LITUANISTICA. 2004. T. 60. Nr. 4. P. 53–62

LIETUVIŲ BENDRINĖS ŠNEKAMOSIOS KALBOS GARSŲ FONDO KŪRIMO PRINCIPAI

Arimantas Raškinis, Gailius Raškinis, Asta Kazlauskienė

Vytauto Didžiojo universitetas, K. Donelaičio g. 58, LT-44244 Kaunas, Lietuva

1. ĮVADAS

1.1. Ir taikomųjų, ir tiriamųjų kalbos inžinerijos projektų sėkmė pirmiausiai priklauso nuo disponuojamų kalbos duomenų išteklių. Dabar tam jau nebeužtenka tik didelės apimties tekstynų ar garsų fondų[1], reikalingos ištisos kalbos išteklių sistemos, sudarytos ne tik iš pačių duomenų, jų anotacinių bylų, bet ir iš galingos programinio aprūpinimo aplinkos (duomenų kaupimo bei anotavimo, kalbos išteklių paieškos, analizės ir įvairaus panaudojimo programinių posistemių). Pastaruoju laiku Europos valstybių kalbos išteklių archyvai taip išaugo, jog iš siauros mokslo tyrimų ar technologijų kūrimo paskirties tapo reikšmingomis tautinį identitetą ir istorinį procesą fiksuojančiomis bei globaliam pasauliui nacionalinius išteklius pristatančiomis saugyklomis. Šiuo metu jau yra kalbama apie galimybę sudaryti visuotinę pasaulio kalbų duomenų saugyklą (Gibbon, 2002: 3–5).

Europos vienijimosi procese akcentuojamas tautinio identiteto išsaugojimo principas, todėl vis labiau ryškėja tendencijos tarpnacionalines komunikacijas grįsti šiuolaikinėmis kalbų technologijomis. Kad Lietuvos atsilikimas kalbų inžinerijoje netaptų mūsų visavertės integracijos stabdžiu, yra svarbus dabarties standartus atitinkančių lietuvių šnekamosios bendrinės kalbos garsų fondo kūrimas.

1.2. Sisteminis Europos kalbų išteklių kūrimas buvo pradėtas 1987–1989 m. ESPRIT (*European Strategic Program for Research and Development in Information Technology*) projektu 1541-SAM (*Speech Assesment Methods*). Jo pagrindu buvo kuriami daugiakalbiai skaitmenų tarimo garsų fondai: EUROM0 (4 diktoriai, 5 kalbos – danų, olandų, anglų, prancūzų, italų, 1989 m.), EUROM1 (60 diktorių, 11 kalbų – prie ankstesnių dar prisidėjo vokiečių, norvegų, švedų (1992 m.) ir graikų, portugalų, ispanų (1993 m.)), jį tęsė Centrinei ir Rytų Europai skirtas BABEL (COPERNICUS) projektas (bulgarų, estų, vengrų, lenkų, rumunų (1996 m.), vėliau – kroatų, rusų, slovėnų kalbos).

[1] Straipsnio autorių nuomone, garsų fondo, rinkinio, bazės reikšme galėtų būti vartojamas ir *garsyno* terminas. Žodyne teikiama garsyno, kaip garsų sistemos, reikšmė išplėstina. Toks terminas labai dera ir greta dabar jau populiaraus *tekstyno*.

Nuo 1995 m. pradedamos dar dvi Europos Sąjungos programos, skirtos telefoninių pokalbių garsų fondams kurti, – CEC-DG-XII programa LRE-RELATOR ir LE-MLAP (*Language Engineering Multilingual Action Plan*). Antrosios pagrindu vykdomi trys projektai: PAROLE (tekstynų), POINTER (terminologijos) ir SPEECHDAT (garsų fondų). Pastarasis realizuojamas kaip ankstesnių projektų tąsa: SPEECHDAT(M) – 1300 skirtingų asmenų telefono pokalbių, SPEECHDAT(II) – 25 telefono pokalbių duomenų bazės po 500–5000 diktorių, SPEECHDAT-CAR – 600 įrašų sesijų automobilyje, SPEECHDAT(E) – Rytų Europai skirtas projektas (2500 rusų diktorių, čekų, slovakų, lenkų, vengrų – po 1000 diktorių).

1995 m. Europos Sąjunga Liuksemburge įsteigė nepelno organizaciją ELRA (*European Language Resources Association*), kurios tikslas – rūpintis centralizuotu Europos kalbų išteklių ir įrankių standartizavimu, platinimu ir vartojimu Europoje. ELRA pati kuria garsų fondus (AURORA projektai, išplečiantys SPEECHDAT). Vėliau ELRA įsteigė ELDA (*Evaluations and Language resources Distribution Agency*), kuri turi sukaupusi ir pardavinėja apie 150 didelės apimties (1000 diktorių) garsų fondų daugiausia Europos tautų kalbomis.

