Šandor Dembitz
Jakov Pavlek
Dejan Stupar

PROBLEM STRANIH IMENA U STROJNOJ TVORBI GOVORA NA HRVATSKOME

Sažetak

U svakom sustavu za automatsku sintezu govora neke dijelove teksta treba pretprocesirati, tj. normalizirati, da bi postali izgovorljivi. To se općenito odnosi na brojeve, kratice, simbole različitih jedinica i strana imena. Hrvatski sustav pisanja je u osnovi fonološki, što olakšava preslikavanje grafema u foneme pri strojnoj tvorbi govora, no strana imena u hrvatskome u pravilu zadržavaju svoju izvornu grafiju. Stoga njih u sustavu za sintezu govora treba transkribirati prema hrvatskim transkripcijskim pravilima. Polazeći od usporedne analize dvaju hrvatskih megakorpusa, u radu se najprije istražuje udio stranih imena u prosječnom hrvatskom tekstu i dinamika njihova ulaska u hrvatski. Nadalje se opisuje postupak za automatsku identifikaciju jezika koji je testiran na uzorku koji čini više od 30.000 stranih imena i njihovih kosih oblika. Polazeći od rezultata ove klasifikacije, programski se pokušavaju transkribirati imena razvrstana kao njemačka ili talijanska. Točnost transkripcije od više od 90%, zajedno s točnošću razvrstavanja imena od oko 90%, ohrabrujuća je za daljnje napore na razvoju sustava.

Ključne riječi: sinteza govora, transkripcija stranih imena, hrvatski jezik

1. UVOD

Ohrabrujući rezultati u sintezi govora na hrvatskome, ostvareni uz pomoć sustava *SPICE*, potaknuli su istraživanja koja će biti opisana u ovom radu. Sustav *SPICE* (engl. ***S****peech* ***P****rocessing – **I**nteractive **C**reation and **E**valuation toolkit for new languages*) javno je dostupni sustav za razvoj govornih tehnologija učenjem iz uzoraka govora i teksta (SPICE, 2009). Razvija se na sveučilištu *Carnegie Mellon* u SAD-u, a voditelji projekta su Tanja Schultz i Alan Black (Kominek i Black, 2005, 2006; Schultz i Black, 2006; Kominek i sur., 2007, 2008; Schultz i sur., 2007).

Prvi pokusi sa sustavom *SPICE*, izvedeni na vrlo skromnom uzorku za uvježbavanje sustava (Šoić, 2008), pokazali su da je moguće bez većeg truda sintetizirati razumljiv govor iz teksta pisanog hrvatskim jezikom. Kako je uzorak bio građen od riječi hrvatskoga općejezičnog fonda, za punu uporabljivost *SPICE-a* potrebno je napraviti sustav za pretprocesiranje teksta koji bi obavljao tzv. normalizaciju, tj. nizove grafema koji ne pripadaju hrvatskom općejezičnom fondu pretvarao bi u nizove koje *SPICE* zna pretvoriti u fonetski oblik. Takav sustav morao bi se sastojati od najmanje triju podsustava: za pretprocesiranje brojeva, za

pretprocesiranje kratica, mjernih i drugih jedinica i sličnog, te za pretprocesiranje stranih imena.

Ne umanjujući važnost drugih podsustava, procijenili smo da je podsustav za normalizaciju stranih imena najveći izazov. Najprije treba prepoznati strano ime, zatim odrediti jezik iz kojeg ono dolazi, da bi se tek onda mogla primijeniti transkripcijska pravila za njegovu normalizaciju.

2. STRANA IMENA U HRVATSKOME

Za određivanje udjela stranih imena u pisanju na hrvatskome poslužili su nam nepreklapajući megakorpusi pisani hrvatskim jezikom te leksičke baze sustava *Hascheck* (HASCHECK, 2009), kao i programski alati razvijeni za potrebe tog sustava (Dembitz i sur., 1999, 2004).

Hascheck (hrvatski mrežni pravopisni provjernik) ima četiri leksičkih baza:

- bazu hrvatskog općejezičnog fonda koja trenutačno ima oko 700.000 različnica;
- bazu hrvatskog posebnojezičnog fonda (vlastita imena, kratice, mjerne i druge jedinice, akronimi, hrvatski pravopisni arhaizmi, riječi iz stranih jezika i njihovi kosi oblici koji se u izvornoj grafiji pojavljuju u pisanju na hrvatskome, itd.), koja trenutačno ima oko 400.000 različnica;
- bazu engleskog jezičnog fonda opsega oko 70.000 različnica koja je stabilna;
- bazu pogrešaka pronađenih u korpusu od 125.000.000 pojavnica koja trenutačno ima oko 900.000 različnica.

Budući da je *Hascheck* učeći sustav, opseg svih baza, osim engleske, raste s obradama.

