Končni super pretvorniki za predstavitev slovarjev izgovarjav pri sintezi govora

Žiga Golob*, Jerneja Žganec Gros*, Simon Dobrišek†

*Alpineon d.o.o.
Ljubljana, Slovenija
{ziga.golob, jerneja.gros}@alpineon.si
†Fakulteta za elektrotehniko
Univerza v Ljubljani, Slovenija
simon.dobrisek@fe.uni-lj.si

Povzetek

Končni pretvorniki predstavljajo kompakten način za predstavitev slovarjev izgovarjav, ki jih potrebujemo pri sintezi govora. V članku je predstavljen nov tip končnih pretvornikov, t.i. končni super pretvorniki, s katerimi lahko slovar predstavimo z manjšim številom stanj in prehodov kot s pomočjo minimalnega deterministčnega končnega pretvornika. Končni super pretvornik ohranja deterministčnost, poleg besed iz slovarja pa lahko dodatno sprejme tudi nekatere druge, neznane besede. Pri tem so lahko oddani izhodni alofonski prepisi za določene neznane besede napačni, vendar se izkaže, da je napaka primerljiva s trenutno najboljšimi metodami za določanje grafemsko-alofonske pretvorbe.

Finite-state super transducers for representing pronunciation lexicons in speech synthesis

Finite-state transducers are well suited for compact representations of pronunciation lexicons used in speech synthesis. In this paper, we present a finite-state super transducer, which is a new type of finite state transducer that allows the representation of a pronunciation lexicon with fewer states and transitions than using a conventional minimized and determinized finite-state transducer. A finite-state super transducer is a deterministic transducer that can, in addition to the words comprised in the pronunciation lexicon, accept some other, unknown words as well. The resulting allophone transcription for these words can be false, but we demonstrate that such errors are comparable to the performance of state-of-the-art methods for grapheme-to-phoneme conversion.

1. Uvod

Ključen del pri sintezi govora je sistem za pretvorbo grafemskega zapisa besed v njihov alofonski prepis. Samodejno določanje alofonskega prepisa v slovenščini temelji na množici kontekstno odvisnih pravil, pri čemer moramo poznati besedni naglas (Gros in Mihelič, 1999). Na žalost pa samodejno določanje besednega naglasa slovenskih besed predstavlja težko nalogo (Golob, 2009), zato je za kvalitetno sintezo govora nujna uporaba obsežnih slovarjev izgovarjav.

Slovar izgovarjav predstavlja preslikavo grafemskih zapisov besed v alofonske prepise. Pri pregibno bogatih jezikih, kot je slovenščina, lahko slovarji vsebujejo več milijonov slovarskih vnosov, zaradi česar je lahko njihova uporaba v pomnilniško manj zmogljivih sistemih, kot so npr. vgrajeni sistemi, problematična. V teh primerih je nujna uporaba postopkov, ki omogočajo pomnilniško učinkovito predstavitev slovarjev.

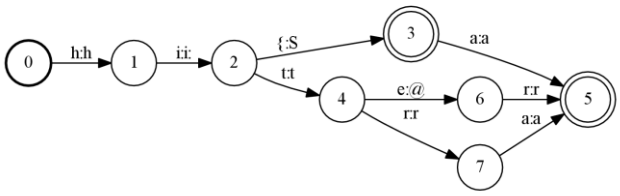
V literaturi je mogoče zaslediti predvsem tri metode, ki omogočajo pomnilniško učinkovito predstavitev slovarjev izgovarjav, in sicer s pomočjo oštevilčenih končnih avtomatov (Lucchesi in Kowaltowski, 1993; Daciuk in Piskorski, 2011), dreves predpon (Ristov, 2005) ter končnih pretvornikov (odslej kratko KP) (Mohri, 1994; Golob at al., 2012). V tem delu bomo predstavili nov način predstavitve s pomočjo končnih super pretvornikov (odslej kratko KSP), ki predstavljajo nekakšno nadgradnjo KP. Poleg manjše predstavitve slovarjev v primerjavi s KP, lahko s KSP z visoko točnostjo določimo alofonski prepis tudi nekaterim neznanim besedam oz. besedam, ki niso vsebovane v izvirnem slovarju.

V članku bomo najprej na kratko predstavili KP ter prikazali, kako lahko z njimi predstavimo slovar izgovarjav. Nadalje bomo pokazali, da zastopanost pregibnih oblik v slovarju močno vpliva na velikost KP. Sledila bo predstavitev t.i. KSP, ki predstavljajo nov način predstavitve slovarjev, nazadnje pa bomo podali še rezultate predstavitve slovarja s KSP ter ocenili napako, ki jo naredimo, če s KSP poskušamo narediti grafemsko-alofonsko pretvorbo za besede, ki niso del slovarja.

