UDK 811.163.42'324
811.163.42'42
Izvorni znanstveni rad
Rukopis primljen 24. IX. 2013.
Prihvaćen za tisak 13. VI. 2014.

Sanja Brangan
Škola narodnog zdravlja „Andrija Štampar"
Medicinski fakultet Sveučilišta u Zagrebu
Rockefellerova 4, HR-10000 Zagreb
skusec@snz.hr

KVANTITATIVNA PROCJENA TEŽINE TEKSTA NA HRVATSKOM JEZIKU

U radu su prikazana dosadašnja istraživanja u području statističke analize teksta, s posebnim naglaskom na razvoj formula čitkosti. Pored teorijskog objašnjenja čitanja, razumijevanja i čitkosti, prikazani su rezultati analize korpusa tekstova na engleskom i hrvatskom jeziku. Na kraju, preporučuje se formula čitkosti za hrvatski jezik, modificirana prema Fleschovoj formuli za engleski jezik, koja se može koristiti kao objektivni pokazatelj za grubu procjenu težine teksta na hrvatskom jeziku.

1. Uvod

Koliko je određeni tekst teško pročitati i razumjeti vrlo često podliježe subjektivnoj procjeni osoba koje nastoje izabrati ili preporučiti tekst ciljnoj populaciji. Tako se događa da npr. neko povjerenstvo za evaluaciju tekstova donese mišljenje o težini određenog teksta imajući u vidu apstraktnog potencijalnog čitatelja, a odluka zapravo predstavlja osobno mišljenje člana ili članova povjerenstva koji su uspjeli pobijediti u međusobnoj raspravi; ili da nastavnici izaberu udžbenike za svoje učenike, koji se kasnije pokažu preteškima za čitanje i razumijevanje, što samo frustrira učenike i dovodi do otežanog procesa učenja (Zakaluk i Samuels 1988: xi).

Težina teksta za čitanje i razumijevanje nije od važnosti samo za područje obrazovanja, gdje tekstovi služe za proces učenja, nego i za druga područja u kojima težina teksta može biti presudna za odluku čitatelja hoće li uopće či-

tati odnosno nastaviti čitati tekst ili može li na temelju pročitanog donijeti dovoljno kompetentnu odluku npr. što učiniti ili kako treba postupiti. Ta područja uključuju novinarstvo, istraživanje, zdravstvenu zaštitu, zakonodavstvo, osiguranje, industriju, vojsku, knjižničarstvo itd. Tako je npr. američko zakonodavstvo, pod utjecajem pokreta za jednostavni jezik (*Plain Language Movement*) 1960-ih godina, odredilo da se u dokumentima za javne i komercijalne svrhe mora koristiti jednostavan jezik (DuBay 2004: 2). U Hrvatskoj se slična odredba u zakonodavstvu nalazi u *Zakonu o zaštiti prava pacijenata* iz 2004. godine (*Narodne novine* br.169/04), gdje se navodi da „pacijent ima pravo dobiti obavijesti na način koji mu je razumljiv s obzirom na dob, obrazovanje i mentalne sposobnosti". Pri tome se može činiti da su bar dob i obrazovanje jednostavno mjerljive kategorije, ali pokazuje se da se stupanj sposobnosti čitanja i razumijevanja ne može s točnošću procijeniti samo na temelju stupnja obrazovanja osobe. Pri procjeni mentalnih sposobnosti još je teže donijeti imalo objektivnu procjenu u nedostatku valjanih testova, a oni se u svakodnevnoj praksi u Hrvatskoj niti ne primjenjuju. Stoga je cilj ovoga rada preporučiti metodu za kvantitativnu procjenu težine teksta na hrvatskom jeziku, koja će biti objektivnija od osobne procjene nekog procjenitelja, a temeljit će se na rezultatima sličnih istraživanja u svijetu. Prije samog prikaza analize korpusa tekstova, koji čine odabrani paralelni prijevodni tekstovi na engleskom i hrvatskom jeziku, u daljnjem tekstu ovog rada pobliže se objašnjavaju pojmovi *čitkost* i *formule čitkosti* u kontekstu dosadašnjih statističkih istraživanja jezika.

1.1. Čitkost, čitanje i razumijevanje

Pojam *čitkost* može se definirati različito, a neke su od najkraćih definicija „jednostavnost čitanja riječi i rečenica" i „jednostavnost razumijevanja kao posljedica načina pisanja" (DuBay 2004: 3). Čitkost tako obuhvaća i proces čitanja i proces razumijevanja, a za objašnjenje težine određenog teksta potrebno je uzeti u obzir oba aspekta, tj. sam tekst, kako je napisan i koje sve elemente sadržava, te čitalačke sposobnosti osobe koja taj tekst čita. To obuhvaća i objašnjenje samog procesa čitanja: što čitanje jest, koji čimbenici utječu na čitanje te po čemu se razlikuju dobri i loši čitatelji.

Postoje brojne teorije čitanja, što pokazuje kako je čitanje složen proces te je vrlo teško opisati što se događa i što radimo dok čitamo. Pretpostavka je da čitanjem dolazimo do značenja teksta kroz proces interakcije s pisanom riječi. Čitanje obuhvaća vještine prepoznavanja riječi, brzine čitanja odnosno tečnosti, te vještine razumijevanja rješavanjem problema. Za tečno čitanje neophodno je da dolazi do automatiziranog prepoznavanja riječi kako bi se oslobodio kapa-

citet pažnje za razumijevanje, a gruba procjena optimalne brzine čitanja iznosi otprilike 300 riječi u minuti. Dobri čitatelji tako će biti ne samo brzi pri čitanju, nego i precizni (Alderson 2000: 57). Uočene su neke konkretne razlike između dobrih i loših čitatelja s obzirom na proces čitanja i na proizvod čitanja. Tako loši čitatelji sporo čitaju, doslovno shvaćaju pročitane riječi, preskaču neke riječi, ne shvaćaju kontekst, brzo se umaraju čitajući te ne razumiju pročitano. S druge strane, dobri čitatelji tečno čitaju, znaju pronaći pomoć za nepoznate riječi, shvaćaju kontekst, uporni su u čitanju i znaju protumačiti značenje riječi (Doak, Doak i Root: 4). Tečni čitatelji tijekom čitanja procesiraju otprilike 80 % leksičkih i 40 % funkcionalnih, gramatičkih riječi, a ne samo da tečno čitaju, nego su i precizni u čitanju, tj. njihovo je čitanje točno. Poteškoće u čitanju zbog nepoznatih riječi negativno utječu na razumijevanje, pa je potrebno da čitatelji poznaju 95 % riječi nekog teksta kako bi ga adekvatno razumjeli te kako bi mogli predvidjeti značenje nepoznatih riječi iz konteksta. Procjena je za engleski jezik da osoba mora vladati vokabularom od otprilike pet tisuća riječi da bi mogla razumjeti 97 % riječi uobičajenog teksta, a ako vlada sa samo dvije tisuće najučestalijih riječi, razumjet će tek 90 % teksta (Alderson 2000: 35).

Pored navedenih osobina čitatelja, na uspješnost čitanja utječu i dob, spol, zanimanje, inteligencija, društveni sloj, obrazovanje i drugi čimbenici. S druge strane, sam tekst ima određene čimbenike koji olakšavaju ili otežavaju čitanje: organizacija i kohezija teksta, tema i sadržaj, vrsta odnosno žanr teksta, tipografske i lingvističke osobine, čitkost itd. Može se očekivati da će jezici s transparentnom ortografijom biti čitkiji od onih s netransparentnom ortografijom, kakav je engleski jezik (Alderson 2000: 75). Težina vokabulara svakako utječe na razumijevanje, a može se izraziti dužinom i frekvencijom riječi. Dužina riječi, za engleski jezik, ugrubo je povezana s frekvencijom riječi: najfrekventnije riječi su obično kraće (Alderson 2000: 71). Chall, autor jedne od formula čitkosti, navodi da je težina vokabulara zapravo najznačajniji prediktor težine teksta (Alderson 2000: 73). Također, na razini rečenice pokazuje se da su kraće rečenice sintaktički jednostavnije, što znači da će biti i razumljivije. Postoji visoka korelacija između čitkosti teksta mjerene testovima razumijevanja na ispitanicima i mjerene formulama čitkosti (Alderson 2000: 72). Iako neki autori iz ovog područja preporučuju da se za mjerenje težine odnosno čitkosti teksta uzme stručna procjena odabrane skupine stručnjaka, a ako to nije moguće, da se primijeni neka od formula čitkosti (Alderson 2000: 72), vrlo je teško procijeniti tko su dovoljno kompetentni stručnjaci koji bi napravili procjenu težine teksta. Još neobjavljeni podaci za Hrvatsku, ali prezentirani na skupu o zdravstvenoj pismenosti, pokazuju da je raspon procjene završne godine studenata medicine u Zagrebu o tekstu za pacijente toliko velik da obuhvaća 3 – 19 godina školo-

vanja, dok bi formulom čitkosti težina tog istog teksta bila izražena kao 11 godina školovanja (Brangan 2013). Test razumljivosti na pacijentima pokazuje da je taj tekst bio potpuno nerazumljiv za sve ispitanike sa završenom osnovnom školom te za 44 % ispitanika sa srednjim obrazovanjem, ali za samo 4 % ispitanika s visokim obrazovanjem (Kušec 2006: 297). Ti rezultati pokazuju koliko je nepouzdana procjena stručnjaka iz određenog područja, ali i kako adekvatna formula čitkosti korelira s testom razumijevanja te može poslužiti kao objektivna procjena težine teksta.

