UDK 811.163.6'344.2:612.85
Tatjana Marvin
Filozofska fakulteta Univerze v Ljubljani
tatjana.marvin@guest.arnes.si

Saba Battelino
Medicinska fakulteta Univerze v Ljubljani
saba.battelino@kclj.si

Samo Beguš
Fakulteta za elektrotehniko Univerze v Ljubljani
samo.begus@fe.uni-lj.si

Jure Derganc
Medicinska fakulteta Univerze v Ljubljani
jure.derganc@mf.uni-lj.si

PORAZDELITEV FONEMOV V SLOVENŠČINI IN IZDELAVA MATRIČNEGA TESTA ZA GOVORNO AVDIOMETRIJO

V članku je predstavljen postopek izbire besed za slovenski matrični stavčni test, ki se uporablja za preizkus slušnega razumevanja pri govoru. Glavni poudarek prispevka je na določitvi fonemske porazdelitve v jezikovnem gradivu za test, ki se mora čim bolj približati porazdelitvi fonemov v jeziku testa. Ker porazdelitev fonemov v slovenščini še ni raziskana, jo določimo s pomočjo črkovne porazdelitve v korpusu pisne slovenščine ccKres v kombinaciji s fonetično podatkovno bazo v Mihelič (2006) za tiste primere, kjer črkovni zapis ne ustreza fonemskemu. Na osnovi ugotovljene fonemske porazdelitve nato predlagamo besede za slovenski matrični test.

Ključne besede: slovenščina, matrični stavčni test, fonem, govorna avdiometrija

This paper presents a word selection process in a Slovenian matrix sentence test for speech intelligibility measurements. We focus on phonemic distribution in the test, which should be approximated as closely as possible to distribution in the language. We establish a phonemic distribution for Slovenian by combining the orthographic distribution in the corpus ccKres and the phonetic distribution in Mihelič (2006) for cases where the orthographic record does not correspond to the phonetic one. The result is a proposal of a phonemically balanced matrix test for Slovenian.

Keywords: Slovenian, matrix sentence test, phoneme, speech audiometry

V članku je predstavljen postopek izbire besed za slovenski matrični test, ki se bo v klinični govorni avdiometriji uporabljal za preizkus sluha. Izdelava matričnega testa je pomembna tudi z jezikoslovnega in slovenističnega vidika. Mednarodne smernice za matrični test namreč zahtevajo, da se mora porazdelitev fonemov v jezikovnem gradivu za test čim bolj približati porazdelitvi fonemov v jeziku preizkusa. Ker porazdelitev

fonemov v slovenščini še ni raziskana, je glavni poudarek pričujočega prispevka vzpostavitev fonemske porazdelitve za slovenščino, rezultat pa doprinos h glasoslovnim raziskavam na področju slovenskega jezika. Fonemsko porazdelitev za slovenščino določimo na osnovi črkovne porazdelitve v korpusu pisne slovenščine ccKres, ki jo dopolnimo s podatki o porazdelitvi fonemov v primerih, kjer črkovni zapis ne ustreza fonemskemu. Pri tem si pomagamo s fonetično podatkovno bazo v Mihelič (2006). Na osnovi ugotovljene fonemske porazdelitve nato predlagamo besede za slovenski matrični test. V članku je najprej na kratko predstavljen koncept govorne avdiometrije (razdelek 1), splošna sestava matričnega testa (razdelek 2) in predstavitev zahtev glede priprave jezikovnega gradiva (razdelek 3). Sledi osrednji del, v katerem je opisan proces določanja pogostosti pojavljanja fonemov v slovenščini (razdelek 4). V zadnjem delu je predstavljen predlog matričnega testa za slovenščino, osnovan na splošnih navodilih za matrične teste in na glasovnih lastnostih slovenskega jezika (razdelek 5). Sledi zaključek (razdelek 6).

1 Govorna avdiometrija za slovenski jezik

Govorna avdiometrija je ena izmed standardnih metod za diagnozo tipa izgube sluha in za oceno sposobnosti sporazumevanja bolnika, saj z njo testiramo nivo razumevanja slišanega in sposobnost ponovitve slišanega (Musiek idr. 2011). V ta namen je bil v slovenski jezik leta 1968 preveden in adaptiran nemški besedni test, ki ga je razvil Hahlbrock leta 1953 in 1960, znan kot Freiburški test z enozložnicami in Freiburški numerični test (Pompe 1968). Test je bil prenovljen leta 2016 (Marvin, Derganc in Battelino 2016). Poleg besednih testov poznamo v govorni avdiometriji tudi teste s povedmi; le-ti bolje odražajo vsakodnevno rabo jezika in so se tako izkazali kot koristna in natančna diagnostična orodja v več jezikih. Na splošno se uporabljata dva tipa – testi, ki vsebujejo povedi z različno skladenjsko zgradbo iz vsakodnevne komunikacije (npr. Plomp & Mimpen 1979) ter t. i. matrični testi, kjer imajo vse povedi enako skladenjsko zgradbo z nepredvidljivo kombinacijo posameznih besed ter besednih pomenov (Hagerman 1982; Wagener 1999a, b, c; Ozimek idr. 2010; Hochmuth idr. 2012; Warzybok idr. 2015).