1.1. KP ter predstavitev slovarja izgovarjav

KP sestavljajo stanja ter prehodi med stanji. Vsak prehod ima vhodno in izhodno oznako. Ko se na vhodu KP pojavi določen vhodni niz, se ta nahaja v začetnem stanju. KP nato po vrsti sprejema vhodne simbole. Pri vsakem sprejetju vhodnega simbola odda izhodni niz simbolov, ki ga določa izhodna oznaka pripadajočega prehoda, ter se premakne v naslednje stanje. Če za poljuben vhodni simbol v trenutnem stanju ne obstaja prehod, ki ima vhodno oznako enako temu simbolu, pravimo, da KP vhodnega niza ne sprejema. Če se KP po prejetju vseh simbolov vhodnega niza nahaja v končnem stanju, pravimo, da vhodni niz sprejema, pri tem pa postane oddan izhodni niz veljaven. Omenimo še to, da je lahko vhodna ali/in izhodna oznaka enaka praznemu simbolu oziroma nizu.

KP, ki imajo v poljubnem stanju največ en prehod z določeno vhodno oznako, pravimo deterministični KP. Za takšne KP je hitrost pretvorbe vhodnega niza v izhodni niz zelo hitra in ob primerni izvedbi odvisna samo od dolžine vhodnega niza. Druga prednost determinističnih KP je ta, da obstajajo učinkoviti algoritmi za njihovo minimizacijo. Tako dobimo minimalni KP, ki ima najmanjše število prehodov in stanj med vsemi ekvivalentnimi KP (Mohri, 1997), torej KP, ki za poljuben sprejet vhodni niz oddajo enak izhodni niz.

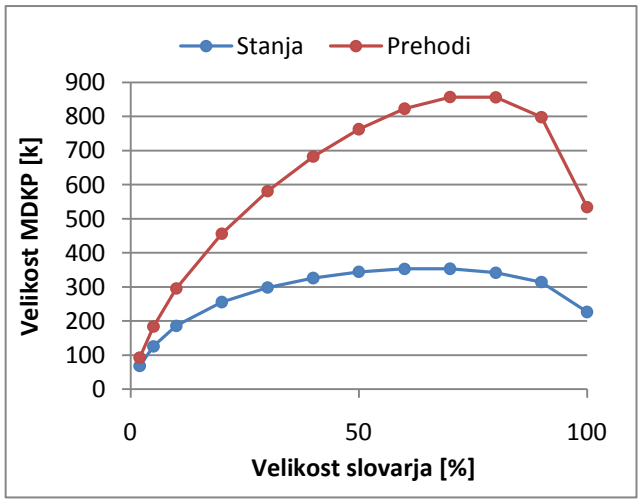


Slika 1: Primer KP, ki predstavlja slovar izgovarjav za tri slovenske besede. Krogi predstavljajo stanja, puščice pa prehode med stanji. Vsak prehod je označen z vhodno in izhodno oznako, ki sta ločeni z dvopičjem. Začetno stanje je označeno z odebeljenim krogom, končna stanja pa z dvojnim krogom.

Vseh KP ni mogoče determinizirati, saj imajo deterministični KP manjšo izrazno moč kot nedeterministični (Hellis, 2004). KP, ki predstavlja slovar izgovarjav, lahko vedno determiniziramo, če iz slovarja odstranimo enakopisnice. Slika 1 prikazuje primer minimiziranega in determiniziranega KP (odslej kratko MDKP), ki predstavlja slovar za štiri slovenske besede.

2. Vpliv velikosti slovarja izgovarjav na velikost KP

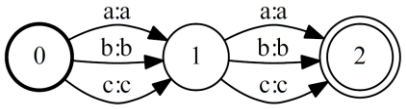
V tem eksperimentu smo želeli preveriti odvisnost velikosti KP od velikosti slovarja, ki ga želimo predstaviti. Na voljo smo imeli slovar SI-PRON za slovenski jezik, ki vsebuje več kot milijon različnih slovarskih vnosov (Žganec-Gros, Cvetko-Orešnik, Jakopin, 2006). Z naključnim izbiranjem slovarskih vnosov smo zgradili 11 pod-slovarjev različnih velikosti in za vse pod-slovarje zgradili MDKP. Rezultate števila stanj in prehodov pridobljenih MDKP prikazuje graf 1.