1.2. Teorijske postavke o jeziku i komunikaciji i statistička analiza jezika kao polazište za razvoj formula čitkosti

Jezik je sredstvo komunikacije i međusobnog razumijevanja među ljudima. Ali jezik je i mnogo više: on je instrument mišljenja, sredstvo bilježenja činjenica, izražavanje identiteta, ali i kontrole stvarnosti. Jezik određuje način našeg razmišljanja, a također je točno da se ljudi bolje sjećaju onih stvari koje mogu izraziti poznatim riječima. Jezik isto tako određuje našu pripadnost određenoj skupini – jezičnoj, ali i društvenoj. Međutim, iako u svijetu postoji mnoštvo različitih jezika, međusobno razumijevanje ipak je moguće uspostaviti.

Postoje brojne definicije jezika, od kojih neke sagledavaju jezik kao općeniti pojam, a druge vrlo specifično. Enciklopedija jezika tako navodi sljedeće definicije jezika (Crystal 1997: 400): „skup rečenica ograničene dužine sastavljenih od ograničenog broja elemenata" (Chomsky); „sustavno sredstvo komuniciranja ideja ili osjećaja korištenjem konvencionalnih znakova, zvukova, gesta ili oznaka, koje imaju razumljiva značenja" odnosno „sposobnost verbalnog izražavanja i korištenja riječi u odnosu među ljudima" (Webster's Third New International Dictionary); „potpuno ljudska neinstinktivna metoda komuniciranja ideja, osjećaja i želja putem dobrovoljno proizvedenih simbola" (Sapir). Prema hipotezi Sapir–Whorfa, jezik je „oblikovač ideja" (Hudson 1996: 96).

Psiholingvistika pruža brojne dokaze da je svaki govornik po svojoj psihičkoj konstituciji i po svojem intelektualnom potencijalu različit od drugog govornika, da nijedan ne vlada cjelinom jezičnog sustava, niti svim njegovim upotrebnim mogućnostima. To znači da je svaki čovjek u slobodi upotrebe jezičnog sustava ograničen i osobinama svoje ličnosti, a ogleda se u ograničenosti jezične i komunikacijske kompetencije pojedinca (Škiljan 1988: 69–78). Tu možemo govoriti o 'stilu', koji jedni definiraju kao način upotrebe jezika, a drugi kao obilježje govora pojedinca (Olsson 2004: 34). Stilistika, kao disciplina na raskrižju znanosti o jeziku i znanosti o književnosti, bavi se suštinom izraza te govori o stilu kao o „individualnoj upotrebi jezika" (Malmberg 1979:

240), „skupu razlikovnih obilježja po kojima možemo identificirati osobe, mjesta, predmete ili razdoblja", a više pjesnički rečeno, stil je „ruho misli" (Crystal 1997: 66). Pritom se statistička lingvistika definira kao „područje koje istražuje ne samo razlike među uzorcima ili tekstovima, već i obilježja koja su zajednička uzorcima, a time i cijelim jezicima odnosno svim jezicima, kao dio potrage za jezičnim univerzalijama" (Crystal 1997: 67). Govoreći o teoriji stila, Olsson pokazuje dva različita pristupa odnosno poimanja stila: jedni autori shvaćaju stil kao rezultat lingvističkog izbora nekog govornika odnosno pisca, koji se može svjesno primjenjivati, dok drugi smatraju stil rezultatom nesvjesnih lingvističkih navika koje nas odaju kao govornike. Može se zaključiti da je stil ipak višedimenzionalan proizvod koji je izvan potpune kontrole pojedinca, a uvjetovan je žanrom odnosno registrom i kontekstom, te je vlasništvo šire zajednice, a ne pojedinca. Ono što je specifično za nekog pojedinca, njegov 'idiolekt', bit će uvjetovan obilježjima jezika unutar nekog društva, tj. 'sociolektom', ali i obratno (Olsson 2004: 34).

Govoreći o slobodi govora, Škiljan je definira kao „mogućnost da se bilo koji jezični sadržaj pojavi bilo u kojem kanalu javne komunikacije, i to u svakom kontekstu" (Škiljan 1988: 69–78). No, brojna su ograničenja slobode govora: jezična, komunikacijska, kontekstualna, socijalna, psihološka, čak i etička. Strukture jezika ipak nisu tako neprikosnovene, one su u izvjesnom smislu elastične, neka su odstupanja dozvoljena, samo je pitanje koliko veliko se odstupanje može tolerirati i koja mu je svrha. To znači da jezični znak stječe svoju realnu znakovnost jedino ako prihvatimo zajedno s njim i društveno determiniranu konvenciju o njegovoj vrijednosti (Škiljan 1988: 69–78). Društveno determinirane konvencije usko su vezane za društvo u cjelini ili za pojedine skupine u društvu. Npr., kod komunikacije koja uključuje profesionalne i laičke sugovornike, postavlja se pitanje koliko konvencija odnosno koje konvencije ti sugovornici dijele, a koje ne. To nas vodi do problema znanja i onoga što je zajedničko tim dvjema skupinama. Istraživanje zdravstvene komunikacije u Hrvatskoj (Kušec 2007: 319–328) pokazuje jasne razlike u jeziku između liječnika i pacijenata s obzirom na odabir termina, sintagme, frekvenciju riječi, dužinu riječi itd. To pokazuje koliko je teško npr. u usmenoj komunikaciji u kratkom vremenu i na osnovu šturih informacija procijeniti razinu znanja i prihvatljivost jezika sugovornika. S druge strane, kada se radi o masovnoj komunikaciji, kakvu često susrećemo u pisanom obliku, postavlja se pitanje kako odrediti što je prosjek, što je ono 'opće' širem društvu ili pojedinoj skupini. Uzmemo li u obzir različitu dostupnost i izvore znanja te socijalnu, geografsku i drugu raznolikost jezika, još će teže biti odgovoriti na pitanje o opsegu društveno determiniranih konvencija.

Potvrdu o poimanju jezika kao čimbenika ograničavanja čovjekovih mogućnosti u jezičnoj djelatnosti nalazimo u mnogih mislilaca 20. stoljeća. U svojem djelu *Tečaj opće lingvistike* iz 1949. godine Ferdinand de Saussure govori i o razlici između jezika i govora: govor je individualni čin, pojedinac ne može u jeziku ništa promijeniti, a jezični sustav kojim vlada 'masa koja govori' (*masse parlante*) uvijek prethodi pojedinačnom iskazu i određuje jezično djelovanje govornika. Jasno je da svaki pojedinac ipak malo drukčije strukturira svoje iskaze, ali time ustvari dokida mogućnost prave komunikacije s drugima, pa se potvrđuje da je govor podređen jeziku (Škiljan 1988: 69–78).

Komunikacija je, općenito gledano s aspekta teorije komunikacije, „prenošenje obavijesti kojim se želi postići neki efekt" (Škiljan 1988: 69–78). Ona je, dakle, funkcija jezika. Tradicionalni lingvisti, govoreći o stilistici, kažu da se funkcija jezika očituje iz specifičnog, svrsi prilagođenog izbora jezičnih jedinica u iskazu. To znači da komunikacija može biti uspješna samo onda ako je poruka, i na planu izraza i na planu sadržaja, adekvatna cilju koji govornik želi postići. Može se reći da „ništa što u jeziku postoji a nije primjereno svrsi nekog konkretnog iskaza, ako se jezikom želi komunicirati, nije dozvoljeno, ili bar nije dobro izabrati" (Škiljan 1988: 69–78). Odabir za koji se autor poruke odluči između različitih mogućnosti koje mu jezik stavlja na raspolaganje tada postaje njegovim stilom. Sa stajališta stila kaže se da je izraz neke poruke obojen u odnosu na primatelja poruke kada je on pod određenim dojmom, odnosno učinkom, a takav obojeni izraz poruke može biti rezultat psihološkog cilja govornika odnosno autora poruke (Molinie 2002: 22). Može se reći da je u govoru svakog pojedinca prisutno njegovo individualno, ali i kolektivno iskustvo.

Proučavanja o jeziku obuhvaćaju dva glavna područja: proučavanje strukture jezika i proučavanje uporabe jezika. Tradicionalno su lingvističke analize naglašavale važnost strukture (npr. morfemi, riječi i izričaji, gramatičke strukture itd.) i opisivale na koji se način manje jedinice mogu kombinirati i tvoriti veće gramatičke jedinice. Drukčiji je pristup jeziku proučavanje uporabe jezika: kako se kod govorenja i pisanja koristimo jezičnim resursima, a pritom nije važno što je teoretski u jeziku moguće, nego kako jezik u stvarnosti izgleda (Biber 1998: 1).

Pri analizi uporabe jezika možemo postaviti različita pitanja kao polazišta istraživanja: npr. zašto u jeziku postoje različite strukture za slična značenja i gramatičke funkcije, preferiraju li usmeni i pisani jezik različite strukture, upotrebljavaju li se preferirane strukture za specijalizirana značenja, kakav je jezik nekog teksta ili skupine ljudi koji nešto govore odnosno pišu, razlikuje li se jezik muškaraca i žena itd. Svako znanstveno opisivanje jezičnih pojava pretpostavlja bar nekakav oblik statističke obrade prikupljenih podataka, ali da bi re-

zultat bio znanstveno prihvatljiv, mora se temeljiti na statistički značajnim razlikama u podacima mjerenja. Tako se i u području jezika i komunikacije često upotrebljavaju statistički proračuni i matematičke metode za analizu sistematiziranih korpusa jezičnih materijala (Malmberg 1979: 213–231).

