UPORABA ŠESTIH MER SKLADENJSKE KOMPLEKSNOSTI ZA PRIMERJAVO JEZIKA V GOVORNEM IN PISNEM KORPUSU

Luka TERČON, [1,2]

[1]Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva cesta 2, Ljubljana, Slovenija

2Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, Ljubljana, Slovenija

Obstajajo številne metode za merjenje skladenjske kompleksnosti v digitaliziranih bazah jezika. Jezikovni korpusi, posebej takšni, ki vsebujejo skladenjske oznake, nam omogočajo, da analize in primerjave skladenjske kompleksnosti izvedemo avtomatsko in učinkovito. V tem prispevku predstavljam metodo za avtomatsko primerjavo dveh korpusov – korpusa pisnih besedil in korpusa govorjenih besedil – s pomočjo šestih uveljavljenih mer skladenjske kompleksnosti. Rezultati kažejo, da je skladenjska sestava jezika v pisnem korpusu nekoliko kompleksnejša kot v govornem korpusu. Razlike so najbolj izrazite predvsem pri dolžini povedi in globini skladenjskih dreves. Analiza korelacije med različnimi merami nakazuje na to, da nekatere od uporabljenih mer podajo precej drugačno informacijo o skladenjski sestavi neke povedi kot druge.

Ključne besede: skladenjska kompleksnost, pisni korpus, govorni korpus, mere kompleksnosti, študentski prispevek

1 UVOD

S čedalje večjo digitalizacijo jezikovnih virov se pojavlja vedno več možnosti za avtomatsko analizo besedil. Pojav skladenjsko označenih korpusov omogoča hitro analizo skladenjskih vzorcev, ki se pojavljajo v slovenščini, tako v govorjenih kot tudi pisnih besedilih. Prav jezikovna zvrst v smislu prenosnega medija (i. e. ali imamo opravka z govorjenim ali pisnim jezikom) je eden od glavnih dejavnikov, ki v veliki meri vpliva na skladenjsko kompleksnost jezika (Ehret in sod., 2023). Lintunen in Mäkilä (2014) v svoji študiji pokažeta, da obstaja neka razlika v skladenjski kompleksnosti med govorjenim in pisnim jezikom, vendar ta ni nujno višja v pisnih besedilih, kot je pogosto predpostavljeno, pač pa lahko na končni rezultat vplivajo številni dejavniki, kot so vrsta izbrane mere in način

segmentiranja besedil na povedi oz. druge osnovne enote. Razlika v skladenjski kompleksnosti med govorjenim in pisnim jezikom torej ni tako jasno opredeljena.

Slovenska korpusa SSJ-UD (Dobrovoljc in sod., 2017) in SST-UD (Dobrovoljc in Nivre, 2016) vsebujeta jezikoslovne oznake, med katerimi so tudi skladenjske odvisnostne relacije po sistemu Universal Dependencies,[1] zato sta dobra vira za raziskovanje skladenjskih značilnosti besedil v slovenščini. SST-UD vsebuje transkripte govora, SSJ-UD pa pisna besedila, zato lahko SST-UD obravnavamo kot reprezentativen korpus govorjene slovenščine (Dobrovoljc in Nivre, 2016, str. 1566), SSJ-UD pa kot reprezentativen korpus pisne slovenščine (Dobrovoljc in sod., 2017, str. 34).

V svoji raziskavi preučujem razlike na ravni skladenjske kompleksnosti med tema dvema korpusoma s pomočjo šestih uveljavljenih mer skladenjske kompleksno-sti. Predmet raziskave so tako potencialne razlike v skladenjski kompleksnosti kot tudi razlike med samimi merami, ki jih uporabljam za primerjanje. Številne študije izpostavljajo večplastnost pojma kompleksnost v jezikoslovju in poudarjajo, da tudi znotraj iste domene (skladnja, oblikoslovje, itd.) ene prave stabilne definicije kompleksnosti ni (Berdicevskis in sod., 2018; Bentz in sod., 2023; Ehret in sod., 2023; Jagaiah in sod., 2020). Posledično različne uveljavljene mere, ki se uporabljajo za merjenje skladenjske kompleksnosti, lahko podajo precej različne informacije o obravnavanem jeziku. V luči tega sem oblikoval dve raziskovalni vprašanji, ki naslavljata tako razlike med obema jezikovnima zvrs-tema kot tudi razlike med uporabljenimi merami, ko jih apliciramo na slovenski jezik:

1. Ali se skladenjska kompleksnost v govorjenem in pisnem jeziku na kakšen način razlikuje?

2. Se izbrane mere skladenjske kompleksnosti med seboj kako razlikujejo?

Sistematičnih raziskav skladenjske kompleksnosti v slovenskih besedilih je relativno malo. Gaberščekk (2018) v svoji raziskavi za analizo pisnih besedil, ki so jih ustvarili osnovnošolci, uporabi tako novonastale metode za merjenje skladenjske kompleksnosti, ki jih oblikuje na podlagi štirih identificiranih meril kompleksnosti (merilo dolžine, strukturne kompleksnosti, pogostnosti in ra-

[1] https://universaldependencies.org/

znolikosti), kot tudi že uveljavljene, starejše metode. Martinc in sod. (2021) se v svoji študiji ukvarjajo predvsem z načini merjenja berljivosti v angleških in slovenskih besedilih, vendar med tradicionalnimi pristopi omenjajo tudi mere skladenjske kompleksnosti, kar nakazuje na določeno mero prekrivanja med obema konceptoma. Kolikor se zavedam, je moja študija prva, ki preučuje razlike v skladenjski kompleksnosti med govorjeno in pisno slovenščino z uporabo skladenjsko razčlenjenih korpusov.