Graf 1: Odvisnost velikosti MDKP od velikosti slovarja izgovarjav, pri čemer so vnosi v slovar izbrani naključno iz prvotnega slovarja. Opazimo lahko obrat trenda rasti števila stanj in prehodov MDKP.

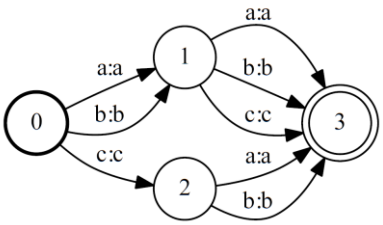
Iz rezultatov lahko razberemo, da velikost MDKP doseže vrh pri 70% do 80% velikosti prvotnega slovarja. Z drugimi besedami, velikost MDKP začne pri določeni velikosti z dodajanjem novih besed oz. slovarskih vnosov iz slovarja padati.

Da bi si ta pojav lahko lažje predstavljali, poglejmo minimalni primer, ki prikazuje mehanizem tega zmanjšanja velikosti MDKP. Kot primer vzemimo izmišljen slovar, katerega ključi[1] so sestavljeni iz vseh možnih izborov dveh črk od treh možnih, npr. črk *a*,*b* in *c*. Na ta način dobimo 9 različnih ključev, in sicer: *aa*, *ab*, *ac*, *ba, bb*… Zaradi enostavnosti naj bodo pripadajoče vrednosti enake ključem. MDKP za ta slovar prikazuje slika 2.



Slika 2: MDKP za izmišljen slovar, katerega ključi so sestavljeni iz vseh možnih izborov dveh črk od treh možnih – *a*, *b* in *c*. Pri tem so vrednosti enake ključem.

Sedaj iz našega izmišljenega slovarja odstranimo slovarski vnos *cc : cc* ter ponovno zgradimo MDKP. Rezultat prikazuje slika 3.



Slika 3: MDKP za enak slovar, kot ga predstavlja MDKP na sliki 2, pri čemer mu manjka slovarski vnos *cc : cc*.

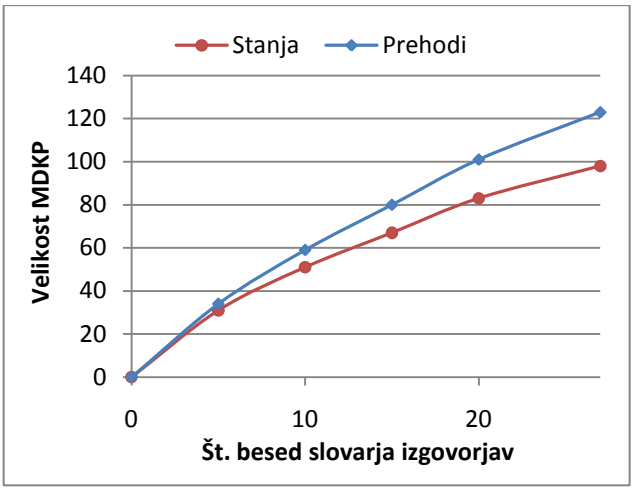
Opazimo lahko, da se je pri odstranitvi slovarskega vnosa iz slovarja kompleksnost MDKP povečala, saj je za predstavitev slovarja potrebno eno dodatno stanje ter dva dodatna prehoda.

V naslednjih poglavjih bomo podrobneje raziskali vzroke, ki vplivajo na zmanjšanje MDKP pri predstavitvi slovarja pri dodajanju novih slovarskih vnosov v slovar.

2.1. Vpliv množičnosti pregibnih oblik na velikost slovarja izgovarjav

Preverili smo vpliv množičnosti pregibnih oblik lem besed iz slovarja na velikost MDKP. Pri tem z množičnostjo pregibnih oblik mislimo na število različnih pregibnih oblik za določeno lemo. Za primer smo vzeli besedo *skopati* ter v slovarju poiskali vse slovarske vnose, katerih grafemski zapisi predstavljajo pregibne oblike leme izbrane besede. Dobili smo 27 različnih slovarskih vnosov, iz katerih smo s pomočjo naključnega izbiranja vnosov tvorili še štiri različno velike pod-slovarje. Za vsak pod-slovar smo zgradili MDKP. Rezultate prikazuje graf 2.

[1] Slovarski vnosi so sestavljeni iz para ključ, vrednost. Pri slovarju izgovarjav tako grafemski zapis predstavlja ključ, alofonski prepis pa vrednost.



Graf 2: Odvisnost velikosti MDKP od števila besed v slovarju izgovarjav. Vse besede slovarja pripadajo isti lemi.

Iz rezultatov je razvidno, da hitrost naraščanja velikosti MDKP z večanjem slovarja rahlo pada, vendar pa ni opaziti obrata trenda povečevanja MDKP.

