UDK811.163.42'374
811.163.42'42
Stručni rad

Sanda Martinčić Ipšić, Mihaela Matešić, Ivo Ipšić
Filozofski fakultet, Rijeka
Hrvatska

KORPUS HRVATSKOGA GOVORA

SAŽETAK

U radu je predstavljen korpus hrvatskoga govora Odsjeka za informatiku Filozofskoga fakulteta u Rijeci. Korpus se sastoji od triju dijelova: radijskih i telefonskih vremenskih prognoza te televizijskih vijesti. Prikupljen je govor 250 različitih govornika u ukupnom trajanju od gotovo 20 sati. Prikupljen je čitani i spontani govor. Prikazani su struktura korpusa, njegova organizacija i osnovni statistički parametri. Opisani su postupci snimanja govora i transkripcije. U radu su predstavljeni korišteni alati CSLU SpeechView, Transcriber i HTK, rječnik, koji sadržava sve riječi govornoga korpusa i njihov fonetski zapis te postupak validacije govornoga korpusa. U zaključnom dijelu predstavljeni su rezultati automatske segmentacije na fonetskoj razini.

Ključne riječi: *govorni korpus, hrvatski govori, segmentacija govora, validacija govora*

UVOD

Postupci se za računalnu obradu govora, poput raspoznavanja ili sinteze, temelje na podacima, dakle govoru, prikupljenom i zapisanom u govornom korpusu. Govorni je korpus baza podataka u kojoj su pohranjeni govorni signali, pripadajuće tekstne transkripcije tih signala te metapodaci. Metapodaci opisuju postupak prikupljanja govora, značajke prikupljenoga govora te značajke govornika. Postupak se izgradnje govornoga korpusa sastoji od nekoliko ključnih aktivnosti (Schiel i sur., 2004), od kojih su najvažnije: snimanje, transkripcija, segmentacija i validacija korpusa.

Govorni korpus može biti višejezični ili jednojezični, općenit ili usredotočen na određenu problemsku domenu ili specijaliziran za rješavanje određenoga zadatka (npr. gradnja informacijskoga sustava za rezervaciju zrakoplovnih karata). Korpus nadalje može sadržavati različite vrste govora, poput: spontanoga govora, čitanoga govora, diktata, razgovora, emotivnoga ili neutralnoga govora, govora koji sadržava određene unaprijed definirane fonetske jedinice (npr. za korpusnu sintezu govora), a može sadržavati i kombinacije različitih tipova govora. Izabrana vrsta govora utječe na broj govornika i dužinu govora, a time i na veličinu korpusa. Približna procjena ukupne veličine i opsega korpusa bitna je da se unaprijed pripreme i rezerviraju svi potrebni resursi za njegovu izgradnju. Značajke korpusa utječu i na tehničke odluke o načinu snimanja (npr. predviđeni format snimljenoga govora, način snimanja). Prije početka snimanja potrebno je odrediti akustičke uvjete snimanja, tip opreme koja će se standardno koristiti, sve tehničke uvjete snimanja, poput: frekvencije uzorkovanja, formata zapisa, broja kanala, formata datoteka govora i transkripcija.

Nakon što su svi potrebni parametri određeni, pristupa se testnom snimanju. Ako su rezultati testa primjereni, nastavlja se s redovitim snimanjem, obradom snimljenih signala te transkripcijom. U tom se procesu kontinuirano provjerava snimljeni i transkribirani materijal. U zaključnoj fazi izgradnje izrađuje se fonetski rječnik te izvodi fonetska segmentacija korpusa.

U ovome će radu biti predstavljen govorni korpus Odsjeka za informatiku Filozofskoga fakulteta u Rijeci. Opisat će se njegova namjena i struktura, proces prikupljanja, anotiranja, validacije. Bit će predstavljen i fonetski rječnik te rezultati automatskoga segmentiranja.

GOVORNI KORPUS

Govorni je korpus nastao snimanjem radijskih vremenskih prognoza i televizijskih vijesti tijekom 2002. i 2003. godine i sadržava govor u ukupnom trajanju od 19 sati i 24 minute. Sastoji se od nekoliko dijelova: vremenskih prognoza i izvješća o stanju na moru, koje čitaju i izgovaraju profesionalni spikeri u radijskim vijestima, vremenskih izvješća, koje telefonski prezentiraju profesionalni meteorolozi na radiju te televizijskih dnevnika. Korpus je tematski

sastavljen od dvaju dijelova: prvi dio VEPRAD (VrEmenske Prognoze-RADio) obuhvaća govor tematski vezan uz vrijeme i vremensku prognozu, dok drugi dio TV HRBCN (TeleVizijski HRvatski BroadCast News) obuhvaća širok spektar tema televizijskih vijesti (Graff, 2002). Naziv TV HRBCN slijedi standardu projekta COST "Spoken Language Interaction in Telecommunication - Broadcast News Transcription".

