-gramov, ki so tako v referenčnem kot v generiranem stavku in številom n-gramov v generiranih stavkih. Zatem lahko izračunamo vrednost $F1$, ki je harmonična sredina priklica in točnosti in leži na intervalu med 0 in 1. Njena formula je $F1 = \frac{2*P*R}{P+R}$, končna vrednost ROUGE pa je $F1$ pretvorjen v odstotek.

Primer izračuna:

Referenčni stavek: Marsikdaj so ljudje napleli pogovor o Ljubljani.

Generiran stavek: Ljudje so se pogovarjali o Ljubljani.

Unigram (so, ljudje, o, Ljubljani): $R = \frac{4}{7}$, $P = \frac{4}{6}$, $F1 = 0.6154$, ROUGE-1= 61.54.
Bigram (o Ljubljani): $R = \frac{1}{7}$, $P = \frac{1}{6}$, $F1 = 0.1538$, ROUGE-2= 15.38.
LCS (ljudje o Ljubljani): $R = \frac{3}{7}$, $P = \frac{1}{2}$, $F1 = 0.4615$, ROUGE-L= 46.15.

Metrika ROUGE v splošnem dokaj dobro oceni uspešnost generiranih stavkov, vendar se moramo zavedati, da je za uspešnost modela smiselno uporabiti še druge avtomatske mere. Šibka točka metrike ROUGE je, da temelji zgolj na tem, koliko besed se hkrati pojavi v generiranem in referenčnem stavku, ne upošteva pa smiselnosti celotnega stavka. Če generiramo stavek iz popolnoma istih besed, kot se nahajajo v referenčnem stavku, bomo dobili rezultat 100.0, čeprav so besede morda naključno premešane in kot celota nimajo smisla.

3.3.1.2 Avtomatska mera BERTScore

Tudi BERTScore podobno kot ROUGE za svoj izračun potrebuje vrednosti točnosti P in priklica R med vsemi pari besed, vendar ne zahteva popolnega ujemanja med besedama. Pravimo, da računa oceno podobnosti med vsako besedo v referenčnem in generiranem stavku, pri čemer za izračun podobnost med njima uporabi kontekstne vektorske vložitve na podlagi kosinusne podobnosti. Ta tip vložitve za vsak stavek, v kateri nastopa beseda, ustvari ločeno vektorsko predstavitev in tako loči besede enakopisnice po različnih pomenih. Npr. besedi pót – hoja in pôt – znoj obravnava ločeno. Posamezna beseda ima več

vektorskih predstavitev, odvisno v kakšnem kontekstu se pojavi. Vrednosti BERTScore ležijo med 0 in 1.

Za računanje metrike BERTScore uporabljamo jezikovni model BERT, ki temelji na arhitekturi transformer. Vektorske vložitve se pri BERT modelu pridobijo z maskiranjem besed v stavkih, ki jih model skuša napovedati. Na voljo imamo več različic modela BERT glede na jezik: angleško, kitajsko in večjezikovno inačico. Modeli se ločijo tudi po globini oz. številu nivojev. Za našo evalvacijo smo uporabili večjezikovni model (mBERTScore), ki zajema slovenščino, z 12 nivoji. Uporabili smo tudi model SloBERTa (Ulčar in Robnik-Šikonja, 2021), ki je enojezični model, narejen za slovenski jezik. Uporablja arhitekturo RoBERTa, ki je različica BERT modela.

Izračunajmo vrednosti BERTScore:
Referenčni stavek: Marsikdaj so ljudje napleli pogovor o Ljubljani.
Generiran stavek: Ljudje so se pogovarjali o Ljubljani.
mBERTScore: $P = 0.8854$, $R = 0.6967$, $F1 = 0.7798$.
SloBERTa: $P = 0.9114$, $R = 0.7599$, $F1 = 0.8288$.

3.3.1.3 Izbira hiperparametrov na podlagi ROUGE in BERTScore

Pri iskanju najboljših vrednosti hiperparametrov, smo spreminjali po en parameter naenkrat in iskali optimalno kombinacijo, ki generira čim boljše rezultate vseh zgeneriranih stavkov glede na skupino metrik ROUGE in BERTScore. Vrednosti hiperparametrov velikost paketa, gradient akumulacije in število epoh smo ročno nastavljali, tako da smo jih eksponentno večali. Vrednosti parametrov smo nastavljali na 1, 2, 4, 8, 16, 32 in 64. Zaradi časovne zahtevnosti smo se po zgledu avtorjev Raffel idr. (2021) ustavili pri 64 epohah. Nismo testirali vseh kombinacij trojic velikost paketa, gradient akumulacije, število epoh. Opazili smo, da pri nizkih vrednostih vseh parametrov, npr. 2, 2, 2 dobimo slabe rezultate vseh ROUGE metrik, vseh BERT metrik in tudi zgenerirani stavki so večinoma nesmiselni. Najboljše rezultate smo dobili pri trojicah 2, 8, 64 in 8, 4, 64. V tabeli 1 prikažemo vrednosti metrik in izgube za obe najboljši trojici za testno in evalvacijsko množico na velikosti vzorca 96.

Vrednosti vseh mer so pri obeh izbirah trojic zelo podobne, podobne so med sabo tudi vrednosti mer za testno in evalvacijsko množico. Za končno izbiro hiperparametrov smo zato še ročno pregledali primere zgeneriranih stavkov. Model s parametri 2, 8, 64, torej 1. konfiguracija, je deloval na vtis nekoliko boljši, zato so to naše končne vrednosti hiperparametrov.

Opazili smo, da so vrednosti ROUGE in BERTScore visoke in nam dajejo dobro napoved delovanja našega sistema. Za primerjavo lahko vzamemo angleški model tipa T5-small (Raffel idr., 2020) za nalogo abstraktivnega povzemanja dnevnih novic iz CNN. Model je dosegel vrednosti 41.12 za ROUGE-1, 19.56 za ROUGE-2 in 38.35 za ROUGE-L. Uporabni povzemalniki vračajo vrednosti ROUGE nekje od 20 naprej, najboljši za angleščino pa okoli 40. Npr. slovenski povzemalnik, ki ga je izdelal Zidarn (2019) doseže vrednosti 23,77 za ROUGE-1, 7,97 za ROUGE-2, 23,95 za ROUGE-L, 0,670 za BERTScore, drugi slovenski povzemalnik, ki ga je izdelal Žagar (2020) doseže vrednosti 24,97 za ROUGE-1, 7,43 za ROUGE-2, 21,50 za ROUGE-L in 0,679 za BERTScore, eden izmed trenutno najuspešnejših angleških povzemalnikov pa dosežejo vrednosti 44,17 za ROUGE-1, 21,47 za ROUGE-2 in 41,11 za ROUGE-L (Žagar, 2020).

	1. konfiguracija	2. konfiguracija
Velikost niza	2	8
Gradient akumulacije	8	4
Število epoh	64	64
Izguba za testno množico	2.79	2.78
ROUGE-1 za testno množico	32.03	32.41
ROUGE-2 za testno množico	13.77	14.44
ROUGE-L za testno množico	29.83	30.74
ROUGE-LSum za testno množico	29.91	30.60
Izguba za evalvacijsko množico	2.44	2.45
ROUGE-1 za evalvacijsko množico	32.76	32.22
ROUGE-2 za evalvacijsko množico	14.06	14.24
ROUGE-L za evalvacijsko množico	29.44	28.94
ROUGE-LSum za evalvacijsko množico	29.97	29.41
BERTScore z mBERTScore	0.81	0.78
BERTScore s SloBERTo	0.89	0.83

Tabela 1 – ROUGE in BERTScore metrike za različne hiperparametre.

3.3.2 Predstavitev človeške evalvacije sistema

Poleg avtomatske evalvacije smo se odločili za človeško evalvacijo, ker omogoča boljši vpogled v uspešnost poenostavljanja. Visoka ocena ROUGE ali BERTScore namreč ne zagotavlja nujno smiselnih in razumljivih poenostavljenih besedil. Predstavili bomo, katere so objektivne, formalne lastnosti stavkov, ki so pomembne za razumljivost, ter predstavili vprašalnik.

3.3.2.1 Predstavitev objektivnih lastnosti stavkov

V avtomatsko generiranih besedilih bomo sami pregledali objektivne lastnosti stavkov oz. ali so zastopana priporočila za pisanje v lahkem branju (Haramija in Knapp, 2019), tako da bomo identificirali jezikovne napake ter preverili, če je zgradba stavkov poenostavljena, ali so povedi razbite na krajše, pretežno enostavčne povedi, pri morebitnih večstavčnih povedih pa bomo preverili, če so enostavne s preprostimi priredji in podredji. Poleg tega bomo za nikalne povedi preverili, če vsebujejo trpnik, če je znotraj besedila za isti pomen vedno uporabljena ista beseda, preverili bomo, da ni tujk, večmestnih števil in posebnih simbolov, kot so odstotki, podpičja, tropičja, narekovaji ipd. S tem bomo zajeli kriterija razumljivosti leksikalna in skladenjska preprostost, s katerima si bomo lahko pomagali pri iskanju odgovora na vprašanje, v kakšni meri bodo stavki, zgenerirani z našim sistemom, razumljivi bralcem.

3.3.2.2 Predstavitev vprašalnika

Za evalvacijo uspešnosti našega modela na nivoju besedil smo izdelali vprašalnik smiselnosti in razumljivosti avtomatsko generiranih odstavkov, ki smo ga dali v reševanje naravnim govorcem slovenščine. Izhajali smo iz jezikoslovnih modelov razumljivosti in se naslonili predvsem na delo Ferbežar (2006). Vprašalnik je kombinacija vprašalnika za

jezikoslovce, ki ga je izdelala Ferbežar (2006) in splošnega vprašanja o razumljivosti, ki smo mu dodali vprašanje o razlogih morebitne nerazumljivosti. Tu smo podali več opcij, iz katerih so lahko anketiranci izbirali razloge. Ker naša ciljna skupina niso jezikoslovci, smo vprašalnik priredili z bolj opisnimi vprašanji.

V vprašalniku smo najprej vprašali po nekaj splošnih demografskih podatkih, kot so spol, starost in izobrazba. Nadalje smo vprašalnik razdelili na dva dela. Povsod smo odgovore ponudili na štiristopenjski Likertovi lestvici. Čeprav so pri generativnih nalogah bolj pogoste pet-, dvo- ali tristopenjske lestvice (Lee, Gatt, Miltenburg, Wubben in Krahmer, 2019) ter najbolj primerne sedemstopenjske lestvice (Lee, Gatt, Miltenburg, Wubben in Krahmer, 2019), smo se skušali čim bolj držati originalne raziskave, kot jo je izvedla Ferbežar (2006). S štiristopenjsko lestvico se izognemo, da ocenjevalci izberejo srednjo vrednost, kadar so negotovi v svoj odgovor in jih s tem spodbudimo k globljemu razmisleku.

V prvem delu vprašalnika smo anketirancem prikazali samo avtomatsko poenostavljen odstavek. V tem delu smo preverjali, kako anketiranci dojemajo avtomatsko generirano besedilo kot končen samostojen izdelek. S tem smo želeli postaviti besedilo v čim bolj naravno situacijo in se (vsaj deloma) izogniti umetni testni situaciji. Iz testne množice smo izbrali dva odstavka in njune ročno generirane in avtomatsko generirane poenostavitve z najboljšim modelom. Izbrali smo en kratek in en daljši odstavek. Kratek odstavek se nam je zdel primeren, ker se hitro prebere in je zato za bralca manj časovno zahteven ter ker je naš sistem ta odstavek zgeneriral brez slovničnih napak. Po drugi strani ima daljši zgenerirani odstavek več slovničnih napak. Pri tem odstavku smo pričakovali, da čeprav je (izvirno) besedilo izrezano iz celote, poda bralcu bolj popolno sliko vsebine. Zato se nam zdi, da sta izbrana odstavka reprezentativna in da bi za drugo izbiro besedil dobili podobne rezultate. Zdi se nam, da sta besedili sami brez konteksta (torej originalnega besedila) dokaj razumljivo napisani. Večji problem je, da ne ohranjata bistva. Ker gre za poenostavljeni besedili, nas je predvsem zanimalo, če sta zapisani z enostavnim besediščem in ali sta logično povezani. O leksikalni preprostosti smo zato spraševali z vprašanjem: Ali je izbira besed v poenostavljenem besedilu primerna (vsebuje enostavne in pogoste besede)? Koherenco odstavka smo naslovili z vprašanjem: Ali je poenostavljeno besedilo logično povezano?

