RĒZEKNES TEHNOLOĢIJU AKADĒMIJA
INŽENIERU FAKULTĀTE

ANSIS ATAOLS BĒRZIŅŠ

DABISKO VALODU AUTOMATIZĒTA SALĪDZINĀŠANA

DISERTĀCIJAS KOPSAVILKUMS
zinātnes doktora (PhD) grāda
iegūšanai
informācijas tehnoloģijas nozarē

Rēzeknē, MMXX

Disertācija izstrādāta:

Rēzeknes Tehnoloģiju akadēmijas
(Rēzeknes Augstskolas)
Inženieru fakultātē,

Latvijas Universitātes
Datorikas fakultātē
(Fizikas un matemātikas fakultātes Datorikas nodaļā)
un Matemātikas un informātikas institūtā

laika posmā no 2003. līdz 2020. gadam.

Zinātniskais vadītājs:

Artis Teilāns
Dr. sc. ing.,
Rēzeknes Tehnoloģiju akadēmijas
Inženieru fakultātes profesors

Zinātniskie konsultanti:

Pēteris Grabusts
Dr. sc. ing.,
Rēzeknes Tehnoloģiju akadēmijas
Inženieru fakultātes profesors

Matīss Blumberģis
Stokholmas Karaliskās tehniskās augstskolas
Runas, mūzikas un dzirdes centra
pensionēts pētnieks

Lūkass Burģets
Dr. sc. ing.,
Brinnes Tehniskās universitātes
Informācijas tehnoloģijas fakultātes
Skaitļotājgrafikas un multimēdiju nodaļas
Runas grupas vecākais pētnieks

NB Šis ir tikai kopsavilkums! Tiem, kas tiešām vēlas pētījumu izprast, iesakām lasīt DISERTĀCIJU.

Iesniegta Rēzeknes Tehnoloģiju akadēmijas Informācijas tehnoloģijas promocijas padomē 2020. gadā.

Saturs

[1] Kopsavilkumā saglabāta disertācijas numerācija.

Priekšvārdi

Mūsu ceļu uz darba tēmu un tās risināšanu esam plašāk atspoguļojuši ievadā, savukārt sausās, formālās ziņas iznesām priekšvārdos.

Jāatzīmē, ka, kaut disertācija ir iesniegta aizstāvēšanai inženierzinātnēs, tā ir izteikts robežzinātnes jeb starpnozaru zinātnes pētījums un tikpat labi varētu tikt aizstāvēta arī datorzinātnēs, lietišķajā matemātikā vai valodniecībā. Šādu starpnozaru darbu klasificēšana un iedalīšana kādas nozares pārziņā parasti ir atkarīga no lokālās akadēmiski-vēsturiskās tradīcijas.

Darba specifika ir citās skaitļotājvalodniecības jomās pielietotu metožu pielāgošana lingvometrijas jeb dialektometrijas jomai.

Darba tēmas aktualitāti nosaka tās pārlaicīgums: spekulācijas valodu tuvības un no tās izrietošo politisko konsekvenču jautājumos ir bijušas aktuālas visā cilvēces attīstības posmā, kurā tautas izjutušas nacionālu pašapziņu un tās sasaisti ar valodu. XX gs. sākumā, līdz ar nacionālu valstu veidošanās procesa lielo izplatību, šis jautājums kļuvis īpaši aktuāls un XXI gs. tā aktualitāte nav mazinājusies.

Dotā **darba mērķis** ir tādu metožu izstrāde, kas ļautu skaitliski novērtēt dabisko valodu tuvības pakāpi. Šim novērtējumam veiksmīgas realizācijas gadījumā būtu jāatbilst noteiktām prasībām:

- tam ir jābūt metrikai, resp., jāizpildās simetrijas, trijstūra un identiskuma jeb nedeģenerētības aksiomām;
- tam jāatpas sabiedriski-intuitīvajiem priekšstatiem par valodu tuvību (kas vairumā gadījumu atbilst arī valodnieciski-analītiskajiem uzskatiem);
- tam jābūt pielietojamam pēc iespējas lielākam valodu lokam, vēlams – visām cilvēces dabiskajām runas valodām;
- ieejas datu sagatavošanai jābūt iespējami resursneietilpīgai, it īpaši manuāla darba nozīmē.

Mērķa sasniegšanas labad tika uzstādīti un risināti šādi **uzdevumi**:

- ieejas datu formātu atrašana un izvēle, datu sagatavošanas prasību definēšana;
- metožu algoritmu izvēle un pielāgošana;
- izmēģinājuma datu vākšana un sagatavošana;
- metožu izmēģināšana un novērtēšana.

Disertācija **sastāv** no ievada un trijām daļām: pirmā veltīta darbam ar fonētiskajām transkripcijām, otrā – darbam ar runas ierakstiem, savukārt, trešā – rezultātu pārbaudei. Pirmās divas daļas iedalītas vairākās nodaļās, vairums nodaļu – vairākās apakšnodaļās.

Paveiktā zinātniskā **darba galvenie rezultāti**:

- izvēlēti atbilstoši ieejas datu veidi, nodefinētas tehniskās prasības, kurām tiem jāatbilst;
- pašrocīgi savākts un izveidots izlokšņu runas korpuss;
- izstrādātas sešas neatkarīgas, dažādas metodes attāluma noteikšanai:
    - n-grammu metode veseliem, neparalēliem tekstiem (skat. 3. nod.);
    - rediģēšanas attālumu metode paralēliem vārdu komplektiem (skat. 5. nod.);

- fonēmu atpazinēju metode (skat. 8. nod.);
- slēpto Markova modeļu metode nedalītām fonogrammām (skat. 9. nod.);
- i-vektoru metode nedalītām fonogrammām (skat. 10. nod.);
- Gausa maisījuma modeļu metode nedalītām fonogrammām (skat. 11. nod.);
- pielāgota un izstrādāta programmatūra šo metožu iedzīvināšanai;
- veikta rezultātu pārbaude un salīdzināšana;
- iegūti nozīmīgi zinātniski starprezultāti, piemēram, izstrādāta fonēmu telpa (skat. 4. nod.) un hierarhisko izvēļu metrika (skat. 12. nod.).

