UDK 81’23
81’324
Izvorni znanstveni rad
Rukopis primljen 16. IX. 2024.
Prihvaćen za tisak 10. IX. 2025.
https://doi.org/10.31724/rihjj.51.1.10

Marijan Palmović
Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu
Odsjek za logopediju
Laboratorij za psiholingvistička istraživanja
Borongajska cesta 83f, HR-10000 Zagreb
https://orcid.org/0000-0002-4587-4348
marijan.palmovic@erf.unizg.hr
Kristina Cergol
Učiteljski fakultet Sveučilišta u Zagrebu
Katedra za obrazovanje učitelja engleskog jezika
Savska cesta 77, HR-10000 Zagreb
https://orcid.org/0000-0002-5440-0312
kristina.cergol@ufzg.hr

ANALIZA PODATAKA DVOJEZIČNOG KORPUSA ČITANJA S MJERENJEM POKRETA OKA: USPOREDBA FREKVENCIONISTIČKOG I BAYESOVA PRISTUPA

Korpus čitanja nova je metodologija u psiholingvističkim istraživanjima, s pomoću koje se podatci o pokretima oka prikupljaju iz eksperimentalno nemanipuliranih podražaja, tj. tekstova. U radu se raspravlja o prikladnoj metodi statističke analize za tu vrstu istraživanja. Predlaže se model logističke regresije, frekvencionistički ili Bayesov, jer se takav model upotrebljava za predviđanje kategoričkih ciljnih varijabli, kakva je većina lingvistički relevantnih kategorija. Pokazana je superiornost takve analize u odnosu na prethodne analize u potvrđivanju hipoteze implicitne prozodije, a modeli su logističke regresije potvrdili da zaista nesvjesno pratimo naglasnu strukturu teksta, čak i pri čitanju u sebi. Analiza je provedena na paralelenom hrvatsko-engleskom korpusu čitanja. Nisu pronađene velike razlike između frekvencionističkog i Bayesova modela, iako je Bayesov model pokazao neke prednosti u interpretaciji rezultata.

1. Uvod

Eksperiment je tradicionalno temeljna metoda provjere hipoteza u psiholingvistici. Tipičan se eksperiment sastoji od vizualno ili slušno predstavljenih podražaja i mjernog uređaja koji mjeri vrijednosti zavisnih varijabla s obzirom na eksperimentalnu manipulaciju podražajima. Čest je prigovor takvim istraživanjima njihova slaba ekološka valjanost iako se pojednostavnjivanje situacije i manipulacija samo jednom varijablom ističu kao nedvosmislena prednost eksperimentalne metode u proučavanju automatskih procesa u jezičnoj obradi. S druge strane, tko zaista jezik upotrebljava tako da sjedi pred zaslonom računala dok mu se pred očima izmjenjuju nizovi slova, koji su katkad riječi, a katkad nisu? Mjerenje pokreta oka nije iznimka. Eksperimentalna kontrola podražaja često mora biti i stroža nego u, primjerice, mjerenju vremena reakcije jer na izmjerene vrijednosti mogu utjecati i mnoge ometajuće varijable vezane uz okulomotorne ili motoričke procese, pa se eksperimenti čine još manje prirodnima. Stoga se u posljednje vrijeme metoda mjerenja pokreta oka počinje upotrebljavati i drukčije: sudionici u istraživanju moraju pročitati tekstove koji nisu eksperimentalno manipulirani. Često su to novinski članci, kraće priče ili znanstveno-popularni tekstovi. Obično je riječ o desetak stranica teksta, ali katkad ih je i više. Uobičajeno je da ih pročita stotinjak sudionika. Takva metoda prikupljanja podataka dobila je nespretno ime: *prirodno čitanje* (engl. *natural reading*), a podatci se organiziraju u korpuse, *korpuse čitanja* (engl. *reading corpora*), koji sadržavaju uparene podatke o samom tekstu i podatke s uređaja za mjerenje pokreta oka. Ti su podatci najčešće sirovi, tj. sadržavaju podatak o koordinatama fokusa oka u svakom vremenskom odsječku koji uređaj mjeri ovisno o brzini uzorkovanja. Katkad takav korpus sadržava podatke o dvama „događajima" jer uređaj sirove podatke klasificira prema unaprijed zadanim kriterijima. To su sakade, tj. brzi pokreti oka u nekom smjeru i fiksacije, tj. razdoblje kad je oko relativno mirno. Tako organizirani podatci u pravilu su javno dostupni te su u formatu u kojem ih se može pročitati bez obzira na vrstu ili marku uređaja na kojem su prikupljeni. Prvi korpus čitanja bio je GECO (*Ghent EyeTracking Corpus*), nizozemsko-engleski dvojezični korpus u kojem su sudionici pročitali pola romana na nizozemskom, a drugu polovinu na engleskom (Cop i dr. 2016). Danas postoje slični dvojezični korpusi, npr. japansko-engleski TECO (*Tsukuba Eye-tracking Corpus*, Nahatame i dr. 2024), ali i jednojezični korpusi ili korpusi koji imaju neku

posebnu svrhu (Hollenstein, Barret i Björnsdóttir 2022, Scarton i Specia 2016). Velik broj dostupnih podataka osigurava statističku snagu zaključaka, a analiza korpusa danas više nije samo specifična metodologija, nego jezični resurs koji stoji na raspolaganju svim lingvistima.

Korpusi čitanja predstavljaju novu dimenziju psiholingvističkog korpusnog istraživanja. Njima se prikupljaju podatci o jezičnom ponašanju izmjerenom na bihevioralnoj ili fiziološkoj varijabli, kao što je, na primjer, širina zjenice. Tako se mogu dobiti podatci o tome koliko je fiksirana svaka riječ u tekstu za pojedinog sudionika koji ju je pročitao ili o tome kolike su amplitude sakada za svakog sudionika, o tome koliko su i koje riječi govornici preskakali i sl. Takvi korpusi često sadržavaju i metapodatke o jezičnom ili socioekonomskom statusu sudionika. U novije se vrijeme uključuju i drugi psihometrijski podatci, na primjer opseg radnog pamćenja ili kognitivna fleksibilnost (npr. u korpusu koji se prikuplja u sklopu COST-ova projekta *MultiplEye*, https://www.multipleye.eu). Te značajke korpus čitanja kao metodu smještaju između eksperimentalne i korpusne lingvistike.

Priručnici korpusne lingvistike ne daju jasne upute kako takve korpuse analizirati. Jedino udžbenik korpusne lingvistike McEneryja i Hardieja (2011.) spominje psiholingvističku upotrebu korpusa te istraživanja mjerenjem pokreta oka i čitanjem vlastitim tempom, ali samo načelno spominjući korist koju psiholingvist može imati od korpusa u odabiru eksperimentalnih čestica. Priručnici kvantitativnih metoda u korpusnoj lingvistici ne spominju psiholingvističke korpuse (cf. Gries 2009, Brezina 2018). I konačno, statistički priručnici namijenjeni lingvistima ne spominju statističke metode u psiholingvistici, ali daju dobar pregled statističkih metoda koje se mogu upotrijebiti u analizi lingvističkih podataka općenito (Gries 2013, Winter 2020).

U žarištu je ovog rada rasprava o statističkoj metodi pogodnoj za analizu korpusa čitanja. Korpus koji će se tako analizirati paralelni je hrvatsko-engleski korpus čitanja prikupljen radi provjere hipoteze implicitne prozodije (Fodor 2002). Njome se tvrdi da pri čitanju u sebi govornik slijedi prozodijsku strukturu teksta. Ideja o „unutarnjem glasu", koji „čujemo" kao fonetsku realizaciju teksta koji čitamo, stara je više od stoljeća (Huey 1908/1968). Empirijsku potvrdu hipoteze implicitne prozodije do sada čine eksperimenti s mjerenjem neuromišićne veze na glavi i vratu tijekom čitanja u sebi (Orepić 2020) ili eksperimenti u ko-