V tem prispevku predstavljamo postopek izbire besed za test z matrično strukturo, osnovan na slovenskem jeziku. Matrični test – prvi te vrste v Sloveniji – bo uporabljen za natančnejšo oceno sluha pri ljudeh z motnjo sluha, za ocenjevanje razumevanja govora pri ljudeh s centralnimi motnjami sluha in motnjami razumevanja, za ocenjevanje kognitivnih sposobnosti, za ocenjevanje izboljšanja razumevanja govora pri bolnikih z uporabo vsadkov in različnih slušnih pripomočkov in pri bolnikih s tinitusom. Pri izdelavi testa sledimo smernicam Mednarodnega kolegija za rehabilitacijsko avdiologijo (ICRA) (Akeroyd idr. 2015), ki dopolnjujejo standard ISO 8253-3: 2012 (Akustika. Avdiometrične preskusne metode – 3. del: Govorna avdiometrija) z zagotavljanjem korakov, ki so potrebni za izdelavo matričnega preskusa v posameznem jeziku.

2 Smernice za sestavo matričnega testa

Prvi matrični test je leta 1982 sestavil Björn Hagerman (Hagerman 1982); osnovan je bil na švedskem jeziku. Nekoliko spremenjena različica (Wagener idr. 1999a, b, c) je trenutno na voljo v 14 jezikih (npr. angleščina, nizozemščina, nemščina, francoščina, turščina itd.), od katerih sta le dva slovanska (poljščina in ruščina). Test je sestavljen iz povedi, ki vsebujejo 5 besed in so skladenjske oblike *Osebek – Povedek – Predmet*, npr. *Tone kupi pet velikih stolov*. Osebek je vedno enobesedno osebno ime, povedek je enobesedna glagolska zveza, predmet pa samostalniška zveza, katere jedro je samostalnik v množini, kot leva prilastka pa se pojavita številnik ter pridevnik. Obliko povedi lahko poenostavljeno povzamemo kot *Ime – Glagol – Številnik – Pridevnik – Samostalnik*.

Pri pripravi testa je najprej potrebno zbrati jezikovno gradivo, ki sestavlja osnovno matrico; le-ta obsega 50 besed, po 10 besed za vsako od petih pozicij v predpisani obliki povedi: 10 osebnih imen, 10 glagolov, 10 številnikov, 10 pridevnikov, 10 samostalnikov. Iz osnovne matrice je mogoče generirati skupaj 100.000 različnih povedi oblike *Ime – Glagol – Številnik – Pridevnik – Samostalnik* (vsako od desetih imen se kombinira z vsakim od desetih glagolov, vsaka taka kombinacija se nadalje kombinira z vsakim od desetih številnikov itd. Skupno število različnih kombinacij je $10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 100.000$). Za potrebe testiranja se uporablja sezname z desetimi različnimi povedmi, v katerih se vsaka od besed pojavi samo enkrat.

Naslednji korak je snemanje povedi z rojenim govorcem, rezanje povedi v posamezne besede z ohranitvijo koartikulacije ter lepljenje v naključne kombinacije oblike *Ime – Glagol – Številnik – Pridevnik – Samostalnik*. Povedi mora brati rojeni govorec (m/ž), za katerega ni nujno, da ima formalno izobrazbo radijskega napovedovalca. Izgovarjava mora biti knjižna z nevtralno intonacijo in enakomerno jakostjo. Ker je z enim govorcem nemogoče posneti vseh 100.000 različnih povedi, si pomagamo tako, da posnamemo le minimalno količino jezikovnega gradiva, iz katerega nato z rezanjem in lepljenjem besed generiramo vseh 100.000 kombinacij.

Posneto gradivo mora vključevati vse kombinacije dveh besed, ki se lahko pojavita ena za drugo, kar ob nadaljnjem rezanju in ponovnem lepljenju zagotavlja ohranjanje koartikulacije in naravni govor. Tako minimalno jezikovno gradivo obsega le 100 različnih povedi; seznam sestavimo po metodi iz Wagener idr. (1999a). Po snemanju je potrebno posneta zaporedja razrezati v posamezne besede, pri čemer je nujno ohraniti koartikulacijo na koncu vsake izrezane besede glede na besedo, ki ji sledi. Povedi so nato (računalniško) sestavljene s kombiniranjem posnetih besed, zraven pa je lahko dodan še maskirni šum. Preden se test sprejme v klinično uporabo, je potrebno izvesti še optimizacijo, evalvacijo in validacijo posnetega gradiva.

Iz posnetega gradiva lahko računalnik vsakič posebej pri testiranju generira povedi, ki jih nato uporabimo pri govorcih z motnjami sluha. Matrični test je enostaven za testiranje in je zaradi izjemnega števila možnih kombinacij besed v različne povedi koristno diagnostično orodje. Pacient si namreč povedi ne more zapomniti od enega do

drugega testiranja, prednost matričnega testa pa je tudi v medjezikovni primerljivosti rezultatov, v kolikor je za vse jezike uporabljen enak postopek priprave.

3 Priprava jezikovnega gradiva za slovenski matrični test

V pričujočem razdelku je opisan postopek zbiranja jezikovnega gradiva v osnovni matrici, ki je sestavljena iz 50 besed, tj. po 10 besed za vsako od petih pozicij besed povedi oblike *Ime – Glagol – Številnik – Pridevnik – Samostalnik*. Izbrano gradivo mora zadostiti različnim pogojem na pomenski, skladenjski ter glasoslovni ravnini.

Pogoji glede izbora imen narekujejo, da pri tej kategoriji uporabimo 5 moških in 5 ženskih imen. Nadalje se v testu ne smejo pojavljati redke, zastarele ali čustveno zaznamovane besede, pri čemer je potrebno upoštevati, da ne smejo biti čustveno zaznamovane tudi kombinacije posameznih besed, v katerih posamezna beseda sama zase ni čustveno zaznamovana. Izogibati se je potrebno tudi ponavljajočim se kombinacijam, npr. *veliko velikih kamnov*, ali podobnim imenom, npr. *Jana*, *Jasna*.