Lietuva nėra dalyvavusi nei viename iš europinių projektų, todėl lietuvių kalbos garsų fondo nėra minėtuose kataloguose[2].

1.3. Garsų fondai skiriasi paskirtimi, dydžiu, anotacijos detalumu. *Universalūs* garsų fondai turi apimti visus bendruosius kalbos ypatumus. Nemažai vietos garsų saugyklose užima *specializuoti* garsų fondai, atspindintys kalbos vartojimą skirtingose sferose: bendrinės kalbos ir tarmių; vyrų, moterų, vaikų; laisvo pokalbio ir skaitomo teksto; pavienių žodžių, komandų ir rišlaus teksto. Dabar ryškėja trys svarbesnės garsų fondų kūrimo tendencijos: 1) *taikomosios paskirties* garsų fondų apimtys nuolat didėja, jų kūrimui dideles investicijas skiria telefonijos, automobilių pramonės firmos, vienijamos mokslo ir verslo pajėgos; 2) vis daugėja *nacionalinių garsų fondų*, dokumentuojančių kalbą kaip *tautos socialinės kultūrinės aplinkos paminklą*; 3) kuriami gerai anotuoti sisteminiai garsų fondai, skirti išsamiems *kalbos mokslo tyrimo darbams*. Garsų fondai gali būti anotuoti sakiniais, žodžiais, skiemenimis, garsais.

2. BENDRINĖS LIETUVIŲ ŠNEKAMOSIOS KALBOS GARSŲ FONDO SUDARYMO PRINCIPAI

2.1. Lietuvoje jau kelis dešimtmečius kalbos atpažinimo tyrimuose dažniausiai naudojamos nedidelės, konkrečiam uždaviniui spręsti skirtos darbinės garsų įrašų

[2] Tačiau sakyti, kad lietuvių kalbininkai nėra sudarę jokios garsų bazės, tikrai negalima. Prieš dešimtmetį prof. A. Girdenis ir P. Kasparaitis analizavo lietuvių kalbos sintezės galimybes. Tam jie sudarė garsų rinkinį, kurio pagrindu buvo kuriamas lietuvių kalbos sintezatorius. Gaila, kad nei ankstesnis šių mokslininkų darbas, nei neseniai vykdytas tarptautinis projektas, kuriame toliau gilintasi į lietuvių kalbos sintezės problemas, lingvistinėje literatūroje nėra plačiau aprašytas.

duomenų bazės (Rudžionis, Rudžionis, Žvinys, 1999). Nuo 2001 m. Vytauto Didžiojo universitete kuriamas bendrinės lietuvių šnekamosios kalbos garsų fondas[3].

Garsų fondo kūrimo pagrindas – *fonetinių vienetų sistema*, atskleidžianti visą garsinių ypatumų įvairovę. Bendrąjį fondą sudaro dvi dalys: pavieniui tariami žodžiai, kurių kiekvienas skirtas tam tikram fonetiniam vienetui iliustruoti, ir tie patys žodžiai, ištarti trumpame rišliame trijų žodžių sakinyje.

Fonetinių vienetų sistemos sudarymas pirmiausiai remiasi grynaisiais fonologiniais vienetais ir jų pagrindiniais alofonais, kuriuos jau senokai yra nustatę lietuvių kalbininkai (Girdenis, 1995). Tačiau nei rišlaus kalbos srauto sintezei, nei analizei jų neužtenka, todėl ateityje šią bazę reikės papildyti. Dabartinės bazės sudarymo principai tokie.

1. Atskirti balsiai nuo priebalsių (ir artikuliaciniu, ir akustiniu (ir, žinoma, funkciniu) požiūriu skirtingos garsų klasės, tuo nesunkiai galima įsitikinti vien tik žvilgtelėjus į jų oscilogramas).

2. Balsiai lietuvių kalboje gali būti ilgieji ir trumpieji – tai akivaizdus akustinis, artikuliacinis ir funkcinis (t. y. skiria žodžius, pvz., *kas – kąs*) balsių požymis. Trumpieji: *a e i u o*, ilgieji: *ā ē ī ū o ė*.

3. Visi balsiai gali būti nekirčiuoti ir kirčiuoti. Ilgieji kirčiuoti balsiai gali būti dvejopi: tvirtapradžiai – á é í ú ó é ir tvirtagaliai – *ã ẽ ĩ ũ õ ẽ*. Trumpieji – nekirčiuoti ir kirčiuoti: *a e i u o – à è ì ù ò*, ilgieji – nekirčiuoti *ā ē ī ū o ė*.