jima je „prisilna" vokalizacija ometala razumijevanje čitanja, ali ne i slušanja, s obzirom na to da je interferirala sa subvokalizacijom koju prati čitanje u sebi. Od sudionika se, naime, tražilo da čitaju ili slušaju tekst dok naglas broje ili govore *cocacola, cocacola*… (Slowiaczek i Clifton 1980). U engleskom su se u istraživanjima uloge prozodije u tihom čitanju koristili riječima kao što su *convict* ili *abstract*, kod kojih mjesto naglaska ovisi o vrsti riječi, i to u rečenicama u kojima postoji ili ne postoji neslaganje mjesta naglaska s očekivanjem, kao npr. u rečenicama *The brilliant abstract was accepted*... nasuprot *The brilliant abstract the best ideas*..., pri čemu su izmjerene dulje fiksacije na riječi kod kojih su očekivanja govornika iznevjerena (Breen i Clifton 2011). Dulja trajanja fiksacija dobivena su i na ranim mjerama čitanja (trajanje prvog prijelaza preko teksta, ali i na ukupnom trajanju fiksacija, na trajanju drugog prijelaza i na izvedenim mjerama kao što je vjerojatnost fiksacije na riječ od interesa koju čini postotak fiksiranih riječi u odnosu na preskočene). Također, hipoteza implicitne prozodije provjeravala se u engleskom jeziku na riječima sa sekundarnim naglaskom, tj. s naglasnom jekom, npr. *independent* ili *incoherent* (Ashby i Clifton 2005). Na tim su riječima izmjerena dulja trajanja fiksacija, i to na varijabli ukupnog trajanja fiksacija, ali ne i na varijablama koje odražavaju rane procese jezične obrade kao što je trajanje prve fiksacije. Stoga Ashby i Clifton zaključuju da u tihom čitanju prozodija ima ulogu u „postleksičkim" procesima. U toj podjeli na rane i kasne procese rani bi procesi uključivali ortografsko-fonološko preslikavanje (iz grafema u foneme), dok bi kasni, postleksički procesi uključivali pristup značenju i integraciju riječi u rečenični kontekst. U istraživanjima uloge prozodije u tihom čitanju istraživale su se razlike između poezije i proze u njemačkom (Beck i Konieczny 2020). Iznevjerena metrička očekivanja, očekivano, više narušavaju tečnost čitanja poezije nego proze zbog čega Beck i Konieczny zaključuju da čitatelj gradi očekivanja na temelju „slušnog geštalta" (engl. *audible gestalt*) koji vodi čitatelja kroz tekst. Isti zaključak slijedi i iz veće stope preskakanja zadnje riječi u stihu kad je pjesma prikazana na uobičajen način, u stihovima, a ne kao da se radi o proznome tekstu. Slično tumačenje uloge prozodije u čitanju u sebi može se naći i u hipotezi prozodijskog fraziranja (Frazier, Carlson i Clifton 2006, Cumming, Wilson i Goswami 2015) prema kojoj prozodija ima prediktivnu ulogu, tj. sudjeluje u govornikovu stvaranju pretpostavka o onome što u tekstu slijedi, s tim da su u žarištu tih istraživanja intonacija i ritam, a ne naglasak riječi. Na kraju, istraživanja mjerenjem evociranih potencijala u

švedskom također upućuju na prediktivnu ulogu ritma, naglaska i intonacije u tihom čitanju (Söderström 2017).

U ovome je radu pristup drugačiji. Budući da se rad temelji na korpusu čitanja, ne postoji eksperimentalna manipulacija varijablama kao u spomenutim studijama. Neku vrstu eksperimentalne kontrole čini odabir jezika koji se prozodijski razlikuju: hrvatski se klasificira kao jezik slogovnog ritma, a engleski kao jezik naglasnog ritma (Hrvatska enciklopedija 2013 – 2025), tj. u engleskoj terminologiji, hrvatski ima *syllable-framed rhythm*, dok je za engleski karakterističan *time-framed rhythm* (Josipović Smojver 1999, Cummins, Gers i Schmidhuber 1999). Budući da se za razliku od jezika slogovnog ritma u jezicima naglasnog ritma izokronija postiže redukcijom nenaglašenih slogova, može se očekivati da će se čitatelji dulje zadržavati na naglašenim slogovima, i to više u engleskom nego u hrvatskom, u kojem je razlika u duljini između naglašenih i nenaglašenih slogova manja. Tako formulirana hipoteza nije samorazumljiva s obzirom na postojeća istraživanja koja upućuju na slabije eksperimentalne efekte vezane za prozodiju u drugom jeziku (cf. Foltz 2021, Grüter, Rohde i Schafer 2017). Veća bi razlika u zadržavanju pogleda u engleskom snažno govorila u prilog objašnjenju duljeg zadržavanja na naglašenim slogovima (tj. hipotezi implicitne prozodije) koje se temelji na tipološkoj razlici u hrvatskoj i engleskoj prozodiji.

Korpus čitanja prirodni je (ekološki valjan) način dobivanja podataka o tome slijede li govornici naglasnu strukturu teksta prilikom čitanja u sebi jer pokreti oka pri čitanju odražavaju procese pažnje. Tako se otkriva na što se govornici/čitači nesvjesno oslanjaju u procesu jezične obrade. Potvrda za hipotezu implicitne prozodije u ovom korpusu bila bi pronađena razlika između fokusiranja na naglašene i nenaglašene slogove te razlika između jezika. Prethodne analize (temeljene samo na dijelu teksta ili na vrlo malom broju sudionika, Cergol i Palmović 2024a, 2024b) upućuju na takve razlike, ali provedene statističke analize u velikoj su se mjeri razlikovale. U jednom je radu za analizu upotrijebljena linearna regresija s duljinom čitanja kao zavisnom varijablom (2024a), dok je u drugom radu upotrijebljena Bayesova inačica *t*-testa (2024b). Razlozi za različite analize višestruki su: od različitog kodiranja podataka na drukčijoj opremi do spomenutog malog broja sudionika. Za rezultate analize linearne regresije teško je dati jednoznačno tumačenje jer je jedina logična zavisna varijabla bila ukupno trajanje čitanja. Ona, međutim, odražava i vrsnoću u čitanju (kraće trajanje

i veća vrsnoća) i pažnju koju je čitatelj obraćao na čitanje (kraće trajanje možda odražava i čitanje s manje pažnje). Stoga se isti rezultat analize može tumačiti na najmanje dva načina. S druge strane, *t*-test otkriva samo ima li ili nema razlike između fokusiranja na naglašene i nenaglašene slogove, ali ne daje nikakav podatak o psiholingvističkim faktorima koji stoje u pozadini tih razlika. Stoga se kao temeljno pitanje postavlja izbor prikladne metode statističke obrade podataka u istraživanju u kojem prema logici samog istraživanja nema nikakve eksperimentalne kontrole, a koja daje osnovu za objašnjenje psiholingvističkih procesa u pozadini čitanja. Prema spomenutom statističkom priručniku za lingviste (Winter 2020) linearna bi regresija (tj. miješani model s obzirom na različite vrste uključenih varijabli) bila najprimjerenija metoda, a ona se često i upotrebljava u analizama podataka korpusa čitanja (npr. Pai i dr. 2022, Berzak i dr. 2022) pri čemu je trajanje fiksacije zavisna varijabla koja se predviđa iz duljine riječi, predvidljivosti riječi iz rečeničnog konteksta, čestotnosti, stope preskakanja riječi i sl. Oba spomenuta korpusa sastavljena su, međutim, od pojedinačnih rečenica za koje su izračunate „norme predvidljivosti", tj. vjerojatnosti pojavljivanja pojedinih riječi na temelju niza prethodnih riječi u rečenici. Kad govori o linearnim modelima, Winter predlaže izradu modela koji počinje s maksimalnim brojem prediktora, s tim da se oni najmanje značajni odbacuju dok se ne dođe do modela u kojem svi prediktori postižu statističku značajnost. S podatcima mjerenja pokreta oka to nije postupak od dvaju koraka zbog visoke kolinearnosti koja uvelike otežava takvu analizu. Naime, dva događaja, kako ih uređaj klasificira, u kombinaciji s područjima interesa daju četrdesetak varijabli koje imaju različita psiholingvistička tumačenja. Budući da su u njihovoj podlozi ista dva događaja i njihove izmjerene vrijednosti, u pravilu je u rezultatima sadržana visoka kolinearnost koja značajno utječe na rezultate.[1] Teška procjena doprinosa pojedine varijable ili njihov izbor tek na temelju njihova međusobnog odnosa, a ne psiholingvističkog tumačenja, smanjuje upotrebljivost takvih statističkih modela. Nadalje, kako primjećuje McElreath (2020.), takav model može dobro opisivati veze između varijabla, davati dobra predviđanja, ali biti „mehanistički" ili kauzalno pogrešan, baš kao što je to bio, kako kaže, geocentrični