U posebnojezičnom fondu 80% različnica su vlastita imena i njihovi kosi oblici. Formalna analiza, tj. prisutnost digrama, trigrama i tetragrama koji se ne pojavljuju u različnicama hrvatskog općejezičnog fonda, upućuje da je 50% različnica-imena, ili njih oko 160.000, stranoga podrijetla. Izbor kriterija temelji se na iskustvu iz nadgledanja *Haschecḳova* učenja.

2.1. Udio stranih imena u hrvatskom tekstu

Najprije je bilo potrebno utvrditi strukturu prosječnog teksta pisanog hrvatskim jezikom. Za to su nam poslužili *Haschecḳov* korpus obrađen u razdoblju od siječnja 2008. do siječnja 2009. (opseg korpusa je oko 50.000.000 pojavnica) te *Hrvatska jezična riznica* (RIZNICA, 2009), 80-milijunski korpus Instituta za hrvatski jezik i jezikoslovlje čiji je čestotnik sa stanjem u prosincu 2007. (korpus se nakon tog datuma neznatno povećao) prošao kroz sustav *Hascheck*. *Hascheck* je sa sigurnošću od 99,15% obavio razvrstavanje pojavnica u *Riznici*, dok se razdioba nerazvrstanih 0,85% pojavnica temelji na *Haschecḳovoj* procjeni: 0,15% u općejezične, 0,4% u posebnojezične i 0,3% u pogreške. Čestotnik *Riznice* ne sadrži brojeve, iako se oni u

korpusu pojavljuju, tako da je ta rubrika za *Riznicu* ostala prazna (tablica 1). Pojavnice-brojevi inače se iz korpusa izdvajaju trivijalnim postupkom.

Tablica 1. Struktura hrvatskog teksta
Table 1. Structure of Croatian text

Pojavnice	*Hascheck*	*Riznica*
Općejezične	89,5%	93,0%
Posebnojezične	5,9%	5,9%
Engleske	1,6%	0,5%
Brojevi	1,4%	
Pogreške	1,6%	0,6%

Oba korpusa u jednom su suglasni: blizu 6% pojavnica u tekstu su imena, kratice, akronimi, strane neengleske riječi u izvornoj grafiji, hrvatski kosi oblici imena i stranih riječi i slično. Udio od 5% navode Moguš i suradnici (1999).

Imena osoba, tvrtki, toponima, različitih proizvoda i usluga (i druga) te njihovi kosi oblici sudjeluju u posebnojezičnim pojavnicama obaju korpusa s udjelom od oko 90%. Na temelju n-gramske analize za 30% (uzorak iz *Riznice*) do 40% (*Haschecḱov* uzorak) imena s velikom sigurnošću može se reći da su stranoga podrijetla. Slijedom navedenih omjera dolazimo do udjela stranih imena u prosječnom hrvatskom tekstu: od 1,6% do 2,1%. Te brojke također upućuju da podsustavu za normalizaciju stranih imena treba posvetiti primjerenu pozornost jer su strana imena po težini treći konstitutivni element hrvatskoga teksta, nakon općejezičnog i hrvatskog imenskog dijela, a prije brojeva, ostatka posebnojezičnog sadržaja (kratice, simboli jedinica i drugo) i izravnih posuđenica iz engleskoga.

2.2. Ponašanje imena u korpusu

Ovisnost broja različnica o opsegu korpusa iskazuje se za sve prirodne jezike Heapsovim zakonom (Heaps, 1978),

$$V(t) = \alpha \cdot t^{\beta}, \qquad (1)$$

gdje je V broj različnica u korpusu, t opseg korpusa u pojavnicama, dok su α i β parametri ovisni o jeziku korpusa i karakteru korpusa, s tim da parametar β mora biti manji od 1. Heapsov zakon nije ništa drugo do Zipfov zakon u integralnom obliku (Kornai, 2002). *Haschecḱ*, kojemu baza različnica hrvatskoga općejezičnog fonda i baza različnica posebnojezičnog fonda rastu s obrađenim korpusom, dopušta da se za njegov korpus Heapsov zakon iskaže u aditivnom obliku:

$$V(t) = V_W(t) + V_N(t), \qquad (2)$$

Gdje je V_W broj različnica općejezičnog, a V_N broj različnica posebnojezičnog fonda u ovisnosti o veličini obrađenog korpusa. Sve tri funkcije iz (2) ravnaju se po zakonu (1), jasno svaka sa svojim parametrima. Valja također primijetiti da prema (2) V nije broj svih različnica u korpusu, nego broj dobro napisanih neengleskih i nebrojčanih različnica.

U tablici 2 dana je usporedba parametara funkcija iz (2) na razini 10--milijunskog, odnosno 100-milijunskog korpusa.

Tablica 2. Parametri Heapsova zakona za *Hascheckov* korpus
Table 2. Heaps' law parameters for *Hascheck* corpus

	10-milijunski korpus			100-milijunski korpus		
	V	V_W	V_N	V	V_W	V_N
α	207,2440	662,6670	0,43099	162,7710	4.610,6900	0,008444
β	0,4578	0,3716	0,74670	0,4728	0,2682	0,951500

Funkcije vrlo dobro koreliraju s empirijskim podacima; koeficijent korelacije je 0,997 ili viši. To funkcijama daje visoku prediktivnu vrijednost.