V drugem delu vprašalnika smo poleg avtomatsko zgeneriranega besedila prikazali še izvirnik in ročno poenostavljen odstavek. Zanimalo nas je, če avtomatsko generirano besedilo ohranja bistveno vsebino izvirnika in je zato povzemalno ustrezno ter ali je zapisano razumljivo. Želeli smo izvedeti morebitne vzroke nerazumljivosti, tako da smo navedli različne opcije, med katerimi so ocenjevalci lahko izbirali, lahko pa so dopisali tudi nov razlog. Opcije smo vzeli iz slovenskega modela razumljivosti in dodali še povzemalno ustreznost, tako da so bile vse opcije, iz katerih so lahko izbirali naslednje:

- napačni podatki (iz kriterija povzemalne ustreznosti),
- slovnične napake (iz kriterija jezikovna pravilnost),
- neprimerna izbira besed (iz kriterija leksikalna preprostost),
- napačna skladnja (iz kriterija skladenjska preprostost),
- premalo informacij (iz kriterija jedrnatost),

- preveč informacij (iz kriterija jedrnatost),
- nelogična povezanost besedila (iz kriterija koherenca),
- neohranjanje bistva izvirnika (iz kriterija povzemalna ustreznost).

Za konec smo vprašali še, če sta avtomatsko in ročno generirana odstavka res enostavnejša od izvirnika.

Celoten vprašalnik je prikazan v poglavju Priloge. Vprašalnik je v celoti izpolnilo 25 oseb, 32 pa jih je izpolnilo vsa vprašanja samo za prvo besedilo. Ker gre za ocenjevanje dveh ločenih besedil, se nam je zdelo smiselno ohraniti in analizirati odgovore na dveh različno velikih množicah ocenjevalcev. Vsi ocenjevalci so bili stari nad 21 let, med njimi je bilo 7 moških in 25 žensk z različno izobrazbo, od končane srednje šole do končane univerzitetne izobrazbe.

Iz zbranih rezultatov ocenjevalcev smo za vse spremenljivke izračunali mediano, povprečje in standardni odklon.

S pomočjo vprašalnika smo želeli odgovorili na vprašanja, ali lahko z modelom tvorimo smiselne poenostavljene slovenske stavke in besedila, v kakšni meri bodo stavki in besedila, zgenerirani z našim sistemom, razumljivi bralcem ter kateri so kriteriji za razumljivost avtomatsko generiranih poenostavljenih besedil v slovenščini. Odgovor na prvo vprašanje nam bosta dali vprašanji "Ali je besedilo logično povezano?"ter Še vam zdi "poenostavljeno"besedilo res enostavnejše od izvirnika?". S prvim vprašanjem sprašujemo po smiselnosti celote zgeneriranega odstavka, z drugim pa po poenostavitvi odstavka v primerjavi z izvirnikom.

Odgovor na drugo vprašanje nam bo dalo vprašanje "Kako razumljivo je poenostavljeno besedilo?". Tu bomo lahko testirali še ničelne hipoteze, da sta prvo avtomatsko in prvo ročno besedilo enako razumljivi, da sta drugo avtomatsko in drugo ročno besedilo enako razumljivi ter da sta prvo in drugo avtomatsko besedilo enako razumljivi. Za testiranje ničelnih hipotez bomo zaradi ordinalnih spremenljivk uporabili Wilcoxonov test predznačenih rangov (angl. Wilcoxon signed rank test). Če z X označimo razumljivost prvega besedila in z Y razumljivost drugega besedila, velja da so za razliko $X - Y$ rangi pozitivni, če je $X > Y$ in negativni, če je $X < Y$. Če je povprečna vrednost pozitivnih rangov za $X - Y$ večja od negativnih rangov, to nakazuje, da so vrednosti X v povprečju višje od Y in torej, da je prvo besedilo manj razumljivo od drugega in obratno, če je povprečna vrednost negativnih rangov za $X - Y$ večja od pozitivnih rangov, to nakazuje, da so vrednosti Y v povprečju višje od X in torej, da je drugo besedilo manj razumljivo od prvega. V tabeli 2 podajamo vrednost z za zavrnitev ničelne hipoteze pri najbolj pogosto uporabljenih p $p \leq 0.05$ in $p \leq 0.01$ in različnih velikostih vzorca n.

Pri vprašanju, kateri so kriteriji za razumljivost avtomatsko generiranih poenostavljenih besedil v slovenščini, smo predpostavili, da so pomembni kriteriji naslednji: jedrnatost, jezikovna pravilnost, leksikalna preprostost, skladenjska preprostost, koherenca in povzemalna ustreznost. Prvih pet kriterijev je vzetih iz slovenskega modela razumljivosti (Ferbežar, 2006), preostalih dveh (komunikativnost besedila in upoštevanja naslovnika) ne bomo preverjali, ker zgenerirano besedilo ni končno besedilo, ki bi bilo stilsko urejeno in oblikovno prilagojeno za lahko branje in ker smo kriterije razumljivost poenostavljenih besedil preverjali pri "običajnih"govorcih slovenščine. Da bi preverjali kriterij upoštevanje naslovnika, bi morali dati zgenerirana besedila v branje ciljni skupini, ki so

n	p = 0.05	p = 0.01
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68
26	98	75
27	107	83
28	116	91
29	126	100
30	137	109

Tabela 2 – z vrednosti za Wilcoxonov test predznačenih rangov

jim poenostavljena besedila namenjena. Ker gre pri poenostavljanju besedil tudi za povzemanje, smo predpostavili, da so lahko le ustrezno povzeta besedila razumljiva in zato dodali še ta kriterij. V namen določanja pomembnih kriterijev razumljivosti smo v vprašalniku spraševali po morebitnih razlogih za nerazumljivost besedila. Možni odgovori so napačni podatki, slovnične napake, neprimerna izbira besed, napačna skladnja, premalo informacij, preveč informacij in nelogična povezanost besedila. Ti so zajeli kriterije iz slovenskega modela razumljivosti. Odgovor neohranjanje bistva izvirnika je zajel kriterij povzemalne ustreznosti. Pri tem smo predpostavili, da oba ročno generirana odstavka upoštevata vse naštete kriterije razumljivosti.

Da bi ugotovili, če je povzemalna ustreznost smiseln in pomemben kriterij za poenostavljena besedila, smo izračunali korelacijo med razumljivostjo in povzemalno ustreznostjo pri avtomatsko in ročno generiranih besedil. Po povzemalni ustreznosti smo spraševali z vprašanjem ”Ali poenostavljeno besedilo ohranja bistveno vsebino izvirnika?”. Po razumljivosti smo spraševali z vprašanjem ”Kako razumljivo je poenostavljeno besedilo?”. Pričakovali smo, da bo njuna korelacija pozitivna in visoka in da bomo na podlagi tega sklepali, da sta med seboj povezana. Ker so naše spremenljivke ordinalne in ne nujno normalno porazdeljene, smo za izračun korelacije izbrali Kendallov tau, ki lahko zavzame katero koli vrednost med -1 in 1. Če zavzame vrednost 0, pomeni, da med spremenljivkami

ni nobene povezanosti. Če zavzame vrednosti blizu ničle, tako pozitivne kot negativne, pomeni, da je korelacija majhna. Bolj kot se vrednosti bližajo 1 oz -1, večja je povezanost med spremenljivkama. Če je vrednost negativna, gre za negativno korelacijo, kar pomeni, da če se vrednost ene spremenljivke veča, se vrednost druge zmanjša, če pa je vrednost pozitivna, gre za pozitivno korelacijo, kar pomeni, da če se vrednost ene spremenljivke veča, se veča tudi vrednost druge. Velja, da če je ena spremenljivka del vzroka druge, sta dani spremenljivki vzročno povezani, če pa sta spremenljivki neodvisni, sta nekorelirani (Zemljak, 2019).

Predpostavili smo, da je razumljivost poenostavljenih besedil odvisna od več dejavnikov oz. da obstaja več kriterijev, ki jo določajo. Predvidevali smo, kar nam že samo ime pove, da je za poenostavljena besedila med najpomembnejšimi kriteriji enostavnost besedišča, poleg tega pa še koherenca oz. logična povezanost besedila. Zato bomo izračunali multiplo korelacijo med odvisno spremenljivko razumljivostjo besedil in dvema neodvisnima spremenljivkama logična povezanost besedila in enostavnost besed v besedilu. S tem bomo proučevali odvisnost razumljivosti poenostavljenih besedil od ostalih dveh dejavnikov hkrati. Z multiplo korelacijo bomo ugotavljali delež variance, ki ga pojasnujeta spremenljivki logična povezanost besedila in enostavnost besed v besedilu oz. njun vpliv na dano spremenljivko razumljivost generiranih besedil. Uporabili bomo formulo: $P_{X.YZ} = \sqrt{\frac{\rho_{XY}^2+\rho_{XZ}^2-2*\rho_{XY}*\rho_{XZ}*\rho_{YZ}}{1-\rho_{YZ}^2}}$.

Vrednost multiple korelacije leži na intervalu od 0 do 1. Namesto Pearsonovega koeficienta ρ bomo uporabimo Kendallov tau.

Pomembnost preostalih kriterijev bo možno določiti glede na njihovo pojavnost pri razlogih za nerazumljivost besedila. Pri leksikalni preprostosti smo predpostavili, da se odgovor slovnične napake ne bo pojavil med razlogi za slabšo razumljivost pri odstavku, ki teh napak ne vsebuje, in da bo zastopanost tega odgovora med obema besediloma določalo, kako pomemben je ta kriterij za razumljivost besedil.

4 Rezultati in diskusija

V tem poglavju predstavimo dva avtomatsko generirana odstavka z najboljšim modelom ter njuni ročni poenostavitvi, na katerih izračunamo vse avtomatske metrike ROUGE in BERTScore. Odstavka v nadaljevanju evalviramo še s pomočjo analize objektivnih lastnosti stavkov iz odstavkov in s pomočjo vprašalnika. Rezultate vprašalnika prikažemo po takšnem vrstnem redu, kot si sledijo v vprašalniku, nato pa skupne rezultate evalviramo preko zastavljene hipoteze in podvprašanj. Zaključimo z diskusijo prednosti in slabosti naše raziskave.

4.1 Evalvacija sistema z avtomatskimi merami

Spodaj sta prikazani obe ročno in avtomatsko poenostavljeni besedili za izbrani izvirni besedili in njihove ROUGE in BERTScore metrike.

Izvirnik 1

Neki večer je prišel oče pozno domov. Stopil je k postelji, kjer je mama ležala. Nato je Breda videla, kako so se mu razprostrli prsti in se bližali materinemu vratu. Slišala je, kako je mati zaječala in kako so šinile njene roke nekam v zrak. Ubil jo je. Zjutraj so prišli policisti, ki so uklenili očeta. Takoj po pogrebu so Bredo oddali v bolnišnico za otroke, ki so izgubili starše.

Ročno poenostavljeno besedilo 1

Neki večer, ko je mama spala, je oče prišel domov in mater ubil. Očeta so odpeljali policisti. Mater so pokopali. Bredo pa so poslali v bolnišnico za otroke, ki so izgubili starše.

Avtomatsko poenostavljeno besedilo 1

Neki večer je šel k postelji. Zvečer so prišli policisti. Po pogrebu so Bredo oddali v bolnišnico.