[1] Na primjer, ako pet sudionika ima ukupan broj fiksacija 10, 14, 16, 17 i 20, samo dio njih (tj. podskup ukupnog broja fiksacija) pripadat će varijabli prve fiksacije, tako da će broj prvih fiksacija u području interesa biti npr. 9, 12, 13, 15 i 16. Te su dvije varijable u visokoj korelaciji ($r = 0,98$) i varijacija u jednoj objašnjava se varijacijom u drugoj u visokom postotku ($R^2 = 0,96$), što otežava procjenu doprinosa svake pojedine varijable kao prediktora.

model Sunčeva sustava. Posljedica može biti da struktura statističkog modela izgleda proizvoljno, a često i neinterpretabilno. U prethodnim analizama tog hrvatsko-engleskog korpusa poseban je problem takvog modela i to što zavisna varijabla odražava vrsnoću u čitanju općenito. To dovodi do zbrke jer ako je duljina fiksacija veća, bit će veće i ukupno vrijeme čitanja. Veća duljina fiksacija baš na naglašene slogove zapravo loše razlikuje čitače koji se dulje fokusiraju na naglašene slogove jer su vrsni čitači (koji automatski obraćaju pažnju na ono što je najvažnije za razumijevanje) od onih koji se dulje fokusiraju na naglašene slogove jer su lošiji čitači koji jednostavno trebaju više vremena za obradu.

Pristup Bayesove statistike nema taj problem, a logika je same analize drukčija. Prvo, kao što je dobro poznato, ona testira vjerojatnost hipoteze, a ne vjerojatnost podataka pod uvjetom da je nulta hipoteza istinita. Drugim riječima, prema frekvencionističkom pristupu zaključuje se kako je malo vjerojatno da će biti izmjereni podatci kakvi su izmjereni, a pritom je nulta hipoteza istinita, pa se na temelju toga prihvaća radna hipoteza. U Bayesovoj statistici shvaćena je kao stupanj uvjerenja u propoziciju, a ne kao ono što se u frekvencionističkoj statistici naziva graničnom frekvencijom (engl. *long-run frequency*), koja poistovjećuje vjerojatnost s čestotnosti u velikom broju ponavljanja (npr. kad bismo novčić bacali milijun puta, vjerojatno bismo dobili učestalost glave od 50 %, tj. vjerojatnost glave iznosila bi 0,5). Stupanj uvjerenja predstavlja prethodnu, tj. apriornu vjerojatnost (engl. *prior*), a to je veličina za koju istraživač ima neke prethodne podatke. U slučaju prikupljenog korpusa čitanja ti podatci temelje se na broju naglašenih slogova u hrvatskom i engleskom. Iz apriorne vjerojatnosti se prema Bayesovu teoremu može izračunati aposteriorna vjerojatnost (engl. *posterior*), koja predstavlja vjerojatnost hipoteze, ono što želimo utvrditi.[2] Do sada je na korpusu koji se ovdje analizira provedena takva analiza, kao što je spomenuto (Cergol i Palmović 2024b), ali je zbog malog broja sudionika bio

[2] Aposteriorna se vjerojatnost (p(H|D) računa prema Bayesovu teoremu (p(H|D) = (p(D|H)*p(H))/p(D)) kao umnožak izglednosti (eng. likelihood, p(D|H), tj. vjerojatnosti podataka uz danu hipotezu, i vjerojatnosti hipoteze (tj. apriorne vjerojatnosti, p(H)) podijeljenog s vjerojatnosti podataka (p(D), tj. vjerojatnosti podataka bez obzira na hipoteze). Često se koristi u zaključivanju u medicini: neka bolest pogađa, na primjer, 1 % populacije (p(B) = 0,01, tj. p(~B = 0,99)). Ako osoba ima bolest, test će biti pozitivan u 99 % slučajeva (p(T|B) = 0,99); ako je nema, test će biti pozitivan u 5 % slučajeva (p(T|~B) = 0,05). Pitamo se (i to je aposteriorna vjerojatnost) ako je test pozitivan, koja je vjerojatnost da osoba zaista ima bolest. Nazivnik (vjerojatnost pozitivnog testa) čini zbroj svih mogućnosti bez obzira na hipotezu (ima li ili nema osoba bolest: p(T) = p(T|B)*p(B)+p(T|~B)*p(B) = 0,059. Brojnik je umnožak p(T|B) i p(B) i podijeljen s p(T) daje 0,17, dakle 17 % vjerojatnosti da osoba stvarno ima bolest pod uvjetom da je test pozitivan.

moguć samo *t*-test koji je na dvjema varijablama pokazao veliku razliku između očekivane vrijednosti na temelju smjene naglašenih i nenaglašenih slogova u tekstu i posteriorne izmjerene distribucije fiksacija na naglašene slogove u čitanju. Sudionici su se, dakle, mnogo više fokusirali na naglašene slogove nego što bi to slijedilo iz same smjene naglašenih i nenaglašenih slogova u tekstu. To potvrđuje hipotezu implicitne prozodije. Također, rezultati su se razlikovali za hrvatski i engleski. Za hrvatski su dobiveni podatci predstavljali „anegdotalno do umjereno" povećanje vjerojatnosti istinitosti hipoteze (Bayesov faktor je bio oko 4, ovisno o varijabli), dok je za engleski posteriorna vjerojatnost bila vrlo velika s Bayesovim faktorom višim od 1000 (što čini hipotezu implicitne prozodije četiri puta vjerojatnijom od odgovarajuće nulte hipoteze u hrvatskom i 1000 puta vjerojatnijom u engleskom). Nulta bi hipoteza u ovom slučaju bila da se govornici fokusiraju na naglašene i nenaglašene slogove proporcionalno njihovu omjeru u tekstu. Dobivena se razlika u spomenutom radu tumačila razlikama u prozodijskoj strukturi između dvaju jezika (Josipović Smojver 1999). Međutim, ograničenje je takva pristupa to što ne znamo što je to što čini da se na naglašene slogove fokusiramo dulje. Koji su to psiholingvistički ili kognitivni mehanizmi (npr. mehanizmi usmjeravanja pažnje) u pozadini dobivene razlike, a o kojima se može zaključivati na temelju različitih varijabla pokreta oka s obzirom na to da one odražavaju različite vidove jezične obrade. Jednostavno rečeno, *t*-test kaže samo da je razlika veća od očekivane.

U ovome će radu biti provedene obje (frekvencionistička i Bayesova) analiza korpusa, s podatcima prikupljenim na većem broju sudionika nego što je to bilo u prethodnim studijama. Cilj je prvo provesti analizu koja daje uvid u narav procesa u pozadini dulje fiksacije na naglašene slogove s obzirom na različita tumačenja varijabli koje omogućuje praćenje pokreta oka. Ako je prozodija dio „postleksičke" obrade u smislu Ashby i Cliftona (2005), očekuju se dulje fiksacije na „kasnim" mjerama pokreta oka (npr. ukupno trajanje fiksacija u području interesa, broj regresija), a ako je dio ranih procesa fonološke obrade, dulje se fiksacije očekuju na varijablama kao što je prva fiksacija u području interesa ili prvi prijelaz preko područja interesa (zbroj svih fiksacija do prvog napuštanja područja interesa). Drugo, na temelju tih podataka raspravit će se o prednostima i nedostatcima dvaju pristupa analizi u toj specifičnoj metodologiji u psiholingvističkim istraživanjima. Pri tome se provjeravaju iste hipoteze kao i u navedenim analizama ovoga korpusa (da će se pogled govornika dulje zadr-

žavati na naglašenim slogovima, i to više u engleskome) s obzirom na to da je za to i prikupljen.

2. Metodologija

Analize su provedene na paralelnom hrvatsko-engleskom korpusu čitanja koji se temelji na priči *Pripovjedač* (engl. *Storyteller*) engleskog pisca Hectora Munroa Sakija (1984., 1993.), koju su sudionici čitali vlastitim tempom, stranicu po stranicu (na sljedeću stranicu prelazili su pritiskom na razmaknicu). Prije same priče za uvježbavanje su pročitali kratku Ezopovu basnu. Tekst priče i na hrvatskom i na engleskom jeziku činilo je 11 stranica s 12 redaka teksta napisanog fontom Courier New veličine 16. Taj je font odabran jer svako slovo zauzima jednaki razmak, tj. zato što je neproporcionalni font (eng. *monospaced*). Tekst se prikazivao na zaslonu računala veličine 21''. Upotrebom stola podesive visine postignuto je to da su svi sudionici vidjeli tekst pod istim kutom, a i kut između kamere i sudionika bio je jednak za sve sudionike.

