Jaka Čibej

Statistična analiza izgovora črke *l* v *Slovenskem oblikoslovnem leksikonu Sloleks*

Cobiss: 1.01
https://doi.org/10.3986/JZ.31.1.03

Dvoumnost izgovora črke *l* v položaju pred soglasniškim grafemom (*polža*, *alge*, *volilca*) predstavlja problem v grafemsko-fonemski pretvorbi za slovenščino in kljub večkratni obravnavi v slovenskih jezikovnih virih še vedno ni razrešena. Zaradi pomanjkanja empiričnih strojno berljivih podatkov o izgovoru rojenih govorcev in govork smo ročno označili množico približno 6.000 leksemov (in njihovih pregibnih oblik) iz *Slovenskega oblikoslovnega leksikona Sloleks* glede na izgovor (/l/, /ụ/ ali variantno) ter opravili statistično analizo, ki razkrije najbolj problematične točke. Izsledki bodo v pomoč pri razvoju modela, ki napoveduje izgovor črke *l* pred soglasniškim grafemom.
Ključne besede: izgovor črke *l*, grafemsko-fonemska pretvorba, oblikoslovni leksikon, slovenščina, statistična analiza

A Statistical Analysis of the Pronunciation of the Letter *l* in the *Sloleks Morphological Lexicon of Slovene*

The ambiguous pronunciation of the letter *l* before consonant graphemes (e.g., *polža*, *alge*, *volilca*) poses a problem for grapheme-to-phoneme conversion for Slovenian. Despite having been addressed in multiple Slovenian language resources, the problem is still unresolved. Because of a lack of empirical and machine-readable data on the pronunciation of native speakers, a dataset of approximately six thousand lexemes (and their inflected forms) was manually annotated from the *Sloleks Morphological Lexicon of Slovene* according to their pronunciation (/l/, /ụ/, or both) and a statistical analysis was performed to reveal the most problematic points. The findings will be useful in developing a model to predict the pronunciation of the letter *l* before a consonant grapheme.
Keywords: pronunciation of the letter *l*, grapheme-to-phoneme conversion, morphological lexicon, Slovenian, statistical analysis

Jaka Čibej ▪ Univerza v Ljubljani, Filozofska fakulteta ▪ jaka.cibej@ff.uni-lj.si ▪
https://orcid.org/0000-0002-3037-6848

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS), ter projekta *Razvoj slovenščine v digitalnem okolju* (RSDO), ki sta ga med letoma 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Iskrena hvala tudi vsem označevalcem in označevalkam.

1 Uvod

S hitrim razvojem jezikovnih tehnologij se pojavljajo vedno večje potrebe po strojno berljivih podatkih o izgovoru slovenskih besed, zlasti za govorna orodja, kot so razpoznavalniki (npr. Dobrišek idr. 2023) in sintetizatorji govora (Šef 2006; Žganec Gros idr. 2020). Plastičen primer je denimo v letu 2024 zaključeni pilotni projekt *Online Notes*[1] (Bajec idr. 2023; Šoltes idr. 2024), katerega cilj je bila gradnja sistema za sprotno prevajanje slovenskih univerzitetnih predavanj iz slovenščine v angleščino. *Online Notes* ima v svoj cevovod implementiran tudi razpoznavalnik govora; ta se med drugim zanaša na nabor fonetičnih transkripcij besed (tudi npr. terminoloških izrazov in lastnih imen, uporabljenih med predavanji), ki izhaja (tudi) iz *Slovenskega oblikoslovnega leksikona Sloleks* (Čibej idr. 2022), največje odprto dostopne zbirke s strojno berljivimi informacijami o slovenskih besedah. Ta zbirka je podlaga tudi za raziskavo v tem prispevku (več o tem v razdelku 3). Ročno dopolnjevanje in vnašanje izgovorov je zamudno delo, zlasti če upoštevamo, da je treba fonetične transkripcije eksplicitno zapisovati tudi za vse pregibne oblike posameznega leksema.

Z vključevanjem vedno novega besedišča v leksikon (npr. iz spremljevalnega korpusa *Trendi*; Kosem 2022) ročno dodajanje fonetičnih transkripcij ni vzdržno. Strojne metode lahko posodabljanje leksikona znatno pospešijo – za tuje jezike so bili v ta namen že uporabljeni generatorji transkripcij v mednarodni fonetični abecedi IPA na podlagi nevronskih mrež (za francoščino gl. npr. Marjou 2021; za angleščino npr. Reichel idr. 2008). Grafemsko-fonemska pretvorba s strojnim učenjem je bila preizkušena tudi za slovenščino (Križaj idr. 2022b),[2] istočasno pa je bil za posodabljanje leksikona *Sloleks* razvit *Pregibalnik*,[3] odprto dostopno orodje za strojno generiranje pregibnih in naglašenih oblik ter fonetičnih transkripcij. Grafemsko-fonemski pretvornik *Pregibalnika* naglašene oblike pretvarja tako v mednarodno fonetično abecedo IPA kot v njen ekvivalent X-SAMPA, deluje pa na podlagi jezikoslovno podprtih pravil. Slovenščina je namreč za razliko od npr. francoščine in angleščine pravopisno bolj plitek oz. transparenten jezik (Pogačnik 2012; Erbeli – Pižorn 2017; Schüppert idr. 2017), v katerem je izgovor na podlagi grafemov lažje napovedljiv (Zorman 2007; Filipič 2016), ne pa povsem – fonetični zapis se npr. pojavlja pri samoglasnikih *a*, *i* in *u*, ne pa tudi pri samoglasnikih, ki so zapisani z *e* in *o* (polglasnik celo z *r* – *smrt*, *čmrlj*); o pravopisnih načelih v

---

1 Spletna stran projekta *Online Notes*: https://www.cjvt.si/online-notes/.

2 Omeniti je treba, da je grafemsko-fonemska pretvorba s strojnim učenjem (Križaj idr. 2022b) evalvirana prav na *Slovenskem oblikoslovnem leksikonu Sloleks 3.0*, pri katerem so problemi z izgovorom črke *l* še nerazrešeni.