Darbs ir **zinātniski novatorisks**, jo:

- lai arī valodu tuvības jautājums kā tāds ir kužinājis cilvēku prātus jau izsenis, un beidzamajos gadu desmitos ir bijuši arī zinātnieku mēģinājumi to risināt ar skaitļotāja palīdzību, tomēr no mūsu puses tā risināšanai ir piedāvāts līdz šim nebijis un neizskanējis uzstādījums – runas, nevis rakstu, valodas ņemšana par pamatu, tādējādi būtiski paplašinot aplūkojamo valodu loku un arī padarot metodes objektīvākas, neatkarīgas no nosacītām pieraksta sistēmām;
- ir rasti jauni pielietojumi citās skaitļotājvalodniecības nozarēs (tekstu klasifikācija, runas atpazīšana u.c.) izmantotām metodēm, piem., pilnīgi jauna pieeja ir runas atpazīšanā izmantoto statistisko modeļu ģenerēšana no pilnām informantu fonogrammām, nevis atsevišķiem konkrētiem vārdiem, tādējādi izveidojot modeļus, kuri raksturo doto valodu kopumā;
- ieviests un aprakstīts jauns jēdziens – fonēmu telpa, nodefinētas tās dimensijas, kurās piešķirtas koordinātes visām latviešu un latgaliešu izlokšņu fonēmām, tanī definēta anatomiski un fonētiski pamatota metrika (līdz šim neviens nav salicis gan patskaņus, gan līdzskaņus vienotā telpā ar attālumiem, kuri atbilst intuitīvai izpratnei par fonēmu tuvību);
- iegūti ne tikai gaidīti, bet arī negaidīti eksperimentu rezultāti, kurus iespējams zinātniski izskaidrot un pamatot (skat., piem., secinājumus 5. nodaļas beigās);
- ieviesta jauna – hierarhisko izvēļu – metrika (pierādīta tās atbilstība metrikas aksiomām), kura ļauj skaitliski salīdzināt hierarhiskās kategorizēšanas rezultātā iegūtos bināros kokus.

Darba **rezultāti praktiski realizēti** kā:

- PERL CGI skripti fonēmu fonētiskā attāluma tabulas aprēķināšanai un grafiskai izvadei;
- PERL skripts n-grammu metodes pielietošanai;
- PERL skripti Levenšteina un Vāgnera-Fišera attālumu metožu pielietošanai;
- PERL skripti fonēmu atpazinēju *PhnRec* un *Sphinx* pielietošanai un to rezultātu analīzei;
- PERL skripts slēpto Markova modeļu pielietošanai, izmantojot pakotni HTK;
- PERL skripti i-vektoru metodes pielietošanai;
- PERL skripti rezultātu kategorizēšanai;
- Pitona skripts kosīnus novērtējuma pielietošanai;
- Pitona skripts Kuļbaka-Leiblera diverģences pielietošanai;
- MatLab skripts Gausa sadalījumu modeļu no pilnām fonogrammām, bez MAP adaptēšanas, veidošanai;
- MatLab skripts Gausa sadalījumu modeļu ar MAP adaptēšanu veidošanai;
- MatLab skripts dažādu (Eiklīda, L2, Kuļbaka-Leiblera, Žordāna) attālumu aprēķināšanai, pirms tam pārveidojot matricas par supervektoriem;

- MatLab skripts īpašas Mahalanobja distances Gausa sadalījumu pārim aprēķināšanai;
- MatLab skripts īpašas Eiklīda distances Gausa sadalījumu pārim aprēķināšanai;
- PERL skripts hierarhisko izvēļu metrikas aprēķināšanai;
- u.c. mazākas programmatūras vienības.

Laika posmā no 2004. līdz 2019. gadam zinātniskā darba **rezultāti apspriesti**:
- starptautiskās zinātniskajās **konferencēs**:
  - «Корпусная лингвистика», Pēterpilī, Krievijā, 2004. g.;
  - «Identification des langues et des variétés dialectales par les humains et par les machines», Parīzē, Francijā, 2004. g.;
  - «Диалог», Maskavas apgabalā, Krievijā, 2006. g.;
  - J. Endzelīna konferencē, Rīgā, Latvijā, 2007. g.;
  - «Megaling», Kijevā, Ukrainā, 2009. g.;
  - „ქართული ენა და თანამედროვე ტექნოლოგიები“, Tiflisā, Gruzijā, 2011. g.;
  - «Диалог», Maskavā, Krievijā, 2016. g.;
  - «Artificial Intelligence and Natural Language», Tērbatā, Igaunijā, 2019. g.;
- **doktorantūras skolās**:
  - Rīgas Tehniskās universitātes Tālmācības studiju centra doktorantūras skolas ziemas sesijā, Īves pag., 2011. g.
  - Rīgas Tehniskās universitātes Doktorantūras skolas semināršacensībās „Research Slam“, Rīgā, 2015. g.
- **semināros**:
  - Latvijas Universitātes Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijas semināros, 2004.-2007. g.
  - Latvijas Universitātes Fizikas un matemātikas fakultātes Datorikas nodaļas doktorantu seminārā, 2008. g.
  - Rīgas Tehniskās universitātes Tālmācības studiju centra semināros, 2009.-2013. g.
  - Rēzeknes augstskolas Inženieru fakultātes semināros, 2014.-2016. g.
- **stažēšanās** laikā:
  - Karaliskās Tehniskās augstskolas Runas, mūzikas un dzirdes centrā, Stokholmā, Zviedrijā, 2008. g.
  - Brinnes Tehniskās universitātes Informācijas tehnoloģijas fakultātes Skaitļotājgrafikas un multimēdiju nodaļā, Runas apstrādes pētnieciskajā grupā, Brinnē, Morāvijā, Čehijas Republikā, 2015. g.
  - Leišu valodas institūtā, Viļņā, Lietavā, 2015. g.
  - Mehiko Nacionālās autonomās universitātēs Inženierzinātņu fakultātes Runas apstrādes laboratorijā, Mehiko, Meksikā, 2016. g.

Disertācijā ietvertā darba rezultātus esam apkopojuši **9** zinātniskajās **publikācijās**:

¤ Берзиньш А.У. Сравнение балтийских языков методом n-грамм // Труды международной коференции «Корпусная лингвистика - 2004». СПб.: Издательство С.-Петербургского университета, 2004.

¤ Berzinch A.A. La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes // Identification des langues et des variétés dialectales par les humains et par les machines. Paris: École National Supérieure des Télécommunications, 2004.

¤ Берзинь А.У. Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, 2006.

¤ Bērziņš A.A., Grigorjevs J. Latviešu izloksnēs sastopamo fonēmu telpa // Linguistica Lettica XVIII. R.: Latviešu valodas institūts, 2008.