Engleski je tekst sadržavao 1962 riječi i 2696 slogova od kojih je 1835 bilo naglašeno. Hrvatski se prijevod sastojao od 1746 riječi s 3614 slogova od kojih je 1266 bilo naglašeno. U prosjeku engleske su riječi bile duge 1,4 sloga, dok su hrvatske bile dulje, prosječno 2,1 slog. Ta je razlika posljedica tipoloških razlika između jezika. Hrvatski padeži često dodaju slog, npr. u imenica muškog roda. Tako, na primjer, engleskom *to the boy* odgovara hrvatsko *dječaku*, tj. trima engleskim jednosložnim riječima odgovara jedna hrvatska trosložna riječ. Konačno, engleski tekst sadržava 8933 slova, dok u hrvatskom prijevodu ima 8588 slova. Engleska je riječ tako prosječno duga 4,6, a hrvatska 5 slova. Ta sličnost u duljini, bez obzira na razliku u broju slogova, može se tumačiti različitom ortografskom dubinom u engleskom i hrvatskom, koja se izražava kao omjer grafema i fonema. Točnije, izražava se kao omjer korespondencija grafema i fonema podijeljen s brojem grafema (GPC/g, Gontijo, Gontijo i Shillcock 2003). Za hrvatski taj je omjer 1,1 : 1, dok je za engleski 2,4 : 1. Drugim riječima, sličan broj slova u hrvatskom i engleskom tekstu prosječno predstavlja manji broj fonema u engleskom.

Sudionici su bili studenti s razinom poznavanja engleskog B2 ili više. Znanje stranog jezika nije se posebno testiralo za potrebe istraživanja jer su odabrani sudionici na nacionalnom ispitu postigli odgovarajuću razinu, a za potrebe studija svakodnevno se služe engleskim jezikom. Pedeset i jedan sudionik pristupio je istraživanju, a za ovaj su se rad analizirali podatci za njih 45 (Ž = 42, M = 3) koji su u oba mjerenja imali visoku kvalitetu izmjerenih podataka, tj. koji su na kalibraciji imali grešku < 0,35 ° vidnog kuta, što odgovara veličini slova. Također, bilo je potrebno zadržati tu razinu greške tijekom cijelog mjerenja. Svaki je sudionik dva puta dolazio na mjerenje. Pola sudionika (tj. njih 22) u prvom je dolasku čitalo tekst na engleskom, a pola na hrvatskom (tj. njih 23). Razmak između dvaju mjerenja bio je najmanje dva tjedna.

Za mjerenje pokreta oka upotrijebljen je *EyeLink Portable Duo* (*SR Ressearch*) s uzorkovanjem od 1 kHz, uz upotrebu oslonca za bradu. Mjerili su se pokreti desnog oka sudionika. Sudionici su sjedili na 50 cm od zaslona, pročitali upute, a zatim je slijedila kalibracija na 9 točaka te njezina verifikacija. Sudionici su pritiskom na razmaknicu prelazili na sljedeću stranicu, s tim da je prije svake stranice bio izračunat i ispravljen pomak smjera pogleda (*drift*) kako bi se osigurala točnost kalibracije tijekom cijelog mjerenja na više stranica (kad bi pomak bio veći od 0,45 ° vidnog kuta, pristupilo bi se ponovnoj kalibraciji tijekom čitanja). Nakon završetka mjerenja sudionici su odgovorili na tri pitanja kako bi se provjerilo jesu li pažljivo pročitali i razumjeli tekst. Pitanja su se odnosila na zapamćivanje detalja iz priče (npr. koje je medalje dobila djevojčica iz pripovjedačeve priče, Bertha) i razumijevanje poante priče (Bertha je u priči stradala jer je bila dobra). Nije bilo sudionika koji nisu znali odgovoriti na pitanja, pa zbog pogrešnih odgovora nije isključen nijedan sudionik. Pitanja se nisu dalje upotrebljavala u analizi.

Nakon završetka mjerenja podatci su izvučeni iz uređaja s obzirom na područja interesa koja su bila definirana oko naglašenih i nenaglašenih slogova.[3] Ti su podatci predobrađeni za statističku obradu pomoću programa R (R Core Team

[3] Područje interesa definira se u samom programu za kontrolu eksperimenta, *Experiment Centre*. U tekst se umetnuo odgovarajući znak (*) između svakog sloga unutar riječi (Bi*lo je to*plo po*pod*ne, u že*ljez*ni*čkom va*go*nu bi*lo je za*guš*lji*vo…) i odabrala opcija prema kojoj taj znak označava granicu područja interesa i ne prikazuje se sudioniku. Područje interesa obuhvaća područje do pola razmaka između retka iznad i retka ispod tako da osim margina nema područja u tekstu koje nije uključeno u područje interesa. Po tome što omogućuje i takvo definiranje područja interesa, *EyeLink PortableDuo* razlikuje se od uređaja upotrijebljenog u ranijim studijama (SMI iView) u kojima su područja interesa u tekstu bila

2016) i paketa tidyverse (Wickham i dr. 2019). Predobrada podrazumijeva da se sirovi podatci s uređaja za mjerenje pokreta oka pretvore u tablicu pogodnu za daljnju statističku obradu za koju se također koristio R 4.4.3 (paketi lme4 i brms, Bates i dr. 2015, Bürkner 2017). Za statističku analizu upotrijebljena je logistička regresija s kategoričkom zavisnom varijablom naglašenosti sloga s dvjema razinama, *naglasak* i *bez_naglaska*, i kontinuiranim prediktorima ukupnog trajanja fiksacije u području interesa, trajanju fiksacije u prvom prijelazu preko teksta i broj regresija u područje interesa. Također, u obzir su uzete i druge varijable koje se nisu pokazale statistički značajnima, pa su isključene iz drugog koraka analize (npr. tzv. „prelijevanje", tj. produženo trajanje fiksacija na sljedeće područje interesa). Različite su varijable odabrane zbog njihova različitog tumačenja imajući na umu rane i kasne procese u čitanju. Varijabla ukupnog trajanja fiksacija u području interesa odražava kasne, postleksičke procese, a broj regresija iznevjerena očekivanja, tj. predviđanja čitatelja i njegovu potrebu za ponovnom analizom rečenice. Trajanje prve fiksacije u području interesa odražava rane procese, preslikavanje grafema u foneme i prepoznavanje riječi. Taj odabir omogućava tumačenje rezultata u smislu različitih psiholingvističkih procesa u pozadini čitanja uz zadržavanje malog broja prediktora.

Analiza je provedena na podatcima za svaki slog u svakom tekstu (za razliku od prethodnih analiza provedenih na podatcima uprosječenim po sudioniku). Logika je sljedeća: možemo li na temelju poznavanja parametara pokreta oka predvidjeti je li slog, na koji se sudionik fokusirao, bio naglašen ili nije? Ako možemo, tada naglasna struktura objašnjava razlike u dobivenim mjerama pokreta oka. Dvije analize, frekvencionistička i Bayesova, razlikuju se prema tome što Bayesova uzima kao prethodne vjerojatnosti za svaki od prediktora (tj. varijabli mjerenja pokreta oka) vrijednosti kakve bi slijedile samo iz naglasne strukture teksta, tj. smjene naglašenih i nenaglašenih slogova. Takav postupak omogućuje uvid u to što bi slijedilo iz same lingvističke strukture, dok se aposteriorna vjerojatnost, tj. vjerojatnost hipoteze računa na temelju stvarno izmjerenih vrijednosti tih varijabli. Za svakog se sudionika uzima, na primjer, ukupno trajanje fiksacija u hrvatskom tekstu i taj se broj pomnoži s 0,35 zato što hrvatski tekst sadržava 35 % naglašenih slogova. Očekuje se, dakle, da će se 35 % trajanja svih

definirana samo usko oko samog teksta, pa je dio pogleda poslije nemoguće povezati s odgovarajućim dijelom teksta.

fiksacija odnositi na fiksacije na naglašenim slogovima. Na temelju svih sudionika izračunava se distribucija temeljena na srednjoj vrijednosti i standardnoj devijaciji. Takva se prethodna vjerojatnost smatra prethodnom vjerojatnošću niske informativnosti.

3. Rezultati

Modeli logističke regresije izračunati su u dva koraka za svaki jezik posebno. U prvom koraku u model su uključeni svi prediktori koji od velikog broja mogućih varijabli imaju smisla te je za svaki jezik formula izgledala identično:

naglašenost ~ trajanje fiksacije + trajanje prve fiksacije + trajanje prvog prolaza + broj regresija + prelijevanje + (1 | sudionik).