Korpus možemo definirati kao „pisani ili govorni jezični resurs, koji je prikupljen i obilježen u cilju: analize jezika kojom se utvrđuju njegova svojstva, analize ljudskog ponašanja u određenim situacijama, obuke sustava kako bi se njegovo ponašanje prilagodilo specifičnim jezičnim okolnostima, empirijske provjere neke jezične teorije, izrade teksta za neku jezično inženjersku tehniku ili pak primjene kojom se utvrđuje njeno funkcioniranje u praksi" (Tadić 2003: 156). Kraća definicija kaže da je korpus „reprezentativni uzorak jezika, koji se prikuplja u svrhu lingvističke analize" (Crystal 1997: 414), a u novije vrijeme taj se pojam odnosi na „zbirke tekstova koje se pohranjuju i kojima se pristupa elektronički" (Hunston 2002: 2). Lingvistička analiza koja se temelji na korpusu ima sljedeće karakteristike: 1) ona je empirička, jer analizira stvarne obrasce upotrebe jezika u prirodnim tekstovima; 2) za polazište svog istraživanja koristi se velikom zbirkom prirodnih tekstova koju nazivamo korpusom; 3) za analizu se iscrpno koristi računalom, pomažući se automatskim i interaktivnim postupcima; 4) oslanja se na kvantitativne i kvalitativne analitičke postupke (Biber 1998: 4).

Upotreba računala u korpusnoj lingvistici je neizbježna. Računalo omogućuje utvrđivanje i analizu složenih obrazaca jezične upotrebe, pohranjivanje i analizu velikih baza podataka prirodnog jezika te dosljednu i pouzdanu analizu. Međutim, i prije pojave osobnih računala radile su se analize korpusa kao analize statističkih obrazaca uporabe jezika odnosno različitih jezičnih obilježja, što je postalo dio statističke lingvistike, odnosno lingvističke stilistike (Crystal 1997: 67). Jedan od pionira statističke obrade jezičnih podataka bio je George Kingsley Zipf, koji je još 1929. godine otkrio neka opća pravila učestalosti, kao npr. da je složenost nekog fonema obrnuto proporcionalna učestalosti toga fonema te da su bezvučni glasovi dvaput češći od zvučnih u jezicima gdje oni postoje. Svoje je 'načelo najmanjeg napora' primijenio ne samo na foneme, nego i na druge elemente jezika, naročito riječi. Otkrio je da su frekvencija riječi, izražena brojem pojavljivanja te riječi u danom tekstu, i njezin rang, tj. broj u poretku napravljenom prema učestalosti riječi, konstantni. To znači da za svaki sustav znakova postoji optimalna distribucija, koja omogućuje prenošenje maksimuma informacija uz minimum utrošene energije (Malmberg 1979: 214–215). Poslije se ipak pokazalo da taj odnos frekvencije riječi i njezina frekvencijskog ranga ne vrijedi za riječi najviše i najniže frekvencije te da je veličina korpusa ključni faktor, ali ipak ta 'standardna krivulja' frekvencija riječi pokazuje izrazito zanimljiv jezični obrazac (Crystal 1997: 87).

Uslijedile su analize učestalosti slova, a u okviru analize strukture riječi, broja slogova. Važnost dobivaju proučavanja korištenja rječnika neke grupe ili zajednice, za potrebe učenja jezika izrađuju se popisi riječi po učestalosti kako bi se u nastavi najprije učile najobičnije riječi, a tek zatim manje obične odnosno manje učestale riječi. Tako krajem 19. stoljeća W. F. Kaeding pokazuje na materijalu od 11 milijuna riječi da 15 najčešće korištenih riječi predstavlja 25 % od cjelokupnog broja riječi u tekstu, da 66 najfrekventnijih riječi predstavlja 50 %, a 320 najfrekventnijih riječi 72 % ukupnog broja riječi (Malmberg 1979: 217). Ti su omjeri kasnije potvrđeni i istraživanjima na drugim jezicima te se ukratko može reći da će 50 najfrekventnijih riječi činiti oko 45 % bilo kojeg teksta (Crystal 1887: 86). Pierre Guiraud je dodatno pokazao 1954. godine da u bilo kojem jeziku i u bilo kojem tekstu 100 najfrekventnijih riječi čini 60 % svih riječi tog teksta, 1000 najfrekventnijih riječi čini 85 %, a 4000 najfrekventnijih riječi čini 97,5 % teksta (Malmberg 1979: 219). No, kako na rezultate utječe dužina i vrsta teksta, katkad se 25 % svih riječi može postići i sa samo 12 najfrekventnijih riječi, kao u korpusu Lancaster–Oslo/Bergen (Crystal 1997: 87). Guiraud također pokazuje kako u svakom tekstu vrlo malen broj riječi čini njegov najveći dio te da vrlo malen broj odgovarajućih riječi može pokriti velik dio bilo kojeg teksta. Također, promatrajući odnos dužine riječi izražene brojem fonema i frekvencije te riječi zaključuje da su najfrekventnije riječi kraće (Malmberg 1979: 220). Računajući dužinu riječi izraženu brojem slogova u riječi, i Kaeding slično pokazuje na korpusu njemačkih tekstova da najčešće riječi imaju manje slogova, odnosno da je čak 50 % svih riječi jednosložno (Crystal 1997: 87).

Danas su najčešće citirani korpusi: Brown (sveučilišta Brown, korpus američkog engleskog jezika), London–Lund (korpus govornog engleskog jezika), Lancaster–Oslo/Bergen (LOB) (korpus britanskog engleskog jezika), COBUILD (noviji korpus suvremenog engleskog govornog i pisanog jezika), British National Corpus (BNC) (britanski nacionalni korpus) i dr. U novije vrijeme pojavljuju se nacionalni korpusi pojedinih manjih zemalja odnosno jezika, a ideja za Hrvatski nacionalni korpus, dostupan na stranici http://www.hnk.ffzg.hr, koji su radili autori s Filozofskog fakulteta Sveučilišta u Zagrebu i na temelju kojega je nastao i *Hrvatski čestotni rječnik* (Moguš, Bratanić i Tadić 1999), pojavila se 1996. godine (Tadić 1997: 392).

Utvrđivanje ili pripisivanje autorstva, katkad nazvano i ‘atribucijska stilistika’ (Molinie 2002: 26), koristi se rezultatima analiza na velikim korpusima kako bi se uočile sličnosti i razlike pojedinih tekstova ili njihovih dijelova. Najranije ideje o istraživanju i pripisivanju autorstva sežu u 18. stoljeće, s pojavom većeg broja knjiga, a predmet razmatranja bili su tekstovi u *Bibliji*. Kasni-

ja važnija kontroverzna mišljenja o autorstvu vežu se za Shakespearea i njegova djela. No, prvi spomen statističke primjene u analizi teksta veže se za Augustusa de Morgana, koji je dužinu riječi mjerenu brojem slova uzimao kao mjeru, a zatim za T. C. Mendenhalla, koji je proučavao autore s obzirom na različite dužine korištenih riječi i zaključio da postoji karakteristična krivulja nekog pisca (Olsson 2004: 11–12). Dužina riječi može se mjeriti brojem slova, ali i brojem slogova. Osim toga, pokazuje se da je dužina riječi vrlo važna mjera leksičkog bogatstva, a duže se riječi u pravilu češće pojavljuju u pisanim tekstovima nego u usmenom govoru.

Sredinom 80-ih godina 20. stoljeća pojavljuje se nova disciplina, računalna lingvistika, zahvaljujući pojavi jakih računala (Olsson 2004: 13). Zbog toga se u posljednjih 15-ak godina vrlo često za analizu stila i pripisivanje autorstva upotrebljavaju multivarijatne metode. Te su metode vrlo česte u prirodnim i društvenim znanostima, a odnedavno se koriste i u jezičnim analizama o autorstvu. Temelje se na pretpostavci da su frekvencije visokofrekventnih riječi dosljedne u pojedinih autora, kao i da su razlike između pojedinih autora dosljedne. To se može proširiti i na različite likove kojeg romana i različite idiolekte (Hoover 2003: 341, 358). Polazište pretpostavki da u različitih autora postoji različita frekvencija najfrekventnijih riječi jest stajalište da su te najfrekventnije funkcijske ili gramatičke riječi izvan svjesne kontrole pojedinaca, što potvrđuju rezultati neurofizioloških istraživanja brzine i lokacije procesiranja tih riječi u mozgu, koji pokazuju kako nakon desete godine života dolazi do automatskog procesiranja funkcijskih riječi. Tako je analizom tih duboko usađenih stilističkih navika pojedinaca moguće doprijeti do njihova tzv. 'otiska riječi', koji se uspoređuje s 'otiskom prsta' u forenzici, jer ovisi o osobnosti autora, a ne o temi o kojoj se govori ili o kutu gledanja (Hoover 2001: 422). Na taj način možemo reći da je riječ o stilu autora odnosno o 'stilskom otisku' određene osobe ili njezinom lingvističkom otisku (Olsson 2004: 31). Kod pojedinog stila sva stilska obilježja zajedno tada kreiraju 'stilistički profil', iz kojega se onda može otkriti 'stilistički otisak'. Takav pristup naglašava i poznata Buffonova definicija stila: „stil, to je sam čovjek" (Molinie 2002: 25), kao i de Morganova postavka da se manje razlikuju dva različita teksta istog autora nego tekstovi dvaju autora o jednoj te istoj temi (Crystal 1997: 68). Iako se i bogatstvo vokabulara koristi u analizi stilističkog obrasca kao pokazatelj distinktivnog obilježja autora, Hoover pokazuje da je frekventnost riječi pouzdaniji pokazatelj. Npr. čak pet najfrekventnijih riječi to razlikuje, a predstavljaju tek 20 % ukupnog broja riječi (Hoover 2002: 157). Na razini svjesnijeg izbora stila od frekvencija riječi stoji i dužina rečenice, koja također može biti pokazatelj autorstva, ali nije nužno uspješan stilistički pokazatelj za sve autore (Mannion i Dixon 2004: 507).