¤ Берзинь А. Возможности применения статистических методов распознавания речи для определения близости языков // Прикладна лінгвістика та лінгвістичні технології, Megaling-2009. Київ: «Довіра», 2009.

¤ ბერზინი ა. ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის / Принципы сбора информации для автоматизированного анализа фонограмм // ქართული ენა და თანამედროვე ტექნოლოგიები - 2011. თბილისი: „მერიდიანი“, 2011.

¤ Берзинь А.У. Применение распознавателей фонем для автоматического определения уровня близости языков // Труды международной конференции «Диалог 2016». М., 2016.

¤ Bērziņš A.A. Usage of HMM-based Speech Recognition Methods for Automated Determination of a Similarity Level between Languages // AINL Proceedings. «Springer», 2019.
*Indeksēta **Scopus** un **Web of Science**.*

¤ Берзинь А.У. Применение i-векторов для автоматизированного определения уровня близости языков // Труды Института системного программирования РАН. Том 31, № 5. М.: ИСП РАН, 2019.
*Indeksēta **Российский индекс научного цитирования** (РИНЦ) un **КиберЛенинка**.*

Vēl **1** disertācijā izmantotie rezultāti **publicēšanai** iesniegti, pie tam pirmais no tiem arī pieņemts:

¤ Berziñch A.A., Chavarría Amezcua M.A. El espacio de los alófonos del español. Iesniegts publicēšanai «Lengua y Habla», 2020.

Tāpat disertācijas izstrādes laikā esam publicējuši **8** zinātniskās publikācijas citās nozarēs, tai skaitā: 1 – mašīntulkošanā, 2 – datorleksikogrāfijā, 3 – salīdzinošajā valodniecībā, 1 – etnomuzikoloģijā un 1 – salīdzinošajā folkloristikā, kā arī iesnieguši publicēšanai **1** zinātnisko publikāciju terminoloģijā.

Ievads

Ir dažādas hipotēzes, kad cilvēkveidīgas būtnes sākušas runāt. Piemēram, cilvēkpērtiķi australopiteki jau staigājuši stāvus, tādējādi pārnesot elpošanas un runas aparātu runāšanai piemērotākā, ērtākā stāvoklī. Tāpēc pētnieki pieļauj, ka elementāra runa bijusi jau pie tiem, kaut mūsdienu cilvēkveidīgo pērtiķu saziņu par runu vai valodu neuzskata – tā esot sazināšanās atsevišķiem saucieniem, bet runas (un tātad arī valodas) priekšnosacījums esot domas klātesamība. Ir arī konstatētas anatomiskas attīstības pazīmes, kuras varētu būt saistītas tieši ar runāšanu, piemēram, ašeliešiem. Tiek uzskatīts, ka neandertāliešu sabiedrībā jau ticis runāts. Katrā ziņā, mūsdienu cilvēka jeb *homo sapiens sapiens* pastāvēšanas laikā runa kā saziņas līdzeklis droši vien pastāvēja jau no sākta gala.

Saskaņā ar valodniecības klasiķi F. Sosīru, runa ir individuālas izpausmes veids, savukārt valoda jau ir norunu sistēma, kuru to pārzinošs indivīds var izmantot savā runā. Tātad faktiski mūs interesējošajā jautājumā mēs runu no valodas varam neatdalīt, un uzskatīt, ka tās iet tikpat kā roku rokā – nebūs tak cilvēkveidīgam radījumam jēgas runāt, ja viņu neviens nesapratīs, respektīvi, veikt runas aktu, kurš neatbilst kādām sabiedriskām valodas norunām.

Par valodu izcelsmi pastāv dažādas hipotēzes. Piemēram, monoģenēzes teorija vēsta, ka visas valodas laika gaitā cēlušās no vienas pirmvalodas. Ja *homo sapiens sapiens* tik tiešām nāk no vienas vietas Āfrikā, kur radušies mutācijas rezultātā, tad šāds pieņēmums būtu loģisks. Taču iespējami arī citi varianti. Lai nu kā tur bija, tomēr skaidrs, ka kolīdz parādījās valodu atšķirības, tā radās arī to tuvības problēma: tuvākas varēja saprast, tālākas – nevarēja, un ļaudis tās kaut vai šādi pie sevis klasificēja. Tātad šī problēma, lai arī intuitīvā līmenī, ļaužu prātos rūga jau pirms daudziem tūkstošiem gadu.

Pirmās zināmās publikācijas par valodu līdzību Eiropā parādījās XVI gs., bet zinātniskā līmenī šo jautājumu XVIII gs. pacēla Viljams Džonss.

Vēsturiski-salīdzinošā valodniecība paģēr tikai radniecīgas izcelsmes, t.i., tādu, kurām ir kopīga vēsture, valodu salīdzināšanu. Lai arī šāda pieeja ir likumsakarīga un meklē atbildes uz būtiskākajiem valodu tuvības jautājumiem, tomēr, mūsuprāt, šobrīd tā jau ir novecojusi un būtu paplašināma. Daudzas neradniecīgas izcelsmes valodas ir garus gadus līdzāspastāvējušas, viena no otras ietekmējušās, tāpēc valodu tuvība vai tālība ir ne tikai attiecīgo lapu izvietojuma izcelsmes kokā salīdzinājums, bet visu valodas attīstības posmu kopīgā pienesuma rezultāta atšķirību vērtējums. Piemēram, zinot lībiešu valodas ietekmi latviešu valodā, jautājums par igauņu, somu u.c. somugru valodu tuvību latviešu valodai nav nemaz tik bezjēdzīgs, kaut izcelsmes ziņā tās ir dažādu virssaimju valodas (protams, vācu valodas ietekme latviešu un igauņu valodās un krievu valodas ietekme latviešu un somu valodās šo interesantumu tikai palielina). Respektīvi, raugoties no mūsdienu skatupunkta, svarīgs ir ne tikai vēsturiski-salīdzinošs, bet arī tīri salīdzinošs novērtējums.

200 gadu laikā vēsturiski-salīdzinošā valodniecība savu pamatuzdevumu ir veikusi – vairums zināmo pasaules valodu ģenealoģiski saklasificētas (kaut, protams, kā jau jebkurā nozarē, darbu dziļumā un smalkumā neaptrūksies nekad). Tomēr tā nepiedāvā šīs tuvības skaitlisku novērtējumu, lai arī sabiedriski procesi reizēm šādu novērtējumu pieprasa.