Sudionici su bili slučajna varijabla, što se vidi iz odgovarajuće formule. Razlika je između trajanja prve fiksacije i trajanja prvog prijelaza u tome što prva varijabla mjeri samo prvu fiksaciju u prvom prijelazu oka preko nekog područja interesa, dok druga mjeri ukupno trajanje svih fiksacija u prvome prijelazu čitatelja preko teksta.

U hrvatskom i engleskom nisu dobiveni isti statistički značajni prediktori te su u drugom modelu ostavljeni samo oni prediktori koji su se pokazali statistički značajnima u pojedinom jeziku. U hrvatskom su se značajnima pokazale varijable ukupno trajanje fiksacija u prvome prijelazu i broj regresija u područje interesa, dok se trajanje prve fiksacije nije pokazalo statistički značajnim. Uz to, trajanje prve fiksacije visoko korelira s varijablom trajanja prvog prijelaza preko područja interesa (– 0,754) pa je isključeno iz daljnje analize zbog kolinearnosti. Zato je izračunat novi model sa samo dvama prediktorima, ukupnim trajanjem fiksacija u prvom prijelazu i brojem regresija. Međutim, nijedan od prediktora nije se pokazao statistički značajnim, što se vidi u tablici 1 koja prikazuje izlazne podatke modela. Preostala dva prediktora ne koreliraju visoko (0,035).

Tablica 1. Sažetak logističkog modela za hrvatski.

Fiksni efekti:	**Procjena**	**Std. pogreška**	**z vrijednost**	**Pr(>|z|)**	
(Intercept)	0,32261	0,02402	13,428	< 2e – 16	***
Trajanje prvog prijelaza u području interesa	– 0,02098	0,02228	– 0,942	0,346	
Broj regresija u područje interesa	– 0,04456	0,03272	– 1,362	0,173	

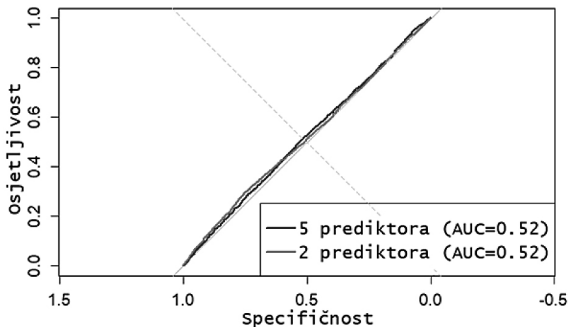
Razine značajnosti: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Dakle, dva prediktora sama po sebi nisu značajna za varijablu *naglašenost*. Usporedba modela temelji se na Akaikovu kriteriju obavijesnosti (engl. *Akaike Information Criterion*, AIC, Akaike 1998) pri čemu manji AIC označava bolji model, tj. model koji bolje objašnjava podatke, s time da „kažnjava" velik broj parametara. U slučaju prvog, maksimalnog modela i drugog modela, sa samo dvama prediktorima rezultati su prikazani u tablici 2.

Tablica 2. Akaikov kriterij obavijesnosti za oba modela za hrvatski.

Model	**AIC**	**Razlika u AIC**
Dva prediktora	11775,74	0,548
Puni model	11775,19	0,0

Budući da manji AIC označava bolji model, čini se da je „puni model", tj model s pet prediktora, samo neznatno bolji. Mala razlika, međutim, upućuje na veliku sličnost između modela. No oba modela imaju vrlo slabu prediktivnu snagu, tj. iz razlika u parametrima pokreta oka mala je vjerojatnost da predvidimo pada li pogled na naglašen ili nenaglašen slog u hrvatskome. To znači da postoje i druge varijable koje mogu utjecati na proces čitanja, ne na analiziranoj razini sloga. Slaba se prediktivna moć modela lako može uočiti na Slici 1, na kojoj su prikazane krivulje radne karakteristike primatelja, tj. ROC krivulje (*Receiver Operating Characteristic*) i površina ispod krivulje (AUC, *Area Under Curve*) kao jedinstven parametar prediktivne snage modela. Što je krivulja bliže lijevom gornjem uglu, to je prediktivna snaga modela veća. Kao što se vidi, AUC je na razini slučaja (AUC > 0,5 i < 0,7 znači slabu prediktivnu snagu; 0,5 je predviđanje na razini slučaja).



Slika 1. ROC krivulja za modele logističke regresije za hrvatski (prediktivna snaga modela).

Za engleski su se tekst u modelu s pet prediktora tri prediktora pokazala statistički značajnima: ukupno trajanje fiksacija u području interesa (tj. u naglašenim i nenaglašenim slogovima), broj fiksacija u prvom prijelazu kroz tekst i trajanje prve fiksacije u području interesa. Zadnja dva prediktora, međutim, pokazuju visoku kolinearnost (0,62) pa je u jednostavniji model uvršten samo drugi prediktor. Tako se drugi model, s dvama prediktorima, donekle razlikuje za hrvatski i engleski. Sažetak je prikazan u tablici 3.

Tablica 3. Sažetak modela s dvama prediktorima za engleski.

Fiksni efekti:	**Procjena**	**Std. pogreška**	**z vrijednost**	**Pr(>|z|)**	
(Intercept)	0,255240	0,019350	13,190	< 2e – 16	***
Trajanje fiksacija u području interesa	– 0,252630	0,010250	– 24,647	< 2e – 16	***
Trajanje prve fiksacije u području interesa	– 0,001361	0,008357	– 0,163	0,871	

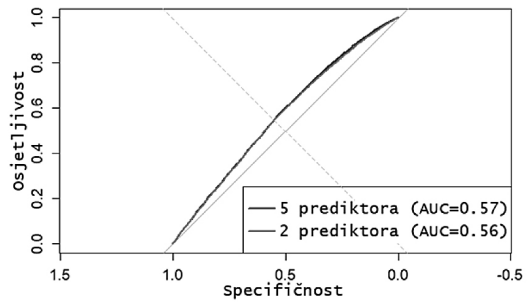
Razine značajnosti: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Kao što se vidi, u modelu s dvama prediktorima samo se ukupno trajanje fiksacija u području interesa pokazalo značajnim. Drugi prediktor u tom modelu ima vrlo malu ulogu. Slično kao i za hrvatski Akaikov kriterij obavijesnosti pokazuje da je puni model ipak bolji i da je razlika značajna (tablica 4), što dobro odražava činjenicu da na usmjeravanje pogleda pri čitanju utječu mnogi čimbenici.

Tablica 4. Akaikov kriterij obavijesnosti za oba modela za engleski.

Model	AIC	Razlika u AIC
Dva prediktora	30018,88	45,97
Puni model	29972,91	0,0

Nešto je bolja prediktivna snaga, što se vidi iz ROC krivulja prikazanih na Slici 2.



Slika 2. Prediktivna snaga modela logističke regresije s dvama prediktorima za engleski.

Kao što se vidi, oba modela imaju slabu prediktivnu snagu, ali ipak veću od slučajnosti. Za model s pet prediktora AUC je 0,57, dakle pogađa o kojem se slogu radi s 57 %, a model s dvama prediktorima pogađa s 56 %, tj. AUC je 0,56. Dakle, oba modela imaju statistički značajne prediktore naglašenosti sloga, ali na čitanje utječu i mnogi drugi faktori. Stoga se samo na temelju tih dvaju modela može govoriti o prediktivnoj snazi od 57 %.