Govorni korpus VEPRAD sastoji se od vremenskih prognoza koje govore profesionalni spikeri na radiju (VEPRAD radio) i vremenskih izvješća meteorologa, koje prezentiraju telefonski (VEPRAD telefon). VEPRAD radio sadržava 3566 izraza[1], koje izgovara 11 muških i 13 ženskih profesionalnih spikera u trajanju od 6 sati i 17 minuta. Korpus VEPRAD radio sadrži 57896 pojavnica, od kojih je 1161 različita. Relativno malen broj različitih riječi proizlazi iz ograničenosti problemske domene (vrijeme i prognoza). Spikeri u bazi VEPRAD radio pretežito čitaju izgovoreni tekst.

Korpus VEPRAD telefon sastoji se od 158 vremenskih izvješća, koje izgovara 7 ženskih i 5 muških meteorologa telefonom, u trajanju od 5 sati i 6 minuta. VEPRAD telefon sastoji se od 48629 pojavnica u 2493 izraza. Ukupan je broj različitih riječi 1717. Meteorolozi svoja izvješća pažljivo planiraju, no neki i spontano dodaju osobne intervencije u govor, pa možemo reći da ovaj dio baze sadržava poluspontani govor, koji ima sporadične improvizacije u odnosu na unaprijed pripremljen predložak toga govora. O unaprijed pripremljenom predlošku i njegovu izgledu moguće je govoriti na temelju tekstova vremenskih prognoza objavljenih na internetskim stranicama pojedinih institucija (npr. Državnoga hidrometeorološkoga zavoda, Hrvatske radiotelevizije). Dakle, izgovoreni tekst nije izravno čitanje objavljenog teksta nego verzija prilagođena govornome stilu, raspoloženju te osobinama pojedinih meteorologa. Na taj je način uvedena određena spontanost u govor pa bi se baza mogla okarakterizirati kao djelomično spontani govor. Drugu značajku baze čine pozadinski zvukovi, koji često prate meteorologov govor i koji su u bazi posebno označeni.

Treći dio korpusa nazvan TV HRBCN sadržava 6 dnevnika nacionalne televizije, u ukupnom trajanju od 3 sata i 28 minuta. Za svaki TV dnevnik postoji audio i video zapis te odgovarajuća transkripcija. Omjer je ukupnoga broja riječi i različitih riječi 1:2, što upućuje na širok spektar obuhvaćenih tema (više od 160). Ukupan je broj riječi 18632, od kojih je 9326 različitih. Pojavilo se 217 govornika. Za svakoga je govornika poznat spol, ime i je li izvorni govornik hrvatskoga jezika. Obuhvaćeni materijal u većem je dijelu popraćen pozadinskom bukom. Govor u televizijskim dnevnicima u rasponu je od čitanoga do spontanoga – naime, u "žanrovima" televizijskoga novinarskoga funkcionalnoga stila varira omjer spontanoga i nespontanoga govora, pa dok je npr. čitana vijest potpuno nespontani govor, dijalozi se u televizijskim dnevnicima najčešće mogu svrstati u spontani i poluspontani govor. Dijalog u npr. anketi između izvjestitelja i slučajnih prolaznika na ulici ima izrazita

[1] o nociji ovdje rabljenoga termina *izraz* v. poglavlje *Transkripcija*

obilježja spontanoga govora, jer izvjestiteljev sugovornik nije pripremljen, a i izvjestitelj nerijetko nakon anketnoga/anketnih pitanja postavlja i pitanja ili potpitanja koja nisu predviđena nego ih generira konkretna komunikacijska situacija. Dijalog između voditelja dnevnika i izvjestitelja na terenu ima manje obilježja spontanosti, jer voditelj svoja pitanja upućena izvjestitelju gotovo isključivo čita iz studija, a izvjestitelj, na drugoj lokaciji, u tzv. "stand-upu", unaprijed pripremljene odgovore izvodi s više ili manje spontanosti (količinu spontanosti odaje struktura izvedenoga govora – ponajviše sintaktičke osobitosti, mimika, govorne pogreške itd.).

Statistika transkribiranoga dijela hrvatskoga govornoga korpusa prikazana je u tablici 1.

Tablica 1. Statistika govornoga korpusa
Table 1. The speech corpora statistics

	Trajanje Duration [min]	Broj / Number				Govornici Speakers	
		Izvješća Reports	Izraza Expressions	Pojavnica Tokens	Različitih riječi Different words	M/M	Ž/F
VEPRAD							
radio radio	392	653	3566	57896	1161	11	13
telefon telephone	344	158	2493	48629	1717	5	7
Ukupno Total VEPRAD	736	811	6059	106725	2160	16	20
TV HRBCN							
Dnevnik / News	208	6		18632	9326	217	
Ukupno / Total HRBCN	208	6		18632	9326	217	
UKUPNO TOTAL	944	817		125357		253	

PRIKUPLJANJE I PRIPREMA PODATAKA

Govorni je korpus snimljen uz pomoć PC računala i HAUPAGE WINTV/RADIO kartice. Snimke su uzorkovane s frekvencijom 16000 Hz i spremljene u 16-bitnom PCM mono wav formatu.