3 Pregibalnik je odprto dostopno na voljo na platformi Github (https://github.com/clarinsi/SloInflector) ali kot javni API (https://orodja.cjvt.si/pregibalnik/docs).

slovenskem jeziku podrobneje piše Unuk (2009). Pri večini besed je izgovor v standardni slovenščini (izjema so besede, ki jih je treba izgovarjati po drugih pravilih grafemsko-fonemske pretvorbe, npr. *sommelier*, *Shakespeare*; gl. npr. Čibej 2024) mogoče napovedati na podlagi grafemov v naglašeni obliki. Po preliminarnih evalvacijah pretvornik *Pregibalnika* dosega 98-odstotno točnost (Čibej 2023: 9), saj tudi v standardni slovenščini obstajajo nebijektivna razmerja med črkami in glasovi: tak primer je npr. izgovor soglasniškega grafema *l* v položajih pred soglasniškim grafemom (*polža* /ˈpɔːu̯ʒa/ vs. *alge* /ˈaːlgɛ/ vs. *volilca* /ʊɔˈliːltsa/ ali /ʊɔˈliːu̯tsa/).[4] Dilema je bila do zdaj večkrat obravnavana v slovenskih jezikovnih virih in jezikoslovnih raziskavah (gl. razdelek 2), še vedno pa je problem pomanjkanje empiričnih podatkov o izgovoru oz. vsaj percepciji standardnega izgovora rojenih govorcev in govork. Trenutna različica pretvornika vse izgovore v tovrstnih položajih obravnava kot /l/, kar je neustrezno. Da bi orodje in leksikon pri tem izboljšali, smo ročno označili množico približno 6.000 leksemov (in njihovih pregibnih oblik) iz *Sloleksa* glede na izgovor (/l/, /u̯/ ali variantno) ter opravili statistično analizo, da bi empirično odkrili najbolj problematične točke oz. potencialne izsledke, ki bi lahko pomagali pri razvoju modela, ki napoveduje izgovor črke *l* pred soglasniškim grafemom.

Prispevek je strukturiran na naslednji način: po kratkem pregledu sorodnih raziskav (razdelek 2) predstavljamo vir podatkov (razdelek 3) in metodo označevanja (razdelek 4). Označene podatke (ki so odprto dostopni na repozitoriju CLARIN.SI) statistično analiziramo (razdelek 5) in članek sklenemo s poglavitnimi ugotovitvami in koraki za nadaljnje delo (razdelek 6).

2 Sorodne raziskave

Izgovor črke *l* je že precej dolgo predmet jezikoslovnih raziskav, kot eno od predlaganih rešitev za dilemo izgovora pa lahko med drugim npr. omenimo pravilo iz *Slovenskega pravopisa* 1935, ki razlikuje med domačimi uveljavljenimi besedami (*bralec -lca* /ˈbɾaːu̯tsa/) in novejšimi besedami, ki se uporabljajo večinoma le v knjižnem jeziku (*metalec -lca* /mɛˈtaːltsa/). Na dilemo izgovora[5] črke *l* v slovenščini je opozoril tudi Rigler (1980) pri obravnavi problematike izgovora v *Slovarju slovenskega knjižnega jezika*, v kateri prav tako omenja pomanjkanje empiričnih podatkov in težave pri njihovem pridobivanju. V *Slovenskem pravopisu* 2001 je npr. navedeno, da načeloma v vseh izglagolskih izpeljankah s priponskimi obra-

4 Problem predstavlja tudi npr. soglasniški grafem *l* v končnem položaju – *pol* /ˈpɔːu̯/ *(potice)* vs. *(severni) pol* /ˈpoːl/ –, a se v tem prispevku osredotočamo na položaj pred drugimi soglasniškimi grafemi.

5 Omeniti je treba, da je dilema v nekaterih primerih prisotna tudi na nivoju zapisa (npr. *volivka*/*volilka*); za korpusno obravnavo te problematike gl. npr. Sever 2023.

zili *-lc-*, *-lk-*, *-lsk-*, *-lstv-* (ter v nadaljnjih izpeljankah iz njih), ko se nanašajo na živega vršilca dejanja, pričakujemo izgovor z /u̯/: *bralca*, *bralka*, *bralski*, *bralstvo* [bráu̯ca, bráu̯ka, bráu̯ski, bráu̯stvo]; izgovor z /l/ pa npr. v SP 2001 ni naveden, čeprav je v SSKJ predviden, a običajno manj priporočljiv (npr. *kopalka -e* [u̯k in lk]). V vmesnem času je z razvojem spleta in digitalnih metod v jezikoslovju anketiranje (gl. npr. Mirtič 2019) postalo lažje (a še vedno časovno zahtevno) – nekatere rezultate perceptivnih testov izgovora črke *l* denimo opisuje Mirtič (2021: 57), a je bila raziskava omejena le na okrog 20 samostalnikov, ki so bili do tedaj leksikografsko obravnavani med pripravo eSSKJ. Raba je kompleksna in nakazuje, da izgovor ni lahko določljiv na podlagi kategorij na pomenski ravni, kot je npr. vršilec dejanja.

Kljub zahtevnosti zbiranja podatkov o izgovoru pa za slovenščino že obstajata dve odprto dostopni množici s strojno berljivimi informacijami o izgovoru: poleg že omenjenega leksikona Sloleks velja omeniti tudi *Slovar izgovarjav OptiLEX* (Žganec Gros idr. 2022), ki ponuja podatke v zapisu MRPA, osnova zanj pa so podatki o slovenskih iztočnicah s pregibnimi oblikami iz druge izdaje *Slovarja slovenskega knjižnega jezika* (SSKJ2). Pregled nekaterih problematičnih iztočnic sicer razkrije, da je izgovor črke *l* tudi v OptiLEXu obravnavan nekoliko nekonsistentno – npr. varianten izgovor za obliko *čistilcu* (/tSist“i:ltsu/, /tSist“i:Utsu/) in zgolj izgovor /u̯/ za obliko *čistilk* (/tSist“i:Uk/); perceptivni testi (Mirtič 2021: 57) denimo kažejo na prevladujoči izgovor /l/ prav pri leksemu *čistilka*.

Izgovor črke *l* obravnava tudi v času pisanja tega prispevka še nastajajoči *Pravopis 8.0*, ki med drugim navaja, da lahko izgovor črke *l* z /u̯/ v položaju pred soglasniškim grafemom pričakujemo na mestu nekdanjega zlogotvornega *ł*, zlasti v kombinaciji *ol* (npr. *bolha*, *solza*). Varianten izgovor je po drugi strani značilen za posamezne samostalnike (npr. *del*, *kolk*) ter pri izglagolskih samostalniških in pridevniških izpeljankah s priponskimi obrazili *-lc-*, *-lk-*, *-lsk-*, *-lstv-*, ne glede na to, ali se nanašajo na živega vršilca dejanja ali ne (*gledalca*, *poslušalka*). Oba načina izgovora sta navedena tudi za pridevnike, ki so tvorjeni iz teh samostalnikov (*tožilčev*, *tožilkin*). *Pravopis 8.0* opozori tudi na izjeme, pri katerih je mogoč zgolj izgovor z /l/ (*gasilci*), besede, pri katerih podstava ni izglagolska (z /l/: *Brazilci*, *Portugalke*, *metalci* v smislu *navdušenci nad metalom*), in posebne primere, ki izhajajo iz lastnih (osebnih ali zemljepisnih) imen, kjer je izgovor odvisen od rodbinskih navad (*Gal* [gál/gáu̯], *Budal* [budál/budáu̯]) oz. od lokalne rabe (*Kalše* z /l/).

V nadaljevanju predstavljamo označevanje in statistično analizo množice označenih leksemov, ki jo statistično analiziramo in na podlagi izsledkov preverjamo tudi nekatere trditve iz *Pravopisa 8.0*.