Izvirnik 2

Nekega dne je Vester spet videl, da gre nekdo v stolp. Ko sta se dva človeka začela povzpenjati navkreber, je v moškem spoznal Huberta, trgovskega pomočnika. Vester je videl, da je mlada kmetica iz okolice, kjer je že tudi on dobil kaj dobrega od nje, vsa utrujena in da težko hodi navkreber. Komaj sta bila v stolpu, je Hubert priskočil in kmetico udaril v obraz. Nato jo je še sunil v trebuh s svojim težkim škornjem. Kmetica je zastokala in padla na tla. Hubert pa je kakor ris skočil nanjo in jo začel obdelovati s pasjim bičem, ki ga je takrat potegnil izza pasa, in vpil: ”Boš vstala, kurba partizanska!”Ker je kmetica neznansko vekala, je pretepač režal nanjo: ”Le vpij, tukaj lahko vpiješ, ker te nihče ne sliši.”Kmetica ni več poskušala vstati, bila je omotena, ker ji je Hubert z obema nogama skočil na vrat in jo teptal, kar se je dalo. Micka je glasno zastokala: ”Hubert, prizanesi mi zaradi otroka, saj veš, da nisem sama.Žaradi njega, da bi ti prizanašal, vlačuga partizanska. ”Pod posteljo si skrivala bandita.Koga si skrivala pod posteljo?”Nič ne vem! je jokala Micka. Spet se je zakadil vanjo in jo stoječo začel obdelovati z bičem in škornji. Čuješ, nisem te prej klicala po imenu zaradi sebe, ampak zaradi otroka! ”Ti si zverina.”Ubij me, zverina, ki si. Hubert je tedaj vzel iz žepa dolgo, tanko vrv in Micki zataknil zanko na vrat, drugi konec vrvi je vrgel čez klin nad njo in začel na vso moč vleči. Brez pomisleka je Vester zagrabil gada za vrat in ga zagnal v globino proti Hubertu. Ko je Hubert videl kačo, je zakričal kakor obseden, spustil vrv in blazno odskočil. Zdaj je udrl naravnost skozi vhod in planil po bregu navzdol. Vester je takoj splezal na tla in odšel v žrelo k Micki. Micka je ležala

trda na zemlji. Naj jo je Vester tresel, kolikor je hotel, ni mogel dobiti od nje nobenega življenja.

Ročno poenostavljeno besedilo 2

Nekega dne je Vester spet videl, da gre nekdo v stolp. Bil je trgovski pomočnik Hubert. Micka je bila mlada kmetica. Vester je videl, da je utrujena. Ko jo je Hubert pripeljal v stolp, jo je udaril v obraz in brcnil v trebuh. Kmetica je padla na tla. Hubert jo je tepel s pasjim bičem in kričal: Boš vstala, kurba partizanska. Le vpij, tukaj te nihče ne sliši. Hubert ji je z obema nogama skočil na vrat in jo teptal. Micka ga je prosila: Hubert, nehaj, saj veš, imam otroka. Hubert jo je spraševal: Koga si skrivala pod posteljo? Micka je jokala: Nič ne vem. Hubert jo je tepel in brcal. Rekla mu je še: Ti si zverina. Ubij me. Hubert je hotel Micko obesiti. Vester je proti Hubertu vrgel strupeno kačo. Hubert je zakričal in zbežal, ker se je ustrašil kače. Vester pa je hitel k Micki. Micka je bila mrtva.

Avtomatsko poenostavljeno besedilo 2

Vester je videl, da gre nekdo v stolp. Huberta sta spoznala trgovskega pomočnika. Vester je videl, da je mlada kmetica zelo utrujena. Hubert je bil v stolp. Kmetica je padla na tla. Hubert je rekel: Boš vstala, kurba partizanska! Ko je bila pijanka, je bila zelo žalostna. Hubert je bila omotena. Micka je glasno rekla: Hubert, prizanesi mi zaradi otroka. Pod posteljo si skrivala bandita. Micka je jokala Micka. Spet se je zapel. Ti si zverina. Ubi me, zverina, ki si si. Hubert je vzel iz žepa dolgo vrv. Vester je bil zelo vesel. Hubert je videl kačo. Vester je odšel k Micki. Micka je ležala trda na zemlji.

Mera	Besedilo 1		Besedilo 2	
	Ročna poenostavitev	Avtomatska poenostavitev	Ročna poenostavitev	Avtomatska poenostavitev
ROUGE-1	50.42	37.25	51.63	41.45
ROUGE-2	25.64	28	29.4	25.32
ROUGE-L	38.66	37.25	39.3	38.46
ROUGE-LSum	43.70	37.25	50.9	40.60
BERTScore z mBERTScore	0.78	0.78	0.75	0.73
BERTScore s SloBERTo	0.86	0.83	0.86	0.85

Tabela 3 – Mere ROUGE in BERTScore za poenostavljena besedila.

Brez zadržkov lahko rečemo, da sta ročno poenostavljeni besedili veliko boljši od avtomatskih, toda visoke vrednosti mer ROUGE in BERTScore pri obeh poenostavitvah, ki so prikazane v tabeli 3, delujejo zavajajoče. Čeprav vrednosti ROUGE-1 in ROUGE-LSum pokažeta razliko med ročno in avtomatsko poenostavitvijo za obe besedili, sta ti vrednosti za avtomatsko poenostavljeni besedili še vedno zelo visoki in dajeta dobro napoved delovanja našega sistema.

4.2 Človeška evalvacija modela

Za boljši vpogled v uspešnost delovanja našega sistema uporabimo človeško evalvacijo, in sicer na nivoju stavkov oz. povedi. Na zgeneriranih odstavkih pregledamo objektivne lastnosti, smiselnost in razumljivost zgeneriranih odstavkov pa bolj podrobno analiziramo s pomočjo vprašalnika.

4.2.1 Rezultati analize objektivnih lastnosti stavkov

Na obeh zgoraj avtomatsko generiranih besedilih smo z analizo objektivnih, formalnih lastnosti avtomatsko poenostavljenih odstavkov, ki so pomembne za razumljivost, ugotovili, da ne uporabljajo nikalnih povedi, trpnika, sinonimov, tujk, prenesenih pomenov, večštevilčnih števil ali posebnih znakov. Vse povedi so pri prvem besedilu enostavčne, pri drugem pa jih je večina enostavčnih, večstavčne pa vsebujejo preprosta podredja.

Med prvim in drugim avtomatsko generiranem besedilu se pojavljajo razlike glede na jezikovno pravilnost, in sicer so pri prvem vse povedi slovnično pravilne, drugo besedilo pa vsebuje veliko slovničnih napak, od napačnega sklanjanja in spreganja, napačne uporabe slovničnega in glagolskega spola, napačne izbire besedne vrste in podvajanja besed do napačno zapisanih besed. Spodaj smo zapisali napačne povedi in njihove pravilne različice.

Huberta sta spoznala trgovskega pomočnika.	Hubert je spoznal trgovskega pomočnika.

V besedilu se pojavi napačno slovnično število, dvojina namesto ednine.

Hubert je bil v stolp.	Hubert je bil v stolpu.

Beseda stolp je v napačnem sklonu.

Ko je bila pijanka, je bila zelo žalostna.	Ko je bila pijana, je bila zelo žalostna.

Izbrana je napačna besedna vrsta, namesto samostalnika pijanka bi bil primernejši glagol biti pijan.

Hubert je bila omotena.	Hubert je bil omoten.

Uporabljen je napačen spol v glagolski zvezi biti omoten. Še bolje bi bilo uporabiti sodobnejši izraz, da je bil omotičen, ki ima v korpusu Gigafida 2.0 večjo frekvenco kot samostalnik omotica, iz katerega je pridevnik izpeljan.

Micka je jokala Micka.	Micka je jokala.

Podvoji se beseda Micka.

Ubi me, zverina, ki si si.	Ubij me, zverina, ki si.

Pojavi se napačno zapisana beseda ubij in podvoji se beseda si. Ta stavek je primer, kako je starejše prozno literarno delo samo po sebi izziv za razumevanje, zaradi vrste jezikovnih specifik.

Na nivoju povedi zgornja avtomatsko generirana odstavka upoštevata večino priporočil za pisanje v lahkem branju (Haramija in Knapp, 2019) ter s tem izpolnjujeta kriterij

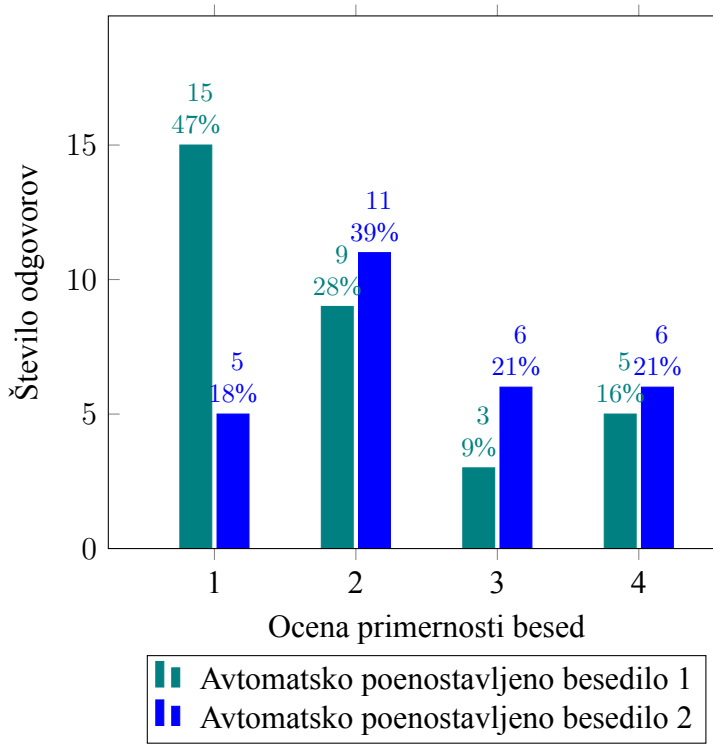
skladenjska preprostost. Ne izpolnjujeta pa v celoti kriterija jezikovne pravilnosti. Predvidevamo, da bi z dodanimi učnimi primeri to vrsto napak lahko odpravili. Poenostavljanje na nivoju povedi je solidno, veliko slabše je za daljša besedila.

4.2.2 Rezultati vprašalnika

Povprečna ocena primernosti izbora besed je v prvem avtomatsko generiranem besedilu na lestvici od 1 do 4, kjer 1 pomeni primerna izbira besed in 4 neprimerna izbira besed, 1.9 s standardnim odklonom 1.11, v drugem avtomatsko generiranem besedilo pa 2.5 s standardnim odklonom 1.04. Rezultati so prikazani na grafu 1 in v tabeli 4. Iz tega lahko zaključimo, da je izbira besed v prvem besedilu večinoma primerna, da vsebuje enostavne in pogoste besede. V 75 % primerov so ocenjevalci to lastnost ocenili z oceno 1 ali 2.

V primerjavi s prvim besedilom, je pri drugem besedilu povprečna ocena slabša, mediani pa sta sicer enaki za obe besedili in imata vrednost 2. Ocenjevalci so pri drugem besedilu zelo različno ocenjevali primernost izbire besed. Domnevamo, da so nekateri pri tem ocenjevali, ali so besede enostavne in pogoste, drugi pa tudi, ali je izbira besed ustrezna, dobra oz. prava. V slednjem primeru lahko sklepamo, da je ustreznosti besed slabša zaradi jezikovnih napak.

Graf 1 – Ocena primernosti/enostavnosti besed za avtomatsko poenostavljena besedila.