Baza vremenskih prognoza VEPRAD nastala je snimanjem vremenske prognoze na nacionalnom radiju četiri puta na dan. Snimane su vremenske prognoze koje su bile emitirane u sklopu vijesti Hrvatskoga radija 1. programa u 13, 15, 17 i 19 sati. Svaka od prognoza sastoji se od dviju cjelina. Prognoza u 13 sati sadržava prognozu za poslijepodne te izvješće o vodostaju rijeka. Prognoza u 15 sati sadržava opsežno vremensko izvješće (VEPRAD telefon), koje prezentiraju profesionalni meteorolozi Državnoga hidrometeorološkoga zavoda te izvješće o stanju na moru do 13 sati koje priprema Pomorski meteorološki centar u Splitu. Prognoze u 17 i 19 sati sadržavaju prognozu za sljedeći dan te ponekad i za kraće sljedeće razdoblje. U drugome se njihovu dijelu donosi stanje na moru pripremljeno u 13, odnosno u 19 sati. Dva dijela svake prognoze razlikuju se i po spikeru koji čita vijesti, tako da pojedini dio govori ženski, a zatim muški spiker, odnosno profesionalni spiker i dežurni meteorolog. Za sljedeće emitiranje prognoze spikeri najčešće zamijene međusobno tekst tako da je u bazi čest slučaj da isti tekst izgovaraju različiti spikeri.

Baza TV HRBCN nastala je snimanjem televizijskih dnevnika na 1. programu Hrvatske televizije.

Struktura baze

U govornome korpusu nalaze se datoteke koje sadrže signale govora (.wav) te pripadajuće transkripcijske datoteke (.txt ili .trs). Samo za HRBCN postoji još i odgovarajući videozapis (.avi). Parovi datoteka imaju isti naziv ali različite ekstenzije. VEPRAD radio i VEPRAD telefon organizirani su prema identifikacijskoj oznaci govornika. Prvi znak označava spol govornika. Za radijske signale korištena je oznaka m ili z, dok je za telefonske h ili f, čime je omogućena jednostavna detekcija tipa snimke. Dakle, naziv svake datoteke u korpusu sastoji se od identifikacije i spola govornika (mxx ili zxx ili hxx ili fxx), datuma prognoze i rednog broja unutar dana (ddmmggr) te rednog broja izraza unutar prognoze (yy).

Za razliku od VEPRAD-a, HRBCN je organiziran samo prema datumu emitiranja dnevnika, a ne prema govorniku, jer se u svakom dnevniku pojavljuje prosječno pedesetak različitih govornika. Stoga naziv svakoga dnevnika izgleda ovako: HRTdnevnikddmmgg .

TRANSKRIPCIJA

U prvoj su fazi transkribirane vremenske prognoze koje izgovaraju profesionalni spikeri jer je za njih postojao predložak objavljen na internetskim stranicama Državnoga hidrometeorološkoga zavoda (DHMZ, 2003). Zatim je izvedena transkripcija televizijskih dnevnika, a naposljetku i telefonskih vremenskih izvješća. Izvješće dežurnoga meteorologa nije praćeno javno objavljenim tekstom. Dodatni je napor pri transkripciji izvješća zahtijevala i nešto slabija kakvoća telefonskih signala, često popraćena šumom i pozadinskom

bukom poput govora, lupanja vratima, automobilskim zvukom, zvonjenjem telefona i slično.

Tekst vremenske prognoze i izvješća o stanju na moru objavljen na internetskim stranicama nije potpuno jednak izgovorenom izvješću, ali ga u znatnoj mjeri slijedi. Zato predstavlja dobru osnovu za transkripciju. Spikeri pri izgovaranju često prilagode tekst predloška vlastitome stilu i raspoloženju. Na nemalo mjesta dolazi i do zabuna, koje govornik "u hodu" ispravlja ili one pak prođu neopaženo te do pogrešaka koje rezultiraju izgovorom nepostojećih i/ili neovjerenih riječi. Stoga je pri prvom koraku u transkripciji preuređen postojeći preuzeti tekst i usklađen sa snimkom, odnosno zapisan je tekst za slučajeve kada unaprijed nije postojao predložak.

U sljedećem su koraku signali ponovno preslušani, pronađena je stanka u govoru (izostanak funkcionalnoga govornoga signala) i na tome je mjestu signal razrezan u manje dijelove – izraze. U relevantnoj se literaturi navodi termin *uterance*, koji se u hrvatskoj lingvističkoj literaturi prevodi kao *iskaz*, no jedinica dobivena opisanim postupkom ne odgovara toj lingvističkoj definiciji iskaza (bitno je obilježje iskaza, naime, njegova aktualiziranost, tj. on je komunikacijski uključena rečenica) budući da se početak i kraj ove jedinice određuje sasvim fizički: ondje gdje dovoljno dugo nije ostvaren funkcionalni govorni signal (najčešće je to na mjestu stanke, no u profesionalnih spikera često udah i izdah sadržavaju funkcionalni glas). Čini se najprimjerenijim tako dobivenu jedinicu nazvati izrazom, jer se do nje dolazi neovisno o planu sadržaja, a izborom toga naziva još se uvijek čuva veza s terminom *uterance*, koji opravdanost uporabe uvelike crpi iz činjenice da ovdje imenuje govorno ostvarenu jedinicu.