Na globalnoj razini (parametri za V u tablici 2) nastupile su najmanje promjene. Parametar β (nagib funkcije) za 100-milijunski korpus tek je 3% veći od istog parametra za 10-milijunski korpus. Nešto veće odstupanje dogodilo se s parametrom α koji je pao za 22%. Unatoč tome, razlika u izračunu vrijednosti V po obje formule na razini 100-milijunskog korpusa je svega 3,5%. To pokazuje da u Heapsovu zakonu parametar β dominantno određuje ponašanje funkcije, što ćemo upotrijebiti u analizi trendova funkcija V_W i V_N. Za predikciju nisu toliko važni apsolutni iznosi funkcija koliko su važne derivacije funkcija iz (2) jer one govore koliki opseg učenja po jedinici korpusa trebamo očekivati na dosegnutoj razini obrade.

Nagib funkcije V_W (parametar β) značajno pada s porastom korpusa. Pad se tek djelomice kompenzira porastom parametra α, no ta kompenzacija bit će nedovoljna kad se β jako približi nuli. Naše očekivanje je da će to stanje nastupiti kad korpus naraste za red veličine, tj. kad bude između milijardu i deset milijardi pojavnica jer će tada hrvatski općejezični fond ući u zasićenje (na 100.000 novih pojavnica bit će manje od deset novih općejezičnih različnica). Kao ilustraciju navodimo da je u siječnju 2009. iz korpusa od 5,7 milijuna pojavnica naučeno svega 8.547 općejezičnih različnica, što znači svega 150 novih na 100.000 obrađenih pojavnica. Zasićenje očito nije daleko. Prema zasićenju će vjerojatno težiti i hrvatski imenski fond jer na to ukazuje analiza u sljedećem odsječku.

Nagib funkcije V_N pokazuje suprotni trend; ta se funkcija linearizira (β teži prema 1). Kompenzacija se događa putem smanjenja parametra α, no on, ma koliko mali bio, nikad neće biti jednak nuli. Svaki prirodni jezik, pa tako i hrvatski, stalno stvara nova imena i druge posebnojezične elemente, ili ih preuzima iz drugih jezika.

U ovom segmentu jezik je mnogo produktivniji, odnosno podatniji za primanje, nego u općejezičnom dijelu. Te odnose dobro ilustrira i učenje posebnojezičnih različnica u siječnju 2009. Od 15.427 naučenih različnica posebnojezičnog fonda njih blizu 90%, ili točno 13.722, bila su imena, od čega je 9.275 (blizu 70% svih imena) bilo stranih, tj. preuzeta su u hrvatski s grafijom iz drugih jezika. U apsolutnom iznosu broj novih stranih imena i njihovih oblika već nadmašuje broj novih različnica općejezičnog fonda, makar je opseg korpusa iz kojih se one crpe 50:1 u korist općejezičnog fonda. Može se očekivati i da će udio stranih imena u novonaučenom imenskom fondu stalno rasti jer je imena lakše posuđivati nego ih stvarati, što potvrđuje svakodnevica hrvatskih ulica i natpisa na njima.

Svi ovi podaci nedvojbeno govore da se stranim imenima u hrvatskome treba sustavno baviti. Za potrebe strojne tvorbe govora potrebno je pratiti njihov ulazak u hrvatski jezik, za što određenu podlogu daje sustav *Hascheck*, te razviti postupke za njihovo automatsko razvrstavanje prema jezicima iz kojih dolaze, kako bi se ona znala transkribirati i prilagoditi hrvatskom izgovoru.

3. NORMALIZACIJA STRANIH IMENA

Proces normalizacije stranih imena za potrebe sinteze govora dijeli se u tri faze:

- lematizacija stranih imena, tj. određivanje osnovnog oblika riječi;
- određivanje jezika iz kojeg riječ dolazi;
- sama normalizacija, odnosno fonetizacija riječi.

Za ispitni uzorak uzeli smo listu od 30.786 imena-različnica zajedničkih *Riznici* i *Haschecku*. Kriterij odabira bio je da različnica iz uzorka ima barem jedan trigram koji ne postoji u hrvatskom općejezičnom fondu, čime se argumentira potreba za normalizacijom jer su takve različnice u izvornom grafijskom obliku neizgovorljive hrvatskom sintetizatoru govora. Odabrani uzorak tvori 0,6% sadržaja *Riznice*.