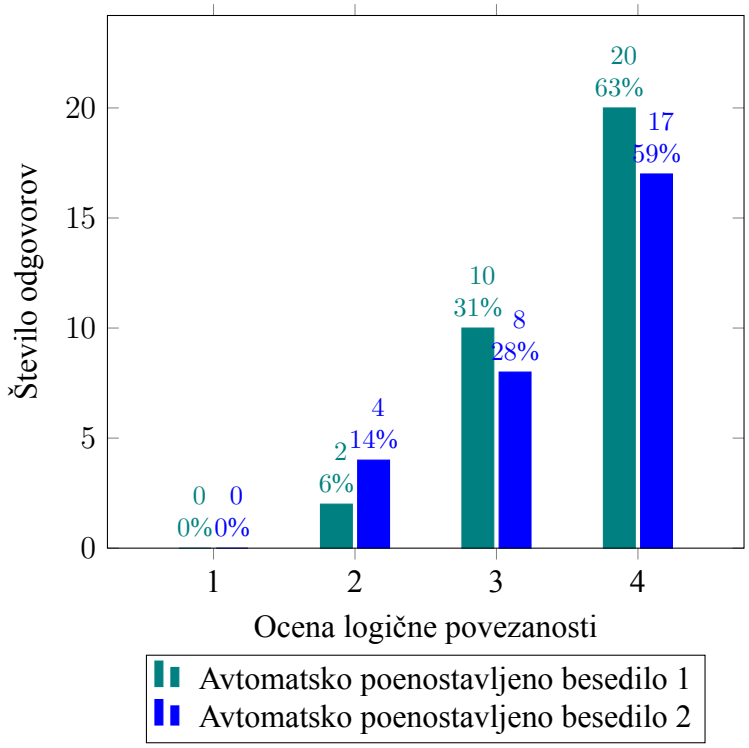


Povprečna ocena logične povezanosti besedila na lestvici od 1 do 4, kjer 1 pomeni v celoti logično povezano besedilo in 4 nelogično povezano besedilo, je za prvo avtomatsko generirano besedilo 3.6 s standardnim odklonom 0.62 in za drugo avtomatsko generirano besedilo 3.4 s standardnim odklonom 0.74. Mediana je v obeh primerih 4. Rezultati so prikazani na grafu 2 in v tabeli 5. Iz tega lahko sklenemo, da ocenjevalci nobeno od besedil niso dojemali kot logično povezano. Z oceno 3 ali 4 je to lastnost za prvo besedilo ocenilo kar 94 % ocenjevalcev in za drugo besedilo 87 % ocenjevalcev.

Ali je izbira besed v besedilu primerna (vsebuje enostavne in pogoste besede)?			
	Mediana	Povprečje	Standardni odklon
Avtomatsko poenostavljeno besedilo 1	2	1.9	1.11
Avtomatsko poenostavljeno besedilo 2	2	2.5	1.04

Tabela 4 – Ocena primernosti/enostavnosti besed za avtomatsko poenostavljena besedila.

Graf 2 – Ocena logične povezanosti za avtomatsko poenostavljena besedila.



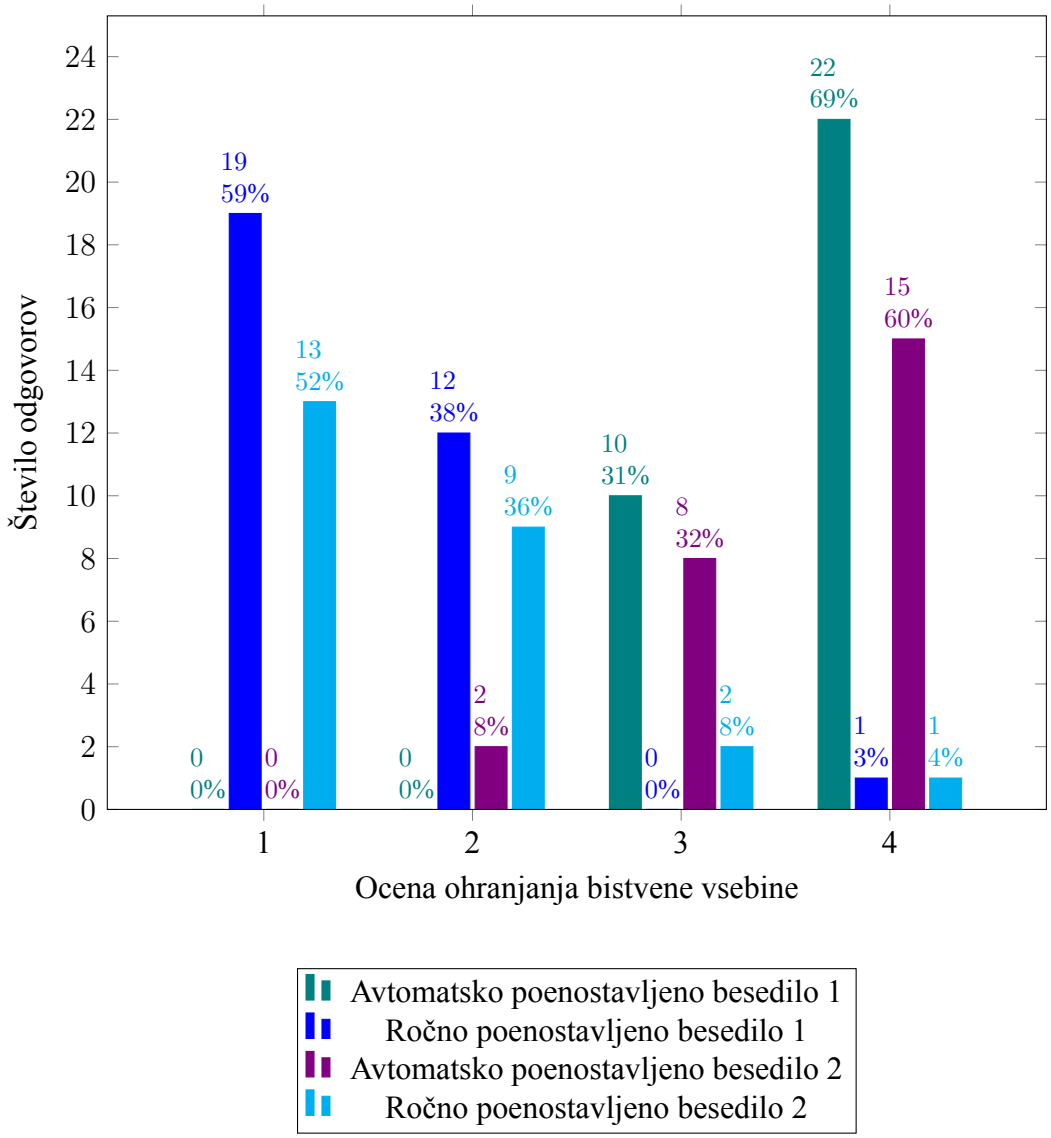
Ali je besedilo logično povezano?			
	Mediana	Povprečje	Standardni odklon
Avtomatsko poenostavljeno besedilo 1	4	3.6	0.62
Avtomatsko poenostavljeno besedilo 2	4	3.4	0.74

Tabela 5 – Ocena logične povezanosti za avtomatsko poenostavljena besedila.

Povprečna ocena ohranjanja bistvene vsebine izvirnika na lestvici od 1 do 4, kjer 1 pomeni, da jo ohranja v celoti in 4, da je ne ohranja, je 3.7 s standardnim odklonom 0.47 za prvo avtomatsko generirano besedilo in 3.5 s standardnim odklonom 0.65 za drugo av-

tomatsko generirano besedilo. Iz tega lahko sklenemo, da avtomatsko genrirani besedili ne ohranjata bistvene vsebine izvirnika. Pri prvem avtomatsko generiranem besedilu so vsi ocenjevalci to lastnost ocenili z oceno 3 ali 4, pri drugem besedilu pa 92 % ocenjevalcev. Po drugi strani je povprečna ocena ohranjanja bistvene vsebine izvirnika za prvo ročno generirano besedilo 1.5 s standardnim odklonom 0.67, za drugo ročno generirano besedilo pa 1.6 s standardnim odklonom 0.81. Mediani za obe avtomatsko generirani besedili je 4, za obe ročni pa 1. Rezultati so prikazani na grafu 3 in v tabeli 6. To potrjuje našo domnevo, da ročno, s strani strokovnjakov, generirano besedilo ohranja ali večinoma ohranja bistveno vsebino izvirnika. Pričakovali bi sicer še večji konsenz med ocenjevalci oz. da ne bo nihče ocenil ročnih poenostavitev z oceno 3 ali 4.

Graf 3 – Ocena ohranjanja bistvene vsebine



Povprečna ocena razumljivosti besedil na lestvici od 1 do 4, kjer 1 pomeni, da je razumljivo in 4, da je nerazumljivo, je za prvo avtomatsko generirano besedilo 3.3, s standardnim odklonom 0.86 in za drugo avtomatsko poenostavljeno besedilo 3.5 s standardnim odklonom 0.59. Mediana je 3.5 za prvo avtomatsko generirano besedilo in 4 za drugo. Prvo besedilo je z oceno 3 ali 4 ocenilo 88 % ocenjevalcev, drugo pa vsi ocenjevalci razen

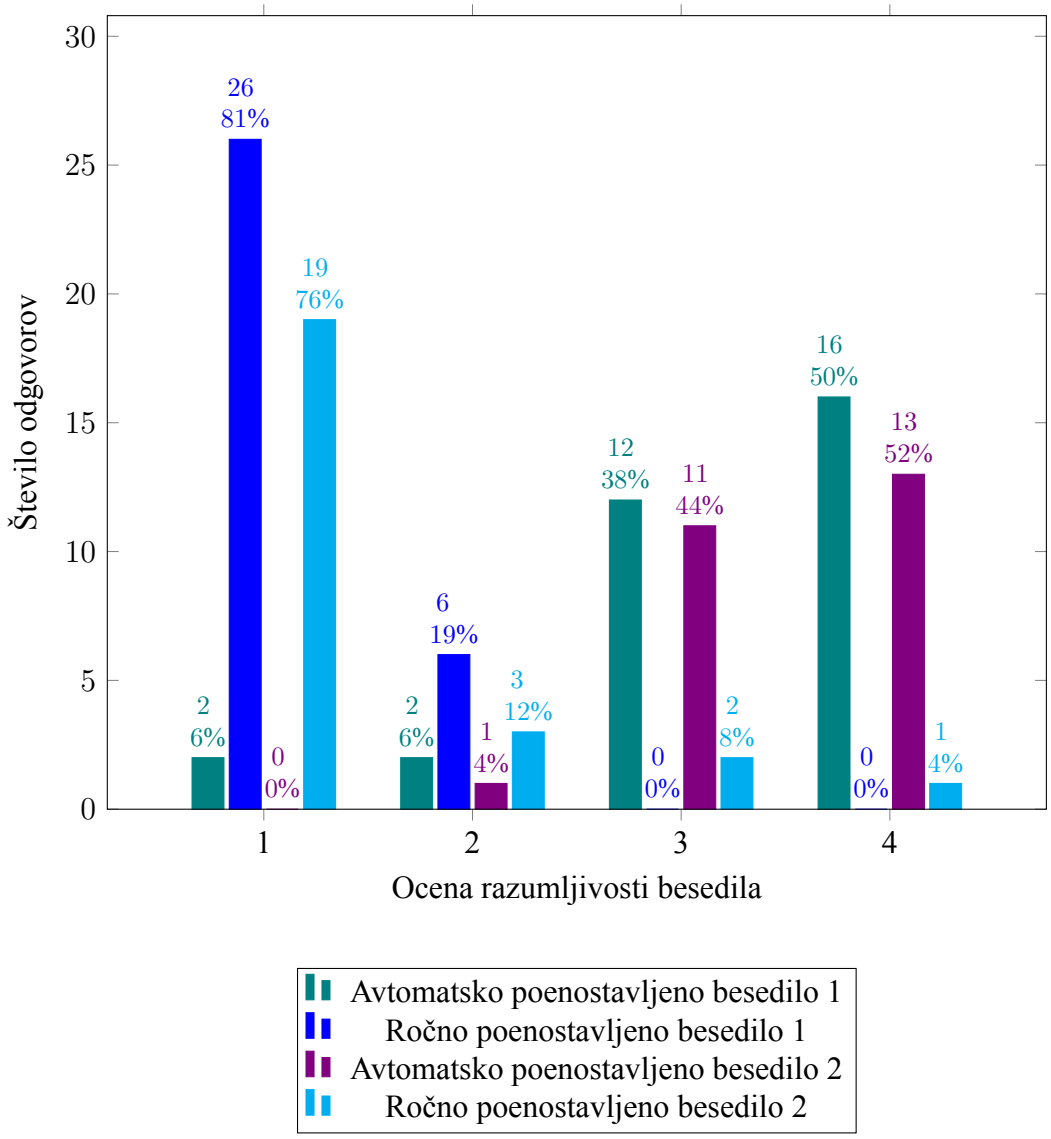
Ali poenostavljeno besedilo ohranja bistveno vsebino izvirnika?			
	Mediana	Povprečje	Standardni odklon
Avtomatsko poenostavljeno besedilo 1	4	3.7	0.47
Ročno poenostavljeno besedilo 1	1	1.5	0.67
Avtomatsko poenostavljeno besedilo 2	4	3.5	0.65
Ročno poenostavljeno besedilo 2	1	3.7	0.81

Tabela 6 – Ocena ohranjanja bistvene vsebine poenostavljenih besedil.

enega. Rezultati so prikazani na grafu 4 in v tabeli 7. Zaključimo lahko, da sta besedili slabo razumljivi in da je prvo vseeno bolj razumljivo kot drugo. Predvidevamo, da zato, ker je krajše, naš sistem pa solidno poenostavlja na nivoju povedi. Ker je v prvem besedilu manj povedi kot v drugem besedilu, je poenostavitev boljša in posledično bolj razumljiva. Predvidevamo tudi, da je prvo avtomatsko poenostavljeno besedilo bolj razumljivo kot drugo, ker je brez slovničnih napak.