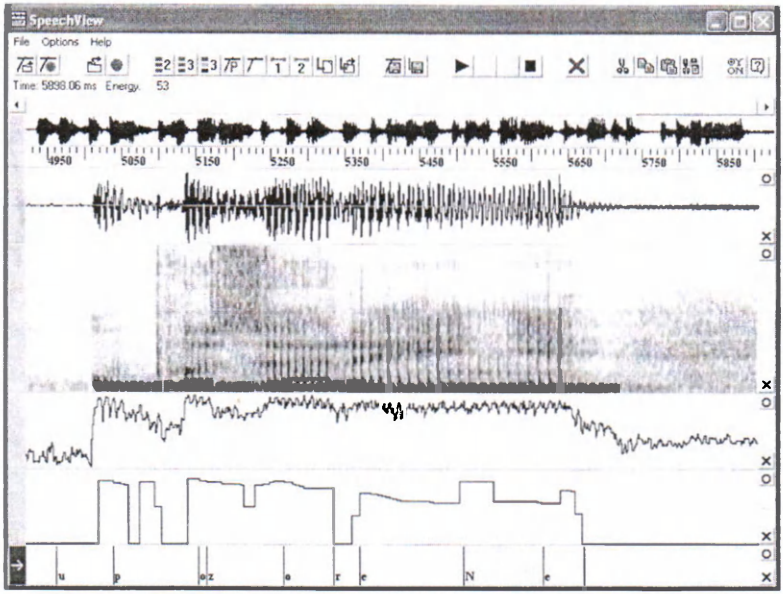
Za svaki izraz u bazi VEPRAD postoji odgovarajuća transkripcija u tekstualnom formatu pohranjena na drugome mjestu ali s istim imenom. Uz izgovorene riječi transkripcija sadržava i oznake posebnih šumova, uzdaha, listanja papira, udarca, stanke te pogrešaka koje su u tekstu jednolično označene unutar posebnih znakova < i > (Graff, 2002). Dakle, govorni je korpus transkribiran na razini riječi i sastoji se od izgovorenih izraza te pripadajućega teksta. Primjer transkripcije drugoga izraza prve prognoze 6. srpnja 2002. godine, koji je izgovorila spikerica z07 prikazan je u tablici 2.

Tablica 2. Transkripcija jednoga izraza u bazi
Table 2. Transcription of one speech uterance

```
z07060702102
postupna naoblaka <uzdah> mjestimice s pljuskovima i
grmljavinom vjetar slab <sil> a na jadranu povremeno
umjeren jugozapadnjak i jugo <uzdah>
```

U postupku transkripcije vremenskih prognoza baze VEPRAD korišten je *SpeechView* alat *CSLU Speech Toolkit* (Sutton i sur., 1998) prikazan na slici 1. Od vrha prema dnu slike prikazani su: signal govora, povećanje jednoga njegova

dijela, spektrogram, energija, ovojnica signala osnovne frekvencije te transkripcija. CSLU (Center for Spoken Language Understanding) Speech Toolkit razvijen je u Oregonu, u Centru za govor i razumijevanje jezika. Sadržava alat namijenjen analizi i rezanju signala, modul za raspoznavanje govora temeljen na prikrivenim Markovljevim modelima – PMM i neuronskim mrežama, alat za sintezu govora, modul za analizu prirodnoga jezika i modeliranje dijaloga, alat za brz razvoj aplikacija iz primjene govorne tehnologije, alat za modeliranje dijaloga i naposljetku Baldi-alat, koji vizualno predočava stvaranje zvuka te se koristi i pri radu s djecom s posebnim potrebama.



Slika 1. Primjer izrezanoga izraza u CSLU Speech View alatu
Figure 1. The CSLU SpeechView tool