Vse možne kombinacije besed, tj. vse povedi, ki jih lahko sestavimo iz teh besed, morajo biti slovnično pravilne. Ta pogoj pomembno vpliva na izbor glagolskega časa ter številnikov v samostalniški zvezi, ki označuje predmet. V matričnih testih za germanske jezike se pogosto uporablja pretekla oblika glagola, v obstoječih matričnih testih za slovanske jezike pa se uporabljata sedanjik ali prihodnjik. V slovenščini moramo v matričnem testu uporabiti sedanjo obliko glagola, npr. *Jana kupi tri velike škatle*. Uporaba preteklika ali prihodnjika bi namreč zahtevala dodatno mesto za pomožni glagol *biti* ter prispevala dodatne zaplete pri ujemanju deležnika na *-l* z osebkom po spolu (poleg ujemanja po številu).

Pri preizkusu se v slovenščini lahko uporabljajo samo številniki od vključno 5 dalje, saj ti brez izjeme zahtevajo, da jim sledita pridevnik in samostalnik v rodilniku množine (npr. *Jana kupi pet*/*šest*/*sedem*/*osem ... velikih škatel*). Številnike od 1 do 4 nadomestimo s kvantifikacijskimi izrazi, ki podobno kot številniki od 5 dalje zahtevajo rodilnik množine, npr. *malo*, *nekaj* (npr. *nekaj velikih stolov*).

Test mora zadostiti tudi nekaterim pogojem, ki spadajo v domeno glasoslovne ravnine. Preizkus je potrebno uravnotežiti glede na število zlogov izbranih besed in sicer tako, da je uravnotežena vsaka posamezna skupina desetih besed v osnovni matrici. Odločimo se za uporabo dvozložnih besed, izjemoma pa enozložnih in trozložnih, ko je taka raba posledica zahtevane fonemske uravnoteženosti. Matrični preizkus mora biti fonemsko uravnotežen na način, da pogostost fonemov v testu odseva pogostost fonemov v jeziku, na katerem je preizkus osnovan. Izpolnitev tega pogoja se je izkazal za najbolj zahtevnega, saj porazdelitev fonemov v slovenščini še ni raziskana. Fonemska uravnoteženost je tako postala osrednja jezikoslovna tema pričujočega raziskovalnega dela. Podrobno je obravnavana v razdelku 4.

4 Določanje porazdelitve fonemov

V slovenski znanstveni literaturi je najti številne raziskave, katerih cilj je določitev porazdelitve črk (npr. Jakopin 1999, Suhadolc 2013, Ključevšek 2016), ne obstajajo pa izračuni porazdelitve fonemov. To je razumljivo, saj je fonem abstraktna enota in zato zahteva ugotavljanje porazdelitve fonemov tudi jezikoslovno analizo, ki za obravnavo črkovnega zapisa v korpusih ni nujno potrebna. V tem razdelku predstavimo postopek določanja fonemske porazdelitve za slovenščino na osnovi črkovne porazdelitve v korpusu pisne slovenščine ccKres (4.1), ki jo dopolnimo s podatki o porazdelitvi fonemov v primerih, kjer črkovni zapis ne ustreza fonemskemu (4.2 in 4.3).

4.1 Izbira korpusa za določitev porazdelitve fonemov

Porazdelitev fonemov v slovenščini smo določili na osnovi korpusa ccKres, ki je največji žanrsko uravnotežen korpus sodobne pisne slovenščine, dostopen na spletnem repozitoriju CLARIN.SI. Vsebuje 10 milijonov besed iz različnih besedilnih zvrsti, od dnevnega časopisja, revij, knjig (leposlovje, neleposlovje, učbeniki) do spletnih strani. Ustreznost izbire smo preverili s primerjavo z dvema korpusoma govorjene slovenščine, ki sta prav tako dostopna v repozitoriju, a sta v primerjavi s korpusom ccKres bistveno manjša in nista enako uravnotežena. Korpus govorjene slovenščine GOS, tj. njegova ortografska transkripcija v knjižni slovenščini, vsebuje milijon besed. Ortografska transkripcija podatkovne zbirke SNABI, natančneje njen del Lingua, vsebuje 910 povedi iz različnih vrst besedil, kot npr. knjige ali časopisi (Kačič idr. 2002).

Izbrane korpuse smo glede na pojavnost črk primerjali z meritvami iz Jakopinovega dela (Jakopin 1999), kjer je prvič analizirana porazdelitev črk v številnih leposlovnih delih v slovenskem jeziku. Izsledki so predstavljeni v Tabeli 1.

Porazdelitev črk v korpusu ccKres se približno ujema tako s korpusoma govorjene slovenščine kot tudi z Jakopinovo analizo. Ob tem opazimo, da ima korpus GOS relativno visoko pojavnost črk "a", "e" in "m", kar bi lahko pripisali uporabi teh črk (oz. glasov, ki jih te črke označujejo) v mašilih v govorjeni slovenščini.

4.2 Fonem, alofon, črka v slovenščini

Za določanje porazdelitve fonemov so ključni pojmi fonem, alofon in črka ter njihova medsebojna razmerja v slovenščini. Prva dva pojma sta pomembna z vidika jezikoslovne analize, zadnji pa z vidika jezikovnih orodij, ki so nam na voljo za določanje pogostosti – korpusi besedil so namreč najpogosteje zapisani s črkami, le redko z alofoni (fonetična transkripcija), s fonemi pa praktično nikoli. Pomembno je torej ugotoviti, kako črkovni in fonetični zapis pretvoriti v fonemski zapis, iz katerega bi potem lahko neposredno razbrali pogostost pojavljanja posameznih fonemov.