V drugem razdelku najprej predstavim oba uporabljena korpusa in shemo skladenjskih oznak, ki jo korpusa uporabljata, ter vsako od šestih izbranih mer skladenjske kompleksnosti. V tem razdelku opišem tudi statistične teste in metode, ki jih uporabljam za preverjanje rezultatov. Tretji razdelek predstavlja glavne rezultate raziskave in ugotovitve statističnih preizkusov. V četrtem razdelku podam svojo interpretacijo rezultatov in odgovorim na zastavljeni raziskovalni vprašanji.

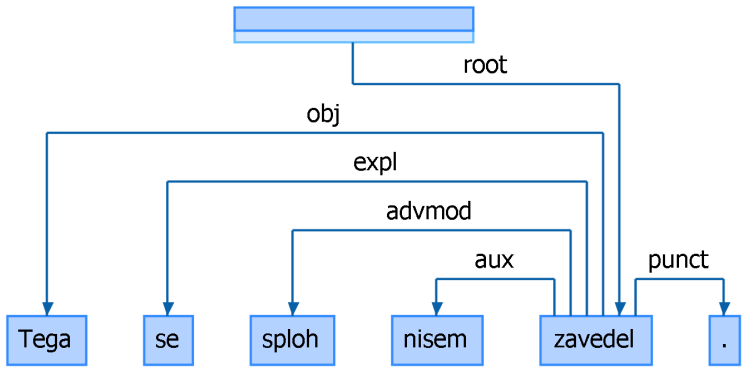
2 RAZISKOVALNI PODATKI IN METODE

2.1 Označevalna shema Universal Dependencies

Korpusa, ki ju v tej študiji uporabljam, vsebujeta oznake, ki sledijo mednarodno uveljavljeni shemi za skladenjsko označevanje Universal Dependencies (de Marneffe in sod., 2021). Označevanje po tej shemi sledi principom odvisnostne skladnje, kar pomeni, da vsaka poved tvori aciklični graf, sestavljen iz usmerjenih odvisnostnih relacij, ki potekajo od nadrejenih besed k podrejenim. Tak skupek razmerij med besedami imenujemo tudi skladenjsko drevo in ga lahko grafično ponazorimo, kot je prikazano na Sliki 1

Vsaka odvisnostna relacija ima po tem sistemu svoj tip (npr. *obj*, *expl*, *advmod*,...), nadrejeni element (oz. glavo) in podrejeni element (oz. cilj). Pri primeru na Sliki 1 na primer med besedama »zavedel« in »Tega« velja relacija *obj* (po slovenskih skladenjskih sistemih to ustreza predmetu, enemu od jedrnih argumentov povedka). Glagol »zavedel« je glava relacije, cilj pa je »Tega«. Vsaka poved mora vsebovati tudi relacijo *root*, katere glava je vedno prazni element, ki ga imenujemo tudi korenski element povedi, cilj pa je jedro povedi – običajno glavni del povedka glavnega stavka.

Slika 1: Primer skladenjskega drevesa po sistemu Universal Dependencies, kot ga zgenerira orodje Q-CAT (Brank, 2023).



2.2 Uporabljena korpusa

Prvi od korpusov, ki sem ju uporabil za primerjavo, je korpus SSJ-UD (Dobrovoljc in sod., 2017; Dobrovoljc in Ljubešić, 2022; Dobrovoljc in sod., 2023), ki vsebuje tako leposlovna kot tudi neleposlovna pisna besedila iz korpusov ssj500k (Krek in sod., 2020) in ELEXIS-WSD (Martelli in sod., 2023) in zato predstavlja uravnotežen vzorec pisne slovenščine. V tej študiji uporabljam različico 2.14,[2] ki šteje približno 260.000 pojavnic oz. približno 13.000 povedi. Del oznak je rezultat avtomatske pretvorbe iz skladenjske označevalne sheme JOS-SYN (Erjavec in sod., 2010), preostanek pa je ročno pregledan.

Drugi korpus je SST-UD (Dobrovoljc in Nivre, 2016), ki vsebuje transkripcije govora, vzete iz referenčnega govornega korpusa Gos (Verdonik in sod., 2023). Korpus vsebuje transkripcije govora, ki so ga sproducirali govorci raznolikih ozadij in je nastal v različnih situacijah, zato ga obravnavam kot reprezentativen vzorec govorjene slovenščine. V svoji raziskavi uporabljam še neobjavljeno različico, ki za razliko od trenutno dostopne različice 2.14[3] vsebuje izboljšane oznake, usklajene z najnovejšimi dodatki v splošne označevalne smernice sistema Universal Dependencies. Ta novejša različica korpusa bo vključena v eno od prihodnjih skupnih izdaj, ki redno izhajajo v okviru projekta Universal Depen-

[2] https://github.com/UniversalDependencies/UD_Slovenian-SSJ
[3] https://github.com/UniversalDependencies/UD_Slovenian-SST

dencies. Po velikosti šteje približno 70.000 pojavnic oz. približno 6.000 povedi. Vse oznake v SST-UD so ročno pregledane.

2.3 Mere skladenjske kompleksnosti

Izbral sem šest mer, ki se v literaturi uporabljajo kot indikator skladenjske kompleksnosti. V grobem jih lahko razdelimo v dve skupini: prva, v kateri so tri mere, ki temeljijo na štetju in razmerjih med osnovnimi skladenjskimi elementi (število besed v povedi, število stavkov v povedi, povprečno število stavkov na T-enoto), in druga, v kateri so tri mere, ki so osnovane na principih odvisnostne skladnje in na sestavi skladenjskih dreves (povprečna dolžina odvisnostnih relacij, normalizirana dolžina odvisnostnih relacij, maksimalna globina skladenjskega drevesa). Za večjo primerljivost sem izbral izključno mere, ki podajo vrednosti za kompleksnost na ravni povedi. Vsaka od mer je podrobneje predstavljena v nadaljevanju.