Neba nu bez iemesla kuldīdznieka Meijera (Makša) Veinreiha 1945. gadā publicētais izteiciens (ko tas dzirdējis no kāda savas lekcijas klausītāja), ka valoda ir dialekts plus armija un flote, šobrīd jau ir nosaucams par spārnotu un vispārzināmu. Šī definīcija trāpīja kā naglai uz galvas, raksturojot politisku procesu ietekmi uz valodnieciskiem jautājumiem. Vai tam tā būtu jābūt? Droši vien – nē, zinātnei, tai skaitā

valodniecībai, būtu jābūt objektīvai, tātad arī neatkarīgai no politiska pasūtījuma. Tomēr līdz pat šai dienai, neskatoties uz zinātnes un skaitļošanas jaudu strauju attīstību kopš 1945. gada, mehānisms, kurš ļautu objektīvu mērījumu rezultātā atspēkot šāda veida politiskas spekulācijas, aizvien vēl nav radīts.

Mēs sajutām pēc tā vajadzību 2000. gadu sākumā, pirms iestāšanās doktorantūrā. Tas arī noteica promocijas darba tēmas izvēli. Vēlāk sastapāmies ar 1980. gada Jahontova publikāciju, kas pamatojuma ziņā gandrīz pilnībā sakrita ar mūsu motivāciju, tomēr arī tanī autora piedāvātās skalas jebkurām valodām un vienas tautas divām valodām dod nevis skaitlisku novērtējumu, bet sakārtotu novērtējumu attiecīgi 5 un 8 pakāpēs. Respektīvi, valodu pāru tuvības atšķirībām esot pietiekami lielām, varam konstatēt, ka pēc šī novērtējuma viens pāris savstarpēji tālāks nekā otrs, bet tiklīdz tie iekļūst vienā „tuvības grupā“, mums nekāda smalkāka nosvēršana nav pieejama.

Mūsu uzmanību piesaistīja arī J. Tambovceva darbi. Tā autors rēķina fonotipoloģisku attālumu starp basku un virkni citu dažādu grupu valodu pēc Eiklīda attāluma formulas astoņdimensionālā telpā, kuras dimensijas ir dažāda veida līdzskaņu sastopamības biežums jeb frekvence. Un lai arī Tambovceva eksperimentu rezultāti ir gaužām interesanti, mūs tie, par nožēlu, neapmierina: autors darbojas tikai ar oficiālām, literārām valodām, piedevām nevis ar runu vai tās fonētisku transkripciju, bet ar rakstītiem avotiem (grāmatām), kas sarakstītas noteiktā ortogrāfiskā tradīcijā un kuras tas pārceļ uz paredzamo skanējumu saskaņā ar formāliem noteikumiem, kā kas jālasa. Nav šaubu, gan pieraksts, gan vēlākā atskaņošana ir pietiekami nosacīti un, ja skatāmies stingri, rada mākslīgu, no patiesās atšķirīgu valodu. Vairāku šādi iegūtu valodu salīdzināšana nav nekorekta, tomēr tā kā mūs interesē arī valodas, kurām rakstu tradīcijas nav, tad mums jāmeklē kādi citi ceļi.

To, ka skaitlisks novērtējums būtu labākais pierādījums, jau 50. gadu vidū saprata arī amerikāņu zinātnieks M. Svodešs. Autors priekš glotohronoloģijas piedāvāja pēc kopīgas valodu pāra (īpaši atlasītas – tā, lai atspoguļotu tieši izcelsmi, nevis vēlākus kontaktus) leksikas izrēķināt to atdalīšanās laiku gadsimtos (proti, skaitu, pirms cik gadsimtiem valodas nošķīrušās), un pēc šī skaitliskā kritērija noteikt valodu iedalījumu (piem., kura kurai cik tuva (sena) rada). Autors pārsvarā strādā ar CVC (līdzskanis-patskanis-līdzskanis) grupu sakritībām un atbilsmēm, pie tam dažādām situācijām (atkarībā no valodu īpašībām un savstarpējām attiecībām) apraksta vairākas metodes. Tomēr šāds novērtējums izmantojams tikai vienas virssaimes valodām, kā arī ņem vērā tikai leksiku, kas ir nepietiekami. Un, lai arī šis novērtējums tīri matemātiski pat ir attālums, respektīvi, atbilst metrikas aksiomām, tomēr tā iegūšanas mehānisma nosacītība un skalas soļa lielums rada pamatotas šaubas, vai tāds būtu pielietojams mūsu gadījumā.

Mūsuprāt, būtu jāņem vērā ne tikai leksika, bet arī fonētika, morfoloģija un sintakse, jo arī tās raksturo valodu pāra tuvību. Skaidrs, ka pie šāda uzstādījuma ir jēga strādāt tikai ar tādiem datiem, kuri šo informāciju satur, pie tam ir fiksēti, valodneatkarīgi, universāli. Teksti ortogrāfiskā pierakstā ir pilnīgi neizmantojami, jo ir nosacīti un neviennozīmīgi – vienai valodai var būt vairākas rakstu tradīcijas, un no šādiem rakstiem mēs nevaram viennozīmīgi iegūt izrunu. Secinājums – jāizmanto vai nu teksti fonētiskā pierakstā (pie tam visām valodām jālieto viena, universāla pieraksta sistēma), vai nu runas skaņu ieraksti jeb fonogrammas. Līdz ar to mēs izvirzām **hipotēzi**, kuru disertācijas izstrādes gaitā centīsimies pārbaudīt: gan runas fonētiskā transkripcija, gan runas ieraksti kā ievades dati ir pietiekami, lai no tiem ar statistiskām metodēm iegūtu skaitlisku valodu tuvības pakāpes novērtējumu.