4. Užpakalinės eilės balsiai *o, u*, vartojami po minkštųjų priebalsių, papriešakėja (skiriasi kokybe, F_2 padėtimi). Todėl dar reikia skirti grynuosius užpakalinius ir papriešakėjusius šiuos balsius. *a* šioje pozicijoje, daugelio kalbininkų nuomone, virsta *e*.

5. Mišrieji dvigarsiai (jie gali būti ir kirčiuoti, ir nekirčiuoti) laikyti savarankiškais fonetiniais vienetais. Skiriami dvibalsiai: *ai au ie eu ei ui uo* – ai au íe éu ei ùi úo – *aĩ aũ iẽ eũ eĩ uĩ uõ* ir dvigarsiai: *al am an ar el em en er ul um un ur il im in ir* – al am an ar el em en er ul um un ur il im in ir – *a*l *a*m *a*n *a*r *e*l *e*m *e*n *e*r *u*l *u*m *u*n *u*r *i*l *i*m *i*n *i*r. Dvigarsio priebalsis gali būti kietasis ir minkštasis. Taigi pridedami dar minkštieji: : *al' am' an' ar' el' em' en' er' ul' um' un' ur' il' im' in' ir'* – al' am' an' ar' el' em' en' er' ul' um' un' ur' il' im' in' ir' – *a*l' *a*m' *a*n' *a*r' *e*l' *e*m' *e*n' *e*r' *u*l' *u*m' *u*n' *u*r' *i*l' *i*m' *i*n' *i*r'. Mišriųjų dvigarsių antrasis dėmuo *n*, vartojamas prieš priebalsius *k, g*, pakeičia tarimo vietą: šioje pozicijoje tariamas ne liežuvio priešakinis dantinis, o liežuvio užpakalinis gomurinis: *a*ŋ *a*ŋ' *á*ŋ *á*ŋ' *a*ɲ *a*ɲ' *e*ŋ *e*ŋ' *é*ŋ *é*ŋ' *e*ɲ *e*ɲ' *i*ŋ *i*ŋ' *í*ŋ *í*ŋ' *i*ɲ *i*ɲ' *u*ŋ *u*ŋ' *ú*ŋ *ú*ŋ' *u*ɲ *u*ɲ'[4].

6. Iš viso parinkti 175 balsiniai vienetai. Pavyzdžiai rinkti tokie, kad fonetinis vienetas būtų vartojamas (jeigu tik įmanoma) keturiose pozicijose: a) žodžio pradžioje, b) žodžio viduje tarp dusliųjų priebalsių, c) žodžio viduje tarp skar-

[3] Straipsnyje aprašomo garsų fondo kūrimas iš dalies finansuotas „Lietuvių kalbos informacinėje visuomenėje 2000–2006 m. programos“ lėšomis pagal sutartį su Valstybine lietuvių kalbos komisija. Autoriai nuoširdžiai dėkoja už finansinę paramą.

[4] Ateityje čia dar reikėtų pridėti dvigarsius, prasidedančius papriešakėjusiu užpakalinės eiles balsiu *u*.

1 lentelė. **Balsio *a* ir dvigarsių su pirmuoju dėmeniu *a* pavyzdžiai**

a	*akmuõ, kapaĩ, kãsa*
à	*àkti, kàpsi, kasà, ràsdavo*
ā	*ąžuolaĩ, grąžõs*
ã	*ãsilas, kãpas, bãdas*
á	*ąžuolas, prakąsto*
ai	*aistrà, skaitaũ, daigaĩ, táikai*
ái	*áiškus, táiko, ráibo, visái*
aĩ	*aĩtrų, skaĩto, baĩdo, takaĩ*
au	*aukà, taukaĩ, daužaĩ, táikau*
áu	*áuksas, káukė, gáudo*
aũ	*aũkuras, kaũkia, daũžo, mataũ*
al	*algà, šaltų̃, baldùs*
al'	*alksniù, kaltì, dalgiù*
ál	*álkanas, páltas, válgo*
ál'	*kálkim, válgė*
aĩ	*aĩpsta, kaĩtas, baĩdų, atgaĩ*
aĩ'	*aĩkis, šaĩtis, daĩgis*
am	*ambasadà, tamsù, rambùs*
am'	*amžinaĩ, tamsiù, nebambėk*
ám	*ámpulė, támsta, bámba*
ám'	*ámžinas, sámtis, grámdys*
am̃	*kam̃pas, ram̃bų*
am̃'	*kam̃ʒtis, dram̃blį*
an	*antrų̃, kantrùs, bandà*
aŋ	*ankstùs, bangà, tvankù*
an'	*antìdė, išbandýt, kantrì*
aŋ'	*ankstì, tankmė̃, bangẽlė*
án	*žándas, diktántas*
áŋ	*ánglas, tánkus*
án'	*ántika, pántis*
áŋ'	*ánkštis, tánki*
añ	*kañtrų, gañdras*
aŋ̃	*añkštas, dañgų*
añ'	*bañdymas, bedañtę*
aŋ̃'	*tañkmę, susirañgė*
ar	*artójas, sartù, žargònas*
ar'	*arklių̃, karčių̃, darbẽlis*
ár	*árka, kártų, gárdų*
ár'	*árklį, kártis, gárvežį*
ař	*kařklą, gařgždo*
ař'	*ařtimą, sařtis, bařnis*