Fonem je v jezikoslovni teoriji definiran kot najmanjša glasovna enota, ki jo lahko izluščimo iz glasovne verige in je pomensko razločevalna znotraj posameznega jezika.

Tabela 1: Pogostost črk (v %) v treh obravnavanih korpusih ter v Jakopin (1999).

	a	b	c	č	d	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž
GOS	12.0	1.7	0.6	1.3	3.5	12.5	0.2	1.2	0.9	8.0	4.9	3.9	4.1	4.2	5.9	8.8	3.6	4.5	4.5	1.2	5.2	1.6	3.3	1.9	0.4
ccKres	10.4	1.8	0.9	1.4	3.5	10.2	0.2	1.5	1.1	9.0	4.3	3.7	4.6	3.1	6.9	9.3	3.5	5.3	4.8	1.0	4.6	2.0	4.1	2.2	0.6
Lingua	9.7	1.9	0.7	1.6	3.5	10.9	0.2	1.4	1.2	8.8	4.8	3.6	4.9	3.5	6.3	9.2	3.6	5.1	4.8	1.2	4.5	1.9	4.2	2.2	0.6
Jakopin	10.5	1.9	0.7	1.5	3.4	10.7	0.1	1.6	1.1	9.0	4.7	3.7	5.3	3.3	6.3	9.1	3.4	5.0	5.1	1.0	4.3	1.9	3.8	2.1	0.7

Pravimo, da glasova pripadata dvema fonemoma, kadar zamenjava enega z drugim menja besedni pomen. V slovenščini denimo kot fonema prepoznamo glasova /p/ in /b/, ki sta pomensko razločevalna v parih besed kot *piti* - *biti*, *brati* - *prati*, itd.[1] Fonemi so abstraktne enote, ki se v govoru udejanjijo glede na glasovno okolje. Pravimo, da je fonem skupek fonetično podobnih različic, ki jih imenujemo alofoni in jih opišemo s pravilom glede na okolje pojavljanja. Na primer, v slovenščini se fonem /n/ udejanja v treh različicah: kot [ŋ] pred glasovi [k, g, x] (*Anka*, *Anglija*, *Anhovo*), kot [nʲ] (pri nekaterih govorcih), ko mu sledi j# ali jC, (*konj*, *konjski*) ter kot [n] povsod drugje (*nos*, *ena*, *dan*). Ker ima v vsakem jeziku fonem vsaj eno različico, v kateri se udejanji, je število alofonov običajno večje od števila fonemov.

Sistemi, ki uporabljajo črkovni zapis, se med seboj razlikujejo – v nekaterih črka v grobem ustreza fonemu, v drugih pa alofonu. V slovenščini je pisava urejena tako, da v večini primerov ena črka označuje en fonem. Na primer, fonem /n/ bo zapisan s črko "n" tako v besedi *nos* kot v besedi *Anka*, čeprav se izgovora fonetično razlikujeta. Preslikava iz črkovnega zapisa v fonemski zapis pa ni enoznačna in na tem mestu izpostavljamo dve skupini težav, na katere naletimo in jih tudi upoštevamo pri analizi.

Najprej je treba pojasniti dejstvo, da imamo v jeziku manj črk kot fonemov, tj. 25 črk v abecedi, a 29 fonemov (glede nabora fonemov sledimo slovnici Toporišič (2000: 45)).[2] Razlogov za takšno stanje je več. V nekaterih primerih se ena črka uporablja za zapis več različnih fonemov. Tako označuje črka "e" tri foneme: ozki /e/ v besedi *led*, široki /ɛ/ v besedi *žep* ali polglasnik /ə/ v besedi *pes*, črka "o" pa dva fonema: ozki /o/ v besedi *nos* ali široki /ɔ/ v besedi *noga*.[3] Nadalje se lahko en fonem zapiše z dvema črkama – tak je npr. fonem /dʒ/, ki se v slovenščini zapiše z "dž" (npr. *džip*).

Z navedeno razlago pojasnimo obstoj štirih fonemov, ki jih črke ne zaobjamejo, vendar se s tem kompleksnost preslikave črka-fonem še ne zaključi. Odnos med njima zamegljuje tudi obstoj primerov, ko se fonem izgovori, a ni zapisan z nobeno črko. Tak je npr. polglasnik /ə/ v besedah *vrt*, *rt*, *država*. Dobimo pa tudi obratno situacijo, tj. ko črka ne označuje fonema, temveč le nakazuje različico fonema, ki se izgovori v nekem okolju. Tako npr. v pisavi za besedi *konj* in *konjski* črka "j" nakazuje palatalizirano različico [nʲ] fonema /n/ in ne označuje fonema /j/.[4, 5] Navedeni primeri vsi potrjujejo, da je črka le približek fonema in da je potrebno pri prehodu iz črkovnega

[1] Za zapisovanje fonemov uporabimo poševne oklepaje (t. i. fonemski zapis), za zapisovanje alofonov pa oglate oklepaje (t. i. fonetični zapis). Glasove zapisujemo v sistemu IPA (Mednarodna fonetična abeceda).

[2] Jurgec (2011) utemeljuje, da ima slovenščina devet in ne osem samoglasnikov. Dodatni samoglasnik je nizek sredinski [ʌ], npr. v besedah *čas*, *brat*, itd. V tem članku privzamemo tradicionalen samoglasniški sestav kot v Toporišič (2000: 48).