Još je krajem 19. stoljeća, i mnogo prije uporabe računala za analizu tekstova, Lucius Adelno Sherman uočio da su pojedini autori tekstova dosljedni u prosječnoj dužini rečenica koje se upotrebljavaju te je upravo ta dosljednost postala osnova za korištenje uzoraka tekstova za procjenu čitkosti, a ne cijelih tekstova. Sherman se prvi koristio statističkom analizom u svrhu analize čitkosti zaključivši da je pisani tekst moguće statistički analizirati, da kraće rečenice i konkretni pojmovi povećavaju čitkost teksta, da je usmeni govor učinkovitiji od pisanog te da tijekom vremena pisani govor postaje učinkovitiji što više nalikuje na usmeni govor. Njegovo je važno otkriće i opažanje da tijekom vremena rečenice ne samo da postaju kraće, nego i jednostavnije i manje apstraktne, što je posljedica procesa korištenja jasnih i jakih rečenica u usmenom govoru kroz tisućljeća radi usavršavanja učinkovitog komuniciranja (DuBay 2004: 10).

Dvadesetih godina 20. stoljeća nastaju znanstveni alati za proučavanje i objektivno mjerenje edukacijskih problema. Jedan od takvih alata opsežan je popis riječi engleskog jezika po učestalosti, psihologa Edwarda L. Thorndikea, gdje se uočava da su češće riječi poznatije i jednostavnije, a odrastajući i učeći izgrađujemo naš vokabular i lakše svladavamo duže i složenije rečenice (DuBay 2004: 12). Do pojave računala za procjenu težine tekstova za čitanje najčešće su se upotrebljavali popisi riječi po učestalosti, pa je tako Thorndikeovo istraživanje postalo osnova za razvoj prvih formula čitkosti.

U najnovije vrijeme za usavršavanje procjene težine teksta koriste se različiti pristupi kojima se procjenjuje jednostavnost čitanja, testiraju hipoteze o procesima mišljenja te izrađuju smjernice za izradu teksta. Cilj tih pristupa je utvrditi čimbenike u tekstu koji utječu na učenje i pamćenje, a njima se bave stručnjaci iz područja čitanja, retorike, lingvistike, psihologije i umjetne inteligencije. Npr. kohezijska analiza nudi više informacija o vokabularu i sintaktičkoj složenosti teksta nego što to mogu formule čitkosti (Zakaluk i Samuels 1988: 98–106). Autori jednog rada o kohezijskoj analizi koriste suvremene tehnike računalne lingvistike i obrade diskursa te izrađuju vlastiti računalni instrument kojim analiziraju više od 200 obilježja kohezije, jezika i čitkosti (Graesser i dr. 2004). Međutim, iako je analiza automatizirana i time pojednostavljena za korisnika, pretpostavka je tekst za unos napisan na engleskom jeziku. Testovi razumijevanja, kao npr. test Cloze, shvaćaju čitanje isključivo kao međudjelovanje teksta i čitatelja (Zakaluk i Samuels 1988: 107–109), ali istraživanja također pokazuju da formule čitkosti dobro koreliraju s testovima razumijevanja (DuBay 2004), kao što to može potvrditi istraživanje provedeno u Hrvatskoj u kojem je taj instrument korišten za testiranje razumijevanja pacijena-

ta (Kušec 2004). Složenost vokabulara pojedini autori proučavaju tehnikama strojnog učenja te nude klasifikacijske algoritme temeljene na Bayesovim teoremima za automatiziranu procjenu težine dokumenta (Leroy i dr. 2008). Drugi autori kombiniraju strojno učenje i sofisticirane mogućnosti tehnike obrade prirodnog jezika, ali analizom korpusa tekstova na francuskom jeziku zaključuju da, iako nova formula ima bolje izvedbene mogućnosti od obične tradicionalne formule čitkosti i nudi informativnije rezultate, ona nije povećala snagu dobivenih informacija, te autori ipak preporučuju kombinaciju tradicionalnih i novih formula (Francois i Miltsakaki 2012).

1.3. Formule čitkosti

Formula čitkosti zapravo je matematička jednadžba dobivena regresijskom analizom, a predstavlja procjenu težine odnosno kompleksnosti teksta u prvome redu za čitanje, ali i za razumijevanje. Razina čitkosti izražava se ili brojem koji predstavlja težinu samog teksta ili brojem koji predstavlja stupanj obrazovanja koji je potreban da se razumije tekst, uz vrlo visoku preciznost predviđanja, ili opisno na ljestvici od npr. ‘vrlo lako’ do ‘vrlo teško’.

Još 1920-ih godina pedagozi su otkrili način kako korištenjem težine vokabulara i dužine rečenice predvidjeti razinu težine teksta. Tu su metodu uklopili u formulu koja pokazuje čitkost teksta, a njihov se uspjeh potvrdio primjenom različitih formula sve do današnjih dana. Velik napredak u razvijanju formula čitkosti ostvaren je 1950-ih godina, kada se razvijaju nove formule, pa se danas u literaturi spominje upotreba više od 40 različitih formula čitkosti (Doak, Doak i Root 1996: 44), iako se još 1980-ih godina spominjalo postojanje čak više od 200 objavljenih formula (DuBay 2004: 19). Formule čitkosti nude prilično točan stupanj težine teksta, dobro koreliraju međusobno, iako razlike mogu biti 1 – 2 stupnja među formulama, a isto tako pojedini autori naglašavaju da se i za dobivenu vrijednost indeksa uzme u obzir +/- 1 – 1,5 stupnja pogreške (Osborne 2005: 18). Najčešće validirane formule su one na engleskom jeziku, a postoje modifikacije formula za ostale jezike, kao npr. njemački, francuski, nizozemski, danski, kineski, ruski, švedski, vijetnamski, korejski, hindu, hebrejski itd. (Zakaluk i Samuels 1988: 46–76). Među najčešće spominjanim i korištenim formulama čitkosti mogu se izdvojiti: *Flesch Reading Ease*, *Flesch–Kincaid Grade Level Index*, *SMOG*, *Fog Test*, *Fry Formula*, *Dale–Chall Formula* itd.

Najčešće je korištena računalna formula čitkosti *Flesch Reading Ease* (FRE), nazvana po autoru Rudolfu Fleschu, američkom državljaninu austrijskog po-

rijekla, koji je u ozračju demokratizacije obrazovanja u SAD-u izradio formulu koja je trebala pomoći pri određivanju težine udžbenika za djecu kako bi se djeci olakšalo učenje. Sam autor svoje područje istraživanja naziva 'znanstvena retorika', a svoju formulu jednostavnom mjerom, koju ipak treba smatrati grubom procjenom težine teksta (Flesch 1949: xi). Osim u području obrazovanja, ta je formula doživjela velik uspjeh u području novinarstva, za koje Flesch nudi pregršt savjeta o umijeću čitkog pisanja, s temeljnom idejom da treba pisati onako kako se govori (Flesch 1949: x) kako bi se privuklo šire čitateljstvo. Treba napomenuti da je Flesch svoju prvu formulu čitkosti objavio 1943. godine u svojoj disertaciji *Marks of Readable Writing* o osobitostima čitkog stila, a poslije ju je modificirao uzevši u obzir vrlo visoku čitkost direktnoga govora u pisanim materijalima te izostavivši složene pokazatelje osobnosti, kao npr. vlastita imena i posvojne zamjenice, te izostavivši varijablu sufiksa, zbog kojih je prvotna formula nailazila na kritike. Revidirana je formula pojednostavnjena te je postala najkorištenija formula čitkosti izvan obrazovnog sustava (Zakaluk i Samuels 1988: 20). Formula se temelji na standardnim tekstovima za testiranje čitanja autora McCalla i Crabbsa, uz 75-postotni stupanj razumijevanja. Svedena je na dvije varijable, prosječnu dužinu riječi i prosječnu dužinu rečenice, te glasi (Zakaluk i Samuels 1988: 20):

FRE = 206,835 – 0,846 wl – 1,015 sl

gdje je *wl* dužina riječi izražena brojem slogova (engl. *word length*), a *sl* dužina rečenice izražena brojem riječi (engl. *sentence length*).