2.2. Vpliv zastopanosti pregibnih oblik na velikost slovarja izgovarjav

Poglejmo sedaj, kako na velikost MDKP vpliva zastopanost pregibnih oblik v slovarju, sestavljenem iz večih besed, ki se podobno pregibajo. Iz slovarja SI-PRON smo izbrali 28 grafemskih zapisov besed, katerih pregibne oblike imajo 9 različnih končnic ter pripadajo štirim različnim lemam - *potop, osmod, zasp, natoč*. Izbrane leme ter pripadajoče končnice so prikazane v tabeli 1. Lema *zasp* pri tem predstavlja izjemo, ki se pregiba nekoliko drugače kot ostale tri.

MOŽNE LEME	MOŽNE KONČNICE
potop, osmod, zasp, natoč	iš,im,imo,ite,ijo
potop, osmod,natoč	i+l,i+li
zasp	a+l, a+li

Tabela 1: Tabela prikazuje postopek za tvorjenje vseh besed, ki so vsebovane v slovarju. V levem stolpcu so navedene leme besed, v desnem pa možne končnice.

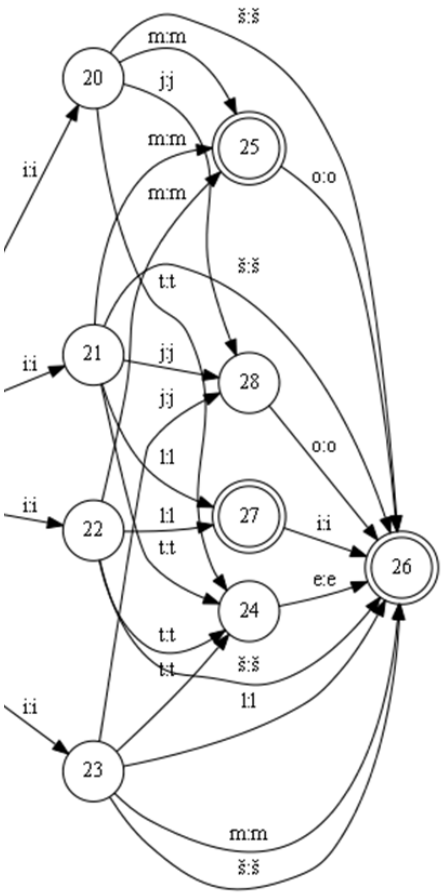
Iz teh besed smo nato tvorili slovar, pri čemer smo zaradi enostavnosti vrednosti ključev izenačili s ključi. Nato smo z naključnim izbiranjem iz tega slovarja tvorili še štiri različno velike pod-slovarje. Za vse tako zgrajene slovarje smo nato zgradili MDKP. Rezultate prikazuje tabela 2.

ŠT. BESED	ŠTEVILO STANJ	ŠTEVILO PREHODOV
5	26	29
9	29	35
17	29	40
23	29	45
28	26	36

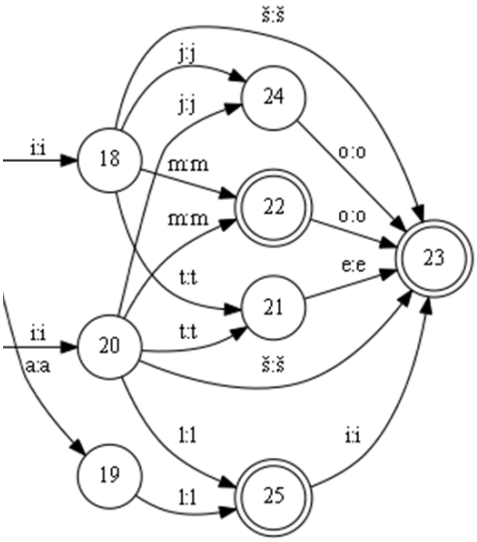
Tabela 2: Tabela prikazuje rezultate števila stanj in prehodov MDKP za vse velikosti pod-slovarjev.

Iz tabele je razvidno, da je velikost MDKP, ki predstavlja vseh 28 vnosov slovarja, manjša od MDKP, ki predstavljata slovarja s 23 in 17 vnosi, število stanj pa je večje celo pri MDKP, ki predstavlja slovar z 9 vnosi. Rezultati nakazujejo, da zastopanost pregibnih oblik močno vpliva na kompleksnost pridobljenega MDKP ter lahko vpliva na obrat trenda rasti velikosti MDKP.