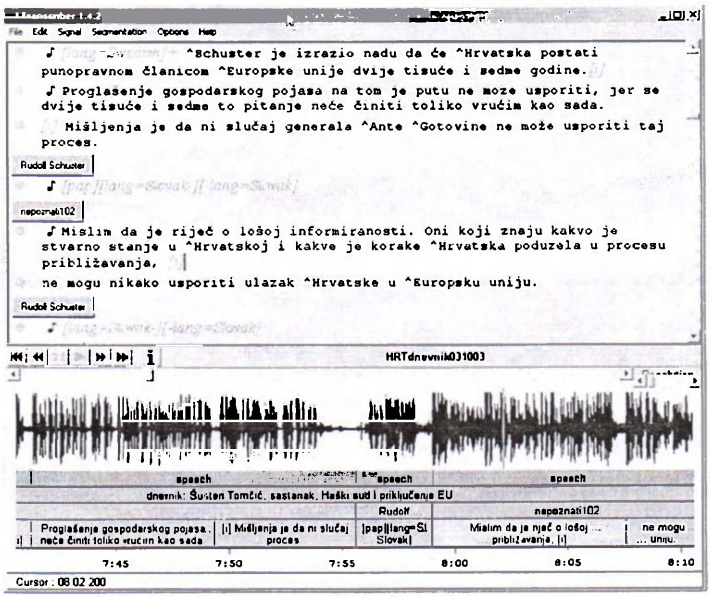
Preliminarni tekst za transkripciju televizijskih dnevnika prikupljan je na internetskim stranicama nacionalne televizije (HTV, 2003). Prikupljeni se tekst u većoj mjeri razlikovao od izgovorenoga jer urednici, odnosno voditelji Dnevnika ponešto modificiraju tekst, a u Dnevnik su često uključeni dijalozi koji su manje ili više izvor spontanoga govora. Pri transkripciji televizijskih Dnevnika korišten je i videozapis, koji je omogućio pravilan zapis imena govornika, a često je poslužio i za točnu identifikaciju tipa i vrste pozadinske buke.

Transkripcija HRBCN dijela korpusa izvedena je pomoću *Transcriber* alata (Barras i sur., 2000) i zapisana je u XML formatu (.trs datoteke u bazi). Za svaki dnevnik postoji jedan audiozapis, jedan videozapis i jedan transkripcijski zapis. Signal u ovom dijelu baze nije rezan na kraće dijelove, nego je svaka

stanka u govoru (fiziološka stanka, psihološka stanka) označena kao poseban događaj u transkripciji. *Transcriber* je specijalizirani alat za izgradnju višejezičnih govornih korpusa. On omogućuje ručnu anotaciju i segmentaciju, označavanje promjena govornika, promjene teme i akustičkih uvjeta. Segmentaciju je moguće izvoditi na više različitih razina: osnovna segmentacija radi dobivanja ortografske transkripcije – kod stanke u govoru, segmentacija prema različitim govornicima – kod promjene govornika te segmentacije s obzirom na sadržaj – pri promjeni teme. Segmentacija je moguća i prema akustičkim uvjetima, odnosno prema pojavi pozadinskoga šuma i njegovoj vrsti. *Transcriber* omogućuje i zapis većega broja metapodataka: atributa koji opisuju govornike (spol, izvorni jezik govornika, tip govora ...), statusa i verzije transkripcije itd. *Transcriber* omogućuje i dobar nadzor nad događajima koji nastupaju u govoru poput: udaha, izdaha, šmrcanja, kašljanja, smijeha, mikrofonije, riječi izgovorenih na jeziku različitom od transkribiranoga itd. *Transcriber* zapisuje u XML standardu i omogućuje *Unicode* potporu za višejezične transkripcije.

Primjer transkripcije u *Transcriberu* prikazan je na slici 2.

U gornjem je dijelu slike transkripcija signala prikazanoga u donjem dijelu slike.



Slika 2. Transkripcija vijesti u *Transcriber* alatu
Figure 2. Croatian BCN in *Transcriber*

Primjer transkripcije u XML standardu:

```
<Event desc="sv" type="language" extent="next"/>
 ^Schuster je izrazio nadu da će ^Hrvatska postati
punopravnom članicom ^Europske unije dvije tisuće i
sedme godine.
<Event desc="i" type="noise" extent="instantaneous"/>
<Sync time="461.287"/>
<Background time="461.287" type="speech"
level="high"/>
Proglašenje gospodarskog pojasa na tom je putu ne može
usporiti, jer se dvije tisuće i sedme što pitanje neće
činiti toliko vrućim kao sada.
<Sync time="469.369"/>
<Event desc="i" type="noise" extent="instantaneous"/>
 Mišljenja je da ni slučaj generala ^Ante ^Gotovine ne
može usporiti taj proces.</Turn>
<Turn speaker="spk11" mode="planned"
startTime="478.997" endTime="490.932" fidelity="high"
channel="studio">
<Sync time="478.997"/>
<Background time="479.037" type="speech"
level="high"/>
Mislim da je riječ o lošoj informiranosti. Oni koji
znaju kakvo je stvarno stanje u ^Hrvatskoj i kakve je
korake ^Hrvatska poduzela u procesu približavanja,
<Event desc="i" type="noise" extent="instantaneous"/>
<Sync time="487.579"/>
ne mogu nikako usporiti ulazak ^Hrvatske u ^Europsku
uniju. </Turn>
<Turn speaker="spk12" mode="spontaneous"
startTime="490.932" endTime="497.175"
fidelity="medium" channel="studio">
<Sync time="490.932"/>
<Background time="491.025" type="speech"
level="high"/>
<Event desc="Slovak" type="language" extent="begin"/>
<Event desc="Slovak" type="language" extent="end"/>
</Turn>
<Turn speaker="spk12" mode="spontaneous"
startTime="475.621" endTime="478.997"
fidelity="medium" channel="studio">
<Sync time="475.621"/>
<Background time="475.647" type="speech"
level="high"/>
```

```
<Event desc="pap" type="noise"
extent="instantaneous"/>
<Event desc="Slovak" type="language" extent="begin"/>
<Event desc="Slovak" type="language" extent="end"/>
</Turn>
```