Povprečna ocena razumljivosti za prvo ročno generirano besedilo je 1.2 s standardnim odklonom 0.40 in za drugo 1.4 s standardnim odklonom 0.82. Mediani sta za obe ročno generirani besedili 1. Rezultati so prikazani na grafu 4 in v tabeli 7. To potrjuje našo domnevo, da sta ročno generirani besedili razumljivi. Obe ročno generirani besedili sta (pričakovano) bolj razumljivi od njunih avtomatskih poenostavitev.

Graf 4 – Ocena razumljivosti poenostavljenih besedil



Kako razumljivo je poenostavljeno besedilo?			
	Mediana	Povprečje	Standardni odklon
Avtomatsko poenostavljeno besedilo 1	3.5	3.3	0.86
Ročno poenostavljeno besedilo 1	1	1.2	0.40
Avtomatsko poenostavljeno besedilo 2	4	3.5	0.59
Ročno poenostavljeno besedilo 2	1	1.4	0.82

Tabela 7 – Ocena razumljivosti poenostavljenih besedil.

Na vprašanje, kateri so glavni razlogi za slabšo razumljivost prvega avtomatsko generiranega poenostavljenega besedila, je odgovarjalo 25 ocenjevalcev. Izmed 32 skupnih

ocenjevalcev jih 7 na to vprašanje ni odgovorilo. Štirje izmed njih so razumljivost besedila ocenili z 1 ali 2, zato so lahko izpustili odgovarjanje na to vprašanje. Kot glavni razlog za slabšo razumljivost so ocenjevalci v 96 % navedli premalo informacij, v 68 % pa nelogično povezanost besedila. V 28 % so navedli kot razlog tudi neohranjanje bistva izvirnika, v 20 % neprimerno izbiro besed in napačno skladnjo ter v 8 % napačne podatke. Nihče ni podal dodatnega razloga za slabšo razumljivost. Rezultati so prikazani v tabeli 8.

Na vprašanje, kateri so glavni razlogi za slabšo razumljivost drugega avtomatsko generiranega poenostavljenega besedila, je odgovarjalo 22 ocenjevalcev. Izmed 25 skupnih ocenjevalcev trije na to vprašanje niso odgovorili. Eden izmed njih je razumljivost besedila ocenil z 2, zato je lahko izpustil odgovarjanje na to vprašanje, trije so za ročno poenostavljeno besedilo podali oceno 3 ali 4, zato ni jasno, za katero besedilo so navajali razloge za slabšo razumljivost. Tega pri sestavljanju vprašalnika nismo predvideli. Kot glavne razloge za slabšo razumljivost so ocenjevalci v 86 % navedli nelogično povezanost besedila, v 73 % premalo informacij, v 55 % napačno skladnjo in slovnične napake, v 50 % neohranjanje bistva izvirnika. V 23 % so navedli kot razlog tudi neprimerno izbiro besed, v 18 % napačni podatki in v 9 % preveč informacij. Zadnji razlog prinaša nekaj dvoumnosti: večini se zdi v besedilu premalo informacij, nekaterim pa preveč informacij. Nihče ni podal dodatnega razloga za slabšo razumljivost. Rezultati so prikazani v tabeli 8.

V obeh besedilih sta bila največkrat izbrana razloga za slabšo razumljivost nelogična povezanost besedila in premalo informacij.

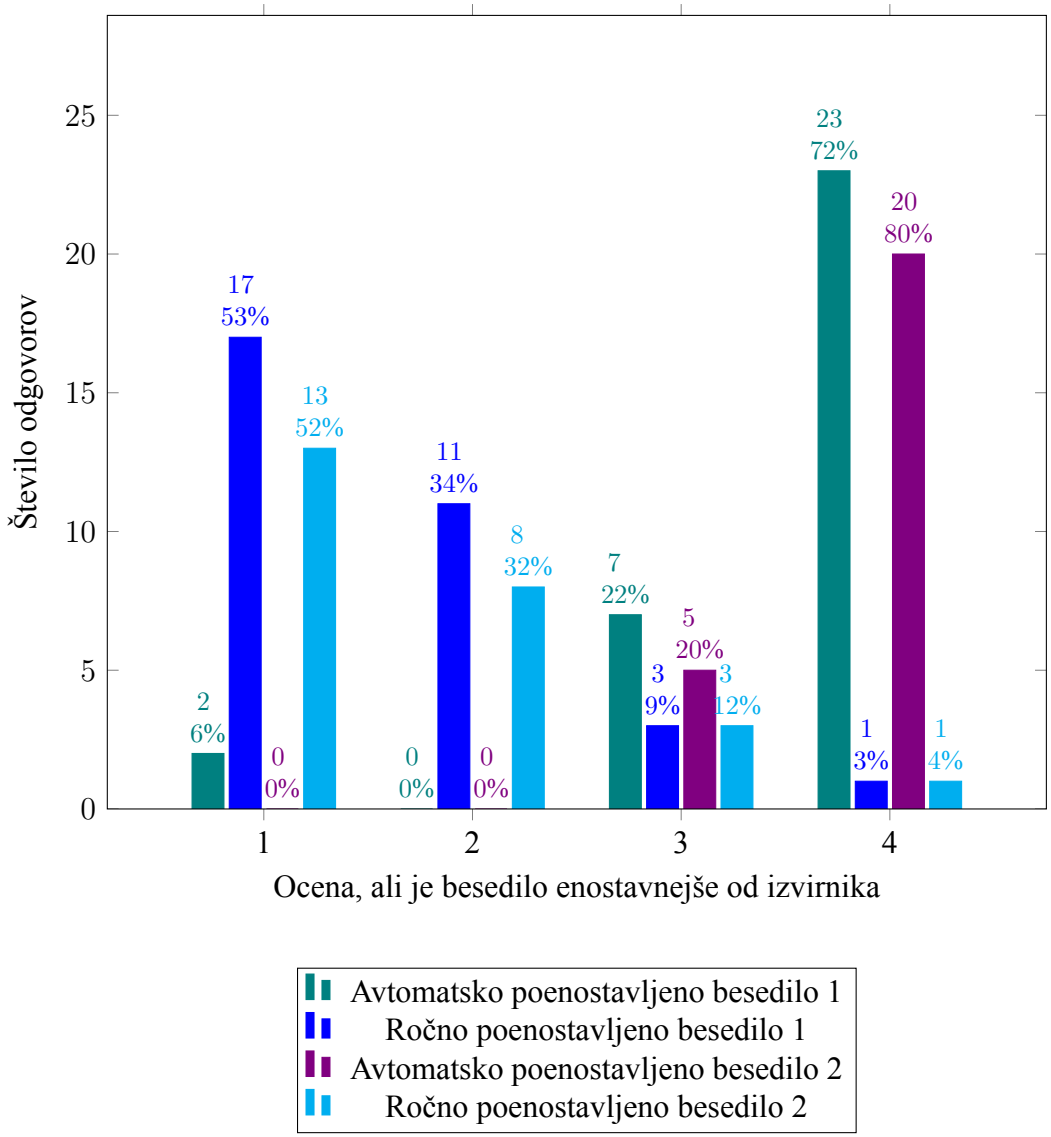
Glavni razlogi za slabšo razumljivost besedila				
	Avtomatsko poenostavljeno besedilo 1		Avtomatsko poenostavljeno besedilo 2	
	Število odgovorov	Odstotek 1	Število odgovorov	Odstotek
Napačni podatki	2	8%	4	18%
Slovnične napake	0	0%	12	55%
Neprimerna izbira besed	5	20%	5	23%
Napačna skladnja	5	20%	12	55%
Premalo informacij	24	96%	16	73%
Preveč informacij	0	0%	2	9%
Nelogična povezanost besedila	17	68%	19	86%
Neohranjanje bistva izvirnika	7	28%	11	50%

Tabela 8 – Glavni razlogi za slabšo razumljivosti avtomatsko poenostavljenih besedil.

Zanimalo nas je, če sta avtomatsko in ročno generirani besedili res enostavnejši od izvirnika. Pri tem ocenjevalcem nismo podali smernic, kaj je enostavnejše, ampak smo se zanašali na njihovo intuicijo. Povprečna ocena prvega avtomatsko generiranega besedila je 3.6, s standardnim odklonom 0.8, drugega pa 3.6, s standardnim odklonom 0.41.

Vrednosti obeh median je 4. Povprečna ocena prvega ročno generiranega besedila je 1.6, s standardnim odklonom 0.79, drugega pa 1.7, s standardnim odklonom 0.85. Vrednosti obeh median je 1. Rezultati so prikazani na grafu 5 in v tabeli 9. Zaključimo lahko, da avtomatsko generirani besedili nista enostavnejši, ter potrdimo predpostavko, da sta ročno generirani besedilo enostavnejši od izvirnika. Ocena za ročno generirana besedila je vseeno nekoliko presenetljiva. Pričakovali bi, da bodo (skoraj) vsi menili, da je to besedilo enostavnejše od izvirnika ali pa vsaj večinoma enostavnejši in da ga ne bo nihče ocenil z oceno 3 in 4.

Graf 5 – Ocena, ali je besedilo enostavnejše od izvirnika



Pri Wilcoxonovem testu predznačenih rangov za prvo ročno in prvo avtomatsko generirano besedilo dobimo z vrednost -4.88 za $p < 0.001$, s čimer lahko ovržemo ničelno hipotezo in zaključimo, da je prvo avtomatsko generirano besedilo manj razumljivo od prvega ročno generiranega besedila.

Za drugo ročno in drugo avtomatsko generirano besedilo dobimo pri Wilcoxonovem testu vrednost $z = -4.28$ za $p < 0.001$, s čimer lahko ovržemo ničelno hipotezo in zaključimo, da je drugo avtomatsko generirano besedilo manj razumljivo od drugega ročno

Sta avtomatsko in ročno generirani besedili res enostavnejši od izvirnika?			
	Mediana	Povprečje	Standardni odklon
Avtomatsko poenostavljeno besedilo 1	4	3.6	0.80
Ročno poenostavljeno besedilo 1	1	1.6	0.79
Avtomatsko poenostavljeno besedilo 2	4	3.8	0.41
Ročno poenostavljeno besedilo 2	1	1.7	0.85

Tabela 9 – Ocena, ali je poenostavljeno besedilo res enostavnejše od izvirnika.

generiranega besedila.

Pri računanju razlike predznačnih rangov med razumljivostjo prvega in drugega avtomatsko generiranega besedila je povprečna vrednost pozitivnih rangov 9.17 in je večja od negativnih rangov, ki je 7.64, kar nakazuje na to, da so vrednosti za razumljivost prvega avtomatsko generiranega besedila v povprečju višja kot za razumljivost drugega avtomatsko generiranega besedila. Wilcoxonov test predznačnih rangov ima z vrednost -0.79 pri $p = 0.43$. Razlika med povprečnima vrednostma ni signifikantna in zato ne moremo zavrniti ničle hipoteze. Rezultati kažejo, da sta oba avtomatsko generirana odstavka enako razumljiva in ne kaže, da bi bilo katero od besedil signifikantno bolj razumljivo.