2.3.1 ŠTEVILO BESED V POVEDI (B/P)

To mero bi lahko imenovali tudi enostavna dolžina povedi. Število besed v povedi (v nadaljevanju *B/P*) je ena najbolj razširjenih mer skladenjske kompleksnosti, ki se uporablja v številnih raziskavah in je pogosto predstavljena kot najosnovnejša (Jagaiah in sod., 2020; Wang, 2022; Lintunen in Mäkilä, 2014; Mylläri, 2020a). V svoji analizi kot »besede« upoštevam pojavnice – osnovne podenote, na katere se deli vsaka poved v obeh obravnavanih korpusih.

2.3.2 ŠTEVILO STAVKOV V POVEDI (S/P)

Število stavkov v povedi (v nadaljevanju *S/P*) je pogosto uporabljena mera, ki je osnovana na razumevanju stavka kot osnovne enote skladenjske strukture neke povedi. Ker v korpusnih raziskavah pogosto naletimo na razne fragmente in druge posebne jezikovne pojave, je lahko dokaj težavno na jasen način zastaviti enoznačno definicijo stavka. Konkretne zamejitve stavkov se lahko zato med seboj precej razlikujejo (Mylläri, 2020b). V tem prispevku se pri določanju mej stavkov opiram na različne tipe skladenjskih relacij, ki obstajajo znotraj sistema Universal Dependencies. Kot število stavkov v povedi tako upoštevam število besed, ki jih uvrščamo med glagole in so cilj ene od relacij *csubj* (osebkovi

odvisniki), *ccomp* (predmetni odvisniki), *xcomp* (odprta stavčna dopolnila), *advcl* (prislovnodoločilni odvisniki), *acl* (prilastkovi odvisniki), *conj* (stavčna priredja), *parataxis* (stavčna soredja), oziroma če so glave vsaj ene relacije *cop* (vezni glagol).

2.3.3 POVPREČNO ŠTEVILO STAVKOV NA T-ENOTO (S/T)

Pri povprečnem številu stavkov na T-enoto (v nadaljevanju *S/T*) se sklicujem na pojem T-enote, ki se pogosto definira kot en neodvisen stavek skupaj z vsemi odvisnimi stavki, ki se vežejo nanj (Jagaiah in sod., 2020).[4] Ta mera v ospredje postavlja podredne strukture, ki razkrivajo drugačno plat skladenjske kompleksnosti kot priredne strukture (Mylläri, 2020b). Vrednost S/T za neko poved sem poračunal po sledeči formuli:

$$S/T = \frac{\text{število stavkov v povedi}}{\text{število T-enot v povedi}} \quad (1)$$

2.3.4 POVPREČNA DOLŽINA ODVISNOSTNIH RELACIJ (PDO)

V številnih študijah se je uveljavila mera povprečne dolžine vseh odvisnostnih relacij v neki povedi (angl. *Mean Dependency Distance*, v nadaljevanju *PDO*). Daljše razdalje med glavo in ciljem neke relacije se povezuje z oteženim procesiranjem jezika, zato se je ta mera pričela uporabljati kot pokazatelj skladenjske kompleksnosti (Futrell in sod., 2015; Lei in Jockers, 2020). Izračunamo jo po spodnji formuli:

$$PDO = \frac{\sum_{i=1}^{n} DO_i}{n} \quad (2)$$

Pri tem je *DO* (dolžina odvisnostne relacije, angl. *Dependency Distance*) za neko besedo absolutna vrednost razlike med indeksom (pozicijo v stavku) te besede in indeksom njenega nadrejenega elementa. V stavku »Oče je šel v trgovino« je torej DO za besedo »Oče« enak 2, ker je ta beseda cilj odvisnostne relacije, ki izvira iz glagola »šel«. DO je enak 2, ker je indeks te besede 1, indeks nadrejene besede (glagol »šel«) pa 3. *n* v zgornji formuli predstavlja skupno število besed

[4] V povedi *Tisti, ki razumejo, so tu, ostali pa so odšli* imamo torej dve T-enoti ((1) *Tisti, ki razumejo, so tu* in (2) *ostali pa so odšli*).

v neki povedi, *i* pa indeks besede. Relaciji *punct* in *root* sta v skladu s pogosto prakso izvzeti iz obravnave (Lei in Jockers, 2020).

2.3.5 NORMALIZIRANA DOLŽINA ODVISNOSTNIH RELACIJ (NDO)

Normalizirana dolžina odvisnostnih relacij (angl. *Normalized Dependency Distance*, v nadaljevanju *NDO*) je podobna mera kot PDO, le da se pri tej v izračunu upošteva še dolžino povedi in pozicijo korena v povedi (Lei in Jockers, 2020). Poračuna se jo po sledeči formuli:

$$NDO = \text{abs}\left(\ln\left(\frac{PDO}{\sqrt{\text{indeksKorena} \times \text{n}}}\right)\right) \tag{3}$$

Pri tem *PDO* predstavlja povprečno dolžino odvisnostnih relacij, *abs* absolutno vrednost, *ln* naravni logaritem, *indeksKorena* je indeks besede, ki je cilj relacije *root*, *n* pa je skupno število besed v povedi.

2.3.6 MAKSIMALNA GLOBINA SKLADENJSKEGA DREVESA (MGD)

Maksimalna globina skladenjskega drevesa (v nadaljevanju *MGD*) je mera, ki nam pove, kaj je največja globina skladenjskega drevesa v neki povedi. Izračunamo jo tako, da začnemo pri neki besedi in se pomaknemo do izvora relacije, ki ima to besedo za cilj. To ponavljamo dokler ne dosežemo korenskega elementa povedi. Število premikov, ki smo jih morali narediti, da smo dosegli korenski element, predstavlja globino skladenjskega drevesa za to besedo. Globine skladenjskih dreves izračunamo za vse besede v povedi in jih primerjamo. MGD je najvišja dobljena vrednost v povedi. V literaturi se ta mera uporablja nekoliko manj pogosteje za merjenje skladenjske kompleksnosti kot ostale predstavljene mere. Xu in Reitter (2016) ugotavljata, da MGD v angleških povedih v veliki meri korelira z dolžino povedi (B/P).