FONETSKI RJEČNIK

Izgradnja fonetskoga rječnika sastoji se od zapisa riječi u fonetskom obliku koji se temelji na SAMPA simbolima (Bakran i Horga, 1996). Rječnik baze VEPRAD radio sadržava 1161 različitu riječ kojima je dodijeljena u bazi pripadajuća fonetska transkripcija. U prvoj fazi izgradnje baze VEPRAD radio (svibanj – lipanj 2002. godine) zbog manjeg opsega snimljenoga materijala, rječnik je sadržavao samo 625 različitih riječi. Gotovo 86%-tno povećanje rječnika proizlazi iz povećanog opsega materijala, koji sadržava i tzv. "zimski" vokabular vremenskih prognoza. Daljnji rad na povećanju opsega baze VEPRAD svakako će biti usmjeren pridobivanju govora od većeg broja govornika te povećanju vokabulara. Rast rječnika smanjit će i problem raspoznavanja riječi izvan rječnika, koji je poseban problem u svim sustavima za raspoznavanje govora.

Rječnik baze VEPRAD telefon sadržava 1717 različitih riječi. Veći broj riječi u bazi VEPRAD telefon u odnosu na bazu VEPRAD radio uvjetovan je samim tipom govora. Kao što je prethodno napisano, VEPRAD telefon odlikuje se djelomice spontanim govorom, koji sadržava jak osobni pečat pojedinih govornika-meteorologa, dok se VEPRAD radio odlikuje "strogim" oblikom izgovorenoga teksta. Ukupan je broj različitih riječi u objema VEPRAD bazama 2160.

Isječak rječnika prikazan je u tablici 3.

Tablica 3. Dio fonetskoga rječnika
Table 3. Part of the phonetic dictionary

Riječ/Word	Fonetski prijepis Phonetic transcription
atlantska	a t l a n t s k a
atmosferskog	a t m o s f e r s k o g
Baranji	b a r a N i
barem	b a r e m
bez	b e z
bi	b i
Biokova	b i o k o v a
bit	b i t
biti	b i t i
blizu	b l i z u
Botovo	b o t o v o
brod	b r o d
brzo	b r z o
Budimpešta	b u d i m p e S t a
Budimpešte	b u d i m p e S t e
bura	b u r a
bure	b u r e
burin	b u r i n
burom	b u r o m
buru	b u r u

SEGMENTACIJA

Budući da je ručna transkripcija izvedena za cijeli korpus na razini riječi, potrebno je napraviti i segmentaciju na fonetskoj razini (Turk, 1992). Inicijalna je segmentacija za fonetsku razinu izvedena automatskim usklađivanjem govornoga signala i transkripcije na razini riječi. Automatsko usklađivanje je izvedeno pomoću monofonskih prikrivenih Markovljevih modela s HTK alatom (Young i sur., 2002).

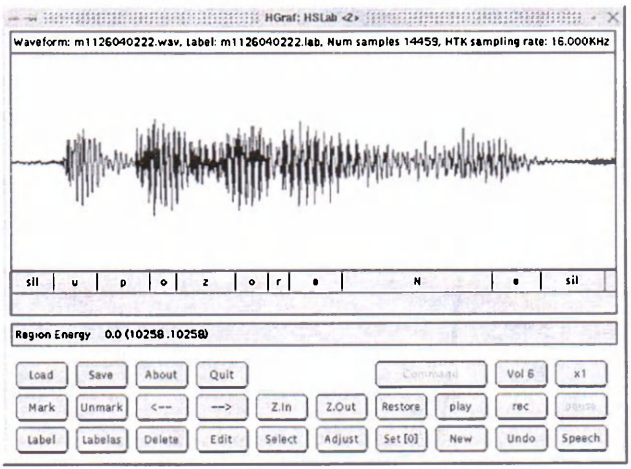
HTK Toolkit (*The Hidden Markov Model Toolkit*) skupina je alata namijenjenih za modeliranje stohastičkih procesa pomoću prikrivenih Markovljevih modela. Programski alat HTK prvenstveno je namijenjen raspoznavanju govora, iako je primijenjen i za sintezu govora, raspoznavanje znakova, DNK sekvencioniranje te za druga problemska područja.