Mūsdienās glotohronoloģijas metode jau skaitās novecojusi, tās vietā tiek izmantotas filoģenētikas metodes, kā ieejas datus lietojot glotohronoloģijā uzdotos Svodeša leksikas sarakstus – papildinātus ar kādām attiecīgajam pētījumam svarīgām leksikas

vienībām vai pat nepapildinātus. Attīstības teoriju un kokus zinātnieki sāka ieviest jau XIX gs., nosaukums „filoģenētika“ parādījās XX gs. 20. gados, bet filoģenētisko koku veidošanu radību sugām 1950. gadā formalizēja vācu entomologs Vilis Henihs, 1965. gadā tas savu ideju publicēja arī angliski. Izrādījās, ka piedāvātā metode ir ļoti universāla, un pamazām to sāka izmantot ari citās zinātnēs un nozarēs, piemēram, ģenētikā (molekulu DNS salīdzināšanai), un, kā var noprast no mūsu intereses, arī valodniecībā. Indoeiropiešu lingvistiskā filoģenētika tiek intensīvi izmantota XX gs. 50.-60. gados, bet kopš 70. gadiem tajā parādās arī statistiskas metodes. Statistiskā filoģenētika tiek izmantota pat ļoti pamatīgās un gluži svaigās publikācijās, kurās, piemēram, ar filoģenētikas metožu palīdzību par pamatotāku tiek pierādīta viena no divām indoeiropiešu valodu izcelsmes hipotēzēm. Interesanti, ka, lai arī pastāv citu autoru radīti vārdu saraksti, kuri atspoguļo arī fonētiskas un/vai morfoloģiskas atšķirības, tomēr publikācijas autori nonāk pie secinājuma, ka racionālāk ir orientēties uz leksiskajām atšķirībām. Publikācijā aprēķinātie koki tik tiešām izskatās ļoti ticami – tie pilnībā atbilst intuitīvajai izpratnei par attiecīgo valodu tuvību. Bet metodes trūkums, tāpat kā glotohronoloģijas gadījumā, ir tas, ka tā ir pielietojama tikai valodām ar kopīgu izcelsmi.

Arī šinī disertācijā definētie attālumi, protams, nedrīkst būt pretrunā ar ļaužu intuitīvo izpratni par valodu tuvību un tālību. Katram kādas valodas runātājam taču ir priekšstats, kādas valodas viņa valodai tuvākas, kādas – tālākas, kuras viņš var saprast gandrīz pilnībā, kuras – daļēji, kurās – tikai atsevišķus vārdus. Protams, šāds priekšstats, atkarībā no cilvēka redzesloka, ir vien par kādu nelielu valodu daļu – pārsvarā par kaimiņu tautu valodām un par lielām, starptautiskām valodām. Lielākoties šī izpratne atspoguļo faktisko tābrīža situāciju valodās, nemēģinot atdalīt izcelsmes radniecību no vēlākiem līdzāsdzīvošanas jaunieguvumiem, kaut, neapšaubāmi, tautas vēsturiskajai atmiņai arī var būt sava nozīme. Tādējādi šī intuitīvā izpratne lielā mērā atbilst arī analītiskas valodu klasifikācijas iespējām: gan augšminētajai ģenealoģiskajai, gan tipoloģiskajai jeb morfoloģiskajai.

Disertācija sakārtota hronoloģiski – tā, kā attīstījās tēmas izstrāde, autoram meklējot metodes, kuras varētu līdzēt problēmas risināšanā. Likumsakarīgi, ka vispirms mēs pieķērāmies darbam ar tekstiem fonētiskajā transkripcijā (faktiski, ne tikai tekstiem, jo viena metode tika izstrādāta paralēliem vārdu krājumiem), jo tie ir uzskatāmāki, tad pārgājām pie fonogrammām, kuru analīze, protams, ir sarežģītāka, tomēr uzdevumu atviegloja tas, ka tām izmantojām jau pieejamas programmatūras pakotnes skaņas statistiskai analīzei.

n-grammu metode veseliem, neparalēliem tekstiem

1994. godā Viljams Kanvārs un Džonis Trenklus lika priekšā tekstu kategorizēšanai izmantot n-grammu biežuma sarakstus. Metodes sāls slēpjas tanī, ka saskaņā ar Cipfa likumu vārdu (vai – kā mūsu gadījumā – grafēmu vai fonēmu virkņu) kopu var sakārtot pēc to lietošanas biežuma. Kanvārs un Trenklus piedāvā sastādīt n-grammu biežuma sarakstus dažādiem tekstiem un, salīdzinot šos sarakstus un ievades teksta n-grammu biežuma sarakstu, noteikt kategoriju, kurai ievades teksts piederīgs. Tādā veidā var sekmīgi automātiski noteikt teksta valodu, kodējumu un pat tematiku. Kas tanī kopīgs ar mūsu uzdevumiem? Lieta tāda, ka n-grammu sarakstu salīdzināšanas labad tiek sarēķināts tekstu līdzības pakāpi raksturojošs skaitlis. Tāpēc likumsakarīga bija jautājuma rašanās: vai varam šo metodi izmantot valodu līdzības pakāpes noteikšanai? Acīmredzot – jā, bet, protams, pie nosacījuma, ka visām aplūkojamajām valodām tiek izmantots kopīgs fonētiskās transkripcijas pieraksts, jo dažādi pieraksti un arī dažādi kodējumi, kuri kategorizēšanas uzdevumam pat iraid noderīgi, šim jautājumam viennozīmīgi skādē, pat padara to bezjēdzīgu. Tāpat jāņem vērā, ka prasības kategorizācijas uzdevuma atrisināšanai ir būtiski vājākas, nekā tuvības noteikšanas uzdevuma gadījumā, tāpēc varam paredzēt, ka pietiekami precīzu rezultātu iegūšanai mums nāksies darboties ar lielāka apjoma tekstiem.

Izmantojamo n-grammu garums (t.i., *n*) tiek izvēlēts empīriski. Mēs nospraudām *n*:*=1..5*, proti, no unigrammām līdz kvintagrammām. Tas, kādas n-grammas būs pārstāvētas biežuma sarakstos, protams, ir atkarīgs arī no saraksta izmēra (ierakstu skaita tanī): to vajag uzdot tā, lai saraksts pēc iespējas labāk reprezentētu tekstu, uz kura radīts, bet tanī pat laikā visos tekstos n-grammu „pietiktu“, t.i., mēs sasniegtu uzdoto saraksta izmēru. Mūsu gadījumā mēs nodefinējām *N*:*=400* un, lai saglabātu visos izmēģinājumos apstākļus vienādus, atstājām to nemainīgu, kas pēc būtības nav pareizi: šī konstante jāpielāgo konkrētā eksperimenta apstākļiem, visbiežāk – datu, resp., ievades teksta, apjomam.