3.1. Prepoznavanje osnovnog oblika (lematizacija)

Prva faza je prepoznavanje osnovnih oblika. Pritom su za prepoznavanje osnovnog oblika različnice korištene paradigme imeničkih i pridjevskih nastavaka za hrvatski jezik te *Haschecko*ve baze posebnojezičnog i engleskog jezičnog fonda. Za svaku različnicu traži se paradigma u koju se sama različnica najbolje uklapa. Pritom se, koliko je moguće, vodi računa o preklapanju nastavaka različitih paradigmi i pokušava se pogoditi paradigma s najvećim brojem potvrda gramatičkih oblika uz uvjet da je osnovni oblik na samom uzorku ili u *Haschecko*voj bazi. Ovisno o složenosti gramatičkog modela, postiže se odgovarajuća kakvoća prepoznavanja oblika. Program za lematizaciju prilično je jednostavan. Nema, primjerice, modul za obradu glasovnih promjena u pismu kojima su podložna i strana imena u hrvatskome te ne tretira neke produktivne pridjevske afikse (npr. sufiks -in) s dopustivim hrvatskim nastavcima, što se odražava i na uspješnost sljedeće faze.

Od ukupnog broja različnica s popisa prepoznate su 17.853 različnice za koje su nađene odgovarajuće gramatičke paradigme i nađena potvrda osnovnog oblika u samom uzorku ili u *Hascheckovoj* bazi. To daje odziv od 58% uz zadovoljavajuću razinu pouzdanosti i te različnice svrstane su u prvu skupinu obrade (tablica 3). Lematizacija u konačnici daje 5.839 osnovnih oblika za 17.853 različnice, što u prosjeku daje 3,06 različnica po osnovnom obliku. Preostale različnice za koje su pretpostavljene određene gramatičke paradigme, no nisu pronađene potvrde osnovnog oblika osim njih samih, raspoređene su u drugu skupinu (tablica 4). Takvih je različnica 12.933 i one su poslužile kao kontrolni uzorak kod određivanja ishodišnog jezika različnica iz prve skupine.

3.2. Određivanje ishodišnog jezika

Druga faza je određivanje jezika iz kojega dolazi različnica. Prvi korak ove faze jest prikupljanje informacija o pretpostavljenim osnovnim oblicima na internetu i njihovo raspoređivanje na temelju domene iz koje dolaze. U tu svrhu napisan je program koji se oslanja na internetski pretraživač *Google Search*. Riječ je o softverskoj izvedenici iz paketa za potporu učenju novih riječi u programskom okruženju sustava *Hascheck* (Pavlek i sur., 2008).

Kako bi se dobili podaci o jeziku iz kojeg riječi dolaze i kako bi do izražaja došli i manji jezici, koristi se ograničenje pretraživanja vršnih internetskih domena sužavanjem na nacionalne vršne domene, tj. isključivanjem svih funkcionalnih domena s popisa svjetskih vršnih domena (*.com*, *.biz*, *.org*, *.net* itd.), njih ukupno 20. Pretraživanje se ograničava na prvih 100 stranica s traženom različnicom prema rangiranju *Googlea*. Rezultati se lokalno spremaju za sljedeći korak analize. Nakon toga obrađuju se spremljeni rezultati i izdvajaju informacije o (približnom) ukupnom broju rezultata za pojedinu različnicu te statistika vršnih domena za prvih 100 rangiranih stranica. Na temelju statistike domena moguće je u velikom broju slučajeva procijeniti izvorni jezik iz kojeg različnica dolazi. U prvoj aproksimaciji, kao izvorni jezik određene različnice uzima se službeni jezik države vršne nacionalne domene s najvećim brojem rezultata u prvih 100 prema *Googleovom* rangu. U tome su prepreka države s više jezika u službenoj uporabi. Na temelju slučajno odabranih poduzoraka radili smo procjenu zastupljenosti jezikâ u pojedinom državnom uzorku. Podrobnija analiza domena može poboljšati kakvoću zaključivanja o izvornom jeziku različnica. Također, postoji mogućnost korištenja metapodataka o jeziku mrežnih stranica, odnosno oslanjanje na *Googleove* algoritme prepoznavanja jezika, kao i druge javno dostupne programe za prepoznavanje jezika (LRS, 2009), no to još treba istražiti te provjeriti rezultate i ocijeniti pouzdanost tih metoda. Može se očekivati da bi kombinirani pristup dao bolji rezultat. Nedostatak takvog pristupa su složeniji algoritmi analize i dulja obrada, pa se u to nismo upuštali.

Tablica 3 donosi raspodjelu različnica prve skupine prema domenama. Vršne IP-domene kodirane su sukladno međunarodnom standardu (ISO 3166, 2009).

5.839 različnica raspoređeno je na 67 različitih domena, od kojih su najzastupljenije nacionalne domene: Velika Britanija (38,86%), Njemačka (14,11%), Hrvatska (9,51%), Italija (4,62%), Poljska (4,42%), Kanada (4,37%), Francuska (3,10%), Mađarska (2,28%), SAD (1,63%), Austrija (1,49%), Nizozemska (1,47%), Španjolska (1,42%), Švedska (1,13%), Slovačka (1,03%), Slovenija (0,98%), Češka Republika (0,96%), Japan (0,75%), Švicarska (0,74%), Indija (0,72%), Rusija (0,63%), što ukupno čini 94,21%, dok su sve ostale domene zajedno (njih 47) zastupljene s ukupno 5,79% različnica.