S obzirom na funkciju HTK, alati su razvrstani u nekoliko skupina: za pripremu podataka, za učenje, za raspoznavanje te za analizu rezultata. Alati za

pripremu podataka namijenjeni su pripremi govornih signala kao i pripadajućih transkripata. Omogućavaju: izračunavanje značajki govornoga signala, pregled izračunanih značajki, transformaciju transkripata u odgovarajući format, prikupljanje i prikazivanje statistike o datotekama labela (transkripcijskim datotekama). Alati za učenje omogućavaju: pripremu prototipnoga PMM-a, koji sadržava grubu topologiju i strukturu modela, postavljanje inicijalnih parametara modela te učenje modela. Alati za raspoznavanje omogućuju izvođenje modificiranoga Viterbijeva algoritma. Raspoznavanje se izvodi pomoću jezične mreže koja opisuje dopuštene sljedove riječi iz rječnika s fonetskim zapisom riječi te skupom naučenih PMM-a. Prilikom raspoznavanja, mreža riječi pretvara se u mrežu fonema te se svakom fonemu dodijeli najvjerojatniji PMM. Alati za analizu rezultata omogućuju mjerenje točnosti i preciznosti izgrađenog sustava za raspoznavanje tako da se raspoznati govor testnog skupa usporedi s unaprijed pripremljenim transkriptima.

Postupak automatske segmentacije sastojao se od izračuna značajki govornih signala, učenja monofonskih modela i raspoznavanja riječi Viterbijevim algoritmom, koje su sastavljane iz monofonskih modela (Martinčić-Ipšić i Ipšić, 2004).

Rezultat automatske segmentacije upotrijebljen je kao ulaz u sustav za raspoznavanje hrvatskoga govora pomoću prikrivenih Markovljevih modela. Rezultat automatske segmentacije riječi *upozorenje* prikazan je na slici 3.



Slika 3. Rezultat automatske segmentacije riječi *upozorenje* u HTK HSLabu

Figure 3. The result of automatic segmentation of the word *upozorenje* in HTK HSLab

Primjer zapisa automatske segmentacije na fonetskoj razini za m01020402103: <sil> najviša temperatura od osamnaest do dvadeset i tri stupnja <sil> prikazan je u tablici 4. Oznaka <sil> na početku i kraju svakoga izraza označava kratku početnu i završnu stanku. Dulje stanke u govoru označene su oznakom <pauza>. Za svaki je fonem automatski izračunat vremenski interval u kojem traje i izražen u vremenskim jedinicama od jedne milisekunde. Na primjer fonem /n/ počinje na 32 [ms], a završava na 104 [ms].

Tablica 4. Rezultat automatske segmentacije
Table 4. The result of automatic segmentation

"*/m01020402103.lab"

0 32 sil	1696 1720 a
32 104 n	1720 1816 e
104 152 a	1816 1848 s
152 192 j	1848 1880 t
192 240 v	1880 1952 d
240 288 i	1952 2096 o
288 408 S	2096 2168 d
408 448 a	2168 2232 v
448 512 t	2232 2312 a
512 576 e	2312 2360 d
576 616 m	2360 2400 e
616 680 p	2400 2512 s
680 720 e	2512 2576 e
720 760 r	2576 2600 t
760 832 a	2600 2624 i
832 912 t	2624 2688 t
912 1040 u	2688 2712 r
1040 1104 r	2712 2824 i
1104 1144 a	2824 2936 s
1144 1272 o	2936 2984 t
1272 1320 d	2984 3072 u
1320 1360 o	3072 3192 p
1360 1488 s	3192 3280 N
1488 1568 a	3280 3376 a
1568 1632 m	3376 3448 sil
1632 1696 n	

VALIDACIJA

Validacija obuhvaća sve postupke provjere pravilnosti i točnosti korpusa tijekom izgradnje. U fazi snimanja provjereni su svi snimljeni signali. S obzirom na to da je snimanje izvođeno automatski, pojedine su snimke sadržavale samo šum ili nepredviđeno tematsko područje te su isključene iz daljnjih postupaka. Snimke koje su odgovarale osnovnim zahtjevima, ali su sadržavale previše šuma (većinom zbog slabih vremenskih uvjeta zimi), označene su i koristit će se pri provjeri točnosti raspoznavanja.

Predvalidacija je izvedena na opsegu podataka, koje je sadržavala baza VEPRAD radio u svojoj prvoj fazi. U toj je fazi zaključen cjelovit postupak izgradnje baze, učenja i raspoznavanja, čime je provjereno jesu li pripremljeni podaci odgovarajući za primjenu u računalnoj obradi govora. Dobri preliminarni rezultati raspoznavanja govora u bazi VEPRAD radio (90,63 %) ohrabrili su daljnju izgradnju korpusa na današnji opseg.

U fazi stvaranja signali su višestruko preslušavani. Najprije se zapisuje transkripcijski tekst, zatim se tijekom drugoga preslušavanja signali i tekst razrezuju te se posebice svraća pozornost na oznake posebnih događaja. Najčešće su ta dva koraka izvodile različite osobe tako da je uveden kriterij uzajamne provjere sadržaja.