džiųjų priebalsių, d) žodžio gale. Kadangi ne visose pozicijose balsiniai vienetai gali būti vartojami (pvz., absoliučiame žodžio gale nėra tvirtapradžių balsių, beveik nevartojami toje pozicijoje ir dvigarsiai), galima sakyti, kad parinkta vidutiniškai po tris pavyzdžius kiekvienam vienetui (balsio *a* ir dvigarsių su pirmuoju dėmeniu *a* pavyzdžius, parinktus kuriamam fondui, žr. 1 lentelėje). Labai preciziškai atsižvelgus į visus akustinius garsų požymius, galima teigti, kad balsinių elementų yra apie 525 vienetus.

7. Parenkant pirminei analizei priebalsinius elementus, pirmiausiai atskirti duslieji nuo skardžiųjų. Dusliųjų priebalsių ir gretimų balsių bei skardžiųjų priebalsių ribos gana nesunkiai nustatomos, todėl šie vienetai imti be didesnės balsio dalies (atsižvelgta tik į jų kietumą ir / ar minkštumą): *c c' ch ch' f f' k k' p p' s s' t t' č č' š š'*. Be balsinės aplinkos imti ir retesni skardieji: *dz dz' dž dž' h h' z z' ž ž*. Iš viso 28 vienetai. Šie elementai tiriamojoje medžiagoje pavartoti: a) žodžio pradžioje, b) žodžio viduje. Stengtasi (kiek įmanoma), kad vienetai būtų vartojami prieš skirtingo pakilimo balsius: a) jeigu kietasis priebalsis, jis vartojamas prieš *a* ir *u*; b) jeigu minkštasis – prieš *e* ir *i*. Atsižvelgę į visas pozicijas, turėtume apie 80 vienetų.

8. Kadangi balsingųjų (arba pusbalsių) bei skardžiųjų priebalsių ir gretimų garsų (ne dusliųjų priebalsių) ribas sunkiau nustatyti, todėl tyrimui imti: a) šie elementai su nemaža balsio *a e i u o ė*

dalimi (pavyzdžiuose jie vartojami žodžio pradžioje, viduryje, gale) ir b) be balsinės aplinkos (kietieji priebalsiai pavyzdžiuose vartojami žodžio pradžioje prieš ilguosius balsius *a, ū, o,* viduryje – greta kito kietojo priebalsio; minkštieji priebalsiai – žodžio pradžioje prieš ilguosius balsius *e, ė, y,* viduryje – greta kito minkštojo priebalsio): *b b' ba be bi bo bu bė d d' da de di do du dė g g' ga ge gi go gu gė l l' la le li lo lu lė m m' ma me mi mo mu mė n n' na ne ni no nu nė r r' ra re ri ro ru rė v v' va ve vi vo vu vė.* Iš viso 72 vienetai (jeigu skaičiuotume skirtingoje aplinkoje pavartotus, – apie 200).

9. Atskirai reikia paminėti priebalsį *j*. Pirminei analizei parinkti elementai *ja, je, ji, jo, ju, jė* pavartoti, kaip ir skardieji priebalsiai, žodžio pradžioje, viduryje, gale. Tačiau jo ir gretimų garsų ribas sunku nustatyti, todėl dar reikia išsiaiškinti galimus tokių garsų junginius, nustatyti jų dažnumą ir akustinius skirtumus.

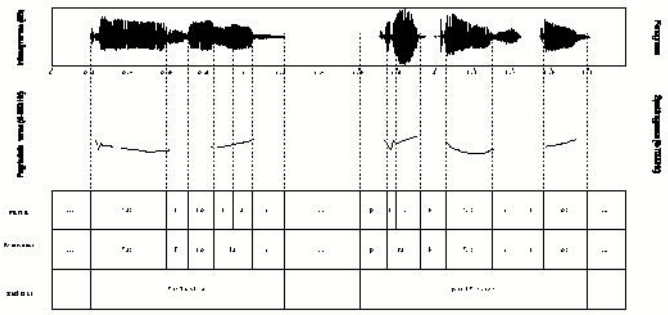
Pirminei analizei sudarėme 106 priebalsinių fonetinių vienetų bazę. Jeigu skaičiuotume visas vartojimo pozicijas, būtų apie 300.