Šinī eksperimentā, kur izmantojām manuāli sagatavotus fonotekstu (tātad tie bija salīdzinoši nelieli), mūsu sarakstos bija visu 5 atļauto izmēru n-grammas, pie tam saraksta pakaļgalā atrodamo n-grammu biežums bija 1, resp., *N* teorētiski bija tuvs iespējamajam maksimumam, bet praktiski – pat nedaudz par lielu. Tas arī noteica to, ka attālumi ir skaitliski lieli, jo ir daudz reti sastopamu n-grammu, kuru varbūtība tikt sastaptām arī citās valodas ir niecīga. Savukārt 8. nodaļā veiktajā eksperimentā, kurā arī pielietojām n-grammu metodi, bet par ievades datiem izmantojām automātiski sagatavotus tekstus (tātad – lielus), mūsu sarakstos bija atrodamas tikai unigrammas, bigrammas un trigrammas (jo kvadragrammas un kvintagrammas sastopamas retāk), bet sarakstu pakaļgala n-grammu biežums bija no vairākiem desmitiem līdz dažiem simtiem, tātad *N* bija pārāk mazs.

Zināmie fonētiska pieraksta datu uzglabāšanas ESM standarti (dažādi lokāli, Unikods, SAMPA, IPA/ASCII) mūs neapmierina, jo tie vai nu nesatur visas skaņas, vai nu tiem ir mainīgs rakstzīmju skaits vienas fonēmas apzīmēšanai (t.i., tie var būt baiti, bet var būt baitu pāri). Mums būtu daudz ērtāk strādāt ar vienizmēra elementu tekstiem, kuros katra skaņa tiek aprakstīta ar vienādu informācijas vienību daudzumu. Tā kā viena baita (256 permutāciju) visu cilvēka runas skaņu aprakstīšanai nepietiek, tad ērtākais veids, acīmredzot, būtu izmantot divus baitus (65536 permutācijas). Protams, teorētiski iespējams arī vēl smalkāks dalījums, piemēram, ja par dažādām skaņām uzskatām ne tikai fonēmas, bet arī to alofonus. Tomēr nesaskatām ieguvumus, ko tas varētu dot, savukārt, noteikti pieaugs transkribētāja kļūdu un individuālās uztveres īpatnību īpatsvars. Pie tam, tā kā jau pastāvošajās fonetiskā pieraksta sistēmās šādas nianses netiek attēlotas, tad nebūs iespējams izmantot jau transkribētus tekstus.

Acīmredzot perspektīvā būtu vēlams izstrādāt speciālus unikoda šriftus, kā arī redaktoru un konvertorus priekš šāda pieraksta teksta. Uz eksperimenta brīdi darbā ar baltu valodu transkripcijām mēs iztikām ar divbaitīgu pseidokodu: katru sastopamo skaņu aprakstījām ar diviem ASCII simboliem saskaņā ar pašu definētu shēmu.

Uz eksperimenta veikšanas brīdi baltu valodu fonotekstu korpusi pieejami nebija, tāpēc mums nācās pašiem tos ievadīt ESM no drukātiem krājumiem. Tā kā to darījām tikai eksperimenta labad, tad ievadījām nelielu skaitu neliela izmēra tekstu: 4 dažādu Latvijas (3 – Latgales un 1 – Kurzemes) izlokšņu paraugus apjomā no 500 līdz 1000 skaņzīmēm. Pēc tam mēs pārveidojām PERLa programmu *TextCat* (kuru 1994. gadā, iespaidojies no Kanvāra un Trenklus publikācijas, uzrakstīja holandietis Hertjānis van Nords) tā, lai tā strādātu ar divbaitu, nevis vienbaita, simboliem, t.i., n-grammas bija 2n baitus garas, nevis n, kā oriģinālajā programmā.

Sākotnēji ieguvām nesimetriskus (lai arī tuvības pakāpes ziņā ticamus) rezultātus. Tā kā zinājām, ka tiem jābūt simetriskiem, tad sākām meklēt kļūdu un programmas *TextCat* algoritmā – tās oriģinālajā pirmkodā – atradām veselas divas, kas pie šīs asimetrijas noveda. Pēc izlabošanas ieguvām simetriskus rezultātus.

Neskatoties uz nelielo ievadīto fonotekstu datu apjomu, mūsu eksperimenta rezultāti pilnībā attaisno mūsu cerības. Tā divas kaimiņos esošas Ziemeļlatgales izloksnes – Baltinavas un Šķilbēnu – ir viena otrai vistuvākās, nākamā tuvākā ir Nirzas, un vienīgā kurzemnieku izloksne izrādās pati tālākā. Nirzas izloksnei pati tuvākā izrādās Baltinavas runa, kas ir pamatoti gan ģeogrāfiski, gan lingvistiski, bet pati tālākā, protams – Džūkstes. Džūkstes izloksnei pati tuvākā izrādās Šķilbēnu, kas arī neizbrīna, jo Ziemeļlatgales izloksnēs ir vairāk kopīga ar pārnovadnieku valodām gan no leksiskā viedokļa, gan no atsevišķu gramatisku formu skanisku noformējuma raugoties, bet pati tālākā – Nirzas, kas tik tiešām ir „dziļāk“ latgaliska gan leksiskā, gan morfoloģiskā ziņā. Baltinavas izloksne, savukārt, izrādās nedaudz tuvāka Nirzas izloksnei nekā Šķilbēnu izloksne; arī tas atbilst analītiskajai izpratnei, jo Baltinavas izloksnē ir jūtama lielāka Kārsavas izloksnes (kas, savukārt, ir tuvāka Nirzas izloksnei) ietekme nekā Šķilbēnu izloksnē.

Tātad ar ļoti nelieliem tekstiem un bināru salīdzināšanas procedūru esam ieguvuši pietiekami (pat ļoti) labus rezultātus. Tāpēc ir pamats uzskatīt, ka šāda n-grammu metode ir izmantojama valodu tuvības pakāpes noteikšanai, jo īpaši pie vēl lielākiem un kvalitatīvākiem fonotekstu korpusiem. Iespējams, ka vēl labākus rezultātus būtu iespējams iegūt, biežuma sarakstu rindu salīdzināšanai izmantojot kādu citu attālumu, kurš raksturotu ne tikai rindu sakrišanu vai nesakrišanu, bet arī līdzības pakāpi, piemēram, Levenšteina vai Vāgnera-Fišera attālumu.

Mēs šī eksperimenta rezultātus publicējām 2004. gadā. Daudzus gadus vēlāk – 2016. gadā – no indiešu kolēģiem uzzinājām, ka viņi 2011. gadā veikuši līdzīgu eksperimentu ar Indijas valodām. Tiesa, viņi nestrādāja ar īstiem fonotekstiem, bet tos emulēja no ortogrāfiskiem tekstiem, konvertējot tos vienotā fonēmu pierakstā – tātad metode ir pielietojama arī īstām fonētiskajām transkripcijām. Atšķirībā no mūsu metodes, viņi veido n-grammu varbūtību modeļus un tad tiem rēķina L_1 un L_2 normas, kā arī Kuļbaka-Leiblera un Rao diverģences. Nākotnē būtu interesanti izmēģināt viņu metodi mūsu datiem un salīdzināt rezultātus.