Sliki 4 in 5 prikazujeta shematski prikaz dela MDKP, ki predstavlja končnice besed, pri čemer slika 4 pripada MDKP za slovar s 23 vnosi, slika 5 pa MDKP za slovar z 28 vnosi. Razvidno je, da je kompleksnost MDKP, ki predstavlja slovar s 23 vnosi, precej večja.



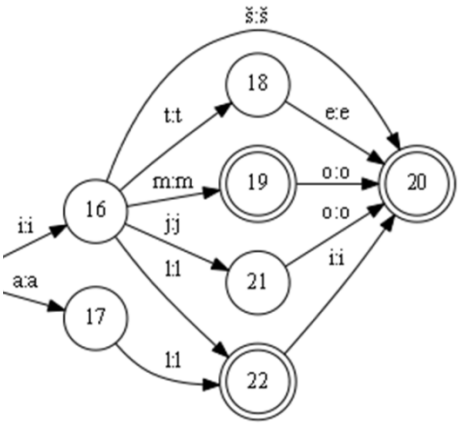
Slika 4: Del MDKP, ki predstavlja slovar s 23 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.



Slika 5: Del MDKP, ki predstavlja celoten slovar z vsemi 28 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.

Smiselno je torej, da so v slovarju, ki ga želimo realizirati s KP, prisotne vse možne pregibne oblike, saj si lahko v tem primeru leme, ki se enako pregibajo, del končnega pretvornika, ki pretvarja končnice, v celoti delijo. Kompleksnost pri tem še vedno povečujejo besede oz. leme besed, ki imajo med pregibnimi oblikami kakšno izjemo, ki se pregiba nekoliko drugače. V našem izmišljenem slovarju je to lema *zasp*, katere dve pregibni obliki imata nekoliko drugačno končnico, in sicer končnico *al* ter *ali* namesto *il* ter *ili*.

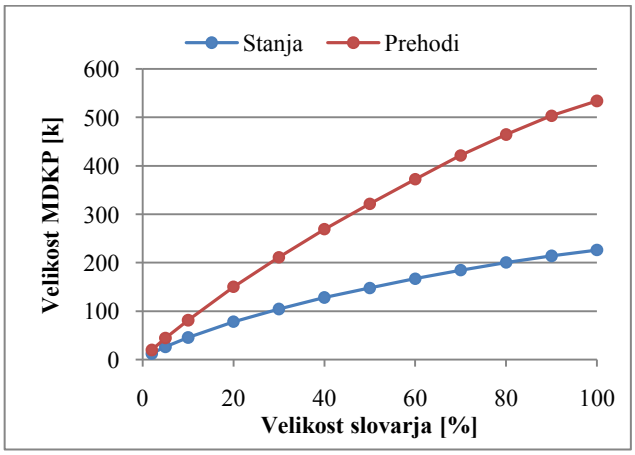
MDKP sprejme samo vnose, ki so vsebovani v slovarju. Če za določeno aplikacijo tako stroga zahteva ni potrebna in je dovolj, da MDKP sprejme vse vnose iz slovarja, ga lahko naprej poenostavimo. Še enostavnejšo obliko bi namreč dobili, če bi za vse štiri leme iz slovarja obstajale pregibne oblike za vseh 9 možnih končnic. V slovar lahko tako dodamo dodatne vnose in sicer vnose z lemami *potop*, *osmod*, *natoč* ter končnicama *al* ter *ali*, ter vnosa z lemo *zasp* in končnicama *il* ter *ili*. Pridobljeni slovar ima tako 36 vnosov, MDKP pa se poenostavi na 23 stanj in 30 prehodov. Shema dela MDKP, ki pretvarja končnice, je prikazana na sliki 6.



Slika 6: Del MDKP, ki predstavlja slovar izgovarjav s 36 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.

Struktura MDKP, prikazana na sliki 6, ki predstavlja slovar z 28 prvotnimi vnosi ter dodatnimi 8 vnosi je torej še nekoliko bolj enostavna kot struktura MDKP, ki predstavlja slovar z le 28 prvotnimi vnosi. Z dodatnimi vnosi smo torej poenostavili strukturo MDKP. S pomočjo KSP, ki ga bomo predstavili v naslednjem poglavju, bomo to idejo posplošili.

Eksperimenti, ki nakazujejo, da na obrat trenda rasti MDKP vpliva predvsem zastopanost pregibnih oblik, so bili izvedeni na poenostavljenem slovarju, katerega vrednosti so bile enake ključem. Da bi pokazali, da podobno velja tudi v primeru dejanskih slovarjev izgovarjav, smo iz slovarja izgovarjav SI-PRON ponovno tvorili 11 različno velikih pod-slovarjev z naključnim izbiranjem, vendar pa smo v tem primeru naključno izbirali le leme besed, nato pa smo vključili še vse pripadajoče pregibne oblike. Za vse pod-slovarje smo nato zgradili MDKP. Rezultate prikazuje graf 3.