2.2. Abi fondo dalis (ir žodžius, ir sakinius) įrašė tie patys 4 neprofesionalūs diktoriai (du vyrai ir dvi moterys, amžiaus vidurkis – 37 m., du iš jų kilę iš vakarų aukštaičių tarmės ploto, vienas – iš rytų aukštaičių uteniškių, vienas – iš Kauno miesto). Įrašai daryti tylioje aplinkoje, bet ne profesionaliose garso įrašų studijose. Buvo naudojama skaitmeninė įrašymo įranga bei specialus mikrofonas. Bendra pirmosios dalies įrašų trukmė – 60,6 min., antrosios – 107,1 min. Fondas skirtas mūsų atliekamiems lietuvių kalbos technologijų tyrimams ir kitiems taikomiesiems darbams. Garsų, esančių norimame kontekste, paieškai ir ištraukimui sukurta originali garsų paieškos programinė įranga *Tescal*.

Kiekvienam įrašui kompiuterio atmintyje sukurtas diktoriaus inicialais pavadintas aplankas, vieno diktoriaus fonetinių vienetų, ištartų pavieniuose žodžiuose, fonograma užėmė 90 MB atmintį. Visi įrašai saugomi PCM 44100 Hz 16 bitų monoformatu.

Tolimesniam naudojimui visa fonograma buvo padalyta į 275 atkarpas. Kiekvienoje atkarpoje buvo visi vienam fonetiniam vienetui atstovaujantys pavyzdžiai (atitinkamas vienetas įvairiose minėtose pozicijose). Atkarpa buvo saugoma atskira „wav“ formato byla, pavadinta to fonetinio vieneto vardu. Prieš pradedant anotavimą, pirmiausiai kiekvieno fonetinio vieneto „wav“ byla redaguojama: iškarpomos klaidos bei pasikartojimai, sutrumpinamos pauzės bylos pradžioje, pabaigoje ir viduryje, nufiltruojamas triukšmas, sunormuojamos pauzės tarp žodžių: prieš pirmą žodį ir po paskutiniojo paliekama po 0,2 sek. tylos ir po 0,4 sek. tylos tarp žodžių.

Anotuotame fonde, be fonetinių vienetų fonogramų, turi būti saugomi ir įrašų tekstai, pateikti fonetine transkripcija bei informacija susiejant teksto elementus su atitinkamais garsų pradžios ir pabaigos laiko momentais. Kiekvienas garso įrašas yra anotuojamas trimis lygmenimis: žodžiais, fonetiniais vienetais ir garsais. Tam yra sukuriamos *žodžių juosta*, *fonetinių vienetų juosta* ir *garsų juosta*, kuriose atitinkamai atskirai saugoma informacija apie žodžių pradžios ir pabaigos laiką, žodį sudarančių fonetinių vienetų pradžios ir pabaigos laiką ir žodį sudarančių garsų pradžios ir pabaigos laiką. Laiko momentai nustatomi rankiniu būdu stebint garso įrašo fonogramos ir spektrogramos vaizdus kompiuterio ekrane, klausant to garso įrašo ir fiksuojant žodžių, fonetinių vienetų bei garsų pradžios ir pabaigos laiką.



1 pav. á pristatymas fonetinių vienetų atlase. Viršuje – žodžių *ąžuolas*, *prakąsto* fonograma ir intensyvumo kreivė, viduryje – spektrograma ir pagrindinio tono kreivė, apačioje – anotavimo juostos.

Kadangi anotavimui buvo naudojama internete laisvai platinama PRAAT programinė įranga (http://www.fon.hum.uva.nl/praat/), todėl ir anotacinės bylos saugomos PRAAT sistemos „TextGrid“ formatu (1 pav.). Visą kuriamą VDU lietuvių bendrinės šnekamosios kalbos garsų fondą dabar sudaro daugiau kaip 3984 fonogramų ir anotacijų bylos. Jo pagrindu VDU buvo parengtas elektroninis „Bendrinės lietuvių šnekamosios kalbos fonetinių vienetų atlasas“.

3. GARSŲ FONDO TEKSTŲ FONETINĖ TRANSKRIPCIJA

Europos Sąjungos garsynų rengimo projektuose tekstų fonetinei transkripcijai naudojama SAMPA (*Speech Assessment Methods Phonetic Alphabet*) kodavimo sistema. Kaip nurodoma SAMPA transkripcijos protokolo tinklapyje (http://www.phon.ucl.ac.uk/home/sampa/home.htm), ši šnekos garsų kodavimo ASCII kodais sistema taikoma 24 kalboms: arabų, bulgarų, kinų, čekų, kroatų, danų, olandų, anglų, estų, prancūzų, vokiečių, graikų, hebrajų, vengrų, italų, norvegų, lenkų, portugalų, rumunų, rusų, ispanų, švedų, tajų, turkų. Nuspręsta ir kuriamame fonde naudoti SAMPA kodus.