3 Podatki

Podatke za označevalni eksperiment smo črpali iz *Sloleksa 3.0* (Čibej idr. 2022), ki vsebuje približno 365.000 leksemov, od katerih je približno 100.800 ročno pregledanih. Iz nabora ročno pregledanih leksemov smo najprej izvozili vse lekseme, pri katerih se je v vsaj eni pregibni obliki pojavila črka *l* neposredno pred soglasniškim grafemom (v nadaljevanju to kombinacijo imenujemo bigram *lC*, tj. črka *l* + *C* kot simbol za soglasniški grafem). Za primer: pridevnik *pepelen* v osnovni obliki nima bigrama *lC*, a se ta pojavi v vseh ostalih pregibnih oblikah (*pepe**ln**ega*, *pepe**ln**emu* itd.). Pri luščenju smo upoštevali tudi število bigramov *lC* v vseh oblikah, v posebno kategorijo pa smo izvozili lekseme s kombinacijo *ll*. Rezultati prvega izvoza so prikazani v preglednici 1, iz katere je razvidno, da ima večina leksemov v pregibnih oblikah le po en bigram *lC*, le okrog 3 % pa po dva ali več. V okrog 2 % izluščenih leksemov najdemo bigram *ll*, večinoma v tujih lastnih imenih (*Hillary*) in svojilnih pridevnikih (*Isabellin*) ter redkih drugih primerih (*polleten*, *pollitrski*).

Preglednica 1: Leksemi z vsaj enim bigramom *lC* v naboru ročno pregledanih leksemov iz *Sloleksa 3.0*

Kategorija	Število leksemov	Delež	Primeri
Leksemi z enim bigramom *lC*	6.650	94,72 %	*absolventka*, *parcialnost*, *pepelen*
Leksemi z več bigrami *lC*	221	3,14 %	*dopolnilen*, *tolkalec*, *multikulturen*
Leksemi z bigramom *ll*	150	2,14 %	*Agnelli*, *bestseller*, *polleten*
∑	7.021	100,00 %	-

Iz logističnih razlogov pri razvoju označevalnega vmesnika (gl. razdelek 4) smo se glede na majhno število primerov z več bigrami *lC* osredotočili na primere z enim samim bigramom *lC*. Zanemarili smo tudi primere z bigramom *ll* ter lastnoimenske samostalnike in svojilne pridevnike, ki se večinoma ne izgovarjajo po pravilih slovenske grafemsko-fonemske pretvorbe. Končni nabor za označevanje je tako zajemal 5.990 leksemov in vse njihove pregibne oblike, ki so vsebovale vsaj en bigram *lC* (skupno 173.419 oblik). Število leksemov in pregibnih oblik glede na njihove osnovne oblikoskladenjske značilnosti (po sistemu Multext East v6)[6] je prikazano v preglednici 2.

6 Oblikoskladenjske specifikacije za slovenščino po sistemu Multext East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

Preglednica 2: Končni nabor leksemov in oblik za označevanje glede na osnovne oblikoskladenjske značilnosti po sistemu Multext East v6

Značilnosti	Leksemi	Delež (%)	Oblike	Delež (%)
Samostalnik, občni, moški	1.641	27,4	28.255	16,29
Samostalnik, občni, ženski	1.614	26,94	28.279	16,31
Samostalnik, občni, srednji	230	3,84	3.530	2,04
Glagol, glavni, dvovidski	46	0,77	1.151	0,66
Glagol, glavni, nedovršni	39	0,65	993	0,57
Glagol, glavni, dovršni	54	0,90	1.344	0,78
Pridevnik, splošni	1.771	29,57	105.968	61,11
Pridevnik, deležniški	54	0,90	3.078	1,77
Prislov, splošni	534	8,91	814	0,47
Prislov, deležje	5	0,08	5	<0,01
Predlog, rodilnik	2	0,03	2	<0,01
∑	**5.990**	**100,00**	**173.419**	**100,00**

Za namen navzkrižnih primerjav pri statistični analizi smo leksemom in oblikam vnaprej pripisali tudi nekatere metapodatke: vrsto bigrama *lC* (preglednica 3), predhodni grafem (preglednica 4) in morebitne relevantne končne besedne dele leksemov, ki nastopajo v oblikah in vsebujejo vsaj en bigram *lC* (preglednica 5).

Preglednica 3: Najpogostejši bigrami *lC* v naboru oblik za označevanje

Bigram *lC*	Število oblik	Delež (%)
ln	91.096	52,53
ls	15.979	9,21
lk	15.656	9,03
lc	13.900	8,02
lt	9.711	5,60
lg	4.728	2,73
lč	4.176	2,41
Drugo[7]	18.173	10,48
∑	**173.419**	**100,00**

V bigramih *lC* se črka *l* daleč najpogosteje pojavlja pred črko *n*. Sledijo še *s*, *k* in *c* s podobnimi deleži oblik, ostali bigrami pa so redkejši. Kombinacije, ki niso navedene v preglednici 3, vsebujejo grafeme *m*, *p*, *z*, *d*, *v*, *ž*, *š*, *b*, *f*, *h*, *y*, *r* in *j*; vse imajo delež manj kot 1,5 %.

7 V nekaterih tabelah v prispevku se zaradi prostorskih omejitev omejujemo na najpogostejše primere. Podrobnejši prerez je na voljo v vnosu na repozitoriju CLARIN.SI: Čibej 2024b.

Preglednica 4: **Najpogostejši grafemi pred bigrami *lC* v naboru oblik za označevanje**

Bigram lC	Število oblik	Delež (%)
a	88.080	50,79
i	32.983	19,02
o	31.354	18,08
e	12.790	7,38
u	7.623	4,40
j	225	0,13
Drugo	364	0,20
∑	**173.419**	**100,00**

V veliki večini primerov (99,67 %) je v obliki neposredno pred bigramom *lC* samoglasniški grafem, najpogosteje *a*, ki zajema več kot polovico primerov. Sledita *i* in *o* z 18–19 %, najredkejša pa sta *e* in *u*.

Preglednica 5: **Najpogostejši končni besedni deli leksemov, ki vsebujejo bigram *lC***

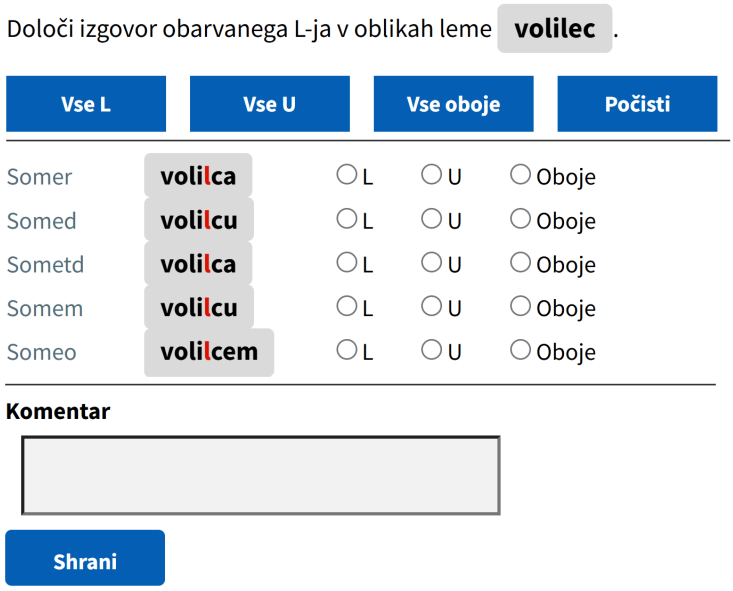
Končni besedni del	Število oblik	Delež (%)
-alen	41.965	24,20
-ilen	16.170	9,32
-alec	9.174	5,29
-alka	7.890	4,55
-alski	5.445	3,14
-alnik	3.774	2,18
Drugo	29.138	16,80
Brez	59.864	34,52
∑	**173.419**	**100,00**