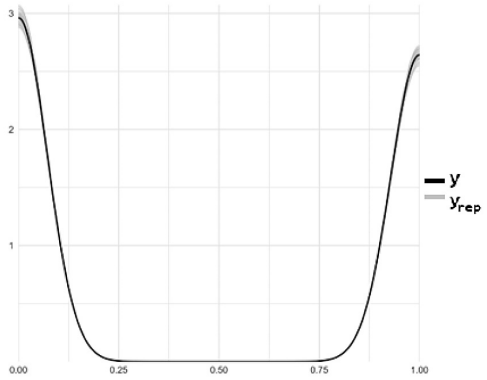
Bayesov model logističke regresije daje vrlo slične rezultate. Budući da p-vrijednost nema smisla u tom pristupu, ono što odgovara statističkoj značajnosti jest interval vjerodostojnosti (IV). Taj interval nije uključivao 0 za dva prediktora u modelu od pet prediktora za hrvatski, za duljinu fiksacije u prvom prijelazu kroz tekst i za regresije u područje interesa. Model ima nešto bolju prediktivnu snagu od frekvencionističkog s AUC = 0,544, što može biti posljedica uzimanja u obzir apriornih vjerojatnosti koje se temelje na smjeni naglašenih i nenaglašenih slogova u tekstu. Model s dvama prediktorima neznatno je bolji, AUC = 0,566. Sažetak je prikazan u tablici 5. Za Bayesove su modele sve vrijednosti varijabla skalirane, tj. izračunate su z-vrijednosti radi brže računalne obrade koja je na velikom skupu podataka (ukupno više od 300 000 redaka u *Excelovoj* tablici)

bila vrlo spora jer paket brms upotrebljava metodu temeljenu na Monte Carlo Markovljevim lancima (MCMC) koja je komputacijski vrlo zahtjevna.

Tablica 5. Sažetak Bayesova modela logističke regresije za hrvatski.

Fiksni efekti:	**Procjena**	**Procijenjena greška**	**Donji 95 % IV**	**Gornji 95 % IV**
(Intercept)	– 0,52	0,03	– 0,60	– 0,46
Trajanje prvog prijelaza	0,06	0,01	0,04	0,07
Broj regresija	0,05	0,01	0,04	0,06

Dakle, isti su prediktori značajni za hrvatski i u frekvencionističkoj i u Bayesovoj analizi. Bayesov model računa i interval vjerodostojnosti za svakog sudionika, dakle za slučajnu varijablu. Za 39 od 45 sudionika taj interval nije uključivao 0, što znači da postoje velike individualne razlike među sudionicima te stoga i različit utjecaj pojedinih sudionika na ukupan rezultat. ROC krivulje izgledaju slično kao i za frekvencionistički model (uz malo veći otklon prema gornjem lijevom kutu, što je jasno iz vrijednosti AUC), pa se zbog sažetosti izostavljaju. Upotrijebljen paket programa R, brms, nudi jasnu vizualizaciju Bayesova modela. Prikazom provjere posteriorne prediktivnosti modela (engl. *posterior predictive check plot*) vizualizira se koliko se predviđanja modela poklapaju s izmjerenim podatcima. Na Slici 3 vidi se jasan binarni rezultat (krivulja oblika slova *U*), što znači da model s dvama prediktorima temeljenima na mjerama pokreta očiju razlikuje naglašene od nenaglašenih slogova u hrvatskom.



Slika 3. Provjera posteriorne prediktivnosti modela s dvama prediktorima za hrvatski (crna linija predstavlja izmjerene podatke (*y*), dok plavo područje oko

nje (y_{rep}) predstavlja stupanj nesigurnosti (engl. *uncertainty*). Vjerojatnost je prikazana na apscisi.

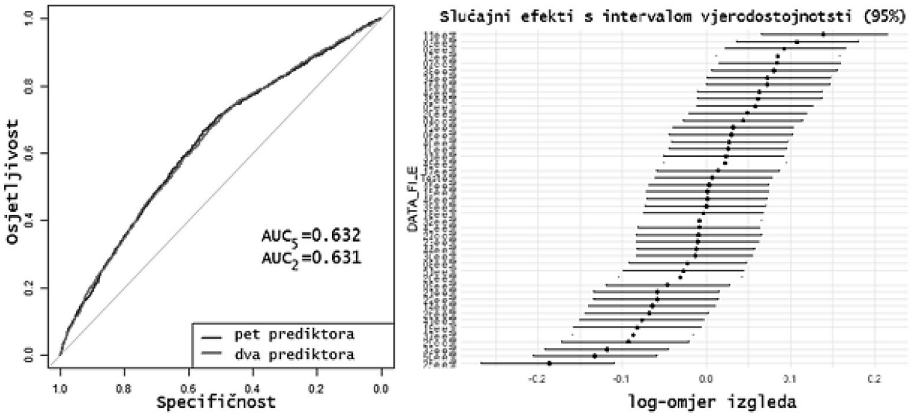
Dobro preklapanje krivulja y i y_{rep} te usko područje y_{rep} oko krivulje y potvrđuju da model pogađa ukupnu distribuciju podataka. To područje također upućuje na to da nema prenaučenosti (engl. *overfitting*). Jednako tako, visoki vrhovi krivulje (oko vjerojatnosti blizu 0 i 1) govore o tome da model sa sigurnošću predviđa naglašenost slogova, ali područje oko 0,5 upućuje na to da je za neke sudionike to predviđanje na razini slučaja. U svakom slučaju, model – kao mjera temeljena na psiholingvističkim varijablama pokreta oka – dobro predstavlja binarnu narav smjene naglašenih i nenaglašenih slogova.

Za engleski su podaci Bayesova modela slični, s time da su se u punom modelu pokazali značajni prediktori kao i u frekvencionističkom modelu logističke regresije, dakle, prediktori različiti od onih za hrvatski. U drugi su model uključeni samo prediktori koji su se pokazali značajnim, tj. kod kojih interval vjerodostojnosti ne uključuje 0, kako je prikazano u tablici 6.

Tablica 6. Sažetak Bayesova modela logističke regresije za engleski

Fiksni efekti:	**Procjena**	**Procijenjena greška**	**Donji 95 % IV**	**Gornji 95 % IV**
(Intercept)	– 0,44	0,01	– 0,45	– 0,43
Trajanje fiksacije	0,06	0,01	0,43	0,46
Trajanje prve fiksacije	0,05	– 0,06	– 0,03	– 0,06

Za razliku od frekvencionističkog modela, u Bayesovu se modelu s dvama prediktorima kao značajan prediktor pokazala i varijabla koja mjeri trajanje prve fiksacije u prvom prelasku preko područja interesa. Uz to, ona je negativno povezana s naglašenošću sloga, tj. što je vrijednost te varijable veća, manja je vjerojatnost da je slog naglašen. Bayesov model logističke regresije za engleski ima bolju prediktivnu snagu što se vidi iz ROC krivulje (slika 4). Neznatno je bolji model s pet prediktora. Bayesov model ipak je ukupno oko 6 % bolji od frekvencionističkog modela u predviđanju naglašenosti sloga iz parametara pokreta oka.



Slika 4. Lijevo: ROC krivulja za oba Bayesova modela logističke regresije za engleski. Desno: prikaz individualnih razlika među sudionicima.

Slučajni efekt razlike između sudionika istraživanja manji je za engleski nego za hrvatski. Samo za 6 sudionika interval vjerodostojnosti ne uključuje 0, što se vidi na Slici 4, desno. Dakle, isti su se sudionici daleko manje međusobno razlikovali u čitanju engleskog nego hrvatskog. Model s dvama prediktorima od kojih je jedan „jači" i pozitivno povezan s vjerojatnošću naglašenog sloga, a drugi „slab" i negativno povezan s vjerojatnošću da se radi o naglašenom slogu može se vizualizirati linearnom povezanošću tih prediktora, kao na Slici 5. Kao što se vidi, za različite standardizirane vrijednosti (z-vrijednosti duljine trajanja pogleda u području interesa) vjerojatnost se kreće od 0,20 pa sve do 0,75 za naglašeni slog (na slici gore), dok se za odgovarajuće vrijednosti prediktora prve fiksacije u području interesa te vrijednosti kreću od 0,55 do 0,70. Dok su za nenaglašeni slog (na slici dolje) vrijednosti varijable prve fiksacije u području interesa slične, model previđa nenaglašeni slog na temelju ukupne duljine fiksacije u područje interesa katkad čak i s više od 80 %.

Slika 5. Efekti prediktora na vjerojatnost naglašenog sloga.