Tablica 3. Raspodjela imena prema globalnim vršnim domenama (prva skupina)
Table 3. Name distribution in country domains – base sample

GB	**DE**	**HR**	**IT**	**PL**	**CA**	**FR**	**HU**	**US**	**AT**
2.269	824	555	270	258	255	181	133	95	87
NL	**ES**	**SE**	**SK**	**SI**	**CZ**	**JP**	**CH**	**IN**	**RU**
86	83	66	60	57	56	44	43	42	37
RO	**FI**	**BE**	**DK**	**NO**	**CN**	**IL**	**ZA**	**TR**	**NZ**
31	30	27	27	22	21	18	17	17	11
GR	**LT**	**PT**	**EE**	**IR**	**KR**	**IE**	**LU**	**COM**	**LV**
10	10	9	7	6	6	6	5	5	4
ID	**KZ**	**BG**	**MX**	**BA**	**CU**	**CL**	**AE**	**AU**	**IS**
4	3	3	3	3	3	3	3	2	2
TH	**AZ**	**VN**	**LI**	**BW**	**MD**	**ZW**	**GE**	**EU**	**LK**
2	2	2	1	1	1	1	1	1	1
PA	**RS**	**TZ**	**AR**	**SG**	**UA**	**UG**			
1	1	1	1	1	1	1			

Razvrstavanje imena iz druge (kontrolne) skupine (tablica 4) u osnovi potvrđuje raspodjelu imena iz prve (osnovne) skupine. Prvih je 20 vršnih domena, koje daju blizu 95% različnica u uzorku, u obje tablice identično. Njihov redoslijed među prvih 20 u tablici 4 nešto je promijenjen u odnosu na redoslijed u tablici 3, no to ne utječe bitno na redoslijed pretežućih jezikâ iz kojih imena neizgovorljiva hrvatskom sintetizatoru govora ulaze u hrvatski jezik: engleski (45%), njemački (16%), talijanski (8%), francuski (6%), poljski (5%), španjolski (4%), mađarski (3%), nizozemski (2%) i latinski (2%). Postoci u zagradama izvedeni su analizom uzoraka višejezičnih vršnih domena i domene HR (Hrvatska), kamo su raspoređeni brojni kosi oblici imena za koja program zbog svoje nesavršenosti nije uspio pronaći točan osnovni oblik. Posebnu skupinu unutar domene HR tvore brojni latiniteti za koje ne postoji vršna domena kao domena jezika-izvorišta.

Tablica 4. Raspodjela imena prema globalnim vršnim domenama (druga skupina)
Table 4. Name distribution in country domains – control sample

GB	**DE**	**HR**	**IT**	**FR**	**CA**	**PL**	**HU**	**US**	**CZ**
3.987	2.143	1.445	844	530	527	392	330	272	214
JP	**ES**	**AT**	**SE**	**NL**	**SI**	**SK**	**RU**	**IN**	**CH**
224	201	162	151	152	138	121	111	98	86
FI	**NO**	**RO**	**DK**	**CN**	**BE**	**GR**	**IL**	**TR**	**COM**
73	55	52	51	52	45	44	42	38	33
ZA	**NZ**	**PT**	**IE**	**EE**	**ID**	**CL**	**IS**	**BA**	**KR**
26	26	23	23	20	17	15	15	12	12
IR	**LV**	**VN**	**EU**	**BR**	**CU**	**LT**	**RS**	**UA**	**AE**
10	9	9	8	7	7	6	6	6	6
LU	**GE**	**TV**	**PH**	**BG**	**MX**	**AZ**	**AM**	**TH**	**KE**
4	4	4	4	4	3	3	3	3	3
MN	**KZ**	**TZ**	**PE**	**MD**	**EG**	**AR**	**FO**	**NU**	**LK**
2	2	2	2	1	1	1	1	1	1
UG	**SZ**	**AD**	**TC**	**CY**	**AU**	**UZ**	**CO**		
1	1	1	1	1	1	1	1		

Očita podzastupljenost vršne domene US (SAD) u obje tablice posljedica je činjenice da Amerikanci ne upotrebljavaju toliko svoju nacionalnu domenu koliko se koriste svojim tradicionalnim funkcionalnim domenama (*.com*, *.biz*, *.job*, *.net*, *.org*, *.edu*, *.gov*, *.mil*, ...). Pojava poslovne domene *.com* u obje tablice (5 različnica u tablici 3, 33 različnice u tablici 4), iako smo je programski željeli eliminirati iz prikaza vršnih domena, posljedica je *Googleove* politike izlistavanja sponzoriranih stranica na vrhu popisa, pa je u ovom slučaju prevagnula snaga novca nad logikom programiranja.