Približno tretjino oblik pokrivata pridevniški *-alen* in *-ilen*, okrog 10 % pa samostalniška *-alec* in *-alka*. Oblike smo delili na 46 različnih besednih delov – med neomenjenimi npr. še *-alnost*, *-elen*, *-elski*, *-ilski*, *-ilec*, *-alnica*, *-ilka*, *-ilnik*, *-ilnost*, *-ilnica*, *-elka* ipd. Ostali deli so redkejši in zajemajo manj kot 0,30 % vseh oblik.

4 Označevanje

Označevanje pripravljenih podatkov je potekalo v okolju PyBossa.[8] Razvili smo poseben vmesnik (slika 1), v katerem je bilo mogoče označiti, ali se podane oblike izgovarjajo z /l/, /ụ/ oz. ali je mogoč varianten izgovor.

8 Dokumentacija za platformo PyBossa: https://docs.pybossa.com/ (dostop: 20. 12. 2024).



Slika 1: Vmesnik za označevanje izgovora črke *l* v okolju PyBossa

Označevalci in označevalke so dobili navodilo, naj oblike označijo glede na to, kateri izgovor bi glede na lastno jezikovno intuicijo pričakovali kot najbolj nezaznamovan v standardni slovenščini, kakršno bi npr. slišali v televizijskih in radijskih oddajah. Pri tem je sodelovalo pet označevalcev in označevalk (štiri označevalke in en označevalec, starostni razpon med 25 in 35 let), ki se na Centru za jezikovne vire in tehnologije Univerze v Ljubljani ukvarjajo z gradnjo digitalnih slovarskih virov. Koncept se zanaša na jezikovno intuicijo podobno kot perceptivni testi s spletnim anketiranjem, ki so bili za podobne dileme v slovenščini uporabljeni že večkrat (gl. npr. Mirtič 2019; Tivadar 2004); v našem primeru zaradi velike količine gradiva nismo pripravljali posnetkov in primerov rabe, zato so označevalci ocenjevali le izolirane leksikonske oblike. Omeniti je treba tudi, da je označevanje potekalo še pred objavo glasoslovnega dela *Pravopisa 8.0*, zato ocenjevanje izgovora še ni odsevalo novih smernic. Poleg tega označevalci v vmesniku niso imeli podanih dodatnih metapodatkov (npr. vrsta bigrama *lC*, končni besedni del), saj s prikazovanjem podatkov nismo hoteli vplivati na njihove odločitve. V označevanje bi bilo smiselno vključiti še dodatne označevalce, tudi tiste, ki so neobremenjeni z jezikoslovnim vprašanjem izgovora črke *l*. Naša raziskava je po gradivu precej obsežna, zato smo dodatne oznake, ki bi ponudile še bolj realno sliko, pustili za prihodnje delo.

Vsak leksem oz. njegove oblike sta označila po dva označevalca v različnih kombinacijah. Preglednica 6 prikazuje vrednosti Cohenove kape (κ; Cohen 1960),[9] koeficient ujemanja med pari označevalcev – pripisano je tudi število oblik, ki so skupne vsakemu paru.

Preglednica 6: Vrednosti Cohenove kape za pare označevalcev

Označevalski par	Število skupnih oblik	κ
A – B	84.810	0,76
A – C	21.697	0,85
A – D	65.924	0,56
A – E	988	0,79

Večina parov glede na koeficient dosega dobro ujemanje, nekoliko nižji koeficient, ki nakazuje srednje ujemanje, je le v primeru označevalcev A in D. Več podrobnosti o ujemanju pri različnih podskupinah obravnavanih oblik (npr. glede na vrsto bigrama *lC* in končni besedni del) opisujemo v razdelku 5.

Označevanje posameznega leksema je v povprečju trajalo 39 sekund (pri polovici manj kot 13 sekund; najmanj približno 3,5 sekunde; največ približno 303 sekunde), za označevanje celotnega nabora podatkov pa je bilo porabljenih približno 130 ur, kar znese v povprečju 26 ur na označevalca. Označevalna kampanja je trajala približno en mesec (od junija 2021 do julija 2021), v povprečju torej manj kot uro označevanja na dan, s čimer smo se izognili utrujenosti in manj zanesljivim odgovorom pri označevanju. Podatki o trajanjih označevanja posameznega leksema so na voljo tudi v končni množici, kar npr. omogoča nadaljnje filtriranje po tistih leksemih, pri katerih so se označevalci odločali najtežje.

5 Analiza

Po končanem označevanju je bilo zbranih 346.838 oznak za posamezne oblike: 247.783 (71,44 %) odgovorov je nakazovalo izgovor /l/, 56.479 (16,28 %) odgovorov izgovor /u̯/, 42.576 (12,28 %) odgovorov pa je dopuščalo varianten izgovor. Če si ogledamo kombinacije odgovorov po parih za posamezne lekseme, je razvidno, da se označevalci v svojih odločitvah ujemajo pri 5.228 leksemih (87,28 %), le pri 762 (12,72 %) so se razhajali. Podrobnejši prerez prikazuje preglednica 7 – pri leksemih, pri katerih pride do razhajanja, je v večini primerov en označevalec, ki

9 Cohenova kapa za razliko od odstotka ujemanja upošteva končno razporeditev odgovorov in verjetnost, da se označevalci ujemajo naključno (zlasti v primerih, ko ena oznaka v množici močno prevladuje v primerjavi z ostalimi). Za interpretacijo se pogosto uporablja naslednje orientacijske vrednosti (gl. npr. McHugh 2012): ≤ 0 (brez ujemanja), 0,01–0,20 (nizko ujemanje), 0,21–0,40 (delno ujemanje), 0,41– 0,60 (srednje ujemanje), 0,61–0,80 (dobro ujemanje), 0,81–1,00 (zelo dobro ujemanje).

dopušča varianten izgovor, drugi pa zagovarja le eno od možnosti. Le v peščici primerov se označevalca nikakor ne strinjata in kot nezaznamovana dojemata povsem različna izgovora.