4. Rasprava i zaključak

Korpusi čitanja metodologija su istraživanja jezične obrade kojom se nastoji iskoristiti prednosti uređaja za mjerenje pokreta oka: sudionik treba sjesti pred zaslon računala i pročitati u sebi tekst, a da se od njega ne traži nikakav odgovor. Kao da lista stranice knjige, pritiskom na tipku prelazi na sljedeću stranicu. Ta metoda, međutim, ima i svoju drugu stranu: potpuni izostanak eksperimentalne kontrole nad varijablama koje uređaj bilježi. Metode statističke analize podataka uobičajene u eksperimentalnoj psiholingvistici stoga nisu primjenjive. Ne postoje faktori s dvjema razinama ili s više njih kojima manipulira eksperimentator. Kako je navedeno u uvodu, u dostupnoj se literaturi uglavnom predlaže regresijski model, ali logika takve analize pretpostavlja da se specifični prediktori (kakvi su upotrijebljeni u ovom radu) uspoređuju s nekom „općom" varijablom, kao što je na primjer ukupno vrijeme čitanja. Takva analiza, također, pretpostavlja prosječne vrijednosti ostalih varijabli po sudionicima i područjima interesa pa se gubi glavna prednost korpusa: velika količina podataka. U ovom radu pokušalo se krenuti drugim putem: ako parametri pokreta oka objašnjavaju procese čitanja i ako je za te procese važno je li slog na koji se čitatelj fokusira naglašen ili ne, onda bi se iz parametara pokreta oka naglašenost sloga trebala moći predvidjeti. Budući da se predviđa vrijednost kategoričke varijable, logistička regresija pokazuje se kao logičan izbor. Rezultati su pokazali da su predikto-

ri temeljeni na varijablama pokreta oka, iako su statistički značajni, slabi – za hrvatski predviđaju na razini slučajnosti, očito zbog mnogih drugih čimbenika koji utječu na čitanje. Bayesova logistička regresija pokazala se prikladnijom zato što kao apriorne vjerojatnosti (engl. *priors*) uzima podatke koji se mogu izračunati iz same strukture teksta (ovdje: smjene naglašenih i nenaglašenih slogova). Takva analiza omogućuje da se za mnoge varijable, koje nudi uređaj za mjerenje pokreta oka, unaprijed izračuna koje se vrijednosti očekuju s obzirom na strukturu teksta. Izmjerene vrijednosti za pojedine varijable upućuju na to fokusira li se sudionik više ili manje na tu strukturu, što se može tumačiti s obzirom na uobičajenu psiholingvističku interpretaciju za svaku pojedinu varijablu (ovdje: „leksički" nasuprot „postleksičkim" procesima). Tako se od prikupljenih bihevioralnih podataka i same lingvističke strukture može ponešto doznati o psiholingvističkim procesima, tj. o procesima jezične obrade.

Također, Bayesova metoda daje bolju sliku individualnih razlika među sudionicima. Činjenica da je raspršenost vrijednosti varijabla veća za hrvatski nego za engleski može objasniti i razliku u dobivenim rezultatima; veća raspršenost u hrvatskome posljedica je različitih čitalačkih i jezičnih vještina sudionika, kao i njihova podrijetla, tj. dijalekta kojim govore. Veća kompaktnost vrijednosti varijabla u engleskom posljedica je sličnosti sudionika u vrsnoći u jeziku; njime su svi morali ovladati barem do razine B2 u barem sličnom procesu učenja jezika kao stranog jezika. Svakako, veća kompaktnost vrijednosti varijabli doprinosi njihovoj boljoj prediktivnoj snazi.

Različiti prediktori koji su se pokazali značajnim u hrvatskom i engleskom upućuju na različite procese koji sudjeluju u razumijevanju pročitanog teksta na materinskom i na stranom jeziku. Za čitanje hrvatskoga teksta važnije su varijable prvog prijelaza preko područja interesa i broj vraćanja. Te su dvije varijable prediktivne za naglašenost sloga (barem u modelu s više prediktora). Varijable prvog prijelaza tumače se kao odraz početne izgradnje strukture rečenice, od preslikavanja grafema u foneme do prepoznavanja riječi. Takav se rezultat razlikuje od rezultata dobivenih u spomenutom istraživanju Ashby i Cliftona (2005.), koje se, ipak, razlikuje po metodologiji jer uključuje eksperimentalnu manipulaciju varijablama, ali i čitanje pojedinačnih rečenica kod kojih je općenito zbog nedostatka konteksta predvidljivost manja, a manju ulogu imaju i drugi leksički čimbenici, na primjer čestotnost. (Rayner i Clifton 2009). Dobiveni su rezultati

– razlika u fiksaciji na naglašene slogove na ranim mjerama pokreta oka – konzistentni s rezultatima koji upućuju na prediktivnu ulogu prozodije u jezičnom razumijevanju, tj. s hipotezom prozodijskog fraziranja (Frazier, Carlson i Clifton 2006, Cumming, Wilson i Goswami 2015), s tim da su Ashby i Clifton u svojem radu istraživali leksički naglasak (kao i u ovom istraživanju), dok su Frazier, Cumming i njihovi suradnici istraživali intonaciju.

Regresije u područje interesa smatraju se indeksom sintaktičke obrade, tj. govornikove ponovne analize rečenice nakon što njegov pogled padne na jednu od sljedećih riječi i ne može je uklopiti u rečenični kontekst. To se događa zato što se za vrijeme fiksacije već planira sljedeća sakada. Sakade se opisuju kao „balističke", tj. jednom kad se sakada pokrene, pogled će pasti na ono mjesto koje je unaprijed planirano i to se mjesto ne može promijeniti (za vrijeme sakade potiskuju se vizualne obavijesti koje oko prima). Činjenica da su se oči sudionika više vraćale na naglašene slogove u hrvatskom teško se može jednoznačno tumačiti: moguće je da su se vraćali na najistaknutiji dio riječi, onaj s kojeg se riječ prepoznaje s najmanjim kognitivnim naporom, ali je moguće da su to većinom bili prvi slogovi pa su se vraćali na početak riječi. Budući da je regresija mjera iznenađenja govornika tekstom koji slijedi, dakle iznenađenja na razini sintakse ili semantike, ona ne mora imati izravne veze s prozodijskom strukturom unutar riječi.

U engleskom je za predviđanje naglašenosti sloga iz podataka pokreta oka u najvećoj mjeri prediktivna ukupna duljina fiksacije u području interesa, tj. na naglašeni slog, što se poklapa s rezultatima gore spomenute studije Ashby i Cliftona (2005.). Drugi prediktori imaju mnogo manju ulogu, koju je teško procijeniti zbog njihove kolinearnosti. Duljina fiksacije tumači se kao varijabla koja upućuje na „postleksičke" procese prepoznavanja značenja riječi i njegova uklapanja u rečenični kontekst, pa, čini se, hrvatskom čitatelju engleskog teksta jednostavno treba nešto više milisekundi da prizove značenje riječi koju čita jer čita na stranom jeziku. Pri tome se manje oslanja na predviđanja onog što u tekstu slijedi. Takvo bi tumačenje bilo u suglasnosti s rezultatima dobivenima u postojećim analizama ovog korpusa (Cergol i Palmović 2024a) gdje se analiza i temelji na ukupnoj duljini čitanja (koja je veća za engleski). U svakom slučaju, rezultati ove studije potvrđuju hipotezu implicitne prozodije. U oba je jezika zabilježeno veće zadržavanje pogleda na naglašenim slogovima, i to više u jeziku

u kojem je naglašeni slog istaknutiji od nenaglašenog. Treba imati na umu da u procesu razumijevanja teksta veliku ulogu imaju i razni drugi čimbenici koji se nisu mogli analizirati u ovoj studiji.

Metoda analize korpusa čitanja upotrijebljena u ovom radu može se upotrijebiti za bilo koju drugu kategoriju (npr. vrstu riječi). Potrebno je samo na odgovarajući način u tekstu označiti te kategorije kao područja interesa. Prikladna je i za bilo koje druge izmjerene varijable s pomoću drugih metoda – na primjer, na korpusu tekstova koji se čitaju metodom čitanja vlastitim tempom (engl. *self-paced reading*), koja se također upotrebljava za prikupljanje podataka koji nisu eksperimentalno manipulirani. Ono što neku varijablu predviđa objašnjava je, to je u najkraćim crtama logika takve analize. Pri tome se, kako je spomenuto, Bayesova logistička regresija pokazala nešto uspješnijom zbog toga što se za apriorne vjerojatnosti uzimaju vrijednosti koje se mogu izračunati na temelju jezične strukture teksta zato što se korpus može podijeliti na područja interesa prema nekom gramatičkom načelu (nažalost, samo prema jednom kriteriju).