Sam autor daje opisno objašnjenje kako izračunati broj slogova, jer je zapravo riječ o broju slogova na 100 riječi, tj. kaže da se ukupan broj slogova podijeli s ukupnim brojem riječi i pomnoži sa 100 (Flesch 1949: 214). Takva se formula danas nalazi u računalnim programima za obradu teksta, pa se čitkost određenog teksta jednostavno može pogledati pod statistikom čitkosti (engl. *readability statistics*) i ne treba se posebno računati, što je dodatno dovelo do popularnosti te formule.

Čitkost pisanih tekstova tom se formulom izražava na ljestvici od 0 do 100, gdje se pojedina kategorija težine objašnjava opisno, te je 0 indeks najveće težine teksta, a 100 indeks jednostavnosti teksta odnosno lakoće čitanja i razumijevanja. Grafički prikazano za svaku kategoriju, u Tablici 1 prikazan je i stupanj obrazovanja koji osoba mora imati da bi razumjela tekst te postotak odrasle populacije koja bi razumjela tekstove određene težine, prema procjeni samog autora formule, koji ujedno napominje da indeks 100 ima značenje težine teksta koji je razumljiv osobama koje su završile četiri razreda, tj. koje su rječnikom popisa stanovništva 'funkcionalno pismene' (Flesch 1949: 225).

Tablica 1: Ljestvica indeksa po formuli FRE (Flesch 1949: 149)

FRE	opis stila	procjena završenog stupnja obrazovanja	procjena udjela odrasle populacije SAD-a (%)
90 – 100	vrlo lako	4 razreda	93
80 – 90	lako	5 razreda	91
70 – 80	donekle lako	6 razreda	88
60 – 70	standardno	7 ili 8 razreda	83
50 – 60	donekle teško	nešto srednje škole	54
30 – 50	teško	srednja ili viša škola	33
0 – 30	vrlo teško	viša škola	4,5

Radi što jednostavnijeg određivanja težine teksta, Flesch je osim formule ponudio i grafički oblik, nomogram, koji se sastoji od ukupno triju stupaca, pri čemu jedan stupac sadržava brojčanu ljestvicu ukupnog broja riječi u rečenici, drugi stupac sadržava brojčanu ljestvicu broja slogova na 100 riječi, a središnji stupac nudi rezultat formule izražen indeksom čitkosti i opisom stila. Nomogram se nalazi na unutrašnjim koricama knjige *The Art of Readable Writing* (Flesch 1949) i zaštićen je autorskim pravima.

Fleschova formula korištena je kao temelj za formulu čitkosti za neke jezike izvan engleskog govornog područja, npr. za talijanski, nizozemski, francuski, španjolski (Zakaluk i Samuels 1988: 46–76), grčki (Kondilis i dr. 2010: 547–552), slovenski jezik (Kaesnik i Kline 2011: 33–40) itd., gdje su modifikacije formule bile nužne zbog uočenih razlika u dužini riječi i dužini rečenica u usporedbi s engleskim jezikom.

Općenito, autori formula čitkosti naročito naglašavaju da se one nipošto ne smiju koristiti mehanički, kao iznimno točne matematičke vrijednosti, ili pri pisanju tekstova za način pisanja ‘za formulu’, nego da su formule čitkosti grubi pokazatelji za procjenu težine teksta, koje su ujedno brza, jednostavna i jeftina metoda (DuBay 2004: 19). Protivnici korištenja formula čitkosti za procjenu težine teksta vrlo često zaboravljaju tu preporuku, kao i određene zakonitosti u jeziku koje se mogu dokazati statističkom analizom tekstova, te često zamjeraju da se formule oslanjaju ‘samo’ na površinska obilježja teksta, pa prema tome ne mogu biti pouzdani pokazatelji. Koliko su dužina riječi i dužina rečenica tek površinska obilježja ili zapravo izraz i pokazatelj kompleksnijih odnosa u jeziku koji utječu na čitanje i razumijevanje, najbolje pokazuju uvodno opisani rezultati statističkih istraživanja jezika.

2. Analiza korpusa tekstova na engleskom i hrvatskom jeziku

Kontrastivna analiza engleskog i hrvatskog jezika, čiji su rezultati prikazani pojedinačno u tablicama 2 – 5, a zbirno u tablici 6, temelji se na prikupljenom kor-

pusu tekstova za potrebe šireg istraživanja razvoja formula čitkosti za hrvatski jezik. Korpus se sastoji od gotovo 100 000 riječi, a sadržava tekstove na engleskom jeziku i njihove prijevode na hrvatskom jeziku objavljene nakon 1995. godine. Ukupno je 90 odlomaka s po 30 rečenica na svakom jeziku, a tekstovi su izabrani iz publikacija očekivano različitih težina kako bi se prikupio što širi raspon indeksa čitkosti. Publikacije su podijeljene u četiri vrste uzoraka te sadržavaju sljedeće tekstove: 1) odlomci 1 – 33 iz književnih djela autora J. K. Rowling, Robina Cooka, Dana Browna, Stephena Kinga, Michaela Crichtona, Deana Koontza, Davida Lodgea i Johna Grishama; 2) odlomci 34 – 57 iz reportaža časopisa *SETimes* (*Southeast European Times*) dostupnog na internetskim stranicama (www.setimes.com); odlomci 58 – 75 iz popularno-znanstvenih djela autora Bryana Sykesa, Richarda Dawkinsa, Billa Brysona, Stevea Jonesa, Mishe Glennyja i Stephena Hawkinga; 4) odlomci 76 – 90 iz znanstvenih radova objavljenih u medicinskom časopisu *JAMA* (*Journal of the American Medical Association*).

Kao što je uobičajeno u istraživanjima čitkosti, odlomci su uzeti s početka, iz sredine i s kraja dužih publikacija, s time da je izostavljena prva rečenica s početka i posljednja rečenica s kraja. Reportaže, koje su kraći tekstovi, analizirane su u cijelosti, s time da je izostavljena naizmjenično prva odnosno posljednja rečenica. Također su izostavljeni naslovi i/ili podnaslovi, a kako je riječ o prijevodnim tekstovima, u slučajevima gdje prijevod rečenica nije bio u omjeru 1 : 1, te su rečenice izostavljene. Kratice i brojevi prilagođeni su za analizu dužine riječi tako što su izraženi kao puni nazivi kratica, odnosno brojevi su navedeni slovima.

Ukupan broj riječi i prosječna dužina rečenica izračunani su računalnim programom za obradu teksta (Word for Windows XP), a ukupan broj slogova i prosječna dužina riječi izražena u slogovima izračunani su programom za izračunavanje čitkosti kojem se besplatno može pristupiti na internetskoj adresi www.wordscount.info. Taj se program pokazao najpouzdanijim prilikom usporedbe s drugim programima i ručnim brojanjem izvornih govornika hrvatskog i engleskog jezika. Testiranje pouzdanosti programa učinjeno je na probnom uzorku od deset odlomaka, a rezultati su prikazani u disertaciji autorice rada (Brangan 2011: 57). Sama formula čitkosti FRE izračunana je korištenjem računalnog programa (Excel for Windows XP) prema formuli koja je navedena u uvodnom dijelu ovog rada. Frekvencije, prosječne i srednje vrijednosti te rasponi indeksa izračunani su statističkim programom za analizu podataka SPSS.

U daljnjem tekstu u tablicama 2 – 5 prikazani su izračuni ukupnog broja riječi i slogova, prosječne dužine riječi i rečenica te indeksi čitkosti po formuli FRE zasebno za engleski i hrvatski jezik te zasebno za svaku vrstu uzorka tekstova.

Tablica 2: Ukupan broj riječi i slogova, prosječna dužina riječi i rečenica te indeks formule čitkosti FRE za odlomke na engleskom i hrvatskom jeziku – uzorak 1

odlomak br.	ukupan broj riječi		ukupan broj slogova		prosječna dužina riječi u slogovima		prosječna dužina rečenica		indeks FRE	
	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.
1	487	460	673	954	1,3819	2,0739	16,2	15,3	73,4	15,8
2	401	369	540	696	1,3466	1,8862	13,4	12,3	79,3	34,8
3	333	324	448	634	1,3453	1,9568	11,1	10,8	81,8	30,3
4	451	389	677	814	1,5011	2,0925	15,0	13,0	64,6	16,6
5	279	233	371	435	1,3297	1,8670	9,3	7,8	84,9	41,0
6	259	219	368	459	1,4208	2,0959	8,6	7,3	77,9	22,1
7	416	373	638	796	1,5337	2,1340	13,9	12,4	63,0	13,7
8	494	444	748	927	1,5142	2,0878	16,5	14,8	62,0	15,2
9	323	288	503	635	1,5573	2,2049	10,8	9,6	64,2	10,6
10	466	390	748	871	1,6052	2,2333	15,5	13,0	55,3	4,7
11	299	255	399	510	1,3344	2,0000	10,0	8,5	83,8	29,0
12	252	216	349	423	1,3849	1,9583	8,4	7,2	81,1	33,9
13	200	191	276	383	1,3800	2,0052	6,7	6,4	83,3	30,7
14	156	140	225	271	1,4423	1,9357	5,2	4,7	79,5	38,3
15	249	244	371	494	1,4900	2,0246	8,3	8,1	72,4	27,3
16	183	164	235	330	1,2842	2,0122	6,1	5,5	92,0	31,1
17	220	220	281	439	1,2773	1,9955	7,3	7,3	91,3	30,6
18	163	171	220	311	1,3497	1,8187	5,4	5,7	87,1	47,2
19	590	481	776	906	1,3153	1,8836	19,7	16,0	75,6	31,2
20	329	270	445	516	1,3526	1,9111	11,0	9,0	81,3	36,0
21	404	325	522	627	1,2921	1,9292	13,5	10,8	83,9	32,6
22	272	240	367	482	1,3493	2,0083	9,1	8,0	83,5	28,8
23	308	285	406	560	1,3182	1,9649	10,3	9,5	84,9	31,0
24	278	250	360	499	1,2950	1,9960	9,3	8,3	87,9	29,5
25	442	395	640	789	1,4480	1,9975	14,7	13,2	69,4	24,5
26	581	551	821	1118	1,4131	2,0290	19,4	18,4	67,6	16,5
27	360	344	452	635	1,2556	1,8459	12,0	11,5	88,4	39,0
28	602	564	928	1256	1,5415	2,2270	20,1	18,8	56,1	-0,6
29	632	559	895	1094	1,4161	1,9571	21,1	18,6	65,6	22,4
30	625	553	937	1183	1,4992	2,1392	20,8	18,4	58,9	7,1
31	356	380	498	748	1,3989	1,9684	11,9	12,7	76,4	27,4
32	265	245	324	426	1,2226	1,7388	8,8	8,2	94,4	51,4
33	174	168	228	313	1,3103	1,8631	5,8	5,6	90,1	43,5