[3] V tej raziskavi dolžina in naglas nista upoštevana.

[4] Nekateri govorci izgovorijo namesto [nʲ] različico [n], v pisavi je tudi pri teh govorcih ohranjena črka "j".

[5] Slovenski pravopis (2014: 75) in Toporišič (2000: 78) navajata, da se zvočnika /n/ in /l/ pišeta z dvočrkjema "nj" oz. "lj", kadar sta malo podaljšana ali zmehčana, tj. na koncu besede ali pred soglasnikom (npr. *konj*, *konjski*, *polj*, *poljski*); v teh primerih jima pred samoglasnikom ustrezata glasovna sklopa [nj] in [lj] (npr. *konja*, *polja*).

zapisa v fonemski zapis upoštevati jezikovna pravila slovenskega jezika. Natančna obravnava je predstavljena v razdelku 4.3.

4.3 Določitev porazdelitve fonemov v slovenščini

Za določitev porazdelitve fonemov se opremo na podatke o porazdelitvi črk v korpusu ccKres, ki jih dopolnimo na mestih, kjer črke ne odražajo neposredno posameznih fonemov (tj. za foneme /e/, /ɛ/, /ə/, /o/, /ɔ/, /dʒ/, /j/).[6] To dosežemo tako, da dodamo foneme, ki se ne pojavijo v pisavi (/ə/), odvzamemo foneme, ki se pojavijo v pisavi, a se ne izgovorijo (/j/) ter s sklicevanjem na razmerja med fonemi, ki so zapisani s črkama "e" in "o". Pri slednjima namreč črka "e" označuje kar tri foneme, /e/, /ɛ/, /ə/, črka "o" pa dva fonema, /o/, /ɔ/. Razmerja med fonemi z istim črkovnim zapisom določimo tako, da sledimo porazdelitvi omenjenih samoglasniških fonemov v delu Mihelič (2006), ki obravnava porazdelitev alofonov v 300.000 fonetično transkribiranih povedih.[7]

Celoten proces se izvede po naslednjih korakih:

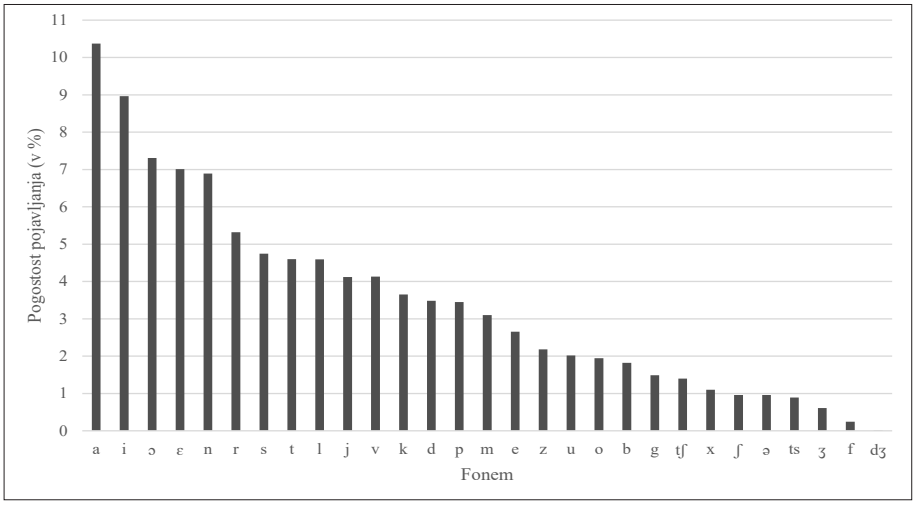
1. Vse črke v korpusu ccKres se pretvorijo v male črke. Vsa naglasna znamenja na črkah "a", "e", "i", "o" in "u" se odstranijo; take črke se nadomestijo z ustreznimi črkami brez znamenj. Izbrišejo se tudi vsi znaki, ki jih ne najdemo v slovenski abecedi (razen "đ" in "ć").
2. Število pojavitev črke "ć" se prišteje k pojavitvam črke "č".
3. Število pojavitev fonema /dʒ/ se določi tako, da se prešteje število kombinacij črk "dž" ter število črke "đ".
4. Število pojavitev fonema /j/ je potrebno prilagoditi tako, da odštejemo primere, ko je izražen v pisavi, a se ne izgovarja, tj. v kombinacijah nj#, njC, lj#, ljC, kjer "C" označuje soglasnik, "#" pa besedno mejo.
5. Število fonemov /o/ in /ɔ/ je določeno tako, da se število pojavitev črke "o" razdeli glede na distribucijo /o/ in /ɔ/ v Miheličevi raziskavi, kjer je ugotovljeno razmerje med fonemoma 21 % proti 79 %, iz česar sklepamo, da 21 % pojavitev črke "o" označuje fonem /o/, 79 % pa fonem /ɔ/.

[6] Do fonemske transkripcije bi lahko prišli tudi tako, da bi izhajali iz fonetične transkripcije. Ker pa v več primerih pride do prekrivanja alofonov različnih fonemov, bi za natančno fonemsko transkripcijo v vsakem primeru potrebovali tudi ortografski zapis. Na primer, fonema /l/ in /v/ se na koncu besede izgovarjata enako, npr. [piu̯] v besedah *pil* in *piv*, pri čemer je fonemski zapis besed različen, /pil/ ter /piv/.