Kao ograničenje ove studije može se spomenuti činjenica da sudionici nisu dodatno testirani u znanju engleskog s obzirom na to da su svi oni prethodno pristupili uniformiranom nacionalnom ispitu iz engleskog te postigli zadovoljavajuću razinu potrebnu za čitanje teksta na engleskom. Kao ograničenje se također može navesti i njihova neujednačenost po dijalektalnom podrijetlu u hrvatskom, što je doprinijelo većoj raspršenosti rezultata u hrvatskom, pogotovo zato što se mjesto naglaska razlikuje u hrvatskim dijalektima. U budućim se istraživanjima svakako treba kontrolirati dijalektalno podrijetlo sudionika. Podatak o tome uključen je u, na primjer, spomenuti višejezični korpus korpus *MultiplEye*. U taj je korpus uključen i velik broj drugih mjera kognitivne obrade (npr. radno pamćenje, kognitivna fleksibilnost, leksičko znanje). Nažalost, ti podatci nisu bili dostupni za analizu ovog korpusa čitanja. Metodologija korpusa čitanja brzo se razvija i brzo prelazi put od velikog skupa podataka dobivenih čitanjem tekstova (kao u ovom radu) do kompleksne baze podataka koja uključuje i razrađen skup podataka o socioekonomskom statusu sudionika, njihovu jezičnom statusu i rezultatima psihometrijskih testova.

Ova je studija izrađena u sklopu hrvatsko-švicarskog projekta HRZZ IPCH-2022-04-3316 *Measurement reliability of individual differences in sentence pro-*

cessing: A cross–linguistic perspective (MeRID) [Pouzdanost mjerenja individualnih razlika u rečeničnoj obradi: Međujezična perspektiva].

Literatura

Akaike, Hirotugu. 1998. Information theory and an extension of the maximum likelihood principle. *Selected papers of Hirotugu Akaike*. Springer New York. New York, NY. 199–213.

Ashby, Jane; Clifton, Charles Jr. 2005. The prosodic property of lexical stress affects eye movements during silent reading. *Cognition* 96. B89–B100.

Bates, Douglas; Mächler, Martin; Bolker, Ben; Walker, Steve. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67/1. 1–48. https://doi.org/10.18637/jss.v067.i01.

Beck, Judith; Konieczny, Lars. 2020. Rhythmic subvocalization: An eye-tracking study on silent poetry reading. *Journal of Eye Movement Research* 13/3. 1–40. https://doi.org/10.16910/jemr.13.3.5.

Berzak, Yevgeni; Nakamura, Chie; Smith, Amelia, Weng, Emily; Katz, Boris; Flynn, Suzanne; Levy, Roger. 2022. CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind* 6. 41–50. https://doi.org/10.1162/opmi_a_00054.

Breen, Maria; Clifton, Charles Jr. 2011. Stress Matters: Effects of Anticipated Lexical Stress on Silent Reading. *Journal of Memory & Language* 64/2. 153–170.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. Cambridge. https://doi.org/10.1017/9781316410899.

Bürkner Paul-Christian. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80/1. 1–28. https://doi.org/10.18637/jss.v080.i01.

Cergol, Kristina; Palmović, Marijan. 2024a. The role of prosodic information in silent reading: An eye–tracking study *Suvremena lingvistika* 50/97. 1–22. https://doi.org/10.22210/suvlin.2024.097.01.

Cergol, Kristina, Palmovic, Marijan. 2024b. The role of stress in silent reading. *ETRA '24: Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*. The Association for Computing Machinery, Inc. New York (NY). Article No. 83. 1–5. https://doi.org/10.1145/3649902.3656492.

Cop, Uschi; Dirix, Nicolas; Drieghe, Denis; Duyck, Wouter. 2016. Presenting geco: an eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods 49/2. 602–615. https://doi.org/10.3758/s13428-016-0734-0.

Cummins, Fred; Gers, Felix; Schmidhuber, Jürgen. 1999. *Comparing prosody across*

many languages. Instituto Dalle Molle di Studie sull'Intelligenza Artificiale, Lugano, Switzerland, Tech. Rep., IDSIA-07-99.

Cumming, Ruth; Wilson, Angela; Goswami, Usha. 2015. Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: a new look at a perceptual hypothesis. *Frontiers in Psychology* 6. 972. https://doi.org/10.3389/fpsyg.2015.00972.

Fodor, Janet Dean. 2002. Prosodic Disambiguation In Silent Reading. *Proceedings of the North East Linguistic Society* 32/1. 113–132.

Foltz, Anouschka. 2021. Using prosody to predict upcoming referents in L1 and L2: The role of recent exposure. *Studies in Second Language Acquisition* 43. 753–780. https://doi.org/10.1017/S0272263120000509.

Frazier, Lyn; Carlson, Keith; Clifton, Charles Jr. 2006. Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences* 10/6. 244–249. https://doi.org/10.1016/j.tics.2006.04.002.

Gontijo, Possidonia F. D; Gontijo, Isa; Shillcock, Richard. 2003. Grapheme-phoneme probabilities in British English. *Behavior Research Methods, Instruments & Computers* 35/1. 136–157.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routlege. New York.

Gries, Stefan Th. 2013. *Statistics for Linguistics with R: A Practical Introduction*. 2nd revised edition. Walter de Gruyter GmbH. Berlin.

Grüter, Theres; Rohde, Hannah; Schafer, Amy J. 2017. Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approach to Bilingualism* 7/2. 199–229. https://doi.org/10.1075/lab.15011.gru.

Hollenstein, Nora; Barrett, Maria; Björnsdóttir, Marina. 2022. The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources. Marseille, France. 1712–1720.

Hrvatska enciklopedija, mrežno izdanje. *Prozodija*. Leksikografski zavod Miroslav Krleža. 2013. – 2025. https://www.enciklopedija.hr/clanak/prozodija (pristupljeno 29. travnja 2025.).

Huey, Edmund Burke. 1908/1968. *The psychology and pedagogy of reading*. MIT Press. Cambridge, MA.

Josipović Smojver, Višnja. 1999. *Phonetics and phonology for students of English*. Targa, Zagreb.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd edition. Taylor & Francis Group. New York.

McEnery, Tony; Hardie, Andrew. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Munro Saki, Hector Hugh. 1993. *The Complete Stories of Saki*. Wordsworth Classics. Ware, Hertfordshire.

Munro Saki, Hector Hugh. 1984. *Sredni Vashtar i druge priče* [*Sredni Vashtar and Other Stories*]. Znanje. Zagreb.

Nahatame, Shingo; Ogiso, Tomoko; Kimura, Yukino; Ushiro, Yuji. 2024. Teco: an eye-tracking corpus of japanese l2 english learners' text reading. *Research Methods in Applied Linguistics* 3/2. 100123. https://doi.org/10.1016/j.rmal.2024.100123.

Orepić, Pavo. 2020. *Dissecting self-voice perception: From bone conduction to robotically induced self-other voice misattribution in healthy listeners*. Doctoral dissertation. École Polytechnique Fédérale de Lausanne. Switzerland.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Rayner, Keith; Clifton, Charles Jr. 2009. Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology* 80/1. 4–9. https://doi.org/10.1016/j.biopsycho.2008.05.002.

Scarton, Carolina; Specia, Lucia. 2016. A Reading Comprehension Corpus for Machine Translation Evaluation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16). European Language Resources Association (ELRA). Portorož, Slovenia. 3652–3658.

Slowiaczek, Maria L.; Clifton, Charles Jr. 1980. Subvocalization and reading for meaning. *Journal of Verbal Learning and Verbal Behavior* 19/5. 573–582. https://doi.org/10.1016/S0022-5371(80)90628-3.

Söderström, Pelle. 2017. *Prosody and prediction in neural speech processing*. Doktorska disertacija. Sveučilište u Lundu.

Wickham, Hadley i dr. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4/43. 1686. https://doi.org/10.21105/joss.01686.

Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. Routlege. New York.

Bilingual Reading Corpus Analysis: A Comparison Between Frequentist and Bayesian Approach

Abstract

The *reading corpus* constitutes a new methodology in psycholinguistic research, a methodology where eyetracking data are collected from experimentally non-manipu-

lated stimuli, i.e. texts. This paper discusses the appropriate statistical analysis method for this type of research. A logistic regression model, either frequentist or Bayesian, is proposed, as such a model is used for predicting categorical target variables, which most linguistically relevant categories are. The superiority of this analysis over previous analyses in confirming the Implicit Prosody Hypothesis has been demonstrated and the logistic regression models confirmed that we unconsciously follow the stress structure of the text even in silent reading. For this analysis the existing parallel Croatian-English corpus has been used. No substantial differences were found between the frequentist and Bayesian models, although Bayesian model has shown some advantages.

Ključne riječi: korpus čitanja, dvojezičnost, hipoteza implicitne prozodije, statistički modeli
Keywords: reading corpus, bilingualism, Implicit Prosody Hypothesis, statistical models