Korelacija med povzemalno ustreznostjo in razumljivostjo je 0.09 za avtomatsko generirano prvo besedilo, 0.38 za ročno generirano prvo besedilo, 0.57 za avtomatsko generirano drugo besedilo in 0.70 za ročno generirano drugo besedilo. Zaključimo lahko, da je za krajša besedila povezanost med razumljivostjo in povzemalno ustreznostjo nižja kot za daljša besedila. Zdi se, da za kratka avtomatsko generirana besedila ni povezanosti med povzemalno ustreznostjo in razumljivostjo besedil ter da je ta nizka za krajša ročno generirana besedila. Pri daljših avtomatsko povezanih besedilih je povezanost med razumljivostjo in povzemalno ustreznostjo zmerna, pri ročnih pa močna.

Za izračun multiple korelacije označimo naše spremenljivke z matematičnimi simboli na naslednji način:

X: ocena razumljivosti poenostavljenega besedila (dobljena z vprašanjem Kako razumljivo je poenostavljeno besedilo?),

Y: ocena enostavnosti besed v besedilu (dobljena z vprašanjem Ali je izbira besed v besedilu primerna (vsebuje enostavne in pogoste besede)?),

Z: ocena logične povezanosti besedila (dobljena z vprašanjem Ali je besedilo logično povezano?).

Multipla korelacija za prvo avtomatsko generirano besedilo je 0.47, za drugo pa 0.15.

Opazimo, da je v obeh primerih multipla korelacija višja od korelacij med posameznimi spremenljivkami posebej (glej tabeli 10 in 11) in da je korelacija med spremenljivkama Y in Z nizka, kar potrjuje, da sta med sabo neodvisni. Pri prvem besedilu imata enostavnost besed in logična povezanost večji vpliv na razumljivost besedila kot pri drugem besedilu. Pri prvem besedilu enostavnost besed in logična povezanost pojasnjujeta skoraj 50 % delež variance pri razumljivosti besedila, pri drugem pa le 15 %. Eden izmed razlogov za veliko razliko multiple korelacije med besediloma bi lahko bila jezikovna pra-

spremenljivka	X	Y	Z
X	1		
Y	0.37	1	
Z	0.34	0.14	1

Tabela 10 – Korelacijska matrika za spremenljivke razumljivost (X), enostavnost besed (Y) in povzemalna ustreznost (Z) za avtomatsko poenostavljeno besedilo 1.

spremenljivka	X	Y	Z
X	1		
Y	0.14	1	
Z	-0.03	0.06	1

Tabela 11 – Korelacijska matrika za spremenljivke razumljivost (X), enostavnost besed (Y) in povzemalna ustreznost (Z) za avtomatsko poenostavljeno besedilo 2.

vilnost prvega besedila v primerjavi s prisotnostjo slovničnih napak v drugem besedilu. Če je temu tako in če so pri drugem besedilu ocenjevalci res ocenjevali izbiro besed tudi glede na slovnične napake, bi pomenilo, da ima tudi kriterij leksikalne preprostosti velik vpliv na razumljivost besedil. Za potrditev te domneve bi morali narediti nadaljnjo raziskavo, kjer bi bolj nedvoumno in ločeno spraševali po oceni enostavnosti besed in po oceni slovnične pravilnosti zapisnih besed.

Za zaključek odgovorimo na zastavljena vprašanja in hipotezo. V povprečju so ocenjevalci odstavka zgenerirana z našim modelom ocenili kot logično nepovezana in nepovezana v smiselno celoto. Odstavka se prav tako nista izkazala kot poenostavitev v primerjavi z izvirnikom. Model zaenkrat še ne tvori smiselnih poenostavljenih besedil.

Z analizo objektivnih lastnosti avtomatsko poenostavljenih odstavkov smo ugotovili, da ti upoštevajo večino priporočil za pisanje v lahkem branju, da pa vsebujejo slovnične napake. Zdi se nam, da bi s povečanjem učne množice te lahko odpravili. Model zaenkrat na nivoju stavkov oz. povedi tvori precej dobre in precej razumljive poenostavitve, na nivoju besedil pa ne tvori razumljivih besedil. Glavna razloga za nerazumljivost besedil sta nelogična povezanost besedila in premalo informacij. Ostali razlogi, ki so jih ocenjevalci navedli, so še napačna skladnja, slovnične napake, neohranjanje bistva izvirnika, neprimerna izbira besed in napačni podatki. Sistem bi se dalo izboljšati z uporabo izboljšanega osnovnega naučenega T5 modela [4], z obsežnejšo podatkovno množico, z uporabo sodobnejših besedil ter s kombiniranjem poenostavljanja s povzemanjem.

Ugotovili smo, da je povzemalna ustreznost pomemben kriterij razumljivosti za poenostavljena besedila, glede na to, da je korelacija med razumljivostjo in povzemalno ustreznostjo pri daljših avtomatsko zgeneriranih besedilih srednje visoka in pri daljših ročno zgeneriranih besedilih visoka. Glede na multiplo korelacijo med razumljivostjo, logično povezanostjo in enostavnostjo besed smo ugotovili, da sta tudi kriterija koherenca in leksikalna preprostost oz. enostavnost besedišča pomembna kriterija razumljivosti. Pri tem ima koherenca velik vpliv pri krajšem in daljšem besedilu, kar lahko sklepamo iz korelacije med razumljivostjo in koherenco ter iz odgovorov ocenjevalcev, ki so kot eden izmed glavnih razlogov za slabšo razumljivost obeh besedil podali odgovor nelogična povezanost besedila. Enostavnost besed ima velik vpliv le pri krajših besedilih, zato je tudi njun skupni vpliv večji pri krajših besedilih, pri daljših pa ugotovimo, da so pomembni še drugi kriteriji. Kot pomemben se izkaže kriterij jezikovna pravilnost, saj je kar 55 % ocenjevalcev ocenilo, da je to razlog za slabšo razumljivost besedila s slovničnimi napakami.

[4] V času po izdelavi vprašalnika se je osnovni naučen T5 model nadgradil in je dosegljiv na `https://huggingface.co/cjvt/t5-sl-small`

Pomemben kriterij je tudi jedrnatost, saj so ocenjevalci ocenili, da je med glavnimi razlogi za slabšo razumljivost premalo informacij. Obe besedili sta ustrezali kriteriju skladenjske preprostosti, kar smo ugotovili sami z analizo skladenjskih struktur besedil. Presenetljivo so ocenjevalci kot razlog za slabšo razumljivost besedil izbrali napačno skladnjo v 20 % za prvo avtomatsko generirano besedilo in v 55 % za drugo avtomatsko generirano besedilo.

Sklenemo lahko, da so vsi predpostavljeni kriteriji razumljivosti (jedrnatost, jezikovna pravilnost, leksikalna preprostost, skladenjska preprostost, koherenca in povzemalna ustreznost) pomembni kriteriji razumljivosti za avtomatsko generirana besedila. Z izboljšanjem teh lastnosti bi bila besedila bolj razumljiva. Glede na to, da ni nihče dopisal drugega razloga za slabšo razumljivost besedil, lahko sklenemo, da smo zajeli vse bistvene kriterije razumljivosti za (avtomatsko) poenostavljena besedila, razen upoštevanje razumevalca in komunikativnost besedila, ki ju nismo raziskovali.

Zaključimo lahko, da je sistem le delno uporaben kot pomoč poenostavljalcem. Metrike ROUGE in BERTScore dajajo obetavne rezultate, ki jih človeška analiza zamaje. Vseeno lahko rečemo, da je uporaben na nivoju povedi, ker so povedi večinoma enostavčne, vsebujejo enostavne besede in spuščajo posebne znake. Na nivoju krajših odstavkov je sistem slabši, na nivoju daljših odstavkov ali besedil pa ni dovolj robusten, predvsem zaradi nelogične povezanosti med stavki in premalo ohranjenih informacij iz originala.

Prednost naše metode je, da smo opravili avtomatsko in človeško evalvacijo. Evalviranje poenostavljenih besedil (in drugih nalog generiranja besedil) namreč ni lahka naloga. Največja težava je, da nimamo vnaprej določeno, kaj je pravilna poenostavitev besedila, niti ni ta enoznačna, vsak posameznik bi besedilo drugače poenostavil. Avtomatske metrike so le deloma uporabne, ker besedila z različnimi napakami (npr. ponavljajoči se ali absurdni stavki, popačena dejstva, slaba izbira relevantne vsebine ipd.), ocenijo enako, kot če napak ne bi bilo (Lee, Gatt, Miltenburg, Wubben in Krahmer, 2019). Za večjo zanesljivost moramo zato poseči po dragi človeški evalvaciji, ki je daleč najbolj zanesljiva evalvacijska metoda za ocenjevanje poenostavitev (Stodden, 2021).

Toda tudi človeška evalvacija je problematična iz več možnih razlogov, ki so se pokazali tudi v našem delu.

1. Eden izmed razlogov je, da se ocene med ocenjevalci lahko zelo razlikujejo in zaradi neizkušenih ocenjevalcev lahko pridemo do napačnih zaključkov (Freitag idr., 2021; Stodden, 2021). Na primer pri ocenjevanju poenostavitev, v primeru, ko je zgeneriran stavek identičen originalnemu stavku, se je izkazalo, da lahko pride med ocenjevalci do različnih interpretacij lestvice (Stodden, 2021). Nekateri so takšne poenostavitve ocenjevali z najvišjo vrednostjo na lestvici, nekateri pa s srednjo vrednostjo (Stodden, 2021). Pri ocenjevanju strojnih prevodov na lestvicah se je izkazalo, da včasih zaradi slabe človeške evalvacije pride do napačnih trditev, da so strojni prevodi enakovredni človeškim (Freitag idr., 2021). Tudi v naši raziskavi je prišlo do razhajanj med ocenjevalci. Pri ročno generiranih poenostavitvah je npr. en ocenjevalec ocenil, da prvo ročno poenostavljeno besedilo ne ohranja bistva izvirnika, in dva ocenjevalca, da drugo besedilo ne ohranja bistva izvirnika. Prav tako so drugo ročno generirano besedilo trije ocenjevalci ocenili kot nerazumljivo. Za obe ročno generirani besedili so štirje ocenjevalci ocenili, da nista enostavnejši od izvirnika. Tu je najverjetnejša razlaga, da so se ocenjevalci v svojem odgovoru zmotili. Morda smo zaradi neizkušenih ocenjevalcev in njihovih morebitnih previsokih

pričakovanj prišli do napačnih zaključkov.

2. Problem pri ocenjevanju avtomatsko generiranih poenostavitev je, da ne obstaja najboljša praksa, ni soglasja, katera vprašanja vprašati ocenjevalca in koliko stopenjsko Likertovo lestvico uporabiti (Stodden, 2021). Za slovenski jezik se nismo mogli nasloniti na nobeno raziskavo o avtomatsko generiranih poenostavitvah, saj te ne obstajajo. Uporabljali smo štiristopenjsko Likertovo lestvico, kjer so nižje ocene pomenile boljše lastnosti besedila, kar je mogoče manj običajno, kot če bi višje ocene pomenile boljše lastnosti. Lestvica tako ni bila nujno najboljša. Nadalje lahko slabo zastavljena vprašanja vodijo k napačnim zaključkom (Stodden, 2021). V našem vprašalniku se je izkazalo, da vprašanje 'Ali je izbira besed v besedilu primerna (vsebuje enostavne in pogoste besede)?' ni bilo najbolje zastavljeno, ker so ocenjevalci pri drugem besedilu zelo različno ocenjevali primernost izbere besed. Bolje bi bilo, če bi zastavili dve ločeni vprašanji 'Ali so v besedilu izbirane enostavne besede?' in 'Ali so besede v besedilu zapisane slovnično pravilno?'. Tudi vprašanje 'Če ste na zgornje vprašanje odgovorili s skoraj nerazumljivo ali nerazumljivo, iz naštetih možnosti izberite glavne razloge za slabšo razumljivost. Lahko tudi dopišete nov razlog.' bi lahko bilo bolje zastavljeno, če bi predvidevalo možnost, da je lahko tudi ročno poenostavljeno besedilo slabo razumljivo.