Graf 3: Odvisnost velikosti MDKP od velikosti slovarja izgovarjav, pri čemer so v slovarju vedno vsebovane vse pregibne oblike. Povečevanje MDKP je tokrat skoraj linearno odvisno od števila vnosov v slovarju. Opaziti je le rahlo upadanje trenda rasti.

Vidimo lahko, da tokrat ne pride do obrata trenda rasti, kar potrjuje našo hipotezo.

3. Končni super pretvornik (KSP)

V prejšnjem poglavju smo pokazali, da lahko s pomočjo dodatnih, izbranih slovarskih vnosov v slovar zmanjšamo kompleksnost MDKP. Problem predstavlja iskanje takšnih slovarskih vnosov, ki bi zmanjšali kompleksnost, še posebej v primeru realnih slovarjev, kot so npr. slovarji izgovarjav, ki so prvič večji, drugič pa se ključ in vrednost posameznih slovarskih vnosov razlikujeta, s čimer je iskanje primernih slovarskih vnosov težja naloga. Problema smo se zato lotili na drugačen način, in sicer tako, da smo združevali določena stanja, pri čemer smo želeli zadostiti naslednjima dvema pogojema:

- Pridobljen KP mora ostati deterministične.
- Pridobljen KP mora sprejemati vse ključe prvotnega slovarja ter za sprejete ključe oddati pravilne pripadajoče vrednosti.

Tako smo lahko združevali samo stanja, ki so imela določene lastnosti. Takšna stanja smo poimenovali

združljiva stanja. Dve stanji sta združljivi, če zadoščata naslednjim pogojem.

- Če je eno od stanj končno stanje, stanji ne smeta imeti izhodnih prehodov s praznimi vhodnimi simboli oz. ε simboli. Rezultat združevanja takšnih stanj je lahko nedeterministični KP.
- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter različnimi izhodnimi simboli.
- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter enakimi izhodnimi simboli, ki prehajajo v različna naslednja stanja, ki so nezdružljiva.

Da bi lahko določili združljiva stanja, je potrebno preveriti zgornje pogoje, kar pa je v praksi lahko problematično, saj je preverjanje združljivosti stanj zaradi rekurzivnosti, ki je lahko ciklična, zahtevno. V ta namen smo zadnji pogoj poenostavili:

- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter enakimi izhodnimi simboli, ki prehajajo v različna naslednja stanja.

Zaradi poenostavitve pogoja za združljivost stanj nekaterih združljivih stanj nismo mogli zaznati.

KSP smo zgradili tako, da smo najprej zgradili MDKP, nato pa smo nadalje združili vsa stanja, ki so združljiva. Za vsako stanje je bilo potrebno preveriti, ali je združljivo s katerim koli drugim stanjem. Ker nekatera stanja postanejo združljiva šele, ko združimo neka druga stanja, je bilo potrebno to storiti v več iteracijah.

4. Predstavitev slovarja izgovarjav s KSP

Za slovar izgovarjav SI-PRON smo najprej zgradili MDKP s pomočjo odprtokodnega orodja OpenFST (Cyril at al., 2007), nato pa smo s postopkom, ki smo ga opisali v poglavju 3, zgradili še KSP. Tabela 3 prikazuje število stanj in prehodov MDKP in KSP.

		MDKP	KSP	Zmanj.
En izhodni simbol	Stanja	226.363	172.833	23.6%
	Prehodi	534.061	428.114	19.8%

Tabela 3: Zmanjšanje števila stanj in prehodov pri gradnji KSP iz MDKP.

Opazimo lahko, da smo velikost MDKP uspeli zmanjšati za približno 20%.

Čeprav lahko s KSP vnose v slovarju predstavimo z manjšim KP kot v primeru MDKP, pri tem izgubimo informacijo o tem, katere besede so vsebovane v slovarju. Tako se lahko zgodi, da KSP sprejme določeno besedo, ki je slovnično pravilna, vendar ni bila vsebovana v slovarju. V tem primeru je lahko oddan alofonski prepis napačen. V naslednjem poglavju smo poskušali oceniti napako, ki jo naredimo, če za predstavitev vnosov slovarja namesto MDKP uporabimo KSP.