2.4 Potek statistične analize

V sledečem razdelku na podlagi poračunanih vrednosti za vsako od šestih zgoraj opisanih mer najprej poročam osnovne vrednosti opisne statistike (povprečje in standardni odklon) in rezultate ponazorim s pomočjo grafikonov kvartilov.

Po pregledu histogramov frekvenčne porazdelitve in t. i. grafov Q-Q (angl. *Q-Q plot*) sem ugotovil, da podatki niso normalno porazdeljeni, zato za preverjanje statistične pomembnosti v raziskavi uporabljam neparametrični test Mann-Whitney U. Za korigiranje pri številnih primerjavah za statistično značilne razlike uporabim metodo Holm-Bonferroni. Za merjenje velikosti efekta uporabljam biserialni korelacijski koeficient.

Da bi raziskal raven ujemanja različnih mer skladenjske kompleksnosti med seboj, poračunam tudi Pearsonov korelacijski koeficient za vsak par mer kompleksnosti. Rezultate prikažem v obliki korelacijske matrike. Tudi tu preverim statistično pomembnost rezultatov in uporabim metodo Holm-Bonferroni za korigiranje.

Ker je korpus SSJ-UD po številu pojavnic precej večji kot korpus SST-UD, sem analizo izvedel hkrati še z manjšim vzorcem naključno izbranih povedi iz korpusa SSJ-UD in celotnim korpusom SST-UD. Število povedi za ta manjši vzorec je bilo enako številu v SST-UD (6.104 povedi), tako da sta bila vzorca po velikosti primerljiva. Vse izvedene analize in statistični testi so pokazali povsem enake tendence kot pri analizi s korpusom SSJ-UD v polni velikosti, zato v naslednjem razdelku poročam le rezultate primerjave obeh korpusov v polnem obsegu.

Za izvedbo statističnih analiz sem uporabil program Jamovi (The jamovi project, 2024), za izris grafov in korelacijske matrike pa Pythonovo knjižnico Seaborn (Waskom, 2021).

3 REZULTATI

3.1 Opisna statistika

Za vsako od mer v Tabeli 1 prikazujem primerjavo aritmetične sredine in standardnega odklona med obema korpusoma. Kaže se jasen vzorec: pri vseh merah skladenjske kompleksnosti je aritmetična sredina višja v pisnem korpusu SSJ-UD kot v govornem SST-UD. Slika 2 prikazuje primerjavo grafikonov kvartilov obeh korpusov za vsako mero. Ker se je izkazalo, da je v podatkih pri večini mer zelo veliko osamelcev, sem se odločil za postavitev končnih ročajev grafa na minimalno in maksimalno vrednost. Pri večini mer je frekvenčna porazdelitev zgoščena pri nižjih vrednostih, hkrati pa je prisotnih tudi veliko višjih vrednosti s

postopoma padajočo frekvenco. Ta pojav imenujemo tudi »dolgi rep« (angl. *long tail*) in ga pri delu z raznimi jezikovnimi podatki pogosto srečamo (Ryland Williams in sod., 2015). Drugačno frekvenčno porazdelitev opazimo predvsem pri meri NDO, ki pa se vseeno ne ujema povsem z normalno porazdelitvijo. Pri tej meri je večina vrednosti zgoščena bolj proti sredini variacijskega razmika kot pri ostalih.

Tabela 1: Pregled aritmetične sredine in standardnega odklona za vse mere kompleksnosti.

		B/P	S/P	S/T	PDO	NDO	MGD
SSJ	Aritm. sredina	19,9	2,28	1,59	2,57	1,14	5,09
	Stand. odklon	12,8	1,45	0,80	0,92	0,51	1,81
SST	Aritm. sredina	12,5	2,11	1,34	2,31	0,90	3,70
	Stand. odklon	14,5	1,95	0,63	0,89	0,48	2,13

3.2 Statistični testi

Tabela 2 prikazuje rezultate statističnega testa Mann-Whitney U za preverjanje razlik med korpusoma SSJ-UD in SST-UD za vsako od obravnavanih mer kompleksnosti. Rezultati kažejo, da so p-vrednosti za vse mere manjše od 0,01. Tudi po apliciranju metode Holm-Bonferroni so razlike med korpusoma pri vseh merah statistično pomembne. Velikosti efekta so večinoma nizke (< 0,3) z izjemo mer B/P in MGD, ki dosegata nekoliko višje vrednosti (> 0,4). To nakazuje, da so pri teh dveh merah razlike med korpusoma precej bolj očitne kot pri ostalih. V splošnem so nizke velikosti efekta pri ostalih merah lahko tudi posledica razmeroma majhne velikosti obravnavanih korpusov. Ena od možnih razlag je, da so te mere v prvi vrsti povezane s skladenjskimi pojavi, ki so v jeziku redkejši, zato pri manjši količini podatkov te razlike ne pridejo do izraza v tako veliki meri.

3.3 Korelacija med merami

Poleg preverjanja razlik med obema korpusoma sem preveril še stopnjo korelacije med različnimi merami kompleksnosti. Slika 3 prikazuje korelacijsko matriko za vse pare mer kompleksnosti. Vse obravnavane mere vsaj do neke mere pozitivno korelirajo, pri čemer je precej visoka korelacija med merama B/P in MGD ter B/P in S/P (> 0,7), nekoliko nižja, a vseeno znatna, pa je med MGD in S/P ter B/P in PDO (> 0,6). Najnižjo stopnjo korelacije dosega par NDO

Slika 2: Grafikoni kvartilov za primerjavo obeh korpusov pri vsaki meri.