Presjek vršnih domena iz tablice 3 (67 domena) i tablice 4 (78 domena) daje 60 nacionalnih domena, jednu komercijalnu *(.com*) i jednu nadnacionalnu domenu (EU, Europska Unija). Zanimljivo je primijetiti da se na popisu vršnih domena iz kojih su *Haschecku* stizali tekstovi na provjeru nalazi 56 nacionalnih domena, od kojih se svega četiri domene, Bahrein (BH), stara domena Srbije i Crne Gore (CS), nova domena Crne Gore (ME) te Malezija (MY), ne pojavljuju u tablicama. Kako je izvor ispitnog uzorka *Riznica*, reprezentativni megakorpus hrvatskog jezika, slijedi zaključak o dvostranosti odnosa stranih imena u hrvatskome i zemalja, slijedom toga i jezikâ, odakle su ona u hrvatski stigla: stroju neizgovorljiva imena dolaze u hrvatski iz zemalja u kojima žive ljudi koji aktivno rabe hrvatski u pismu. To potvrđuje činjenica da su iz tih zemalja hrvatskom pravopisnom provjerniku pristizali tekstovi na obradu.

3.3. Transkripcija imena

Treća faza je transkripcija stranih imena prema izvoru jezika iz kojeg dolaze. Transkripcijski algoritam oslanja se na pravila za transkripciju i transliteraciju stranih imena opisana u *Hrvatskom pravopisu* (Badurina i sur., 2008). U implementaciji su korišteni regularni izrazi sa slijednim izvođenjem pravila za transkripciju, pazeći pritom na redoslijed pravila tako da ona ne interferiraju. Transkripcija engleskih imena, koja su najbrojnija, vrlo je zahtjevna, ponajprije zbog niskog stupnja predvidivosti transkripcije samoglasnika, tako da smo tu skupinu imena morali za sada preskočiti. U nastavku dajemo rezultate transkripcije njemačkih i talijanskih imena, sljedeće dvije skupine prema zastupljenosti u hrvatskome jeziku.

3.3.1. Transkripcija njemačkih imena

U transkripciji imena iz njemačkog jezika s njemačke nacionalne domene od 824 različnice iz prve skupine (tablica 3) pravilno je transkribirana 741, nepravilno su transkribirane 22, a 61 je nepravilno svrstana pod njemački. Time je na danom uzorku postignuta točnost svrstavanja riječi od 92,60% i točnost transkripcije od 97,12%. U skupini *uljeza* učestalošću se izdvaja ime *Winnetou* i njegovi kosi oblici. Ono kulturno pripada njemačkom jezičnom prostoru, no jezik iz kojega ime potječe ipak je engleski. Također se ističu imena *Hüseyin*, *Inçirlik*, *Irtemçelik*, *Sükür*, *Süleyman*, *Öcalan*, *Özdemir*, *Özkan*, *Öztürk*, *Ülker* i *Ümit*, sva aktualna u Njemačkoj, a koja su odraz društvene situacije i izviru iz turskog jezika. Nadalje, javljaju se imena i prezimena, toponimi i strane riječi (pa i posuđenice) iz španjolskog (*Acuña*, *Gonzáles*), francuskog (*Aynaoui*, *Chansonfest*, *René*, *Varieté*), mađarskog (*Deszö*, *Györ*, *Kuranyi*), engleskog (*Dilthey*, *Eurotower*), talijanskog (*Demichelis, Giovanne*, *Biennale*, *Girardelli*), grčkog (*Tsakalidis*), poljskog (*Galoński*), albanskog (*Çeku*, *Krasniqi*, *Rexhep*, *Thaqi*). Iz hrvatskog kulturnog kruga u njemačku domenu u ovom uzorku zalutali su Kovačićev *Illustrissimus*, Krležin *Latinovicz* i zagrebački *Bastardzi*. Primjer njemačke nacionalne domene dobro ilustrira koliko je automatsko razvrstavanje imena po jezicima-izvorištima osjetljiv i zahtjevan posao.

Većina pogrešno transkribiranih imena javlja se kod transkribiranja njemačkih imena tvorenih kao složenice. Tipične pogreške su kod transkribiranja suglasničke skupine *st* i *sp* (*Beckstein/Bekštajn*, *Eppenstein/Epenštajn*, *Generalstäbler/Generalštebler*, *Grosswallstadt/Grosvalštat*, *Neustädter/Nojšteter*, *Schweinsteiger/Švajnštajger*, *Tagesspiegel/Tagešpigel*, *Willstätt/Vilštet*), zatim stapanje suglasnika (*Gottschalk/Gočalk*, *Weltschmerz/Velčmerc*), gubljenje suglasnika *h* u složenicama gdje bi trebao ostati (*Bierhoff/Birof*, *Kunsthalle/Kunstale*, *Lufthansa/Luftansa*) i u tradicionalnim imenima (*Johann/Joan*, *Johannes/Joanes*), *chs->ks* umjesto *chs->hs* (*Friedrichshafen/Fridrikshafen*, *Reichstag/Rajkstag*), pogrešno transkribirani samoglasnici kod imena koja su iznimke od navedenih transkripcijskih pravila (*Matthäus/Matojs*, *Marienplatz/Marinplac*, *Stoiber/Štober*). Problem s njemačkim imenima-složenicama moguće bi bilo riješiti

tako da se posebno pripazi na suglasničke skupine u riječima koje bi mogle biti složenice te uz pomoć baze riječi njemačkog jezika.