Preglednica 7: Prerez leksemov glede na kombinacijo oznak

Par oznak	Število leksemov	Delež (%)	Primeri
/l/ | /l/	3.710	61,94	*dotikalen*, *mobilnik*, *konzultacija*
/u̯/ | /u̯/	889	14,84	*polnoletnik*, *polmrak*, *stolči*
variantno | variantno	629	10,50	*topilec*, *šepetalka*, *trosilec*
variantno | /l/	333	5,56	*zavarovalstvo*, *nebelka*, *soldaški*
variantno | /u̯/	359	5,99	*ščipalec*, *negovalka*, *zaničevalec*
/l/ | /u̯/	70	1,17	*vogalen*, *darovalski*, *spreminjevalček*
∑	**5.990**	**100,00**	**-**

Da bi podrobneje raziskali, kje so najpogostejša mesta razhajanja, smo izračunali ujemanje med označevalci glede na kombinacijo bigrama *lC* in predhodnega grafema – npr. koliko se označevalci ujemajo pri leksemih, ki vsebujejo kombinacije *olf*, *ulc*, *alč* ipd. Od skupno 87 tovrstnih kombinacij jih je bilo 50 povsem neproblematičnih, saj so se označevalci pri njih povsem ujemali, tj. vsi so označili vse relevantne lekseme z eno samo oznako – pri 43 kombinacijah je bil označen izključno izgovor /l/ (npr. *ulk – vulkanski*, *alv – devalvirati*, *ulz – impulziven*, *alz – alzaški*, *ulc – krulec*), pri šestih kombinacijah izključno /u̯/ (npr. *olb – vdolbenost*, *olz – odpolzeti*, *alh – malha*, *olž – molža*) in pri eni kombinaciji varianten izgovor (*rlm* – oblike samostalnika ženskega spola *škrl*; *škrlmi, škrlma*).

Pri 13 kombinacijah slika ni bila povsem enoznačna, a so označevalci kljub temu dosegli visoko raven ujemanja. Kot prikazuje preglednica 8, je pri nekaterih kombinacijah pogostejši izgovor /l/, pri drugih /u̯/, v določenih primerih pa sta pogostosti izgovorov za dano kombinacijo precej izenačeni (tj. približno enako število leksemov s to kombinacijo ima izgovor /l/ in /u̯/). Visoko ujemanje v teh primerih nakazuje, da pri teh kombinacijah v slovenščini glede izgovora ni veliko dvomov – skoraj vse kombinacije bigrama *lC* s predhodnim samoglasniškim grafemom *o* so bodisi povsem neproblematične bodisi izkazujejo zelo visoko ujemanje; nekoliko nižje je le pri *olš* (*golšarica*; 0,63) in *olg* (*dolg*; 0,73).[10]

10 Tudi *Pravopis 8.0* navaja, da je zapis dvoglasniške dvoustnične variante /u̯/ zapisan s črko *l* sredi besede, ko gre za črkovni sklop *ol*, ki zaznamuje zgodovinski zlogotvorni *ł*; zaradi tega je izgovor v tem kontekstu lažje določljiv.

Preglednica 8: Kombinacije črk z zelo dobrim ujemanjem

Kombinacija in primer	/l/	/ụ/	/l/ in /ụ/	Število oblik	κ
ols (*polstiti*)	4.921 (72,47 %)	1.608 (23,68 %)	261 (3,84 %)	3.395	0,91
olk (*premolkniti*)	2.124 (47,82 %)	2.250 (50,65 %)	68 (1,53 %)	2.221	0,97
olp (*kolportažen*)	1.242 (42,62 %)	1.672 (57,38 %)	-	1.457	0,97
olč (*komolčarski*)	254 (6,53 %)	3.430 (88,22 %)	204 (5,25 %)	1.944	0,91
old (*livoldski*)	821 (33,98 %)	1522 (63,00 %)	73 (3,02 %)	1.208	0,88
oln (*monopolni*)	5.707 (36,90 %)	9.502 (61,43 %)	259 (1,67 %)	7.734	0,96
olm (*olmeški*)	1.100 (49,64 %)	1.098 (49,55 %)	18 (0,81 %)	1.108	0,83
olv (*revolverski*)	982 (79,07 %)	260 (20,93 %)	-	621	1,00
olt (*svetlopolt*)	1.381 (30,69 %)	3.119 (69,31 %)	-	2.250	0,94
olh (*bolha*)	90 (11,78 %)	656 (85,86 %)	18 (2,36 %)	382	0,81
olc (*četrtošolci*)	777 (58,51 %)	492 (37,05 %)	59 (4,44 %)	664	0,93
olf (*golf*)	292 (73,00 %)	108 (27,00 %)	-	200	1,00
alp (*predalpski*)	1.722 (97,95 %)	-	36 (2,05 %)	879	1,00

Pri določenih kombinacijah je bila Cohenova kapa zelo nizka (preglednica 9) – odvisno od razporeditve oznak to pomeni neujemanje pri večjem številu leksemov ali pa neujemanje pri izjemah v sicer zelo enoznačni kategoriji. Posebej je treba izpostaviti zelo pogosti kombinaciji *aln* in *iln*, pri katerih je večina oznak v prid izgovora /l/, precej redkeje pa so se označevalci odločali za /ụ/ ali za varianten izgovor, a se pri teh izjemah niso strinjali.

Preglednica 9: Kombinacije črk, pri katerih je ujemanje nizko

Kombinacija in primer	/l/	/ụ/	/l/ in /ụ/	Število oblik	κ
elš (*jelševje*)	805 (88,07 %)	109 (11,93 %)	-	457	-0,03
alč (*vogalček*)	911 (65,63 %)	274 (19,74 %)	203 (14,63 %)	694	0,07
alš (*kalški*)	649 (89,89 %)	-	73 (10,11 %)	361	-0,08
aln (*počivalen*)	101.119 (93,05 %)	256 (0,24 %)	7.297 (6,71 %)	54.336	-0,03
iln (*sprožilen*)	42.816 (95,56 %)	73 (0,16 %)	1.917 (4,28 %)	22.403	0,15

Največ leksemov z neujemanjem je bilo pri kombinaciji *aln* (184 leksemov oz. 11,70 % vseh leksemov s to kombinacijo; *splakovalnik*, *počivalen*) in *iln* (50 leksemov oz. 8,13 %; *škropilnik*, *kosilnica*). Pri ostalih kombinacijah je problematičnih leksemov manj: šest (*elš*; *jelševka*, *jelševina*), 14 (*alč*; *plezalček*, *kanalček*), dva (*alš*; *kalški*, *kanalščina*).

Preglednica 10 prikazuje kombinacije, pri katerih je ujemanje med označevalci delno – med najpogostejšimi izstopata *ilk* in *alc*, ki izstopata tudi po številu leksemov, pri katerih je prišlo do neujemanja: 163 oz. 28,54 % vseh relevantnih leksemov (*alc*; *stiskalec*, *prečiščevalec*) ter 39 oz. 34,51 % (*ilk*; *samomorilka*,

delilka). Pri ostalih besednih delih gre za posamezne izjeme (npr. *belkast*, *pogorelka* za *elk*; *koralda* za *ald*; *rilček*, *rilčkar* za *ilč*; *pogorelček* za *elč*).