3. Ena slabost evalvacije z vprašalnikom je, da se lahko rezultati na Likertovi lestvici različno interpretirajo od raziskovalca do raziskovalca (Stodden, 2021). Tu med raziskovalci predvsem prihaja do razhajanj, ali smatrajo vrednosti na Likertovi lestvici kot ordinalne ali zvezne in posledično, katere statistične teste uporabijo. V naši raziskavi sicer domnevamo, da glede na dobljene odgovore iz vprašalnika, izbira drugih testov ne bi bistveno spremenila zaključkov.

Možne izboljšave naše raziskave so, da bi namesto naključnih ocenjevalcev izbrali strokovnjake, ki se ukvarjajo z lahkim branjem ali ki učijo ciljno publiko, lahko pa bi preverjali razumljivost avtomatsko generiranih besedil kar pri naslovljencih, ki jim je poenostavljeno besedilo prvotno namenjeno. Naslednja možna izboljšava je izboljšava samega vprašalnika ter dodajanje nalog razumevanja. Lahko bi določanje enostavnosti oz. zahtevnost besedil dopolnili še z avtomatskim ocenjevanjem berljivosti besedil s formulami berljivosti, npr. z uporabo aplikacije https://orodja.cjvt.si/berljivost/. Sistem bi se dalo izboljšati tudi s tehničnega vidika, z uporabo izboljšanega osnovnega naučenega T5 modela, ki se je v času po izdelavi vprašalnika nadgradil, z večjim modelom SloT5 (t5-sl-large). Poskusili bi lahko tudi modele mT5 (mT5-large ali mT5-small) ter naš sistem v kombinaciji s povzemanjem. Sistem bi se dalo izboljšati tudi z obsežnejšo podatkovno množico. Zanimivo bi bilo videti, če bi uporaba sodobnejših besedil v podatkovni množici pokazala drugačne rezultate, če bi bila učna množica namesto iz leposlovnih del sestavljena iz npr. novic, če bi izbira drugih odstavkov iz testne množice spremenila rezultate, če bi se rezultati spremenili, če bi namesto dveh odstavkov dali v ocenjevanje več odstavkov ali če bi namesto odstavkov avtomatsko poenostavljali celotno besedilo.

5 Zaključek

V delu smo ustvarili podatkovno množico za poenostavljanje besedil v slovenskem jeziku, ki smo jo uporabili za učenje nevronske mreže arhitekture zaporedje v zaporedje za problem poenostavljanja besedil. To je prva podatkovna množica namenjena poenostavljanju besedil, ki se jo bo lahko uporabilo in nadgradilo v nadaljnjih študijah. Dosegli smo dobre rezultate glede na metriki ROUGE in BERTScore. Za boljšo interpretacijo rezultatov smo opravili tudi človeško evalvacijo.

Na dveh primerih avtomatsko zgeneriranih odstavkov smo raziskovali upoštevanje priporočil za pisanje v lahkem branju, kako logično povezana so besedila, ali ohranjajo bistveno vsebino izvirnika in kako razumljiva so ter identificirali glavne razloge za nerazumljivost besedil. Ugotovili smo, da avtomatski sistem generira enostavčne ali enostavne večstavčne povedi s preprostimi priredji in podredji in ne uporablja trpnika ali posebnih simbolov. Glede na solidne ROUGE in BERTScore mere in glede na uspešnost generiranja preprostih stavčnih struktur ter zmožnosti generiranja večinoma preprostih besed, naš model deluje obetaven. Sistem je uporaben na nivoju povedi, na nivoju krajših odstavkov je sistem slabši, na nivoju daljših odstavkov ali besedil pa s človeško presojo na podlagi vprašalnika ugotavljamo, da so avtomatsko poenostavljeni odstavki slabše razumljivi od ročno generiranih in da model ni preveč uspešen pri tvorjenju razumljivih odstavkov. Odstavki, ki jih generira naš sistem, niso dovolj logično povezani, ne ohranjajo bistvene vsebine izvirnika in niso dovolj razumljivi. Glavna razloga za slabšo razumljivost sta nelogična povezanost besedila in premalo informacij. Razlogi, ki so jih ocenjevalci še prepoznali, so napačna skladnja, slovnične napake, neohranjanje bistva izvirnika, neprimerna izbira besed in napačni podatki. Ugotovili smo še, da avtomatsko generirana besedila niso enostavnejša od izvirnika. Sistem je delno uporaben kot pomoč poenostavljalcem, potencialno pa bi se ga dalo izkoristiti v kombinaciji s povzemanjem za zagotavljanje preprostejšega besedišča in preproste skladenjske strukture.

Raziskovali smo še kriterije razumljivosti za avtomatsko generirana besedila in ugotovili, da so pomembni kriteriji razumljivosti jedrnatost, jezikovna pravilnost, leksikalna preprostost, skladenjska preprostost, koherenca in povzemalna ustreznost. Naš sistem se je najboljše odrezal v kriterijih skladenjske preprostosti in leksikalne preprostosti, najslabše pa v povzemalni ustreznosti, koherenci in jedrnatosti. Določitev kriterijev razumljivosti za avtomatsko generirana besedila je pomemben doprinos k nadaljnjemu razvoju in evalvaciji modelov avtomatskega poenostavljanja besedil, saj omogočajo objektivno oceno razumljivosti takih besedil. S sestavljenim vprašalnikom smo postavili temelj človeške evalvacije razumljivosti in smiselnosti avtomatsko generiranih poenostavljenih besedil v slovenščini.

6 Literatura

Ambrožič, U. (2019). *Kriteriji razumljivosti v izbranem korpusu pravnih besedil. Primer zakonov (Pleonastično zanikanje v pravnem jeziku)* [Diplomsko delo, Univerza v Ljubljani, Filozofska fakulteta]. Repozitorij Univerze v Ljubljani.

Aprosio, A. P., Tonelli, S., Turchi, M., Negri, M. in Di Gangi, M. A. (2019, June). Neural text simplification in low-resource conditions using weak supervision. *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation* (str. 37–44).

Bojanowski, P., Grave, E., Joulin, A. in Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics, 5* (str. 135–146).

Brinovec, T. in Krečan, J. (2010). *Razumevanje navodil za uporabo zdravil* [Diplomsko delo, Univerza v Ljubljani, Filozofska fakulteta]. Repozitorij Univerze v Ljubljani.

Denkowski, M. in Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the ninth workshop on statistical machine translation* (str. 376–380).

Devlin, J., Chang, M. W., Lee, K. in Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (str. 4171–4186). Association for Computational Linguistics.

Dostopna Ljubljana, https://www.ljubljana.si/sl/aktualno/izsla-je-knjizica-dostopna-ljubljana/, pridobljeno januarja 2020.

Ferbežar, I. (2009). *Besedilnost med formalnimi lastnostmi in kognitivnimi procesi* [Doktorska disertacija, Univerza v Ljubljani, Filozofska fakulteta]. Repozitorij Univerze v Ljubljani.

Ferbežar, I. (2012). *Razumevanje in razumljivost besedil.* Znanstvena založba Filozofske fakultete.

Ferš, D. (2019). *Priprava turističnega vodnika za mesto Maribor v lahko berljivi obliki v sodelovanju z osebami in za osebe z motnjami v duševnem razvoju* [Magistrsko delo, Univerza v Ljubljani, Pedagoška fakulteta]. Repozitorij Univerze v Ljubljani.

Freitag, M. idr. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9*, 1460–1474.

Grum, S. (1976). *Deček in Blaznik.* Zbrano delo. Državna založba Slovenije.

Haramija, D. in Knapp T. (2019). *Lahko je brati: Lahko branje za strokovnjake.* Zavod RISA. Podgorje pri Slovenj Gradcu.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A. in Fidler S. (2015). Skip-thought vectors. *Advances in neural information processing systems, 28.* Curran Associates, Inc.

Kosmač, C. (1980). *Tantadruj.* Mladinska knjiga.

Kryściński, W., Paulus, R., Xiong, C. in Socher, R. (2018). Improving abstraction in text summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1808–1817. Association for Computational Linguistics.

Le, Q. In Mikolov, T. (2014, June). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196.

Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S. in Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368.

Leskovec, M. (2009). *Med poljudnostjo in strokovnostjo: razumljivost sodobnih slovenskih poročevalnih besedil* [Diplomsko delo, Univerza v Ljubljani, Fakulteta za družbene vede]. Repozitorij Univerze v Ljubljani.

Lin, C. Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. *Proceedings of text summarization branches out*, 74–81.

Ljubešić, N. in Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0* . Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1204

Magajna, B. (1940). *Breda.* Zaznamovani: knjiga novel, Literarni klub.

Magajna, B. (1940). *Jerinova Lida.* Zaznamovani: knjiga novel, Literarni klub.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. in Joulin, A. (2017). Advances in pre-training distributed word representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nisioi, S., Štajner, S., Ponzetto, S. P. in Dinu, L. P. (2017). Exploring neural text simplification models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 85–91.

Osolnik Kunc, V. (2007). *Strokovno sporočanje z vidika razumljivosti.* Razvoj slovenskega strokovnega jezika. Zbornik Obdobja, 24, 143–152.

Papineni, K., Roukos, S., Ward, T. in Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pennington, J., Socher, R. in Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 conference on Empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters M. idr. 2018. Deep contextualized word representations. *Proceedings of NAACL-HLT 2018.*

Raaijmakers, S. (2019, version 4). *Deep Learning for Natural Language Processing* str. 38–101. MEAP.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. in Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR.*

Risa, 2020a, http://www.risa.si/ Zavod Risa, pridobljeno januarja 2020.

Risa, 2020b, http://www.risa.si/Portals/4/Portali/risa.si/knjiznica/Knjige/romani/ Zavod Risa, pridobljeno januarja 2020.

Saleški Finžgar, F. (1978). *Pod svobodnim soncem.* Mladinska knjiga.

Shakespeare, W. (1974). *Romeo in Julija.* Mladinska knjiga.

Stodden, R. (2021). When the Scale is Unclear-Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification.*Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, 84–95.

Surya, S., Mishra, A., Laha, A., Jain, P. in Sankaranarayanan, K. (2018). Unsupervised neural text simplification. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2058–2068.

Štajner, S. in Saggion, H. (2018). Data-Driven Text Simplification. *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, 19–23.

Tavčar, I. (1987). *Visoška kronika.* Mladinska knjiga.

Tissier, J., Gravier, C. in Habrard, A. (2017). Dict2vec: Learning word embeddings using lexical dictionaries. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Tomažič, N. (2020). *Razumljivost in razumevanje besedil ter svetopisemska govorica; Analiza na podlagi pastirskih pisem slovenskih škofov za postni čas med leti 2010 in 2020* [Diplomsko delo, Univerza v Ljubljani, Filozofska fakulteta in Teološka fakulteta]. Repozitorij Univerze v Ljubljani.

Ulčar, M. in Robnik-Šikonja, M. (2019). High quality ELMo embeddings for seven less-resourced languages. *CoRR.*

Ulčar, M. in Robnik-Šikonja, M. (2021). *SloBERTa: Slovene monolingual large pretrained masked language model.* Slovenian language resource repository CLARIN.SI.

Ulčar, M. in Robnik-Šikonja, M. (2022). *Sequence to sequence pretraining for less-resourced language.* arXiv.

Vaswani, A. idr. (2017). Attention is all you need. Advances in neural information processing systems, *CoRR*.

Verdonik, D. (2011). Teorija mentalnih modelov v raziskovanju pogovora: pot k sociokognitivnemu modelu pogovora?. *Jezik in slovstvo, 3(56)*, 93–110.

Vizjak Pavšič, M. (2006). Razumevanje in reprezentacija znanja. *Philological Studies, 4(1)*, 1–9.

Voranc, P. (1981). *Černjakova terba.* Jamnica: roman soseke. Založba Obzorja.

Voranc, P. (1981). *Stari Grad.* Izbrana dela, Nicina – novele in črtice. Založba Obzorja.