4.1. Ocena verjetnosti napake KSP pri predstavitvi slovarja izgovarjav SI-PRON

Ker pri uporabi KSP izgubimo informacijo o tem, ali je določena beseda vsebovana v slovarju, lahko besede, ki niso del slovarja, pretvorimo napačno v njihov alofonski prepis. Če bi takšno informacijo imeli, bi lahko takšne besede namesto s KSP v alofonski prepis pretvorili s pomočjo kakšnih drugih metod, npr. s pomočjo metod strojnega učenja ali kakšnih drugih statističnih metod.

Da bi ocenili verjetnost napake, ki jo na ta način naredimo, smo slovar SI-PRON naključno razdelili v dva pod-slovarja, pri čemer je prvi vseboval 90% besed, drugi pa preostalih 10% besed in je služil kot testni del. Za prvi del smo zgradili MDKP ter KSP. Rezultate gradnje prikazuje tabela 4.

	MDKP	KSP	Zmanjšanje
Stanja	315.191	190.842	39.5%
Prehodi	800.026	478.315	40.2%

Tabela 4: Rezultati gradnje MDKP in KSP za pod-slovar, ki je vseboval 90% vnosov slovarja izgovarjav SI-PRON.

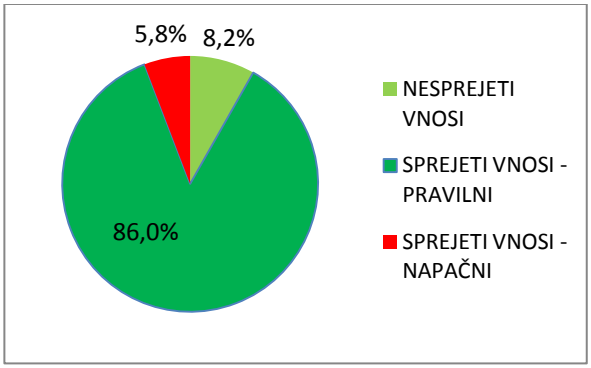
Opazimo lahko, da je tokrat zmanjšanje števila stanj in prehodov precej večje kot v primeru celotnega slovarja in je glede na MDKP približno 40%. Poleg tega je končno število stanj in prehodov manjše kot v primeru gradnje MDKP za celoten slovar. Iz tega lahko sklepamo, da je gradnja KSP še posebej smiselna, ko v slovarju niso vsebovane vse pregibne oblike.

Nadalje smo grafemske zapise 124.099 slovarskih vnosov iz testnega dela slovarja dali na vhod zgrajenega KSP. Za grafemske zapise, ki jih je KSP sprejel, smo spremljali, če se pri tem oddan alofonski prepis ujema z alofonskim prepisom pripadajočega slovarskega vnosa. Rezultate prikazuje tabela 5.

Št. vseh testnih vnosov	124.099
Nesprejeti testni vnosi	10.190
Sprejeti testni vnosi (pravilni)	106.698
Sprejeti testni vnosi (napačni)	7.211

Tabela 5: Tabela prikazuje število sprejetih ter nesprejetih grafemskih zapisov slovarskih vnosov testnega slovarja, ko te damo na vhod KSP. Pravilne sprejete vnose predstavljajo tisti slovarski vnosi, pri katerih je KSP oddal pravilen alofonski prepis.

Bolj pregledno razmerja med posameznimi skupinami slovarskih vnosov prikazuje graf 4.



Graf 4: Razmerja med nesprejetimi ter sprejetimi vnosi, med katerimi nadalje ločimo tiste, za katere oddan alofonski prepis je bil bodisi pravilen bodisi napačen.

Iz rezultatov lahko razberemo, da je odstotek sprejetih vnosov kar 91.8%. Pri tem naj še enkrat opozorimo, da ti vnosi niso bili vsebovani v slovarju izgovarjav, iz katerega smo zgradili KSP. Od sprejetih vnosov je delež tistih, za katere je KSP oddal napačen alofonski prepis, le 6.7%.

5. Zaključek

V članku je predstavljen nov tip KP, ki smo jih poimenovali končni super pretvorniki (KSP), ki poleg želenih besed sprejemajo še nekatere druge z namenom, da lahko pretvorbo želenih besed predstavimo bolj kompaktno.

Pokazali smo, da lahko pri predstavitvi slovarja izgovarjav s pomočjo KSP število stanj in prehodov zmanjšamo za približno 20%, ko so za vsebovane leme v slovarju izgovarjav prisotne tudi vse pripadajoče pregibne oblike besed oz. kar 40% v primeru, ko vse pregibne oblike niso vsebovane.