Preglednica 10: Kombinacije črk, pri katerih je ujemanje delno

Kombinacija in primer	/l/	/u̯/	/l/ in /u̯/	Število oblik	κ
ilč (*rilčar*)	1.930 (94,70 %)	36 (1,77 %)	72 (3,53 %)	1.019	0,32
elč (*vozelček*)	554 (79,37 %)	54 (7,74 %)	90 (12,89 %)	349	0,41
ald (*koralda*)	1.132 (96,92 %)	18 (1,54 %)	18 (1,54 %)	584	0,49
ilk (*dražilka*)	1.376 (34,06 %)	756 (18,71 %)	1.908 (47,23 %)	2.020	0,45
alc (*tipalec*)	2.332 (12,02 %)	5.140 (26,50 %)	11.926 (61,48 %)	9.699	0,47
elk (*kiselkast*)	1.631 (85,39 %)	90 (4,71 %)	189 (9,90 %)	955	0,40

Ker tako naša analiza kot priporočila v *Pravopisu 8.0* kot problematične za izgovor črke *l* navajajo določene končne besedne dele, smo opravili še statistično analizo razporeditve oznak glede na skupno 46 končnih besednih delov, za katere je bilo v leksikonu dovolj gradiva (z dodatno kategorijo za lekseme z bigrami *lC*, ki niso izkazovali nobenega od končnih besednih delov). Pri dvanajstih delih so bile oznake povsem enoznačne in so nakazovale izgovor z /l/: *-elce*, *-elničar*, *-ilničar*, *-elnina*, *-elnost*, *-alnina*, *-ilnosten*, *-elstvo*, *-elsko*, *-alničar*, *-elskost*, *-lka* (npr. *polka*; ne vključuje delov *-ilka*/*-alka*/*-elka*). Na preostalih 35 kategorijah smo opravili test χ^2, ki pokaže statistično pomembne razlike v razporeditvi oznak za izgovor črke *l* (χ^2 = 187.823,655; N = 343.376; df = 68; p ≤ 0,0001; Cramerjev V: 0,52). Poračunali smo Pearsonove preostanke,[11] da bi določili, za katere besedne dele je statistično značilnejši določen izgovor – rezultate prikazuje preglednica 11 (vsaka celica za izgovor vsebuje število oznak, pod njim pa Pearsonov preostanek).

Preglednica 11: Pogostnosti oznak in vrednosti Pearsonovih preostankov za različne končne besedne dele

Besedni del	/l/	/u̯/	/l/ in /u̯/	κ
-ilnost	2.736	0	72	≤ 0
	+16,51	–21,49	–14,80	
-alnost	7.326	0	54	≤ 0
	+28,63	–34,84	–28,47	
-alec	1.368	5.125	11.855	0,41
	–102.29	+38.36	+200,85	

11 Pearsonovi preostanki z absolutno vrednostjo najmanj 1,92 nakazujejo, da je določena lastnost glede na vzorec statistično značilnejša za določeno kategorijo. Vrednosti pod –1,92 kažejo, da je lastnost za določeno kategorijo neznačilna, vrednosti nad +1,92 pa obratno.

Besedni del	/l/	/u̯/	/l/ in /u̯/	κ
	12	0	2	
-ilsko				1,00
	+0,65	–1,52	+0,20	
	265	0	3	
-ilno				≤ 0
	+5,38	–6,64	–5,24	
	162	36	90	
-elček				0,35
	–3,00	–1,65	+9,09	
	54	18	0	
-alskost				≤ 0
	+0,39	+1,79	–2,99	
	270	18	36	
-ilček				0,63
	+2,60	–4,83	–0,66	
	504	0	48	
-elnica				1,00
	+5,61	–9,53	–2,47	
	504	72	144	
-elnik				0,78
	–0,37	–4,27	+5,79	
	2.478	5.058	8.244	
-alka				0,43
	–82,58	+48,34	+142,14	
	1.152	756	1.908	
-ilka				0,42
	–30,00	+5,12	+65,96	
	810	90	72	
-elka				0,75
	+4,50	–5,53	–4,42	
	522	0	54	
-alce				≤ 0
	+5,54	–9,73	–2,06	
	2.556	18	162	
-ilnica				0,16
	+13,81	–20,37	–9,62	
	35	5	10	
-alsko				0,66
	–0,10	–1,12	+1,53	
	343	72	91	
-alček				0,01
	–0,90	–1,23	+3,57	
	6.516	90	942	
-alnik				0,12
	+15,63	–32,68	0,20	
	587	68	153	
-elec				0,50
	+0,50	–5,63	5,28	

Besedni del	/l/	/ụ̯/	/l/ in /ụ̯/	κ
-ilec	909	871	3.022	0,61
	–42,90	+2,89	+99,45	
-alnica	4.218	35	187	0,21
	+18,84	–25,73	–15,49	
-alstvo	294	162	264	0,34
	–9,65	+4,00	+18,49	
-elen	5.500	110	220	0,32
	+20,99	–27,41	–18,70	
-elski	5.335	165	0	0,66
	+22,73	–24,59	–26,11	
-elno	68	1	1	0,49
	+2,58	–3,10	–2,61	
-alen	77.824	111	5.995	≤ 0
	+74,09	–116,55	–43,25	
-ilnik	3.156	0	360	0,22
	+13,08	–24,05	–3,64	
-lec	810	64	34	1,00
	+6,45	–6,98	–7,41	
-alski	5.830	1.045	4.015	0,53
	–21,80	–17,63	+72,52	
-alno	805	2	9	≤ 0
	+9,31	–11,41	–9,16	
-ilen	31.075	0	1.265	0,15
	+53,16	–72,93	–43,35	
-ilski	3.795	220	1.155	0,74
	+1,92	–21,62	+20,30	
-ilstvo	282	6	108	0,14
	+0,01	–7,33	+8,41	
-ilce	720	0	36	≤ 0
	+7,85	–11,15	–5,96	
/	75.500	42.261	1.965	0,93
	–33,19	+160,82	–105,71	

Izgovor /l/ se pokaže kot statistično značilen za besedne dele *-ilnost*, *-alnost*, *-ilno*, *-elnica*, *-elka*, *-alce*, *-ilnica*, *-alnica*, *-elen*, *-elski*, *-elno*, *-alen*, *-ilnik*, *-lec* (*kozolec*; to ne vključuje delov *-ilec*/*-alec*/*-elec*), *-alno*, *-ilen* in *-ilce*; za omenjene dele je obenem statistično neznačilen izgovor /u̯/ oz. varianten izgovor. Po drugi strani je izgovor /u̯/ oz. varianten izgovor statistično značilen za dele *-alec*, *-alka*, *-ilka*, *-ilec*, *-alstvo*; hkrati je zanje neznačilen izgovor zgolj z /l/.