U drugoj skupini (tablica 4) pod njemački su svrstane 2.143 različnice. Od toga su 203 (9,47%) pogrešno svrstane pod njemački, 1.774 su točno transkribirane (91,44%), a 166 pogrešno (8,56%). Vidljivo je da se statistički podaci u obje skupine dobro preklapaju.

Transkripcija njemačkih imena provjerena je i na uzorku iz austrijske nacionalne domene. Od 87 austrijskih različnica iz prve skupine točno je transkribirana 81 različnica, netočno je transkribirano njih 5, a pogrešno je svrstana samo jedna različnica (*Vienna*) i to samo jezično, ne i zemljopisno, što otkriva da je riječ o jednom od najposjećenijih europskih turističkih gradova u globaliziranom svijetu kojim dominira engleski jezik kao jezik međunarodne komunikacije. Na uzorku je postignuta točnost svrstavanja od 98,85%, točnost transkripcije od 94,21%, odnosno pogreške pri transkripciji od 5,81% gdje su tipične pogreške istog tipa kao i za njemačku vršnu domenu (*Burgstaller/Burgstaler*, *Dietrichstein/Ditrikstajn*, *Hofgastein/Hofgastajn*, *Mauthausen/Mautausen* i *Museumsquartier/ Musojmkvartir*). Problem sa samoglasnicima zahtijeva daljnju doradu modula za transkripciju i vjerojatno se može djelomično izbjeći uz navođenje iznimaka.

U drugoj skupini pod austrijsku domenu svrstane su 162 različnice, od toga 8 pogrešno (4,94%), točno je transkribirano 139 različnica (90,26%), a pogrešno njih 15 (9,74%). Značajnijih odstupanja u odnosu na dominantnu njemačku nacionalnu domenu nema.

3.3.2. Transkripcija talijanskih imena

Od 270 različnica prve skupine svrstanih u talijansku nacionalnu domenu 13 ih je pogrešno svrstano u talijanski, 254 ih je transkribirano pravilno, a 3 nepravilno. Na danom uzorku točnost razvrstavanja iznosi 95,19%, a točnost transkripcije iznosi 98,83%. Pogrešno su svrstani *Baquba*, *Berlocq*, *Cayard*, *Courmayeur*, *Dadullah*, *Dudovich*, *Haniyeh*, *Katyna*, *Kronplatz*, *Mitteleurope*, *Propheta*, *Weah*, *Wojtyla*, što su uglavnom strani toponimi i prezimena poznatih osoba vezani uz politički, sportski, kulturni i duhovni život Italije. Nepravilno transkribirana imena su *Iudice* (*Iudiče* umjesto *Judiče*), *Iuliano* (*Iuljano* umjesto *Julijano/Juljano*) i *Triennale* (*Trjenale* umjesto *Trijenale*). Primjeri upućuju da bi transkripcijska pravila za talijanski trebalo upotpuniti pravilom *#iu->ju*.

Od 844 različnice druge skupine iz talijanske domene pogrešno ih je raspoređeno 89, dok je njih 749 transkribirano pravilno, a 6 nepravilno. To daje točnost razvrstavanja za talijanski na zadanom uzorku u iznosu 89,45%, točnost transkripcije 99,21% i pogrešku transkripcije od 0,79%. Karakteristična je nešto veća točnost transkripcije u odnosu na njemački, gdje su na netočnost transkripcije uglavnom utjecale složenice kakvih u talijanskom nema. Netočno razvrstavanje u oba jezika na razini je od oko 10%.

4. ZAKLJUČAK

Istraživanja opisana u ovom radu su, kako sadržajem tako i metodologijom, prva ove vrste obavljena s reprezentativnim uzorkom hrvatskog pisanog jezika. Prikazani rezultati, makar ohrabrujući, upućuju na složenost problema normalizacije stranih imena za potrebe strojne tvorbe govora na hrvatskome. Ostvarenu točnost razvrstavanja imena od oko 90% te sličnu, čak i veću točnost transkripcije točno razvrstanih imena iz njemačkog i talijanskog uzorka doživljavamo ponajprije kao poticaj za nastavak istraživanja. Morfonološki pravopis olakšava strojnu tvorbu glasova iz riječi hrvatskog općejezičnog fonda, no kako hrvatski većinu stranih i stroju neizgovorljivih imena preuzima u izvornoj grafiji – pokazano je da njih hrvatski uglavnom preuzima iz jezika koji rabe latinično pismo – svaki sustav za sintezu govora na hrvatskome mora imati podsustav za pretprocesiranje stranih imena. Takav podsustav ne može se učitati s interneta niti kupiti, njega Hrvati trebaju sami razviti. Nedvojbeno je da naša rudimentarna metodologija za prepoznavanje jezika-izvorišta, koja je ključna za dobru normalizaciju, mora doživjeti nadopunu. U sljedećem koraku namjeravamo testirati određene javno dostupne programe za prepoznavanje jezika, npr. *Polyglot 3000* (POLYGLOT 3000, 2009) te istražiti koliko takvi softverski paketi mogu popraviti našu metodologiju.