Tablica 3: Ukupan broj riječi i slogova, prosječna dužina riječi i rečenica te indeks formule čitkosti FRE za odlomke na engleskom i hrvatskom jeziku – uzorak 2

odlomak br.	ukupan broj riječi		ukupan broj slogova		prosječna dužina riječi u slogovima		prosječna dužina rečenica		indeks FRE	
	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.
34	530	457	840	965	1,5849	2,1116	17,7	15,2	54,8	12,7
35	513	489	815	1093	1,5887	2,2352	17,1	16,3	55,1	1,2
36	716	594	1246	1441	1,7402	2,4259	23,9	19,8	35,4	-18,5
37	528	444	720	875	1,3636	1,9707	17,6	14,8	73,6	25,1
38	628	529	977	1146	1,5557	2,1664	20,9	17,6	54,0	5,7
39	659	597	1141	1358	1,7314	2,2747	22,0	19,9	38,1	-5,8
40	859	810	1550	1839	1,8044	2,2704	28,6	27,0	25,1	-12,6
41	620	547	1075	1324	1,7339	2,4205	20,7	18,2	39,2	-16,4
42	605	560	1096	1272	1,8116	2,2714	20,2	18,7	33,1	-4,3
43	538	483	937	1074	1,7416	2,2236	17,9	16,1	41,3	2,4
44	714	643	1342	1551	1,8796	2,4121	23,8	21,4	23,7	-19,0
45	710	649	1283	1496	1,8070	2,3051	23,7	21,6	29,9	-10,1
46	716	638	1200	1479	1,6760	2,3182	23,9	21,3	40,8	-10,9
47	600	506	1010	1183	1,6833	2,3379	20,0	16,9	44,1	-8,1
48	740	683	1355	1602	1,8311	2,3455	24,7	22,8	26,9	-14,7
49	538	486	897	1144	1,6673	2,3539	17,9	16,2	47,6	-8,7
50	729	642	1344	1495	1,8436	2,3287	24,3	21,4	26,2	-11,9
51	717	650	1258	1541	1,7545	2,3708	23,9	21,7	34,1	-15,7
52	624	571	1136	1330	1,8205	2,3292	20,8	19,0	31,7	-9,5
53	911	862	1492	1926	1,6378	2,2343	30,4	28,7	37,5	-11,4
54	666	555	1162	1331	1,7447	2,3982	22,2	18,5	36,7	-14,8
55	660	596	1159	1422	1,7561	2,3859	22,0	19,9	35,9	-15,2
56	871	817	1635	1960	1,8772	2,3990	29,0	27,2	18,6	-23,8
57	797	680	1428	1568	1,7917	2,3059	26,6	22,7	28,3	-11,2

Tablica 4: Ukupan broj riječi i slogova, prosječna dužina riječi i rečenica te indeks formule čitkosti FRE za odlomke na engleskom i hrvatskom jeziku – uzorak 3

odlomak br.	ukupan broj riječi		ukupan broj slogova		prosječna dužina riječi u slogovima		prosječna duži-na rečenica		indeks FRE	
	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.
58	736	683	1122	1563	1,5245	2,2884	24,5	22,8	53,0	-9,9
59	649	541	1005	1222	1,5485	2,2588	21,6	18,0	53,9	-2,6
60	600	529	983	1204	1,6383	2,2760	20,0	17,6	47,9	-3,6

61	621	546	977	1170	1,5733	2,1429	20,7	18,2	52,7	7,1
62	634	572	1018	1214	1,6057	2,1224	21,1	19,1	49,5	7,9
63	625	522	1087	1222	1,7392	2,3410	20,8	17,4	38,6	-8,9
64	519	426	762	881	1,4682	2,0681	17,3	14,2	65,1	17,5
65	626	557	986	1185	1,5751	2,1275	20,9	18,6	52,4	8,0
66	600	528	876	1100	1,4600	2,0833	20,0	17,6	63,0	12,7
67	677	600	972	1275	1,4357	2,1250	22,6	20,0	62,5	6,8
68	576	492	891	1079	1,5469	2,1931	19,2	16,4	56,5	4,7
69	723	630	1109	1385	1,5339	2,1984	24,1	21,0	52,6	-0,5
70	807	719	1353	1640	1,6766	2,2809	26,9	24,0	37,7	-10,5
71	791	713	1340	1695	1,6941	2,3773	26,4	23,8	36,8	-18,4
72	865	827	1506	1982	1,7410	2,3966	28,8	27,6	30,3	-23,9
73	657	557	947	1142	1,4414	2,0503	21,9	18,6	62,7	14,5
74	834	732	1287	1589	1,5432	2,1708	27,8	24,4	48,1	-1,6
75	656	632	1143	1434	1,7424	2,2690	21,9	21,1	37,2	-6,5

Tablica 5: Ukupan broj riječi i slogova, prosječna dužina riječi i rečenica te indeks formule čitkosti FRE za odlomke na engleskom i hrvatskom jeziku – uzorak 4

odlomak br.	ukupan broj riječi		ukupan broj slogova		prosječna dužina riječi u slogovima		prosječna dužina rečenica		indeks FRE	
	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.	engl.	hrv.
76	683	717	1325	1805	1,9400	2,5174	22,8	23,9	19,6	-30,4
77	715	683	1331	1681	1,8615	2,4612	23,8	22,8	25,2	-24,5
78	762	698	1384	1701	1,8163	2,4370	25,4	23,3	27,4	-22,9
79	681	759	1309	1803	1,9222	2,3755	22,7	25,3	21,2	-19,8
80	680	714	1271	1653	1,8691	2,3151	22,7	23,8	25,7	-13,2
81	924	941	1751	2223	1,8950	2,3624	30,8	31,4	15,3	-24,9
82	678	686	1396	1854	2,0590	2,7026	22,6	22,9	9,7	-45,0
83	796	801	1405	2027	1,7651	2,5306	26,5	26,7	30,6	-34,4
84	650	643	1308	1722	2,0123	2,6781	21,7	21,4	14,6	-41,5
85	769	642	1530	1713	1,9896	2,6682	25,6	21,4	12,5	-40,6
86	715	639	1343	1677	1,8783	2,6244	23,8	21,3	23,7	-36,8
87	864	727	1658	1934	1,9190	2,6602	28,8	24,2	15,3	-42,8
88	634	597	1182	1604	1,8644	2,6868	21,1	19,9	27,7	-40,7
89	764	716	1345	1805	1,7605	2,5209	25,5	23,9	32,1	-30,7
90	611	576	1103	1556	1,8052	2,7014	20,4	19,2	33,4	-41,2

Iz tablica 2 – 5 vidljivo je da indeksi čitkosti po formuli FRE pokazuju očekivanu veću čitkost jednostavnijih književnih tekstova, a manju čitkost složeni-

jih, popularno-znanstvenih i znanstvenih tekstova. To je naravno rezultat uporabe dužih riječi i dužih rečenica u složenijim vrstama tekstova, kako u engleskom, tako i u hrvatskom jeziku. Može se uočiti ne samo da je prosječna dužina riječi i rečenice kraća u književnim tekstovima, nego i da je ukupan broj riječi i slogova najmanji kod jednostavnih književnih tekstova (odlomci br. 1 – 33), da se povećava preko reportaža (odlomci br. 34 – 57) i znanstveno-popularnih tekstova (odlomci br. 58 – 75) te da je najveći kod znanstvenih članaka (odlomci 76 – 90). Konkretno, raspon za dužinu rečenice književnih tekstova iznosi 5,2 – 21,1 za engleski jezik te 4,7 – 18,2 za hrvatski jezik, a za znanstvene članke iznosi 20,4 – 30,8 za engleski jezik te 19,2 – 31,4 za hrvatski jezik, što upućuje na to da su u promatranom uzorku prijevodnih tekstova rečenice na hrvatskom jeziku bile kraće. Jednostavan stil pisanja u promatranim književnim tekstovima uočen je kod autora Dana Browna (odlomci br. 13, 14, 16 – 18), s rasponom od 5,2 – 7,3 riječi po rečenici u engleskom jeziku, a i prijevod na hrvatski odražava taj stil pisanja rasponom od 4,7 – 7,3 riječi po rečenici. Romani Dana Browna odražavaju razgovorni stil, koji za veću čitkost i bolje razumijevanje zagovaraju i pobornici čitkog pisanja, a naročito Rudolf Flesch (Flesch 1949: x), koji navode da treba pisati kao što se govori, što dovodi i do veće čitkosti izražene indeksom čitkosti izračunanoj po formuli.