Pri validaciji već pripremljenih izraza posebna je pozornost posvećena otkrivanju i uklanjanju praznih datoteka – bilo signala, bilo teksta. Za svakoga govornika provjeren je broj izgovorenih izraza i odgovarajućih tekstualnih zapisa. Pogreške pri tipkanju i nepostojeće riječi uklonjene su pri izradi fonetskoga rječnika. Na slučajnome uzorku provjereno je odgovara li prijepis govoru.

Daljnja će primjena podataka u postupcima računalne obrade govora dati objektivnu ocjenu kvalitete korpusa.

Validacija HRBCN dijela baze izvedena je pomoću specijaliziranih alata za provjeru.

ZAKLJUČAK

Izgradnja govornoga korpusa opsežan je zadatak, koji zahtijeva multidisciplinarni pristup. Neki od tipičnih problema poput angažiranja dovoljnoga broja govornika i osiguranja adekvatnih uvjeta za snimanje izbjegnuti su snimanjem emitiranoga programa. Dobrim se pokazalo i to što su jednim snimanjem dobivene i radijske i telefonske snimke koje se odnose na istu temu.

Govorni korpus VEPRAD koristit će se u sustavu za raspoznavanje govora te u postupcima sinteze govora. Telefonski dio baze VEPRAD bit će vrlo koristan u izgradnji informacijskoga sustava za potraživanje vremenskih informacija putem dijaloga, kojem će se moći pristupati i telefonski (Žibert i sur., 2003).

Korpus VEPRAD u nastajanju je, što znači da će biti proširen novim vremenskim prognozama te transkripcijom vremenskih izvješća profesionalnih

meteorologa. Rad na povećanju baze bit će usmjeren povećavanju vokabulara te izgradnji sveobuhvatnog sustava za raspoznavanje govora, prvenstveno uključivanjem telefonskoga govora. Predviđeno je također uređivanje govornoga korpusa u XML formatu prema postojećim standardima (Barras i sur., 2002).

REFERENCIJE

Bakran, J., Horga, D. (1996). SAMPA za hrvatski. *Govor* **XIII**,1-2, 99-104.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2000). Transcriber: use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools* **33**, 1 - 2.

Graff, D. (2002). An overview of Broadcast News Corpora. *Speech Communication* **37**, 1-2, 15-26.

Martinčić-Ipšić, S., Ipšić, I. (2004). Recognition of Croatian Broadcast Speech. U L. Budin, S. Ribarić (ur.), *XXVII. MIPRO 2004, International Convention, Computer in Technical Systems + Intelligent Systems, Hrvatska udruga za mikroprocesorske, procesne i informacijske sustave, mikroelektroniku i elektroniku, MIPRO - HU*, Opatija, Vol. CTS + CIS, 111-114.

Schiel, F., Draxler C., Baumann, A., Ellbogen, T., Steffen, A. (2004). The Production of Speech Corpora, Institut für Phonetik und Sprachliche Kommunikation, Ludwig-Maximilians-Universität München, Version 2.5. http://www.phonetik.uni-muenchen.de/Forschung/BITS/index.html (10.2004.).

Sutton, S., Cole, R. A., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, D., and Cohen, M. (1998) Universal Speech Tools: The CSLU Toolkit. Proc. of the International *Conference on Spoken Language Processing (ICSLP98)*, vol. **7**, 3221-3224.

Turk, M. (1992). *Fonologija hrvatskog jezika*. Rijeka – Varaždin: Biblioteka Dometi.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Vatchev, V. and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*. Cambridge: Cambridge University Engineering Department.

Žibert, J., Martinčić-Ipšić, S., Hajdinjak, M., Ipšić, I., Mihelič, F. (2003). Development of a Bilingual Spoken Dialog System for Weather Information Retrieval. 8th European *Conference on Speech Communication and Technology, Geneva, Switzerland. EUROSPEECH '03. proceedings. ISCA*. Vol. **1**, 1917-1920.

Državni hidrometeorološki zavod, (2003). http://prognoza.hr/prognoze.html (03.2002.-03.2003.).

Hrvatska radiotelevizija, (2003). http://www.hrt.hr/vijesti (03.10.2003.-10.10.2003.).

Sanda Martinčić Ipšić, Mihaela Matešić and Ivo Ipšić
Faculty of Philosophy, Rijeka
Croatia

CROATIAN SPEECH CORPORA

SUMMARY

Speech and language technologies, which perform procedures like speech recognition and speech synthesis are based on speech corpora. Speech corpora contain recorded speech signals together with their annotations, meta data and documentation. The creation of a speech corpus includes procedures for speech signals recording, transcription, segmentation and validation.

In the paper we present the procedures we have performed in order to obtain a domain specific speech database from broadcast news of national programmes. The speech signal acquisition, transcription, validation and segmentation process is described. The Croatian speech corpus contains three parts: radio weather forecast, weather forecast spoken over telephone and TV news. The corpus contains twenty hours of speech spoken by 250 speakers.

In the paper we describe the CSLU SpeechView, Trancriber and HTK tools used to obtain Croatian speech corpus as well as the corpus validation procedure.

Key words: *speech corpus, Croatian speeches, speech segmentation, speech validation*