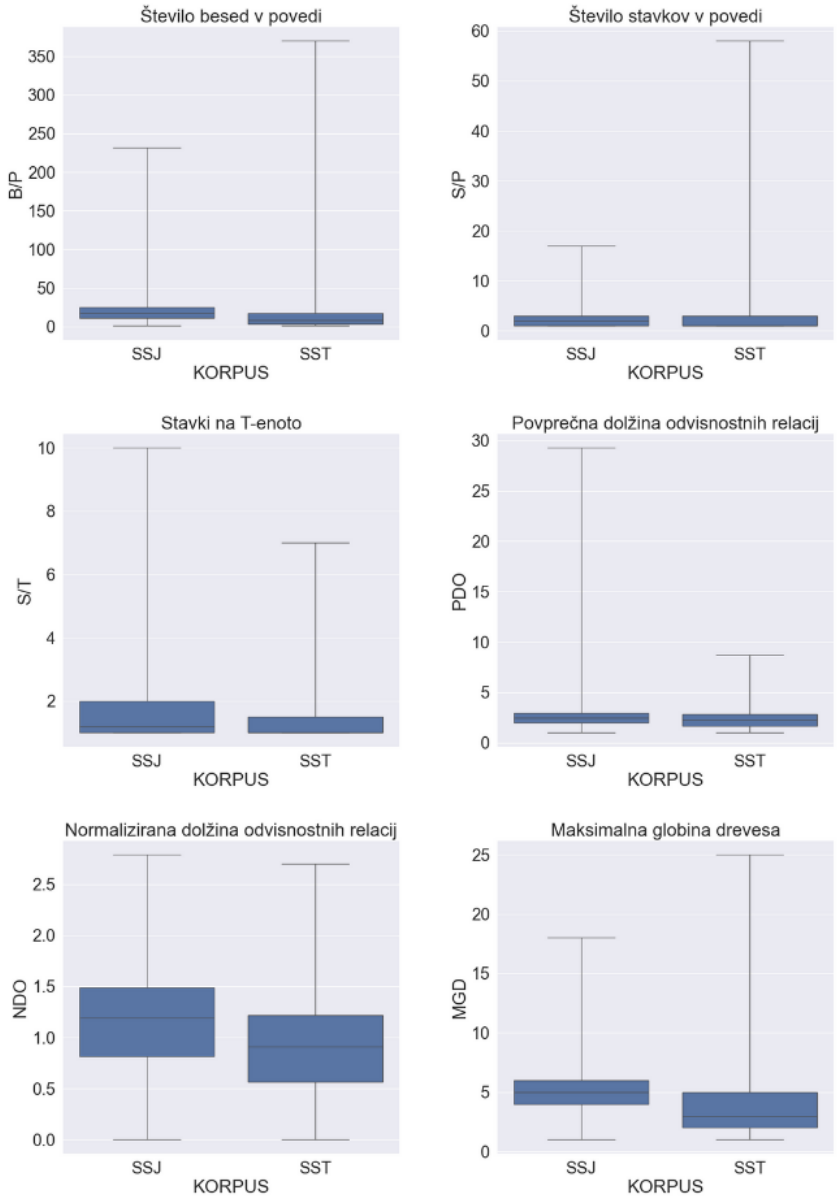
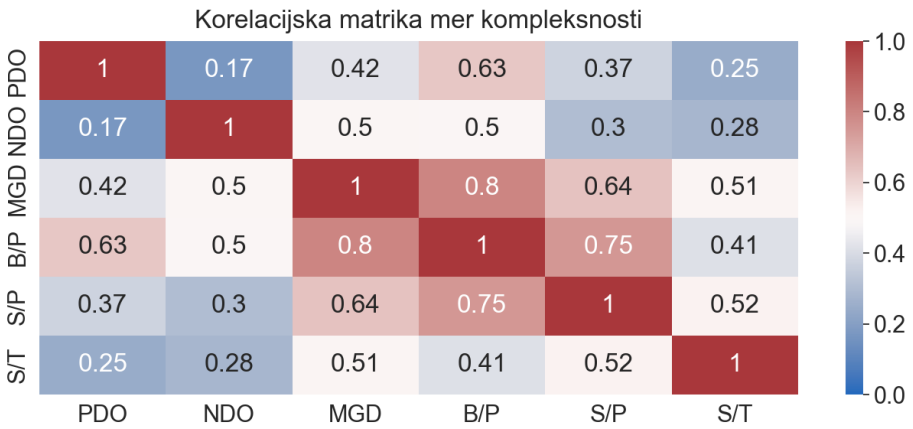


Tabela 2: P-vrednosti za statistični test Mann-Whitney U in velikost efekta (biserialni korelacijski koeficient) za vsako mero skladenjske kompleksnosti.

Mera kompleksnosti	p	Velikost efekta
B/P	<0,01	0,45
S/P	<0,01	0,15
S/T	<0,01	0,19
PDO	<0,01	0,17
NDO	<0,01	0,29
MGD	<0,01	0,43

Slika 3: Korelacijska matrika, ki prikazuje Pearsonov korelacijski koeficient za vsak par mer skladenjske kompleksnosti. Večja mera korelacije je prikazana z rdečo barvo, manjša pa z modro.



in PDO (< 0,2), kar ni presenetljivo, saj je bila mera NDO ustvarjena z namenom zmanjšanja vpliva dolžine povedi in pozicije korena na vrednosti, ki jih vrne mera PDO. Očitno je vpliv teh dveh dejavnikov na PDO tako velik, da je med merama še najmanj linearne korelacije. Največjo skupno korelacijo z ostalimi merami dosega B/P, najmanjšo pa NDO. P-vrednosti za vse pare obravnavanih mer so manjše od 0,01.