[7] Zbirka Lingua vsebuje tudi fonetične transkripcije povedi, ki bi lahko služile kot osnova za izračun razmerij med različnimi fonemi pri črkah "e" in "o". V primerjavi z Miheličevo zbirko, ki vsebuje 300.000 povedi, je Lingua precej manjša – 910 povedi. Razmerja pri črki "o" so podobna kot tista iz Miheličeve zbirke, 75 % črk "o" ustreza fonemu /ɔ/, 25 % pa fonemu /o/ (79 % proti 21 % pri Miheliču). Razmerja pri črki "e" v Lingui – /ɛ/ (51 %), /e/ (45 %), /ə/ (4 %) – pa se bistveno razlikujejo od tistih v Mihelič (2006) (66 % proti 25 % proti 9 %). Razliko lahko pripišemo dejstvu, da je Miheličeva zbirka osnovana na knjižni izgovarjavi, medtem ko Lingua vsebuje transkripcijo pogovornega jezika (večinoma) z območja Štajerske.

6. Število fonemov /e/, /ɛ/ in /ə/ določimo tako, da preštejemo pojavitve črke "e" ter pojavitve polglasnika, ko ta ni zapisan z nobeno črko. Sledimo analizi v Toporišič (2000: 58–59), kjer se fonem /ə/ pojavi v kombinacijah "CrC" (*grd*, *smrt*, *žanrski*), "Cr#" (*žanr*), "#rC" (*rdeč*), "lm#" (*film*), "jm#" (*sejm*), "jn#"(*dizajn*), "vl#" (*favl*), "vn#" (*favn*), "rn#" (*Murn*), "rm#" (*perm*), "lmN" (*filmček*), "jmN" (*sejmski*), "jnN" (*dizajnček*), "vlN" (*favlček*), "vnN" (*favnček*), "rnN" (*Murnček*), "rmN" (*permski*), kjer "C" označuje soglasnik, "N" nezvočnik, in "#" besedno mejo.[8] Nato se število pojavitev razdeli glede na distribucijo fonemov /e/, /ɛ/ in /ə/ v Miheličevi raziskavi, kjer je razmerje /e/ (25 %), /ɛ/ (66 %) ter /ə/ (9 %).

Opisano analizo korpusa ccKres smo izvedli s pomočjo programa Mathematica (Wolfram Research). Ugotovljena porazdelitev fonemov je predstavljena na Sliki 1 ter v Tabeli 2.



Slika 1: Pogostost fonemov v slovenščini (v %, korpus ccKres).

[8] Prešteli smo vse tu naštete kombinacije. Po pregledu seznama besed pod vsako kombinacijo smo se odločili, da kot polglasnike obdržimo samo tiste, pri katerih je bila večina besed slovnično sprejemljiva. Zavržemo pojavnice polglasnika v kombinacijah Cr# (večinoma okrajšave *str*, *dr* in podobno), vl# (večinoma nepravilno zapisani glagoli kot *predstavl*, *spravl*), jnN (večinoma nepravilno zapisane besede *ljublajnskih*, *stojnci*), vlN (npr. *bvlgari*, *živlčjenju*, itd.) ter jmN (npr. *prejmkov*, *jms*). Naštetih pojavitev je 11.754 (11.359 samo za Cr#), kar je zanemarljivo v primerjavi s skupnim številom pojavljanja polglasnika.

Tabela 2: Pogostost fonemov v slovenščini (v %, korpus ccKres).

fonem	ccKres
a	10.37
i	8.97
ɔ	7.31
ɛ	7.01
n	6.89
r	5.32
s	4.74
t	4.60
l	4.59
j	4.12

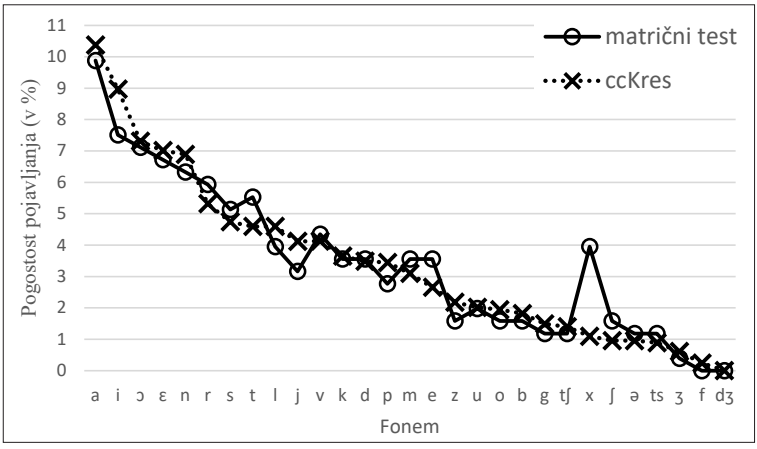
fonem	ccKres
v	4.13
k	3.65
d	3.48
p	3.45
m	3.10
e	2.66
z	2.18
u	2.02
o	1.94
b	1.82

fonem	ccKres
g	1.49
tʃ	1.40
x	1.10
ʃ	0.96
ə	0.96
ts	0.89
ʒ	0.61
f	0.24
dʒ	0.01

5 Matrični test za slovenščino

Na osnovi kriterijev, ki so bili podrobno predstavljeni v razdelku 3, smo matrični test konstruirali ročno, pri čemer smo si pomagali z računalniškim programom, ki je med sestavljanjem matrice sproti izračunaval pogostost fonemov v njej in nas tako usmerjal proti končni želeni porazdelitvi iz Tabele 2. Izdelani slovenski matrični test je predstavljen v Tabeli 3; pod vsako besedo je navedena še fonemska transkripcija s simboli mednarodne fonetične abecede (IPA).