Skirtingai nuo kitų kalbų, kurioms būdinga arba tembrinė koreliacija, arba priegaidės, lietuvių kalbos atveju būtina koduoti ir priebalsių palatališkumą, ir prozodines kirčiuoto skiemens ypatybes. Tai apsunkina kokio nors vieno SAMPA alfabeto pasirinkimą[5] (2 lentelė).

Visi lietuvių kalbos priebalsiai gali būti ir kieti, ir minkšti. Todėl priebalsių minkštumas, kaip siūlo X-SAMPA, žymimas apostrofu (’), t. y. ASCII 39 kodu,

[5] Reikia pasakyti, kad SAMPA nėra vienalytis standartas, o labiau primena kodavimo principų ir gairių rinkinį, kuriam buvo būdingas tam tikras kitimas. Pradėjus SAMPA taikyti skirtingoms (net negiminiškoms) kalboms, atsirado poreikis koduoti vis naujus garsus ir jų požymius, pavyzdžiui, SAMPA alfabete nebuvo numatyta simbolio, žyminčio priebalsio minkštumą. Todėl buvo pasiūlyti išplėstiniai SAMPA variantai: X-SAMPA, kuriuo būtų galima ASCII kodais koduoti daugelį IPA (*International Phonetic Association*) simbolių, ir SAMPROSA – ASCII simbolių sistema, skirta koduoti prozodines šnekos ypatybes (http://www.phon.ucl.ac.uk/home/sampa/x–sampa.htm, http://www.phon.ucl.ac.uk/home/sampa/ipasam–x.pdf, http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm). Visi šie alfabetai ne papildo vienas kitą, o iš dalies persidengia, todėl realiai galima naudotis tik kuriuo nors vienu.

rašomu po priebalsio simbolio. Greta afrikatų rašomas tik vienas minkštumo ženklas.

Vietoje tradicinių trijų kirčio ženklų – (ˋ) – gravio, žyminčio trumpąjį kirčiuotą arba tvirtapradį su pirmuoju trumpuoju mišriojo dvigarsio dėmeniu skiemenį, (´) – akūto, skirto ilgam krintančios skiemens intonacijos (tvirtapradės priegaidės) skiemeniui, ir (ˆ) – cirkumflekso, skirto ilgam kylančios intonacijos (tvirtagaliam) garsui, – naudojami du SAMPA simboliai. Pagrindinio kirčio simboliu (") , t. y. ASCII 34, žymimi trumpieji kirčiuoti ir tvirtapradžiai skiemenys, o simboliu (^), t. y. ASCII 94, – tvirtagaliai skiemenys. Šie simboliai rašomi prieš pat kirčiuotą garsą. Mišriuosiuose tvirtagaliuose dvigarsiuose kirčio ženklas rašomas prieš antrąjį dėmenį, sudėtiniuose ir sutaptiniuose dvibalsiuose – prieš visą dvibalsį.

Balsių ilgumui žymėti naudojamas simbolis (:), t. y. ASCII 58, rašomas po balsio ženklo. Trumpasis, mažai labializuotas ir atviras *o*, vartojamas tik tarptautiniuose žodžiuose, žymimas (O), t. y. ASCII 79.

Priešakinės eilės neaukštutinio pakilimo *e* tipo balsių kodavimo sistema ne visiškai atitinka tikslias IPA rekomendacijas, bet parinkta artimesnė tradicinei lietuviškai transkripcijai. Taigi balsis *ė*, panašiai kaip vokiečių kalboje, koduojamas (E:) (ASCII 69, 58), ilgasis *e* – kaip (e:) (ASCII 101, 58), o trumpasis *e* – paprasčiausiu (e) (ASCII 101)[6].

Dauguma lietuvių kalbos priebalsių koduojami įprastais rašmenimis. Išimtys:

a) liežuvio priešakiniai alveoliniai (š, ž) koduojami atitinkamai didžiosiomis raidėmis (S, Z), o afrikatos žymimos dviem raidėmis (ts, tS, dz, dZ);

b) priebalsis *n* gali būti tariamas dvejopai: kaip dantinis (žymimas (n)) ir kaip gomurinis (N);

c) lietuvių kalbos priebalsis *h* lietuvių bendrinėje kalboje vartojamas kaip *skarduosis gomurinis pučiamasis* ir žymimas jį atitinkančiu SAMPA simboliu (G);

d) priebalsis *ch* pagal SAMPA rekomendacijas žymimas simboliu (x).