4 DISKUSIJA

Vse obravnavane mere kompleksnosti imajo v povprečju višje vrednosti v korpusu SSJ-UD kot v SST-UD. To nakazuje, da je skladenjska sestava pisnega jezika nekoliko kompleksnejša kot pri govorjenem jeziku, vsaj kar se tiče aspektov skladenjske kompleksnosti, ki jih pokriva izbranih šest mer. Višjo razpršenost vrednosti pri nekaterih merah – B/P, S/P in MGD – v korpusu SST-UD v primerjavi s SSJ-UD lahko pripišemo večjemu variacijskemu razmiku za dolžine povedi v govornem korpusu. To namreč odraža zapletenost segmentacije govorjenega jezika, ki jo izpostavljata že Dobrovoljc in Nivre (2016) v opisu govornega korpusa.

Analiza korelacijske matrike razkriva, da, četudi vse mere nakazujejo na precejšnje razlike v skladenjski kompleksnosti med govornim in pisnim korpusom, korelacija med samimi merami ni vedno visoka. To pomeni, da nekatere od mer zaznavajo precej drugačen vidik skladenjske kompleksnosti kot druge. Največjo korelacijo dosega par B/P in MGD, kar se ujema z ugotovitvijo Xu in Reitter (2016), ki opažata, da ti dve vrednosti v jeziku po navadi izkazujeta veliko mero linearne korelacije. Nasploh se zdi, da lahko mere razdelimo v dve skupini: tiste, ki izkazujejo precejšno mero linearne povezanosti s preprosto dolžino povedi B/P (S/P, MGD in PDO), in tiste, ki imajo z dolžino povedi manj linearne povezanosti (S/T in NDO). Meri S/T in NDO nam torej razkrivata nek aspekt skladenjske kompleksnosti, ki je še najmanj odvisen od preproste dolžine povedi, in nam pove več o notranji skladenjski sestavi neke povedi.

To razliko med obema vidikoma skladenjske kompleksnosti lahko ponazorimo s spodnjima dvema povedma iz korpusa SSJ-UD:

(a) *Spomnimo se samo pred leti prenosa žiro računa elektrogospodarstva v Ljubljano, spomnimo se "svežega" sedeža holdinga elektrogospodarstva, v to kategorijo pa spada tudi privatizacija Nove KBM.*

(b) *Kljub majskemu odprtju meja trga tisočerih priložnosti velikega povpraševanja delodajalcev nismo doživeli.*

Pri primeru (a) je vrednost B/P precej visoka (31) glede na povprečje v tem korpusu (19,9), vrednost za NDO pa je precej nizka (0,42) glede na povprečje

(1,14). Prav nasprotno je pri primeru (b), kjer je B/P nižji (13), NDO pa je višji (1,70). Primer (a) torej bolj odraža osnovnejši vidik skladenjske kompleksnosti, ki se v glavnem povezuje z dolžino povedi. V povedi (a) je razvidno, da tak jezik pogosto vsebuje veliko priredij in enostavnih sosledij zaporedno vezanih stavkov. Po drugi strani pa je (b) primer povedi, v kateri se bolj kaže zapletena notranja sestava skladenjskih elementov. Primer (b) predvsem zaznamuje dolg uvajalni element na začetku povedi (*Kljub majskemu odprtju…*), ki sproducira precej visoko dolžino odvisnostne relacije. Posledično ob prvem branju povedi bralec precej težko razbere, kje se začne naslednji skladenjski element in kako se skladenjska struktura povedi nadaljuje.

Ena od pomembnih razlik med obema vidikoma skladenjske kompleksnosti je tudi to, da je mera B/P že po svoji definiciji močno odvisna od prvotne segmentacije besedil v korpusu na povedi, kar posledično pomeni, da so od segmentacije precej odvisne tudi vrednosti MGD, S/P in PDO, ki izkazujejo visoko korelacijo z B/P. V nasprotju sta NDO in B/P, ki ne izkazujeta tako močne linearne povezanosti z B/P, precej manj odvisni od prvotne segmentacije. Povsem verjetno je, da so razlike med korpusi ob uporabi te prve skupine mer predvsem posledica različnih načel segmentacije ob nastanku obeh korpusov in le v manjši meri razkrivajo razlike v notranji skladenjski strukturi povedi. Če se vrnemo k Tabeli 2, lahko razberemo, da je velikost efekta najvišja prav pri B/P in MGD, kar lahko nakazuje na različna načela segmentacije v obeh obravnavanih virih. Tudi prej že omenjen višji variacijski razmik za dolžine povedi v SST-UD govori v prid različnim principom segmentacije v obeh korpusih.

Četudi sem zgoraj uporabljene mere skladenjske kompleksnosti razvrstil v dve grobi skupini, pa to ne pomeni, da razlik znotraj teh skupin ni. NDO in S/T, ki sta edini s precej manjšo korelacijo z mero B/P, hkrati med seboj linearno nista tako močno povezani (r = 0,28). Očitno torej vsaka od obravnavanih mer poda neko svojo posebno informacijo o skladenjski strukturi neke povedi. Skladenjsko kompleksnost moramo tako razumeti kot večplasten koncept, ki se ga ne da povzeti le z eno pravo mero, pač pa nam različni načini merjenja lahko povejo precej različne stvari.

5 ZAKLJUČEK

V študiji na podlagi analize skladenjsko razčlenjenih korpusov odkrivam, da je pisni jezik v slovenščini z vidika skladenjske sestave kompleksnejši kot govorjeni jezik. Vendar pa se ta razlika ne kaže le v enem vidiku skladenjske kompleksnosti, pač pa v večih.