Što se tiče ukupnog broja slogova u promatranu uzorku, rezultati su pokazali, iako ovdje nisu grafički prikazani, da je u engleskim tekstovima on iznosio otprilike jednu trećinu ukupnog broja slova, što odgovara navodima u literaturi (McLaughlin 1974: 377), a u hrvatskim tekstovima otprilike 40 % ukupnog broja slova po odlomku. Ta se razlika ogleda i u prosječnoj dužini sloga izraženoj brojem slova, koja je u promatranom uzorku tekstova iznosila 3 u engleskom i 2,4 u hrvatskom jeziku. Dakle, u hrvatskim tekstovima više je slogova, ali su oni kraći od slogova u engleskom jeziku. Jednaka težina prijevodnih ekvivalenata uz istodobni veći broj slogova u hrvatskom u odnosu na engleski jezik može se objasniti s jedne strane ortografskom transparentnošću hrvatskog jezika. Ortografska transparentnost ubrzava čitanje jer dolazi do bržeg prepoznavanja riječi i točnijeg čitanja. Tako npr. test za brzu procjenu medicinske pismenosti, koji sadržava popis izoliranih medicinskih riječi uzlazne težine na engleskom jeziku, nije moguće pukim prijevodom riječi koristiti u istu svrhu na jeziku s ortografskom transparentnošću, kao što je španjolski (Nielsen-Bohlman i dr. 2004: 48). S druge strane, postojanje gramatičkih nastavaka u hrvatskom jeziku koji općenito povećavaju dužinu riječi, povećava i redundantnost u jeziku, što znači da čitatelj tekstova na hrvatskom jeziku zna što može očekivati u riječima koje slijede, a gramatičke mu kategorije olakšavaju razumijevanje pročitanog. O tome govore i rezultati istraživanja iz područja kognitivne

obrade informacija, gdje se pokazalo da npr. vrijeme reakcije ispitanika ovisi o količini informacije sadržane u određenom morfemu (Kostić 1990: 180). To se podudara i s tvrdnjom (McLaughlin 1974: 380) da je težina teksta u semantičkom i sintaktičnom smislu povezana s ograničenjima pohrane u kratkotrajnom pamćenju, kao i s novijim istraživanjima da je za tečno čitanje neophodno automatizirano prepoznavanje riječi kako bi se oslobodio kapacitet pažnje za razumijevanje (Alderson 2000: 57).

Za korpus prijevodnih tekstova prikupljenih za ovo istraživanje može se reći da publicirani prijevodi engleskih tekstova odražavaju i težinu tekstova na izvornom jeziku te se mogu po težini smatrati ekvivalentima, a razlika u brojčanom indeksu čitkosti zahtijeva modifikaciju izvorne formule validirane za engleski jezik kako bi bila valjana za procjenu težine teksta na hrvatskom jeziku. Stoga je, prije samog prijedloga za modifikaciju formule, u Tablici 6 dan prikaz dobivenih vrijednosti indeksa prema formuli FRE za hrvatski jezik – kroz raspon, prosjek i srednju vrijednost, u usporedbi s opisnom i brojčanom ljestvicom za engleski jezik.

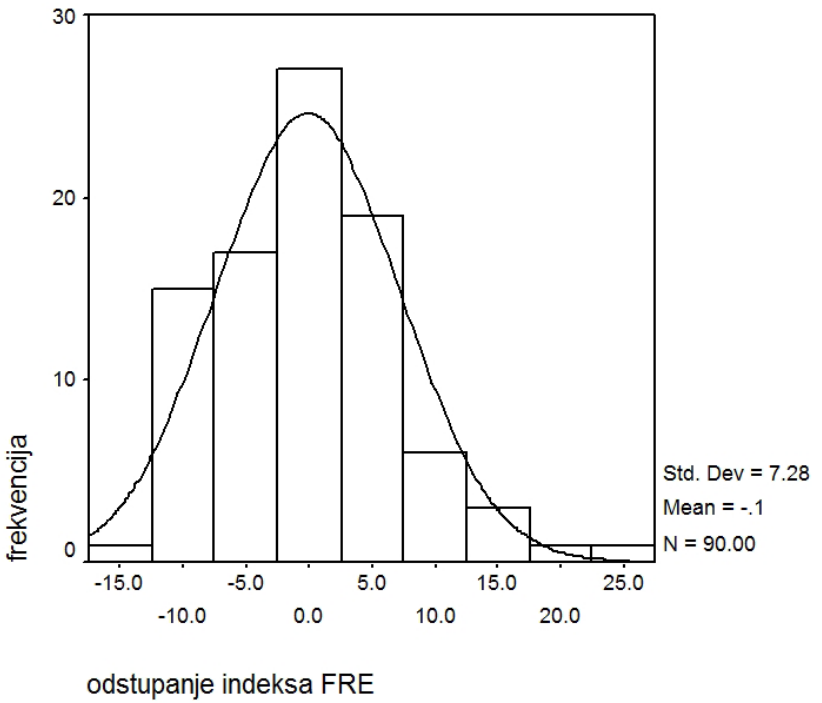
Tablica 6: Indeksi čitkosti odlomaka na hrvatskom jeziku, prema formuli FRE, na temelju opisne ljestvice izvorne formule za engleski jezik

FRE engleski			FRE hrvatski		
raspon	opis	broj odlomaka	raspon	prosjek	srednja vrijednost
90 – 100	vrlo lako	3	31 – 51	37,7	31,1
80 – 90	lako	12	29 – 47	34,1	31,8
70 – 80	donekle lako	7	16 – 35	26,2	27,3
60 – 70	standardno	11	7 – 25	15,6	15,2
50 – 60	donekle teško	12	-10 – 13	3,1	4,7
30 – 50	teško	22	-42 – 2	-13,3	-10,7
0 – 30	vrlo teško	18	-45 – (-11)	-26,5	-24,2

Iz Tablice 6 vidljivo je da se čitkost tekstova na hrvatskom jeziku smanjuje s povećanjem složenosti tekstova, ali se dobiveni rezultati ni u rasponu ni u prosječnim i srednjim vrijednostima ne poklapaju s opisnom ljestvicom definiranom za izvornu formulu FRE za engleski jezik.

Polazeći od indeksa čitkosti prikazanih za pojedinačne odlomke na engleskom i hrvatskom jeziku u tablicama 2 – 5, izračunana je razlika prosječnih vrijed-

nosti indeksa između ovih dvaju jezika, koja iznosi 50, odnosno prosječna vrijednost je za toliko manja u hrvatskom jeziku. Kako bi se dodatno provjerilo može li se izvorna formula modificirati za hrvatski jezik jednostavnim dodavanjem vrijednosti 50, izračunan je raspon odstupanja vrijednosti indeksa dobiven modificiranom formulom u odnosu na vrijednosti indeksa izvornog teksta koji služi kao ekvivalent težine. Odstupanje je izračunano za svaki od 90 odlomaka pojedinačno, a zbirni grafički prikaz na Slici 1 pokazuje prosjek odstupanja od -0,1, sa srednjom vrijednošću 0,0 te koeficijentom asimetrije 0,578.



Slika 1: Odstupanja vrijednosti indeksa FRE prema modificiranoj formuli za hrvatski jezik u odnosu na indeks FRE izvornih odlomaka

Na temelju pravilnosti histograma i uočene prosječne razlike u vrijednostima indeksa FRE između engleskog i hrvatskog jezika predlaže se modifikacija formule FRE za tekstove na hrvatskom jeziku kako slijedi:

postojeća formula FRE, koja za engleski jezik glasi:

FRE = 206,835 – 0,846 wl – 1,015 sl

modificira se za hrvatski jezik da glasi:

FRE = 206,835 – 0,846 wl – 1,015 sl + 50

gdje je *wl* dužina riječi izražena brojem slogova, a *sl* dužina rečenice izražena brojem riječi.

Pri tome se ljestvica indeksa za engleski jezik, kako je prikazana u Tablici 6, treba prilagoditi za dobivene indekse na hrvatskom jeziku kako bi se većina rezultata uklopila sa što manje pogrešaka u opisnu ljestvicu na sljedeći način:

80 – 100 = lako; 60 – 80 = standardno; 50 – 60 = donekle teško; te 0 – 50 = vrlo teško.