Sutaptiniai dvibalsiai ir afrikatos koduojami dviem simboliais, tačiau tam, kad nebūtų painiojami su tokiomis pat garsų, priklausančių skirtingiems skiemenims, sandūromis, naudojamas skirtukas (-) (ASCII 45).

Norėdami atskirti mišriuosius dvigarsius sudarančius balsius ir priebalsius nuo atitinkamų vienalyčių garsų, greta (dešinėje) dvigarsio dėmenų rašome tašką (.), t. y. ASCII 046.

2 lentelė. **Lietuvių kalbos garsų žymėjimas SAMPA simboliais**

Liet. kalba	**SAMPA**	**ASCII**	**SAMPA pavyzdys**	**Lietuvių k. pavyzdys**
p	p	112	p^a:talas	pãtalas
p	p'	112, 39	p'^e:l'ekas	pẽlekas
b	b	98	b":stas	bū́stas
b	b'	98, 39	b'^i:ra	bỹra
t	t	116	p'ieSt"ukas	pieštùkas

Lietuvių kalbos fonetinių ypatybių kodavimo SAMPA kodais sistema jokiu būdu neapima

[6] Lietuvių kalboje raide *ė* žymimas garsas visada ilgas ir, siekiant kuo artimesnio IPA sistemai kodavimo, galėtų būti žymėtinas (e:) simboliais, t. y. ASCII 101, 58. Kiti lietuvių kalbos priešakinės eilės neaukštutinio pakilimo balsiai tada būtų koduotini taip: ilgasis *e* – ({:) , t. y. ASCII 123, 58; trumpasis *e* – (E), t. y. ASCII 69.

2 lentelės tęsinys

Liet. kalba	**SAMPA**	**ASCII**	**SAMPA pavyzdys**	**Lietuvių k. pavyzdys**
t	t’	116, 39	t’”E:vas	tėvas
d	d	100	do:van”a	dovanà
d	d’	100, 39	^a:d’r’esas	ãdresas
k	k	107	k”u:nas	kū̃nas
k	k’	107, 39	r’^E:k’E:	rė̃kė
g	g	103	g^a:ras	gãras
g	g’	103, 39	g’”e.r.’t’i	gérti
c	ts	116, 115	ts”ukrus	cùkrus
c	ts’	116, 115, 39	ts’”i.N.kas	cìnkas
č	tS	116, 83	g’i.^n.tSas	giñčas
č	tS’	116, 83, 39	s’v’^e:tS’es	svẽčias
dz	dz	100, 122	dz^u:kas	dzū̃kas
dz	dz’	100, 122, 39	dz’i.N.g”ul’is	dzingùlis
dž	dZ	100, 90	dZ”aul’is	džáulis
dž	dZ’	100, 90, 39	dZ’^euksmas	džiaũgsmas
f	f	102	f^a:z’E:	fãzė
f	f’	102, 39	f’”iz’ika	fìzika
s	s	115	sak^au	sakaũ
s	s’	115, 39	s’ek’^eu	sekiaũ
š	S	83	Sal’”is	šalìs
z	z	122	z^uik’is	zuĩkis
z	z’	122, 39	z’”ebras	zèbras
ž	Z	90	m^a:Zas	mãžas
ž	Z’	90, 39	Z’”ilas	žìlas
ch	x	120	x”Oras	chòras
ch	x’	120, 39	x’i r”u.r.gas	chirùrgas
h	G	71	G”umOras	hùmoras
h	G’	71, 39	G’i.^m.nas	himnas
j	j	106	jE:g”a	jėgà
v	v	118	g^a:vo:	gãvo
v	v’	118, 39	v’^e:da	vėda
m	m	109	m^u:S’is	mū̃šis
m	m’	109, 39	m’ed’”in’is	medìnis
n	n	110	n^a:ras	nãras
n	n’	110, 39	n’^e:Sa	nẽša
n	N	78	su.N.k”u	sunkù
n	N’	78, 39	p’e.N.’k’”i	penkì
l	l	108	l”u:po:s	lū̃pos
l	l’	108, 39	l’iet”us	lietùs
r	r	114	r^o:g’E:s	rõgės
r	r’	114, 39	r’^e:tas	rẽtas
a	a	97	akm^uo	akmuõ

visos lietuvių sakytinės kalbos įvairovės. Ši sistema orientuota į kalbos technologijų kūrimą, o fundamentalūs moksliniai kalbos darbai ir toliau turėtų būti rašomi naudojantis IPA simboliais arba tradicine lietuvių kalbininkų ištobulinta transkripcija.