Obratno je z delom *-ilski*, ki je značilnejši za /l/ in varianten izgovor, zanj pa je neznačilen izgovor zgolj z /u̯/. Zanimiv je tudi rezultat za lekseme, ki ne izkazujejo nobenega od končnih besednih delov – pri njih se označevalci zelo dobro ujemajo ($\kappa = 0{,}93$), zanje pa je značilen izgovor z /u̯/. Preostala izgovora sta statistično neznačilna. Poudariti je sicer treba, da statistično značilen oz. neznačilen izgovor ne pomeni, da izjeme niso mogoče – gre le za splošne tendence.

6 Zaključek

V prispevku smo predstavili statistično analizo nekaterih vidikov izgovora črke *l* v različnih črkovnih kombinacijah na podlagi leksemov iz Slovenskega oblikoslovnega leksikona *Sloleks*. Rezultat analize je tudi množica označenih oblik, ki je pod imenom ILS 1.0 (*Izgovor črke l v Sloleksu*) odprto dostopna na repozitoriju CLARIN.SI (Čibej 2024b); podatki o približno 173.000 označenih oblikah in njihovih izgovorih so na voljo za nadaljnje jezikoslovne raziskave. Ker gre le za oznake petih jezikoslovcev v parih, podatki ne morejo biti obravnavani kot reprezentativni in dokončni, lahko pa ponudijo nekoliko natančnejši uvid v izgovor posamezne pregibne oblike/leksema oz. seznam problematičnih leksemov z neujemanjem, na katere se je mogoče osredotočiti s poglobljenimi raziskavami.

Rezultati analize potrjujejo tudi nekatere trditve v *Pravopisu 8.0*, npr. pojav izgovora /u̯/ v sklopih bigrama *lC* in predhodnega samoglasniškega grafema *o* (*polh*, *polniti*). Neujemanje in Pearsonovi preostanki kot manj intuitivne za določanje izgovora izpostavljajo npr. končne besedne dele *-alec*, *-alka*, *-ilka*, *-ilec* in *-alstvo*; zanje je značilen varianten izgovor, kot omenja tudi *Pravopis 8.0*. V nadaljevanju bi bilo treba pri njih preveriti še, v kolikšni meri je izgovor odvisen od dodatnih dejavnikov (npr. ali se nanašajo na živega ali neživega vršilca dejanja; ali izhajajo iz lastnega imena; ali imajo izglagolsko podstavo), saj do zdaj ta problem še ni bil obravnavan s statističnimi metodami.

V prihodnje bomo raziskavo ponovili še na gradivu korpusa GOS 2.1 (Verdonik idr. 2023), ki vsebuje zvočne posnetke avtentične govorjene slovenščine v različnih govornih položajih. Izsledke bomo lahko uporabili tudi za učenje preprostega modela, ki bo lahko napovedoval izgovor črke *l* v bigramih lC. Model bomo implementirali v obstoječo infrastrukturo grafemsko-fonemskega pretvornika,

podatki o izgovoru posameznih leksemov pa bodo z usklajevanjem z normativnimi priporočili *Pravopisa 8.0* vključeni v prihodnje različice *Slovenskega oblikoslovnega leksikona Sloleks*.

Literatura

Bajec idr. 2023 = Marko Bajec – Iztok Lebar Bajec – Tjaša Šoltes – Jernej Cvek – Jaka Čibej – Kaja Gantar – Sara Sever – Simon Krek, *Online Notes – A Real-time Speech Recognition and Machine Translation System for Slovene University Lectures*, 2023, 7–10, https://is.ijs.si/wp-content/uploads/2023/11/IS2023_Volume-H.pdf.

Cohen 1960 = Jacob Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20.1 (1960), 37–46, DOI: https://doi.org/10.1177/001316446002000104.

Čibej 2023 = Jaka Čibej, *Leksikon besednih oblik Sloleks: poročilo projekta Razvoj slovenščine v digitalnem okolju: aktivnost DS1.3*, Ljubljana: Univerza v Ljubljani, Center za jezikovne vire in tehnologije, 2023, https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Sloleks_v2.pdf.

Čibej 2024a = Jaka Čibej, Predicting pronunciation types in the Sloleks morphological lexicon of Slovene, v: *Odkrivanje znanja in podatkovna skladišča - SiKDD 2024: zbornik 27. mednarodne multikonference Informacijska družba - IS 2024*, ur. Dunja Mladenić – Marko Grobelnik, Ljubljana: Institut Jožef Stefan, 2024, 23–26, https://is.ijs.si/wp-content/uploads/2024/11/IS2024_Volume-C.pdf, DOI: https://doi.org/10.70314/is.2024.sikdd.2.

Čibej 2024b = Jaka Čibej, *Dataset of Annotated Slovene Words with Pre-Consonant L ILS 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, 2024. http://hdl.handle.net/11356/2025.

Čibej idr. 2022 = Jaka Čibej – Kaja Gantar – Kaja Dobrovoljc – Simon Krek – Peter Holozan – Tomaž Erjavec – Miro Romih – Špela Arhar Holdt – Luka Krsnik – Marko Robnik-Šikonja, *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, 2022, http://hdl.handle.net/11356/1745.

Dobrišek idr. 2022 = Simon Dobrišek – Žiga Golob – Jerneja Žganec Gros, Finite-state super transducers for compact language resource representation in edge voice-AI, *Systems science & control engineering* 10.1 (2022), 636–644, https://www.tandfonline.com/doi/full/10.1080/21642583.2022.2089930, DOI: https://doi.org/10.1080/21642583.2022.2089930.

Erbeli – Pižorn 2012 = Florina Erbeli – Karmen Pižorn, Reading ability, reading fluency and orthographic skills: the case of L1 Slovene English as a foreign language students, *English. Center for Educational Policy Studies Journal* 2.3 (2012), 119–139, https://files.eric.ed.gov/fulltext/EJ1130208.pdf.

Filipič 2016 = Urša Filipič, *Upoštevanje učnega okolja pri prilagajnju* [sic!] *poučevanja angleščine za učence z disleksijo*, diplomsko delo, Univerza v Ljubljani, Pedagoška fakulteta, 2016, https://centerslo.si/wp-content/uploads/2019/10/Obdobja-38_Mirtic.pdf.

Kosem 2022 = Iztok Kosem, Trendi – a monitor corpus of Slovene, v: *EURALEX 2022, Proceedings of the XX EURALEX International Congress*, ur. Annette Klosa, [b. n. m.]: IDS-Verlag, 2022, 230–239, https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p230-239_Kosem.pdf.

Križaj idr. 2022a = Janez Križaj – Jerneja Žganec Gros – Simon Dobrišek, Validation of speech data for training automatic speech recognition systems, v: *EUSIPCO 2022: 30th European Signal Processing Conference (EUSIPCO 2022)*, EURASIP 2022, 2022, 1165–1169, https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0001165.pdf.

Križaj idr. 2022b = Janez Križaj – Simon Dobrišek – Aleš Mihelič – Jerneja Žganec Gros, Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi, v: *Jezikovne tehnologije in digitalna humanistika: zbornik konference 2022*, ur. Darja Fišer – Tomaž Erjavec, Ljubljana: Inštitut za novejšo zgodovino, 2022, 248–251, https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf.