Vrbančič, G. (2021). *Metoda prilagodljivega uglaševanja slojev konvolucijskih nevronskih mrež pri strojnem učenju s prenosom znanja* [Doktorska disertacija, Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko]. Digitalna knjižnica Univerze v Mariboru.

Vu, T., Hu, B., Munkhdalai, T. in Yu, H. (2018). Sentence simplification with memory-augmented neural networks. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 97–85.

Xu, W., Napoles, C., Pavlick, E., Chen, Q. in Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics, 4*, 401–415.

Zemljak, D. (2019). *Korelacijska analiza* [Magistrsko delo, Univerza v Mariboru, Fakulteta za naravoslovje in matematiko]. Digitalna knjižnica Univerze v Mariboru.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. in Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *CoRR*.

Zhang, X. in Lapata, M. (2017). Sentence simplification with deep reinforcement learning. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 584–594.

Zhao, J., Zhou, Y., Li, Z., Wang, W. in Chang, K. W. (2018). Learning gender-neutral word embeddings. *CoRR*.

Zidarn, R. (2020). *Avtomatsko povzemanje slovenskih besedil z globokimi nevronskimi mrežami* [Magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko]. Repozitorij Univerze v Ljubljani.

A Dodatek

V dodatku je prikazan vprašalnik, ki smo ga uporabili v magistrski nalogi.

Vprašalnik

Pozdravljeni! V svoji magistrski nalogi bi rada ocenila uspešnost delovanja sistema za avtomatsko poenostavljanje besedil. Cilj poenostavljanja je iz kompleksnih ustvariti enostavnejša besedila, ki so lažje razumljiva, pri tem pa ohranjajo najpomembnejšo originalno vsebino in pomen. Prosila bi vas, da preberete odlomek besedila, zgeneriranega z našim sistemom in na lestvici od 1 do 4 izberete odgovor na vprašanji. Nadalje bosta poleg avtomatsko zgeneriranega besedila prikazana še izvirnik in njegova ročno poenostavitev. Ponovno vas prosim, da na lestvici od 1 do 4 izberete odgovor na vprašanja v zvezi s poenostavljenima besediloma. Ocene se nanašajo na več lastnosti razumljivosti besedila.

Vprašalnik je sicer anonimen, prosim pa vas, da vseeno odgovorite na nekaj osnovnih demografskih podatkov. Za reševanje vprašalnika boste potrebovali 15-20 minut. Prosim vas, da odgovarjate resno.

Spol:
- ○ Moški
- ○ Ženski
- ○ Ne želim odgovoriti

V katero starostno skupino spadate?
- ○ do 20 let
- ○ 21 - 40 let
- ○ 41 - 60 let
- ○ 61 let ali več
- ○ Ne želim odgovoriti

Kakšna je vaša najvišja dosežena formalna izobrazba?
- ○ Manj kot srednja šola
- ○ Srednja šola
- ○ Višja ali visoka šola
- ○ Univerzitetna izobrazba
- ○ Ne želim odgovoriti

Besedilo 1

Neki večer je šel k postelji. Zvečer so prišli policisti. Po pogrebu so Bredo oddali v bolnišnico.

Ali je izbira besed v besedilu primerna (vsebuje enostavne in pogoste besede)?

	Primerna	Večinoma primerna	Večinoma neprimerna	Neprimerna
Besedilo 1	○	○	○	○

Ali je besedilo logično povezano?

	V celoti	Večinoma	Večinoma ne	Ne
Besedilo 2	○	○	○	○

Izvirnik 1

Neki večer je prišel oče pozno domov. Stopil je k postelji, kjer je mama ležala. Nato je Breda videla, kako so se mu razprostrli prsti in se bližali materinemu vratu. Slišala je, kako je mati zaječala in kako so šinile njene roke nekam v zrak. Ubil jo je. Zjutraj so prišli policisti, ki so uklenili očeta. Takoj po pogrebu so Bredo oddali v bolnišnico za otroke, ki so izgubili starše.

Poenostavljeno besedilo 1

Neki večer je šel k postelji. Zvečer so prišli policisti. Po pogrebu so Bredo oddali v bolnišnico.

Poenostavljeno besedilo 2

Neki večer, ko je mama spala, je oče prišel domov in mater ubil. Očeta so odpeljali policisti.Mater so pokopali. Bredo pa so poslali v bolnišnico za otroke, ki so izgubili starše.

Ali poenostavljeno besedilo ohranja bistveno vsebino izvirnika?

	V celoti	Večinoma	Večinoma ne	Ne
Poenostavljeno besedilo 1	○	○	○	○
Poenostavljeno besedilo 2	○	○	○	○

Kako razumljivo je poenostavljeno besedilo?

	Razumljivo	Delno razumljivo	Skoraj nerazumljivo	Nerazumljivo
Poenostavljeno besedilo 1	○	○	○	○
Poenostavljeno besedilo 2	○	○	○	○

Če ste na zgornje vprašanje odgovorili s skoraj nerazumljivo ali nerazumljivo, iz naštetih opcij izberite glavne razloge za slabšo razumljivost. Lahko tudi dopišete nov razlog.
Možnih je več odgovorov
☐ Napačni podatki
☐ Slovnične napake
☐ Neprimerna izbira besed
☐ Napačna skladnja
☐ Premalo informacij
☐ Preveč informacij
☐ Nelogična povezanost besedila
☐ Neohranjanje bistva izvirnika
☐ Drugo:

Se vam zdi 'poenostavljeno' besedilo res enostavnejše od izvirnika?

	Da	Večinoma	Večinoma ne	Ne
Poenostavljeno besedilo 1	○	○	○	○

Poenostavljeno besedilo 2	○	○	○	○

Besedilo 2

Vester je videl, da gre nekdo v stolp. Huberta sta spoznala trgovskega pomočnika. Vester je videl, da je mlada kmetica zelo utrujena. Hubert je bil v stolp. Kmetica je padla na tla. Hubert je rekel: Boš vstala, kurba partizanska! Ko je bila pijanka, je bila zelo žalostna. Hubert je bila omotena. Micka je glasno rekla: Hubert, prizanesi mi zaradi otroka. Pod posteljo si skrivala bandita. Micka je jokala Micka. Spet se je zapel. Ti si zverina. Ubi me, zverina, ki si si. Hubert je vzel iz žepa dolgo vrv. Vester je bil zelo vesel. Hubert je videl kačo.Vester je odšel k Micki. Micka je ležala trda na zemlji.

Ali je izbira besed primerna (vsebuje enostavne in pogoste besede)?

	Primerna	Večinoma primerna	Večinoma neprimerna	Neprimerna
Besedilo 1	○	○	○	○

Ali je besedilo logično povezano?

	V celoti	Večinoma	Večinoma ne	Ne
Besedilo 2	○	○	○	○

Izvirnik 2

Nekega dne je Vester spet videl, da gre nekdo v stolp. Ko sta se dva človeka začela povzpenjati navkreber, je v moškem spoznal Huberta, trgovskega pomočnika. Vester je videl, da je mlada kmetica iz okolice, kjer je že tudi on dobil kaj dobrega od nje, vsa utrujena in da težko hodi navkreber. Komaj sta bila v stolpu, je Hubert priskočil in kmetico udaril v obraz. Nato jo je še sunil v trebuh s svojim težkim škornjem. Kmetica je zastokala in padla na tla. Hubert pa je kakor ris skočil nanjo in jo začel obdelovati s pasjim bičem, ki ga je takrat potegnil izza pasa, in vpil: "Boš vstala, kurba partizanska!" Ker je kmetica neznansko vekala, je pretepač režal nanjo: "Le vpij, tukaj lahko vpiješ, ker te nihče ne sliši." Kmetica ni več poskušala vstati, bila je omotena, ker ji je Hubert z obema nogama skočil na vrat in jo teptal, kar se je dalo. Micka je glasno zastokala: "Hubert, prizanesi mi zaradi otroka, saj veš, da nisem sama." Zaradi njega, da bi ti prizanašal, vlačuga partizanska. "Pod posteljo si skrivala bandita." "Koga si skrivala pod posteljo?"Nič ne vem! je jokala Micka. Spet se je zakadil vanjo in jo stoječo začel obdelovati z bičem in škornji. Čuješ, nisem te prej klicala po imenu zaradi sebe, ampak zaradi otroka! "Ti si zverina." Ubij me, zverina, ki si. Hubert je tedaj vzel iz žepa dolgo, tanko vrv in Micki zataknil zanko na vrat, drugi konec vrvi je vrgel čez klin nad njo in začel na vso moč vleči. Brez pomisleka je Vester zagrabil gada za vrat in ga zagnal v globino proti Hubertu. Ko je Hubert videl kačo, je zakričal kakor obseden, spustil vrv in blazno odskočil. Zdaj je udrl naravnost skozi vhod in planil po bregu navzdol. Vester je takoj splezal na tla in odšel v žrelo k Micki. Micka je ležala trda na zemlji. Naj jo je Vester tresel, kolikor je hotel, ni mogel dobiti od nje nobenega življenja.

Poenostavljeno besedilo 1

Vester je videl, da gre nekdo v stolp. Huberta sta spoznala trgovskega pomočnika. Vester je videl,

da je mlada kmetica zelo utrujena. Hubert je bil v stolp. Kmetica je padla na tla. Hubert je rekel: Boš vstala, kurba partizanska! Ko je bila pijanka, je bila zelo žalostna. Hubert je bila omotena. Micka je glasno rekla: Hubert, prizanesi mi zaradi otroka. Pod posteljo si skrivala bandita. Micka je jokala Micka. Spet se je zapel. Ti si zverina. Ubi me, zverina, ki si si. Hubert je vzel iz žepa dolgo vrv. Vester je bil zelo vesel. Hubert je videl kačo.Vester je odšel k Micki. Micka je ležala trda na zemlji.

Poenostavljeno besedilo 2

Nekega dne je Vester spet videl, da gre nekdo v stolp. Bil je trgovski pomočnik Hubert. Micka je bila mlada kmetica. Vester je videl, da je utrujena. Ko jo je Hubert pripeljal v stolp, jo je udaril v obraz in brcnil v trebuh. Kmetica je padla na tla. Hubert jo je tepel s pasjim bičem in kričal: Boš vstala, kurba partizanska. Le vpij, tukaj te nihče ne sliši. Hubert ji je z obema nogama skočil na vrat in jo teptal. Micka ga je prosila: Hubert, nehaj, saj veš, imam otroka. Hubert jo je spraševal: Koga si skrivala pod posteljo? Micka je jokala: Nič ne vem. Hubert jo je tepel in brcal. Rekla mu je še: Ti si zverina. Ubij me. Hubert je hotel Micko obesiti. Vester je proti Hubertu vrgel strupeno kačo. Hubert je zakričal in zbežal, ker se je ustrašil kače. Vester pa je hitel k Micki. Micka je bila mrtva.

Ali poenostavljeno besedilo ohranja bistveno vsebino izvirnika?

	V celoti	Večinoma	Večinoma ne	Ne
Poenostavljeno besedilo 1	○	○	○	○
Poenostavljeno besedilo 2	○	○	○	○

Kako razumljivo je poenostavljeno besedilo?

	Razumljivo	Delno razumljivo	Skoraj nerazumljivo	Nerazumljivo
Poenostavljeno besedilo 1	○	○	○	○
Poenostavljeno besedilo 2	○	○	○	○

Če ste na zgornje vprašanje odgovorili s 3 ali 4, iz naštetih opcij izberite glavne razloge za slabšo razumljivost. Lahko tudi dopišete nov razlog.

Možnih je več odgovorov

☐ Napačni podatki
☐ Slovnične napake
☐ Neprimerna izbira besed
☐ Napačna skladnja
☐ Premalo informacij
☐ Preveč informacij
☐ Nelogična povezanost besedila
☐ Neohranjanje bistva izvirnika
☐ Drugo:

Se vam zdi 'poenostavljeno' besedilo res enostavnejše od izvirnika?

Da	Večinoma	Večinoma ne	Ne

Poenostavljeno besedilo 1	○	○	○	○
Poenostavljeno besedilo 2	○	○	○	○

Odgovorili ste na vsa vprašanja v tej anketi. Hvala za sodelovanje.