Straipsnyje aprašytas garsų fondas naudojamas VDU kalbos technologijų tyrimams: šnekai atpažinti, lietuvių kalbos intonacijai tirti, garsų trukmei ir tempui modeliuoti, analizuojama galimybė panaudoti fondą lietuvių kalbos sintezės tyrimams.

e	e	101	l'ed^ai	ledaĩ
i	i	105	t'ik'im'"i:b'E:	tikimỹbė
u	u	117	tur'"E:t'i	turė́ti
o	O	79	Ob'j"ektas	objèktas
a, ą	a:	97, 58	a:Zuol^ai	ąžuolaĩ
e, ę	e:	101, 58	d'^e:da	dẽda
ė	E:	69, 58	r'"i:kS't'E:	rýkštė
y, į	i:	105, 58	t'i:k"us	tykùs
o	o:	111, 58	o:Z^i:s	ožỹs
ū, ų	u:	117, 58	"u:k'is	ū́kis
ai	ai	97, 105	daig^ai	daigaĩ
au	au	97, 117	tauk^ai	taukaĩ
ei	ei	101, 105	v'"eidas	véidas
ie	ie	105, 101	Z'ied^ai	žiedaĩ
ui	ui	117, 105	m^uilas	muĩlas
uo	uo	117, 111	va.n.d^uo	vanduõ

4. IÇVADOS

1. Lietuvių šnekamosios bendrinės kalbos garsų bazės kūrimo patirtis rodo, kad garsų fondas – tai kalbų technologijų produktas, ir, norint kurti tokius produktus, reikia gerai įsisavinti technologijas.

2. Dirbant su garsų fondu (jį segmentuojant, anotuojant, taisant klaidas), išryškėjo kai kurie sudarytos fonetinių vienetų sistemos trūkumai, todėl fondą reikėtų papildyti naujais vienetais.

3. Sąmoningai šiam darbui parinkti pavieniui tariami žodžiai ir sudaryti trumpi sakinukai vis dėlto yra dirbtiniai kalbos vienetai. Jie atspindi ne visas šnekamosios kalbos ypatybes. Tam reikia gerokai ilgesnio rišlaus teksto. Kita vertus, net ir labai ilgame, diktoriaus perskaitytame tekste galime nerasti spontaniškai spalvingai kalbai, įvairiems jos stiliams ir atvejams būdingos kalbos garsinės išraiškos elementų įvairovės. Tam reikia kurti šiek tiek kitokio pobūdžio garsų fondą (tiksliau, papildyti jau turimą), fiksuojantį kuo daugiau realios šnekamosios kalbos variantų.

4. Remiantis garsų fondo duomenimis, buvo sukurta kalbos atpažinimo sistema, naudojanti Paslėptų Markovo modelių atpažinimo metodiką. Taikant šią sistemą atskirai tariami žodžiai atpažįstami 91–97% tikslumu, o rišlios trumpos frazės – 79–94% tikslumu.

Gauta 2004 07 20

Literatūra

D. Gibbon. *Workable Efficient Language Documentation: a Report and a Vision*. Elsnews, 2002, vol. 11, No. 3, p. 3–5.

A. Rudžionis, V. Rudžionis, P. Žvinys. Lietuvių kalbos signalų duomenų bazės LTDIGITS akustinės-fonetinės charakteristikos. *Baltų kalbų fonetikos ir akcentologijos problemos*. St. Peterburgas, 1999 kovo 2–4 d.

A. Girdenis. *Teoriniai fonologijos pagrindai*. Vilnius: Petro ofsetas, 1995.

http://www.elda.fr/index.html

http://www.fon.hum.uva.nl/praat/

http://www.phon.ucl.ac.uk/home/sampa/home.htm

http://www.phon.ucl.ac.uk/home/sampa/x–sampa.htm

http://www.phon.ucl.ac.uk/home/sampa/ipasam–x.pdf

http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm

Arimantas Raçkinis, Gailius Raškinis, Asta Kazlauskienė

THE UNIVERSAL ANNOTATED VDU LITHUANIAN SPEECH CORPUS

S u m m a r y

This paper presents the VDU Lithuanian speech corpus. The corpus has been compiled and annotated by the Center of Computational Linguistics at the Vytautas Magnus University. The corpus aims at providing basic data for Lithuanian language technology researches meant to enable Lithuanian spoken language researchers to use advanced tools provided by today's computer science. The VDU speech corpus contains broadband recordings of 4 speakers (2 males and 2 females), each reading the same set of nearly 7540 isolated words and the same number of word triplets. The corpus includes time-aligned phone-level, phonetic unit-level and word-level transcriptions as well. The VDU Lithuanian speech corpus is universal, *i.e.* its vocabulary has been carefully chosen to include all distinct and independent Lithuanian sounds such as phonemes and phoneme clusters (phonetic units). There have been 275 such phonetic units defined. The paper also describes problems related to the file structure of the corpus and the SAMPAASCII coding of Lithuanian annotations. Some other questions are discussed, such as corpus documentation, validation and standardization. These questions have been addressed in Lithuania for the first time.