Iz rezultatov je razvidno, da so razlike najbolj očitne pri dolžini povedi in pri globini skladenjskih dreves. Odkrivam tudi, da je med uporabljenimi merami veliko razlik, kar potrjuje opažanje številnih študij (Bentz in sod., 2023; Berdicevskis in sod., 2018; Ehret in sod., 2023; Jagaiah in sod., 2020), ki trdijo, da skladenjska kompleksnosti ni enoten pojem, ki bi se ga dalo povzeti oz. izmeriti z eno samo vseobsegajočo mero kompleksnosti, pač pa gre za večplasten koncept, ki ga lahko v jeziku analiziramo na več različnih načinov.

Rezultati kažejo, da se nekatere uporabljene mere skladenjske kompleksnosti v veliki meri ujemajo z dolžino povedi, so pa tudi take, ki tega ujemanja ne izkazujejo in nam zato razkrivajo nek drug vidik skladenjske kompleksnosti. Pri raziskavah skladenjske kompleksnosti je torej potreben poseben premislek o tem, kaj točno želimo izmeriti. Pri raziskavah, ki se osredotočajo na analizo uporabe prirednih struktur, je na primer smiselno osredotočanje na mere, kot je število stavkov v povedi. Če pa bi želeli preučevati skladenjske vzorce, ki ljudem najbolj otežujejo razumevanje besedil, pa bi morali v razmislek vključiti tudi druge mere. Martínez in sod. (2022) na primer ugotavljajo, zakaj pravna besedila v angleščini bralcem pogosto povzročajo težave v razumevanju v primerjavi z ostalimi vrstami besedil. Pri tej analizi uporabijo mere kompleksnosti, ki temeljijo na merjenju dolžine odvisnostnih relacij.

Doprinos te študije je večplasten: (1) v prispevku sem predstavil novo metodo za avtomatsko primerjanje razlik v skladenjski kompleksnosti med različnimi skladenjsko razčlenjenimi korpusi z uporabo šestih različnih mer, (2) preučil sem razlike v skladenjski kompleksnosti med reprezentativnima korpusoma govorjene in pisne slovenščine, (3) preučil sem razlike med šestimi uporabljenimi merami, ki se v literaturi uporabljajo kot merilo skladenjske kompleksnosti.

Pri tem je treba opozoriti na pomembno vlogo skladenjsko razčlenjenih korpusov s kvalitetnimi oznakami po principu odvisnostne skladnje. Takšni korpusi nam

namreč omogočajo uporabo mer, kot so PDO, NDO in MGD, ki sem jih uporabil tudi jaz v tej raziskavi. Stalno izboljševanje, vzdrževanje in nadgrajevanje tovrstnih skladenjsko razčlenjenih virov je torej izjemnega pomena za jezikoslovne raziskave v slovenščini.

Seveda moja študija ni brez pomanjkljivosti. Čeprav sta obravnavana korpusa v primerjavi s številnimi drugimi, ki so vključeni v zbirko Universal Dependencies, kar obsežna, se obenem ne moreta primerjati z velikostjo nekaterih drugih skladenjsko označenih korpusov, kot je npr. PDT (Hajič in sod., 2020). Pri preučevanju skladenjskih pojavov je še posebej pomembna količina preučenih podatkov, zato bo v prihodnosti, ko bo na voljo večja količina podatkov s skladenjskimi oznakami, potrebna ponovna preverba tokratnih ugotovitev. V študiji tudi izhajam iz predpostavke, da je stavčna segmentacija v pisnem korpusu primerljiva s segmentacijo v govornem korpusu, kar pa lahko v določenih primerih vodi do velikih razlik pri rezultatih, ki jih vrnejo določene mere kompleksnosti, posebej pri številu besed na poved. Vsekakor problem določanja ustrezne segmentacije v govornih korpusih ni enostavno rešljiv (Dobrovoljc in Nivre, 2016), zato sem se odločil obdržati izvorno označenost v vseh virih in tako obdržati preprosto zasnovo raziskave.

Rezultati te študije hkrati odpirajo tudi veliko vprašanj za nadaljnje študije. Podobno raziskavo bi bilo zanimivo izvesti z uporabo tujih skladenjsko označenih korpusov in ugotovitve primerjati s temi, ki sem jih predstavil v tem prispevku. V svoji raziskavi sem različne mere skladenjske kompleksnosti primerjal predvsem z vidika linearne korelacije, naslednji korak pa je še natančneje raziskati, na kakšen način se mere kompleksnosti med seboj razlikujejo. Vredno bi bilo na primer na bolj sistematičen in razširjen način identificirati skupine primerov, ki dosegajo visoke vrednosti pri vsaki od mer. Prav tako bi bilo zanimivo podobno raziskavo ponoviti še z drugimi merami, ki se v literaturi povezujejo s konceptom skladenjske kompleksnosti in jih v tem prispevku ne uporabljam.

6 DOSTOPNOST PODATKOV

Vse Pythonove skripte, s pomočjo katerih sem naredil izvoze in poračunal vrednosti za mere, so dostopne prek GitHub repozitorija.[5]

[5] https://github.com/lukatercon/SyntComplex

7 ZAHVALA

Delo, predstavljeno v tem prispevku sta financirala program mladi raziskovalci Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije ter projekt SPOT – Na drevesnici temelječ pristop k raziskavam govorjene slovenščine, ARIS št. Z6-4617.