Ker KSP sprejemajo še druge, neznane besede, za katere lahko oddajo napačen izhodni niz, so KSP uporabni predvsem v aplikacijah, kje ne potrebujemo informacije o tem, katere besede so vsebovane v KP ampak le informacijo o pravilni pretvorbi danih besed oz. besed, iz katerih smo zgradili KSP.

Za slovarje izgovarjav, ki jih uporabljamo pri sintezi govora, si želimo, da pokrivajo čim večji delež besed, saj omogočajo najvišjo stopnjo točnosti pri pretvorbi v alofonski prepis. Vseeno, razen v primeru zaprtih domen, ne morejo vsebovati vseh besed, ki se lahko pojavijo, saj se jezik nenehno spreminja in pri tem stalno nastajajo nove besede. Tako lahko pri uporabi KSP za predstavitev slovarja izgovarjav pride do napake pri pretvorbi v alofonski prepis, ko se na vhodu pojavi neznana beseda. Pokazali smo, da je ta napaka razmeroma majhna, saj je bilo za naš testni slovar od več kot 90% sprejetih besed napačno pretvorjenih le 5.8%. Vidimo torej, da lahko KSP uporabimo tudi kot prepoznavalnik za določanje alofonskega prepisa neznanim besedam, pri čemer je njegova napaka le 6.7%. Tako nizka napaka pa je primerljiva oz. celo manjša od napake namenskih prepoznavalnikov, kjer je ta za slovenski jezik odvisna predvsem od točnosti napovedovanja naglasnega mesta in se giblje nekoliko nad 15% (Golob, 2009).

Pri ocenjevanju verjetnosti napake KSP je bil prvotni slovar SI-PRON naključno razdeljen na testno in učno množico. Tako so bile pregibne oblike besed za določene leme lahko vsebovane tako v učni kot v testni množici. Ker so si pregibne oblike, ki pripadajo isti lemi, med seboj precej podobne, je napovedovanje alofonskega prepisa takšnim besedam iz testne množice, ki so vsebovane tudi v učni množici, nekoliko lažja naloga. V nadaljnjih poskusih bi bilo zato potrebno prvotni slovar razdeliti na učno in testno množico tako, da bi se naključno izbiralo le leme besed, nato pa bi se poleg vključile še vse pripadajoče pregibne oblike. V tem primeru pričakujemo, da bi bila napaka pri pretvorbi neznanih besed nekoliko višja.

6. Zahvala

Raziskovalno delo prvega avtorja je delno financirala Evropska unija iz evropskega socialnega sklada ter sklada za regionalni razvoj v okviru Operativnega programa krepitve regionalnih razvojnih potencialov za obdobje 2007 do 2013, po pogodbi št. P-MR-10/94.

7. Reference

Cyril A., Michael R., Johan S., Wojciech S., Mohri M., 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA 2007). Lecture Notes in Computer Science, Prague, Springer-Verlag, Heidelberg, Germany, 4783: 11-23.

Daciuk J., Piskorski J., Ristov S., 2011. Natural Language Dictionaries Implemented as Finite Automata. Scientific Applications of Language Methods. London : Imperial College Press, World Scientific Publishing.

Golob Ž., 2009. Samodejno določanje mesta besednega naglasa pri sintezi slovenskega govora. Diplomsko delo, fakulteta za elektrotehniko v Ljubljani.

Golob Ž., Žganec-Gros J., Žganec M., Vesnicer B., Dobrišek S., 2012. FST-Based Pronunciation Lexicon Compression for Speech Engines. *International Journal of advanced robotic systems,* 9: 2011.

Gros J., Mihelič F., 1999. Acquisition of an Extensive Rule Set for Slovene Grapheme-to-Allophone Transcription. Proceedings 6th European Conference on Speech Communication and Technology. September 5–9. 1999. Eurospeech 1999. Budapest, 5: 2075–2078.

Hellis T., 2004. On minimality and size reduction of one-tape and multitape finite automata. Doktorska disertacija.

Lucchesi C., Kowaltowski T., 1993. Applications of Finite Automata Representing Large Vocabularies. *Software-Practice & Experience,* 23: 15-30.

Mohri M., 1994. Compact Representations by Finite-State Transducers. 32nd Meeting of the Association for Computational Linguistics (ACL '94). Proceedings of the Conference. Las Cruces. NM, pp. 204–209.

Mohri M., 1997. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 33: 269–311.

Ristov S., 2005. LZ Trie and Dictionary Compression. Jurnual Software-Practice & Experience, pp. 445–465.

Žganec-Gros J., Cvetko-Orešnik V., Jakopin P., 2006. SI-Pron Pronunciation Lexicon: A New Language Resource for Slovenian. Informatica, 30: 447–452.