Modifikacija formule čitkosti FRE za hrvatski jezik napravljena je, dakle, kako na temelju usporedbe indeksa čitkosti u engleskom i hrvatskom jeziku, tako i uvidom u vrijednosti nužnih za pojedine formule, s ciljem što jednostavnije prilagodbe formule za hrvatski jezik, uzimajući u obzir da indeks čitkosti nije precizna vrijednost, nego samo grubi pokazatelj procjene težine teksta. Nadalje, valja spomenuti da se formula čitkosti FRE nalazi u ponekim računalnim programima za obradu teksta, pa nije potrebno pronalaziti posebne programe kojima bi se računali zasebno ukupni i prosječni brojevi slogova i riječi, a jednostavnim dodavanjem vrijednosti 50 za hrvatske tekstove moguće je na brz i učinkovit način dobiti grubu procjenu težine teksta u svakodnevnom radu.

Najbolju potvrdu vrijednosti modificirane formule dala bi praktična evaluacija bilo na nekom novom tekstu za koji je unaprijed poznata težina ili na ispitanicima koji bi pročitali tekst određene težine i zatim bi se testiralo njihovo razumijevanje teksta. Prema saznanjima autorice ovog rada u Hrvatskoj prije ovog istraživanja (Brangan 2011) nije napravljen prijedlog validirane formule čitkosti za hrvatski jezik, pa preostaje retroaktivno provjeriti valjanost formule na rezultatima razumijevanja teksta koji je korišten na uzorku od 75 pacijenata u Zagrebu pri istraživanju razlika profesionalne i laičke terminologije (Kušec 2004). Istraživanje je pokazalo da su ispitanici koji su razumjeli tekst na potpuno zadovoljavajućoj razini, odnosno ispunili test Cloze iznad 60 % točnosti, bili jedino ispitanici s visokim stupnjem obrazovanja te medicinske sestre sa srednjim obrazovanjem (Kušec 2004: 70). Primjena modificirane formule FRE za hrvatski jezik na taj tekst, dostupan autorici rada, daje rezultat čitkosti 42,7 i označava tekst koji je na donjoj granici kategorije „vrlo teško", što po izvornoj klasifikaciji autora formule FRE zahtijeva srednju ili višu školu. Testiranje valjanosti formule u ovakvoj praktičnoj evaluaciji, iako retrospektivnoj, daje potvrdu da je njome na sasvim zadovoljavajući način pružena gruba procjena težine teksta i utvrđena skupina populacije prema obrazovanju kojoj bi tekst bio primjeren po svojoj težini.

3. Zaključak

Kao objektivni pokazatelj težine teksta, koji ipak treba upotrebljavati kao grubu procjenu, može poslužiti primjena neke formule čitkosti. Formula čitkost FRE, koja je prikazana u ovom radu, dostupna je u standardnim računalnim programima za obradu teksta te se vrlo lako može upotrebljavati i za procjenu težine tekstova na hrvatskom jeziku. Nužna je, doduše, modifikacija izvorne formule validirane za engleski jezik zbog razlika između engleskog i hrvatskog jezika u dužini riječi i rečenica, kako pokazuje analiza prijevodnih tekstova u ovom istraživanju. Primjena modificirane formule za hrvatski jezik, u obliku koji se predlaže u ovom radu, olakšat će odabir pisanih materijala svima onima koji moraju prilagoditi korištenje pisanih materijala različitim ciljnim skupinama čitatelja bez obzira na to je li riječ o području obrazovanja, novinarstva, medicine, zakonodavstva i slično. Drugim riječima, primjena formule čitkosti imat će važnu ulogu svugdje gdje čitatelji pisanih materijala moraju na temelju pročitanog s razumijevanjem donositi odluke povezane s različitim aspektima njihova života.

Literatura:

Alderson, J. Charles. 2000. *Assessing reading*. Cambridge University Press. Cambridge.

Biber, Douglas; Conrad, Susan; Reppen, Randi. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge University Press. Cambridge.

Brangan, Sanja. 2011. *Razvoj formula čitkosti za zdravstvenu komunikaciju na hrvatskom jeziku*. Doktorski rad. Medicinski fakultet Sveučilišta u Zagrebu. Zagreb. 113 str. http://medlib.mef.hr/1414 (pristupljeno 25. kolovoza 2013.).

Brangan, Sanja. 2013. *Priprema pisanih materijala za pacijente – doprinos smanjenju zdravstvene nejednakosti u korištenju zdravstvene zaštite*. Medical Information Conference Croatia. Zagreb. http://ark.mef.hr/MICC/MICC9.htm (pristupljeno 25. kolovoza 2013.).

Crystal, David. 1997. *The Cambridge encyclopedia of language*. Cambridge University Press. Cambridge.

Doak, Cecilia C.; Doak, Leonard G.; Root, Jane H. 1996. *Teaching patients with low literacy skills*. J. B. Lippincott Company. Philadelphia.

Dubay, William H. 2004. *The principles of readability*. Impact Information. Costa Mesa, California.

Flesch, Rudolf. 1949. *The art of readable writing*. Harpers & Brothers Publishers. New York.

Francois, Thomas; Miltsakaki, Eleni. 2012. Do NLP and machine learning improve traditional readability formulas? *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations*. Association for Computational Linguistics. Montreal. 49–57. http://dl.acm.org/citation.cfm?id=2390925 (pristupljeno 15. kolovoza 2013.).

Graesser, Arthur C.; McNamara, Danielle S.; Louwerse, Max M.; Cai, Zhiqiang. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers* 36.193–202.

Hoover, David L. 2001. Statistical stylistics and authorship attribution: An empirical investigation. *Literary and Linguistic Computing* 16(4). 421–444.

Hoover, David L. 2002. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing* 17(2). 157–180.

Hoover, David L. 2003. Multivariate analysis and the the study of style variation. *Literary and Linguistic Computing* 18(4). 341–60.

Hudson, Richard A. 1996. *Sociolinguistics*. Cambridge University Press. Cambridge.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge University Press. Cambridge.

Kaesnik, Karin; Kline, Mihael. 2011. Analyzing readability of medicines information material in Slovenia. *Southern Medical Review* 4(2). 33–40.

Kondilis, Barbara K.; Akrivos, Patrick D.; Soteriades, Elpidoforos S.; Falagas, Matthew E. 2010. Readability levels of health pamphlets distributed in hospitals and health centres in Athens, Greece. *Public Health* 124. 547–552.

Kostić, Aleksandar. 1990. Količina informacije kao jedini determinator kognitivne obrade inflekcione morfologije. *SOL* 5. 169–185.

Kušec, Sanja. 2004. *Usklađivanje profesionalne i laičke terminologije u odnosu liječnik–pacijent.* Magistarski rad. Medicinski fakultet Sveučilišta u Zagrebu. Zagreb. 113 str.

Kušec, Sanja; Orešković, Stipe; Škegro, Mate; Korolija, Dragan; Bušić, Željko; Horžić, Matija. 2006. Improving comprehension of informed consent. *Patient Education and Counseling* 60. 294–300.

Kušec, Sanja. 2007. Jezični identitet u biomedicini i zdravstvu. *Zbornik radova Hrvatskog društva za primijenjenu lingvistiku* 3. Hrvatsko društvo za primijenjenu lingvitiku. Zagreb – Split. 19–328.

Leroy, Gondy; Miller, Trudi; Rosemblat, Graciela; Browne, Allen. 2008. A balanced approach to health information evaluation: A vocabulary–based naive Bayes classifier and readability formulas. *Journal of the American Society for Information Science and Technology* 59. 1409–1419.

Malmberg, Bertil. 1979. *Moderna lingvistika*. Slovo ljubve. Beograd.

Mannion, David; Dixon, Peter. 2004. Sentence-length and authorship attribution: The case of Oliver Goldsmith. *Literary and Linguistic Computing* 19(4). 497–508.

McLaughlin, Harry G. 1974. Temptations of the Flesch. *Instructional Science* 2. 367–384.

Moguš, Milan; Bratanić, Maja; Tadić, Marko. 1999. *Hrvatski čestotni rječnik*. Školska knjiga. Zagreb.

Molinie, Georges. 2002. *Stilistika*. CERES. Zagreb.

Nielsen-Bohlman i dr. 2004: *Health literacy: a prescription to end confusion*. 2004. Ur. Nielsen-Bohlman, Lynn; Panzer, Allison M.; Kindig, David A. Institute of Medicine – The National Academies Press. Washington.

Olsson, John. 2004. *Forensic Linguistics*. Continuum. London – New York.

Osborne, Helen. 2005. *Health literacy from A to Z: Practical ways to communicate your health message*. Jones and Bartlett Publishers. Sudbury.

Škiljan, Dubravko. 1988. Sloboda jezika. *SOL* 7. 69–78.

Tadić, Marko. 1997. Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. *Suvremena lingvistika* 43-44. 387–394.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Ex libris. Zagreb.

Zakaluk, Beverly L.; Samuels, Jay S. 1988. *Readability: Its past, present, and future*. International Reading Association. Newark.

Quantitative Assessment of Text Difficulty in Croatian Language

Abstract

This paper presents past research in the field of statistical text analysis, with special emphasis on development of readability formulas. Apart from the theoretical part on reading, comprehension, and readability, the paper also presents results of analysis of a corpus of English and Croatian texts. Finally, a readability formula for Croatian language is suggested, as modified from Flesch Reading Ease for English language, which could be used as an objective indicator for a rough assessment of text difficulty in Croatian language.

Ključne riječi: čitanje, čitkost, formula čitkosti FRE, statistička analiza teksta, korpusna lingvistika, prijevodni i paralelni korpusi

Key words: reading, readability, Flesch Reading Ease readability formula, statistical text analysis, corpus linguistics, translational and parallel corpora