Porazdelitev fonemov v korpusu ccKres v primerjavi s porazdelitvijo fonemov v osnovni matrici za slovenščino je predstavljena na Sliki 2.



Slika 2: Pogostost fonemov (v %) v matričnem testu in v korpusu ccKres.

Tabela 3: Seznam besed s fonemsko transkripcijo v slovenskem matričnem testu.

Ime	**Glagol**	Števnik	**Pridevnik**	**Samostalnik**
Gregor /grɛgɔr/	**kupi** /kupi/	**pet** /pet/	**velikih** /vɛlikix/	**stolov** /stɔlɔv/
Tone /tonɛ/	**dobi** /dɔbi/	šest /ʃest/	**lepih** /lepix/	**copat** /tsɔpat/
Jure /jurɛ/	**najde** /najdɛ/	**sedem** /sedəm/	**novih** /nɔvix/	škatel /ʃkatəl/
Urban /urban/	**skrije** /skrijɛ/	**osem** /osəm/	čudnih /ʧudnix/	**avtov** /avtɔv/
Sašo /saʃɔ/	**vzame** /vzamɛ/	**enajst** /ɛnajst/	**starih** /starix/	**zvezkov** /zvezkɔv/
Branka /branka/	**ima** /ima/	**sto** /sto/	**dobrih** /dɔbrix/	**koles** /kɔles/
Jana /jana/	**pelje** /peljɛ/	**tristo** /tristɔ/	**dragih** /dragix/	**kamnov** /kamnɔv/
Nada /nada/	**nese** /nɛsɛ/	**tisoč** /tisɔʧ/	**modrih** /modrix/	**majic** /majits/
Lara /lara/	**proda** /prɔda/	**nekaj** /nekaj/	**rumenih** /rumɛnix/	**loncev** /lɔntsɛv/
Petra /petra/	**išče** /iʃtʃɛ/	**malo** /malɔ/	**zelenih** /zɛlɛnix/	**nožev** /nɔʒɛv/

Predlagani slovenski matrični test ni edini možen, vendar zadovoljivo izpolnjuje vse različne zahteve iz standarda, njegova fonemska uravnoteženost pa presega uravnoteženost testov za poljščino in ruščino; za primerjavo glej Ozimek idr. (2010) ter Warzybok idr. (2015). Iz grafa je razvidno, da izrazito odstopa pojavnost fonema /x/, ki je v testu prisoten 3,5 krat pogosteje kot sicer v jeziku (podobno velja za poljski in ruski test). To lahko pripišemo lastnostim, ki sledijo iz predpisane skladenjske zgradbe povedi v testu: pridevniki, ki sledijo številnikom oz. kvantifikacijskim izrazom, se pojavijo v rodilniku množine, ki se pri vseh pridevnikih konča na fonem /x/ (*velikih*, *lepih*, *novih* itd.).

6 Zaključek

V članku smo predstavili postopek izbire besed za slovenski matrični stavčni test, ki se bo v klinični avdiologiji uporabljal za preizkus sluha pri govoru. Glavni poudarek prispevka je na določitvi fonemske porazdelitve v jezikovnem gradivu za test, saj se je ta morala čim bolj približati porazdelitvi fonemov v slovenščini. Ker fonemska

porazdelitev za slovenščino še ni analizirana, smo jo v pričujoči raziskavi vzpostavili tako, da smo črkovno porazdelitev v korpusu pisne slovenščine ccKres dopolnili s podatki o porazdelitvi fonemov v primerih, kjer črkovni zapis ne ustreza fonemskemu, pri čemer smo si pomagali s fonetično podatkovno bazo v Mihelič (2006). Na osnovi ugotovljene fonemske porazdelitve smo nato predlagali besede za slovenski matrični test.

Fonemska uravnoteženost je osrednja jezikoslovna tema pričujočega raziskovalnega dela, rezultat pa doprinos k avdiometričnim in statističnim glasoslovnim raziskavam na področju slovenskega jezika. Poleg tega lahko delo služi kot model nadaljnjih, še bolj natančnih določitev fonemske ali fonetične porazdelitve glasov v slovenščini. Ob tem velja – v luči prihodnjih raziskav – poudariti relativno naravo predstavljenih rezultatov. Kot prvo naj izpostavimo teoretična izhodišča, ki zadevajo odnos fonem-alofon-črka. Pri naši raziskavi smo privzeli fonemski sestav kot v Toporišič (2000), obstajajo pa tudi alternativni predlogi – tako Jurgec (2011) npr. že v osnovnem inventarju slovenskih fonemov prepoznava še dodatni samoglasniški fonem, ki ga v naši obravnavi nismo upoštevali. Nadalje je dobljena porazdelitev fonemov odvisna tudi od korpusa, iz katerega črpamo podatke o pogostosti fonemov. V pričujočem članku smo izbrali največji uravnotežen korpus pisne slovenščine ccKres, ki pa seveda ni brez pomanjkljivosti – v njem je najti napačno zapisane besede, besede iz tujih jezikov, ki niso del slovenščine, in podobno. Na področju glasoslovnih raziskav slovenskega jezika je torej odprtih še mnogo raziskovalnih izzivov.

Viri in literatura

Michael A. Akeroyd, Stig Arlinger, Ruth A. Bentler, Arthur Boothroyd, Nobert Dillier, Wouter A. Dreschler, Jean-Piere Gagne, Mark Lutman, Jan Wouters, Lena Wong in Birger Kollmeier, 2015: International Collegium of Rehabilitative Audiology (ICRA) Recommendations for the Construction of Multilingual Speech Tests. *International Journal of Audiology*, Early Online. 1–6.

ccKres, korpusna besedilna zbirka. Na spletu.