Fonēmu telpa. Jēdziens. Realizācija latviešu izloksnēs

Ar vajadzību izvietot visas latviešu (gan baltiešu, gan latgaliešu) izloksnēs sastopamās fonēmas vienā daudzdimensiju koordinātu sistēmā jeb telpā saskārāmies, kad nolēmām mērīt attālumus starp fonētiskajā transkripcijā pierakstītiem izlokšņu vārdiem, izmantojot Vāgnera-Fišera attālumu (tā aprakstīta nākamajā nodaļā): tam bija nepieciešams definēt attālumu starp jebkurām divām pierakstā izmantotajām zīmēm, un bija skaidrs, ka šādam attālumam jābūt fonētiski pamatotam.

Rediģēšanas attālumi, t.sk. Vāgnera-Fišera attālums, tiek izmantoti kā viens no zīmju virkņu līdzīguma noteikšanas paņēmieniem, respektīvi, tie ir mēri, kas raksturo, cik „izmaksātu“ vienas teksta rindas (piem., vārda vai teikuma) pārveidošana otrā. Visvienkāršākais rediģēšanas attālums ir Levenšteina attālums, kurā jebkura burta jeb rakstzīmes (tātad fonētiskajā pierakstā – fonēmas jeb skaņzīmes) dzēšana, pievienošana vai aizvietošana izmaksā vienu vienību. Jau 1995. gadā Kesleris, salīdzinot īru izloksnes, saprata, ka šāda pieeja būtu pārāk neprecīza un jāizmanto Vāgnera-Fišera attālums, respektīvi, dažādu fonēmu aizvietošanai jābūt dažādā vērtē (piem., līdzskaņa aizvietošana ar līdzskani būs „lētāka“ par aizvietošanu ar patskani). Tomēr viņš aprobežojās ar 12 fonētisko parametru izvēli, nemēģinot saprast, cik konsekventa un savstarpēji atbilstoša veidojas šī divpadsmitdimensionālā fonēmu telpa, jo pat šāda pieeja viņam sniedza pietiekoši labus rezultātus īru izlokšņu kategorizācijā. Mūs šāda pieeja neapmierināja, jo vēlējāmies ne tikai iegūt apmierinošus latviešu izlokšņu kategorizācijas rezultātus, bet arī izveidot intuitīvajiem priekšstatiem atbilstošu telpu, pie tam ar mazāko iespējamo asu skaitu (Kesleris līdzīgas dabas raksturlielumus ir sadalījis pa atsevišķām asīm). Tāpēc vērsāmies pie latviešu fonētikas speciālista Jura Grigorjeva pēc padoma, kā rezultātā kopīgiem spēkiem šādu telpu aprakstījām.

Bet vispirms izskaidrosim, kas šajā pētījumā tiek saukts par „fonēmu“. Nevēlamies mainīt tradicionāli pieņemto fonēmas definīciju. Tomēr, ja atzīst, ka fonēma ir tikai konkrētā valodā pastāvoša funkcionālā vienība, kuras tveramā realizācija ir valodas skaņa vai to kopums, nonāk pie tā, ka divu valodu fonēmas vai to sistēmas nav iespējams objektīvi salīdzināt. Šādā gadījumā būtu salīdzināmas tikai fonēmu realizācijas jeb alofoni vienādās fonētiskās apkaimēs pēc to akustiskajām, artikulārajām vai auditīvajām īpašībām. Šim salīdzinājumam būtu jābalstās uz audiomateriālu analīzi un runas orgānu darbības dokumentēšanu ar dažādām ierīcēm. Ja vēlas salīdzināt divas vai vairākas valodas, vai arī vairākas izloksnes vienas valodas ietvaros, parasti mēģina salīdzināt šo valodu vai izlokšņu ideālo skanējumu, kas runātāju un klausītāju apziņā ir veidojies saziņas procesā. Šis ideālais skanējums parasti ir neatkarīgs no individuālām vai sociāli un teritoriāli noteiktām izrunas īpatnībām un ir valodas lietotājam kā standarts, uz ko tas tiecas arī savā izrunā. Šāds standartskanējums tiek attēlots vārdnīcās, kurās fonētiskajā transkripcijā tiek norādīta katra valodas vai izloksnes vārda vēlamā izruna. Tā kā praksē skaņas vai to kompleksi tiek lietoti nozīmes diferencēšanai, var vilkt paralēles starp izrunas attēlojumu vārdnīcās vai citos avotos dotajos izrunas aprakstos šajā plašākajā (fonemātiskajā) transkripcijā, kurā veikta distancēšanās no konkrētu runātāju individuālo izrunas īpatnību detalizēta attēlojuma par labu kolektīvā ideāla izrunai, un valodas vai izloksnes fonēmām. Lai gan katra indivīda runas orgānu uzbūve nedaudz atšķiras, ir skaidrs, ka kopumā visu cilvēku (izņemot tos, kuriem vērojama runas orgānu attīstības pataloģija) runas orgānu anatomija ir ļoti līdzīga. No tā secināms, ka visi pasaules cilvēki teorētiski ir spējīgi izrunāt visas pasaules valodās sastopamās skaņas, kas veido dažādu valodu fonēmu sistēmas. Līdz ar to var pieņemt, ka pastāv teorētiski iespējama universāla pasaules valodu fonēmu sistēma, no kuras katras konkrētās valodas lietotāji izvēlas noteiktus tās elementus. Kopu teorijas jēdzienos to varētu noformulēt šādi: ja apvienojam visu pasaules valodu fonēmu kopas,