LITERATURA

Bentz, C., Gutierrez-Vasques, X., Sozinova, O. in Samardžić, T. (2023). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, *9*(s1), 9–25. Pridobljeno 2024-05-31, https://doi.org/10.1515/lingvan-2021-0054 doi: 10.1515/lingvan-2021-0054

Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., Yan. C., Demberg, V., Lupyan, G., Rama, T., & Bentz, C. (2018). Using Universal Dependencies in cross-linguistic complexity research. V M.-C. de Marneffe, T. Lynn in S. Schuster (Ur.), *Proceedings of the second workshop on universal dependencies (UDW 2018)* (str. 8–17). Brussels, Belgium: Association for Computational Linguistics. https://aclanthology.org/W18-6002 doi: 10.18653/v1/W18-6002

Brank, J. (2023). *Q-CAT corpus annotation tool 1.5.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1844

de Marneffe, M.-., Manning, . D., Nivre, J. in Zeman, D. (2021, June). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://aclanthology.org/2021.cl-2.11 doi: 10.1162/coli_a_00402

Dobrovoljc, K., Erjavec, T. in Krek, S. (2017). The Universal Dependencies treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger in R. Yangarber (Ur.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (str. 33–38). Association for Computational Linguistics. https://aclanthology.org/W17-1406 doi: 10.18653/v1/W17-1406

Dobrovoljc, K. in Ljubešić, N. (2022). Extending the SSJ Universal Dependen-es treebank for Slovenian: Was it worth it? V S. Pradhan in S. Kuebler (Ur.), *Proceedings of the 16th linguistic annotation workshop (law-xvi) within lrec2022* (str. 15–22). European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K. in Nivre, J. (2016). The Universal Dependencies treebank of spo- ken Slovenian. V N. Calzolari in sod. (Ur.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (str. 1566–1573).

European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Dobrovoljc, K., Terčon, L. in Ljubešić, N. (2023). Universal dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni mo- del. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 218–246. https://journals.uni-lj.si/slovenscina2/article/view/12031 doi: 10.4312/slo2.0.2023.1.218-246

Ehret, K., Berdicevskis, A., Bentz, C. in Blumenthal-Dramé, A. (2023). Measuring language complexity: challenges and opportunities. *Linguistics Vanguard*, *9*(s1), 1–8. Pridobljeno 2024-05-31, https://doi.org/10.1515/lingvan-2022-0133 doi: 10.1515/lingvan-2022-0133

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. V N. Calzolari in sod. (Ur.), *Proceedings of the seventh internati- onal conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Futrell, R., Mahowald, K. in Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gaberščček, N. (2018). Merjenje določenih vidikov skladenjske kompleksnosti v pisnih besedilih slovenskih osnovnošolcev. *Jezik in slovstvo*, *63*(2/3).

Hajič, J., Bejček, E., Hlavacova, J., Mikulová, M., Straka, M., Štěpánek, J. in Štěpánková, B. (2020). Prague dependency treebank - consolidated 1.0. V N. Calzolari in sod. (Ur.), *Proceedings of the twelfth language resources and evaluation conference* (str. 5208–5218). Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.641

Jagaiah, T., Olinghouse, N. G. in Kearns, D. M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and wri- ting quality. *Reading and Writing*, *33*(10), 2577-2638. https://doi.org/10.1007/s11145-020-10057-x doi: 10.1007/s11145-020-10057-x

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020). The ssj500k training corpus for slovene language processing. *Jezikovne Tehnologije in Digitalna Humanistika*, 24–33.

Lei, L. in Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, *27*(1), 62–79. https://doi.org/10.1080/09296174.2018.1504615 doi: 10.1080/09296174.2018.1504615

Lintunen, P. in Mäkilä, M. (2014). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language*,

12. doi: 10.1515/rela-2015-0005

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., ... Kolkovska, S. (2023). *Parallel sense-annotated corpus ELEXIS-WSD 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/ 1842

Martinc, M., Pollak, S. in Robnik-Šikonja, M. (2021). Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, *47*(1), 141-179. https://doi.org/10.1162/coli_a_00398 doi: 10.1162/coli_a_00398

Martínez, E., Mollica, F. in Gibson, E. (2022). Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, *224*, 105070. https://www.sciencedirect.com/science/article/pii/S0010027722000580 doi: https://doi.org/10.1016/j.cognition.2022.105070

Mylläri, T. (2020a). Measuring syntactic complexity in learner finnish. *Apples: Journal of Applied Language Studies*, *14*(2).

Mylläri, T. (2020b). Words, clauses, sentences, and t-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association*, *4*(1).

Ryland Williams, J., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M. in Sheridan Dodds, P. (2015). Zipf's law holds for phrases, not words. *Scientific reports*, *5*(1).

The jamovi project. (2024). *jamovi (version 2.5).* https://www.jamovi.org

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ... Rupnik, P. (2023). *Spoken corpus gos 2.1 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/ 11356/1863

Wang, Z. (2022). Dynamic development of syntactic complexity in second language writing: A longitudinal case study of a young chinese efl learner. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.974481 doi: 10.3389/fpsyg.2022.974481

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021 doi: 10.21105/joss.03021

Xu, Y. in Reitter, D. (2016). Convergence of syntactic complexity in conversation. V *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (str. 443–448).

THE USE OF SIX SYNTACTIC COMPLEXITY MEASURES FOR LINGUISTIC COMPARISONS BETWEEN A SPOKEN AND A WRITTEN CORPUS

There are a number of methods for measuring syntactic complexity in digital language databases. Linguistic corpora, especially those containing syntactic annotations, enable researchers to automatically and efficiently conduct analyses and comparisons of syntactic complexity. In this paper, I present a method with which I automatically compare two corpora – one containing written texts and the other containing spoken texts – using six established measures of syntactic complexity. The results of this comparison indicate that the syntactic makeup of the language contained in the written corpus is slightly more complex than in the spoken corpus. The differences are most pronounced in sentence length and in syntactic tree depth. Additionally, an analysis of the correlation between the different measures suggests that some provide quite different information about the syntactic structure of a sentence compared to others.

Keywords: syntactic complexity, written corpus, spoken corpus, complexity measures, student paper