REFERENCIJE

Badurina, L., Marković, I., Mićanović, K. (2008). *Hrvatski pravopis*. Zagreb: Matica hrvatska.

Dembitz, Š., Knežević, P., Sokele, M. (1999). Hascheck – The Croatian academic spelling checker. *Applications and Innovations in Expert Systems VI* (ur. R. Milne i sur.), 184-197. Springer.

Dembitz, Š., Knežević, P., Sokele, M. (2004). Developing a spell checker as an expert system. *Journal on Computing and Information Technology*, **CIT-11**(4), 285-291.

HASCHECK (2009). http://hacheck.tel.fer.hr/ [pristupljeno 12. siječnja 2009].

Heaps, H. S. (1978). *Information retrieval – computational and theoretical aspects*, Academic Press.

ISO 3166 (2009).
http://www.iso.org/iso/english_country_names_and_code_elements
[pristupljeno 15. ožujka 2009].

Kominek, J., Black, A. (2005). Measuring unsupervised and acoustic clustering through phoneme pair merge-and-split tests. *Interspeech 2005*, 689-692. Lisbon, Portugal.

Kominek, J., Black, A. (2006). Learning pronunciation dictionaries: Language complexity and word selection strategies. *Proceedings of the Human Language Technology Conference of the NAACL*, 232-239. New York City, USA.

Kominek, J., Schultz, T., Black, A. (2007). Voice building from insufficient data – classroom experiences with web-based language development tools. *ISCA SSW6*, 322-327. Bonn, Germany.

Kominek, J., Schultz, T., Black, A. (2008). Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. *Workshop on Spoken Language Technologies for Under-resourced Languages*, 63-68. Hanoi, Viet Nam.

Kornai, A. (2002). How many words are there? *Glottometrics 4*, 61-86.

LRS (2009). http://www.word2word.com/identad.html [pristupljeno 15. ožujka 2009].

Moguš, M., Bratanić, M., Tadić, M. (1999). *Hrvatski čestotni riječi*. Zagreb: Školska knjiga.

Pavlek, J., Dembitz, Š., Matasić, M. (2008). Improved methods of word acquisition in developing Hascheck spell checker web service system. *X International PhD Workshop OWD 2008*, Conference Archives PTETIS, Vol 25, 29-34. Poland.

POLYGLOT 3000 (2009). http://www.polyglot3000.com/ [pristupljeno 25. ožujka 2009].

RIZNICA (2009). http://riznica.ihjj.hr/ [pristupljeno 12. siječnja 2009].

Schultz, T., Black, A. (2006). Challenges with rapid adaptation of speech translation systems to new language pairs. *ICASSP 2006*, V.1213-V.1216. Toulouse, France.

Schultz, T., Black, A., Badaskar, S., Hornyak, M., Kominek, J. (2007). SPICE: Web-based tools for rapid language adaptation in speech processing systems, *Interspeech 2007*, 2125-2128. Antwerp, Belgium.

SPICE (2009). http://plan.is.cs.cmu.edu/Spice/spice/index.php [pristupljeno 9. siječnja 2009].

Šoić, R. (2008). Upoznavanje sa sustavom SPICE. Zagreb: Fakultet elektrotehnike i računarstva [završni rad].

FOREIGN NAME PROBLEM IN CROATIAN SPEECH SYNTHESIS

Abstract

For a speech synthesis system a part of a text should be preprocessed and normalized in order to become pronounceable. This refers generally to numbers, abbreviations, different unit symbols and foreign names. Croatian writing system is mostly phonologically based, which makes grapheme-phoneme mapping in machine speech production easier. However, foreign names in Croatian usually follow their original orthography. For the purpose of speech synthesis such names have to be transcribed according to Croatian transcription rules. Based on a comparative analysis of two big Croatian corpora, the paper investigates the relative occurrence of foreign names an average Croatian text, as well as their import dynamics in Croatian. Further, using a sample of more then 30,000 foreign names and their inflected forms, an automatic source language identification procedure is elaborated. Taking into account name classification results, our system tried to transcribe names classified as German or Italian. Average transcription accuracy was over 90%, with classification accuracy around 90%. This is encouraging for further system development.

Key words: speech synthesis, transcription of foreign names, Croatian