Marjou 2021 = Xavier Marjou, GIPFA: Generating IPA Pronunciation from Audio, v: *Zbornik konference eLex 2021*, 2021, 588–597, https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_38_pp588-597.pdf.

McHugh 2012 = Mary L. McHugh, Interrater reliability: the kappa statistic, *Biochemia medica* 22.3 (2012), 276–282. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/.

Mirtič 2019 = Tanja Mirtič, Glasoslovne raziskave pri pripravi splošnega razlagalnega slovarja, v: *Slovenski javni govor in jezikovno-kulturna (samo)zavest*, ur. Hotimir Tivadar, Ljubljana: Znanstvena založba Filozofske fakultete, 2019 (Obdobja 38), 81–90.

Pogačnik 2012 = Aleš Pogačnik, Glasovno domačenje lastnih imen iz nelatiničnih pisav, v: *Pravopisna stikanja. Razprave o pravopisnih vprašanjih*, ur. Nataša Jakop – Helena Dobrovoljc, Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, 2012, 73–83, https://unglue-it-files.s3.amazonaws.com/ebf/f62400a502b74ca89fdf3e6a9fa3282b.pdf.

Pravopis 8.0 = *Pravopis 8.0*: *pravila novega slovenskega pravopisa za javno razpravo*, https://pravopis8.fran.si/.

Reichel idr. 2008 = Uwe Reichel – Hartmut R. Pfitzinger – Horst-Udo Hain, English grapheme-to-phoneme conversion and evaluation, v: *Speech and Language Technology* 11 (2008), ur. Grazyna Demenko – Krzysztof Jassem – Stanislaw Szpakowicz, 159–166, https://www.phonetik.uni-muenchen.de/~reichelu/publications/ReichelPfitzingerHainSASR2008.pdf.

Rigler 1980 = Jakob Rigler, O izgovoru črke l v SSKJ, *Slavistična revija* 28.1 (1980), 114–120, https://srl.si/ojs/srl/article/view/1980-1-0-13.

Schüppert idr. 2017 = Anja Schüppert – Wilbert Heeringa – Jelena Golubovic – Charlotte Gooskens, Write as you speak? A cross-linguistic investigation of orthographic transparency in 16 Germanic, Romance and Slavic languages, *English. From semantics to dialectometry* 32 (2017), 303–313.

Sever 2023 = Sara Sever, *Korpusnojezikoslovna analiza variantnosti obrazil -vec in -lec ter -vka in -lka v sodobni slovenščini*, magistrsko delo, Univerza v Ljubljani, Filozofska fakulteta 2023, https://repozitorij.uni-lj.si/Dokument.php?id=176406&lang=slv.

SP 2001 = *Slovenski pravopis*, Ljubljana: Založba ZRC, ZRC SAZU, 2001.

SSKJ = *Slovar slovenskega knjižnega jezika* I–V, Ljubljana: Državna založba Slovenije, 11970–1991, https://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika, spletna objava 2014.

Šef 2006 = Tomaž Šef, Avtomatsko naglaševanje nepoznanih besed pri sintezi slovenskega govora, *Elektrotehniški vestnik* 73.2–3 (2006), 99–104.

Šoltes idr. 2024 = Tjaša Šoltes – Jan Vasiljević – Marko Bajec, ONLINE NOTES: sistem za razpoznavo govora in strojno prevajanje v realnem času na ravni univerzitetnih predavanj, *Uporabna informatika* 1 (XXXII) (2024), 32–41.

Tivadar 2004 = Hotimir Tivadar, Priprava, izvedba in pomen perceptivnih testov za fonetično-fonološke raziskave (na primeru analize fonoloških parov), *Jezik in slovstvo* 49.2 (2004), 17–36.

Unuk 2009 = Drago Unuk, Pravopisna načela v slovenskem pravopisu, *Revija za elementarno izobraževanje* 2.4 (2009), 27–36.

Verdonik idr. 2023 = Darinka Verdonik – Ana Zwitter Vitez – Jana Zemljarič Miklavčič – Simon Krek – Marko Stabej – Tomaž Erjavec – Tomaž Potočnik – Mirjam Sepesy Maučec – Simona Majhenič – Andrej Žgank – Andreja Bizjak – Lucija Gril – Simon Dobrišek – Janez Križaj – Marko Bajec – Iztok Lebar Bajec – Tjaša Jelovšek – Mitja Trojar – Mitja Bernjak – Naum Dretnik – Gregor Strle – Kaja Dobrovoljc – Nikola Ljubešić – Peter Rupnik, *Spoken corpus Gos 2.1 (transcriptions)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, 2024, http://hdl.handle.net/11356/1863.

Zorman 2007 = Anja Zorman, *Prepoznavanje glasov in spoznavanje njihovih pisnih ustreznic v maternem in drugem oziroma tujem jeziku*, doktorska disertacija, Univerza v Ljubljani, Pedagoška fakulteta, 2007.

Žganec Gros idr. 2020 = Jerneja Žganec Gros – Miro Romih – Tomaž Šef, eBralec 4: hibridni sintetizator slovenskega govora, v: *Interakcija človek-računalnik v informacijski družbi: zbornik 23. mednarodne multikonference Informacijska družba - IS 2020*, ur. Veljko Pejović idr., Ljubljana: Institut Jožef Stefan, 2020, 17–20, http://library.ijs.si/Stacks/Proceedings/InformationSociety/2020/IS2020_Volume_H%20-%20HCI.pdf.

Žganec Gros idr. 2022 = Jerneja Žganec Gros – Tanja Mirtič – Miroslav Romih – Kozma Ahačič, *Slovar izgovarjav OptiLEX*, Ljubljana: Založba ZRC, 2022.

Summary

A Statistical Analysis of the Pronunciation of the Letter *l* in the *Sloleks Morphological Lexicon of Slovene*

The ambiguous pronunciation of the letter *l* before consonant graphemes (e.g., *polža*, *alge*, *volilca*) poses a problem for the otherwise relatively straightforward grapheme-to-phoneme conversion for Slovenian. Despite having been addressed in multiple Slovenian language resources, there is still a lack of empirical and machine-readable data on the pronunciation of native speakers. To rectify this, a dataset of approximately six thousand lexemes (and their inflected forms) from the *Sloleks Morphological Lexicon of Slovene* was annotated by five linguists according to their pronunciation (/l/, /u̯/, or both). A statistical analysis was performed to reveal the most problematic points. The findings confirm some of the premises stated in the new *Slovenian Normative Guide* (*Pravopis 8.0*), such as the prevalence of the /u̯/ pronunciation in *-ol-* bigrams (e.g., *bolha*, *solza*) and the variability of pronunciation in words featuring certain endings (such as *-alec*, *-alka*, *-ilka*, *-ilec*, and *-alstvo*). The dataset is available under open access and, although it is not representative, it provides a good starting point for further linguistic analyses and is also a potential resource for developing models for predicting the pronunciation of the letter *l* in Slovenian.