Gigafida, korpusna besedilna zbirka. Na spletu.

Gos, korpus govorjene slovenščine. Na spletu.

Björn Hagerman, 1982: Sentences for testing speech intelligibility in noise. *Scandinavian Audiology* 11. 79–87.

Karl Heinz Hahlbrock, 1953: Über Sprachaudiometrie und neue Wörterteste. *Arch Ohren Nasen Kehlkopfheilkd* 162. 394–431.

Karl Heinz Hahlbrock, 1960: Kritische Betrachtungen und vergleichende Untersuchungen der Schubertschen und Freiburger Sprachteste. *Zeitschrift für Laryngologie, Rhinologie, Otologie und Ihre Grenzgebiete* 39. 100.

Sabine Hochmuth, Brand Thomas, Melanie A. Zokoll, Franz Jozef Zenker Castro, Nina Wardenga idr., 2012: A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology* 51. 536–44.

Primož Jakopin, 1999. *Zgornja meja entropije pri besedilih v slovenskem jeziku*: Doktorska disertacija. Ljubljana. Na spletu.

Peter Jurgec, 2011: Slovenščina ima 9 samoglasnikov. *Slavistična revija* 59/3. 243–68.

Zdravko Kačič, Bogomir Horvat, Aleksandra Markuš Zögling, Robert Veronik, Matej Rojc, Andrej Žgank, Mirjam Sepesy Maučec in Tomaž Rotovnik, 2002: SNABI Database for Continuous Speech Recognition 1.2. Slovenian language resource repository CLARIN.SI. Na spletu.

Aleksander Ključevšek, 2016: *Statistična analiza slovenskih jezikovnih korpusov*: Magistrsko delo. UL FRI. Ljubljana. Na spletu.

Nataša Logar Berginc in Simon Krek, 2012: New Slovene corpora within the communication in Slovene project. *Prace Filologiczne* 63. 197–207.

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, FDV.

Tatjana Marvin, Jure Derganc in Saba Battelino, 2017: Adapting the Freiburg Monosyllabic Word Test for Slovenian. *Linguistica* 57/1. 197–210.

Aleš Mihelič, 2006: *Sistem za umetno tvorjenje slovenskega govora, ki temelji na izbiri in združevanju nizov osnovnih govornih enot*: Doktorska disertacija. Univerza v Ljubljani.

Frank E. Musiek, Gail D. Chermak, Jeffrey Weihing , Megan Zappulla in Stephanie Nagle, 2011: Diagnostic accuracy of established central auditory processing test batteries in patients with documented brain lesions. *Journal of the American Academy of Audiology* 22/6. 342–58.

Edward Ozimek, Anna Warzybok in Dariusz Kutzner, 2010: Polish sentence matrix test for speech intelligibility measurement in noise. *International Journal of Audiology* 49. 444–54.

Reiner Plomp in A.M. Mimpen, 1979: Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 18. 43–53.

Janko Pompe, 1968: *Razvoj avdiometrije na ORL kliniki v Ljubljani*. Neobjavljen rokopis. Univerzitetni klinični center Ljubljana.

Slovenski pravopis. Elektronska objava, 2014. Ljubljana: Založba ZRC, ZRC SAZU.

Barbara Suhadolc, 2013: *Statistična analiza slovenskih besedil*: Diplomsko delo. UL FRI. Na spletu.

Jože Toporišič, 2000: *Slovenska slovnica*. Maribor: Obzorja.

Kirsten Wagener, Brand Thomas in Kollmeier Birger, 1999a: Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil II: Optimierung des Oldenburger Satztests. *Zeitschrift für Audiologie* 38. 44–56.

Kirsten Wagener, Brand Thomas in Birger Kollmeier, 1999b: Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (in German). *Zeitschrift für Audiologie* 38. 86–95.

Kirsten Wagener, Kühnel Volker in Birger Kollmeier, 1999c: Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests (in German). *Zeitschrift für Audiologie* 38. 4–15.

Anna Warzybok, Melanie Zokoll, Nina Wardenga, Edward Ozimek, Maria Boboshko in Birger Kollmeier, 2015: Development of the Russian Matrix Sentence Test. *International Journal of Audiology* 54. 35–43.

Summary

This paper presents a word selection process for a sentence test with a matrix structure that has been developed for Slovenian. The research is important from the clinical as well as the linguistic point of view. This will be the first matrix test available for Slovenian, and it will be used for a more accurate assessment of hearing in people with a hearing disorder, for assessing the understanding of speech in people with central hearing disorders and comprehension disorders, for assessing cognitive abilities, and for assessing the improvement of speech comprehension in patients using various removable and implanted mechanical and electronic hearing aids and in patients with disturbing tinnitus. In creating the test, we followed the guidelines of the International Collegium of Rehabilitative Audiology (Akeroyd et al. 2015), which provide the steps necessary to create a matrix test in any given language. In this paper, we focus on the preparation of the linguistic material, where the standard procedure crucially requires that the phonemic distribution in the words chosen for the test approximates as closely as possible the distribution in the language of the test. As the phonemic distribution of Slovenian has not yet been analyzed, we derive it from data on letter distribution based on the corpus ccKres (see Logar Berginc and Krek 2012; Logar Berginc et al. 2012 for more information on the corpus) in combination with data on phonetic distribution that is available in Mihelič (2006). The result is a proposal of a phonemically balanced matrix test for Slovenian.