iegūstam „pārvalodas“ fonēmu kopu; un ar šīm „pārvalodas“ fonēmām mēs šinī pētījumā darbojamies. Starptautiskā fonētikas asociācija (*IPA*) ir izveidojusi universālu zīmju sistēmu – starptautisko fonētisko alfabētu (arī *IPA*), ar kuru iespējams grafiski attēlot jebkuras pasaules valodas fonēmas vai konkrētas to realizācijas. Vienotas zīmju sistēmas un tās piemērošanas principu lietojums paver iespējas salīdzināt dažādu valodu vai izlokšņu fonēmas, ja to sistēmas ir pietiekami aprakstītas valodnieciskajā literatūrā. Ja šāda teorētiska apraksta nav, tad ir iespējams salīdzināt valodu vai izlokšņu skaņu sistēmas pēc audiomateriāla, izrunas pieraksta vārdnīcās vai rakstu pārveidojuma fonemātiskajā transkripcijā. Tā kā mūsu nolūks bija izveidot rīku, kas ļautu salīdzināt valodas vai to apakšsistēmas, šis rīks tika nosaukts par „fonēmu telpu“. Ar to saprotam universālu rīku, kas ir stāvošs pāri konkrētām valodām un sakņojas cilvēka runas aparāta uzbūvē un runas spējā. Pašreizējā fonēmu telpas modelī nav iespējams pozicionēt dažas pasaules valodās retāk lietojamas skaņu grupas (klikšķi, svilpieni, implozīvie slēdzeņi u. c.), jo tādas skaņas nav vērojamas valodās, ar kurām latviešu valodai ir radniecība vai ciešāki sakari. Nepieciešamības gadījumā šo trūkumu var novērst, papildinot izveidoto fonēmu telpu ar attiecīgām dimensijām. Lai arī šajā nodaļā tiek runāts par fonēmu telpu, izstrādātā metode tikpat labi ir izmantojama arī dažādu valodu vai izlokšņu skaņu salīdzināšanai. Ja par salīdzināšanas pamatu tiek izmantoti precīzi indivīdu izrunas apraksti, tad mūsu piedāvātajā „fonēmu“ telpā var salīdzināt arī dažādu indivīdu izrunas faktus.

Lai izveidotu fonēmu telpu, ir jāformulē koordināšu sistēma (tās dimensijas un šo dimensiju nulles punkti), kurā aprēķināmie Eiklīda attālumi atbilstu izrunas un uztveres faktiem vai šo procesu subjektīvai uztverei. Tā kā pēc akustiskās fonētikas teorijas katras skaņas kvalitāti visvairāk nosaka sašaurinājuma apjoms un novietojums rezonatorā, sākotnēji par galvenajām asīm tika izraudzītas artikulārā atvēruma un artikulācijas vietas ass. Jau 2006. gadā J. Grigorjevs savā referātā „Alternatīvs latviešu valodas skaņu klasifikācijas modelis“ akadēmiķa Jāņa Endzelīna 133. dzimšanas dienas atceres starptautiskajā zinātniskajā konferencē „Valodas struktūra un valodas vienību funkcijas“ izmantoja tieši šīs dimensijas, lai vienotā sistēmā skatītu gan patskaņus, gan līdzskaņus.

Ideja par šādu sistēmu autoram radās pēc iepazīšanās ar M. Bolla rakstu „Teaching Vowels in Practical Phonetics: The Auditory or Articulatory Route?“, kurā patskaņu sistēmas modelis tika veidots atkarībā no mēles augstākā punkta novietojuma attiecībā pret aukslēju loku un rīkles pakaļējo sienu.

Tā kā patskaņu kvalitāti nosaka nevis mēles augstākā punkta novietojums, bet gan rezonatora sašaurinājuma vieta un apjoms, Grigorjevs izveidoja alternatīvu patskaņu klasifikācijas modeli. Šāds patskaņu sistēmas modelis atbilst ne tikai to artikulārajām pazīmēm, bet ir saistāms arī ar R. Piotrovska izstrādātajiem, uztverē balstītajiem patskaņu tonalitātes un bemolitātes indeksiem.

Nedaudz vēlāk J. Grigorjevs izveidoja tabulu, kurā pēc artikulārā atvēruma apjoma un novietojuma rezonatorā var loģiski izkārtot gan patskaņus, gan līdzskaņus. Sākotnējās fonēmu telpas dimensijas tika veidotas, pamatojoties uz tanī atspoguļotajiem principiem. Par artikulārā atvēruma ass nulles punktu tika izraudzīta vērtība „slēdzenis“, jo slēguma gadījumā ceļš gaisa plūsmai tiek bloķēts, un atvērums ir vienāds ar 0. Palielinoties atvērumam, vērtības uz šīs ass tika definētas ar noteiktu kvantitatīvu soli un nosauktas par „spraudzenis berzenis“, „vairāk atvērts spraudzenis“, „šaurs patskanis“ un „atvērtāks patskanis“.

Tomēr, izkārtojot fonēmas šādā veidā, vairākos gadījumos attālumi starp tām fonēmām, kuras tradicionāli pieņemts uzskatīt par tuvām, bija lielāki par attālumiem starp tām, kuras pieņemts uzskatīt par tālām (piem., fonēmai /n/ tuvāka izrādījās /l/, nevis /ņ/), kā arī attālumi, kuri intuitīvi šķita vienādi, būtiski atšķīrās (piem., attālums starp /n/ un /ņ/ sanāca pusotru reizi lielāks, nekā starp /l/ un /ļ/). Šāda neatbilstība norādīja uz to, ka divas

izmantotās dimensijas fonēmu atšķirības raksturo nepietiekami un acīmredzot dimensiju skaits ir jāpalielina. Pēc rūpīgas izpētes sapratām, ka jāveido trīsdimensiju telpa, agrāko artikulācijas vietas asi sadalot divās: „ne mēles muguras“ vietas asī (neitrālas-alveolāras-dentālas-labiālas) un „mēles muguras“ jeb „mīkstuma“ asī (faringālas-uvulāras-velāras-palatalizētas-palatālas). Tas saistīts ar to, ka līdzskaņiem, ko neizrunā ar mēles muguru, iespējama papildartikulācija ar mēles muguru, piešķirot tiem „gaišāku“ vai „tumšāku“ skanējumu. Tātad – mēles mugura tiek izmantota patskaņu izrunai un daļas līdzskaņu izrunai – gan spraugas vai slēguma radīšanai, gan arī rezonatora tilpņu formas modifikācijai citu artikulāciju gadījumā. Papildus jāņem vērā arī apstāklis, ka pasaules valodās ir līdzskaņi ar dubultu artikulāciju, kuriem parasti viena artikulācijas vieta saistīta ar mēles muguras veidoto, bet otra – ar mēles gala vai lūpu veidoto, tāpēc jādod iespēja arī šādu līdzskaņu apzīmēšanai izstrādājamajā sistēmā. Pēc šāda pārveidojuma mūsu izstrādātais fonēmu telpas modelis pietuvinājās Stokholmas Universitātē un Karaliskajā tehniskajā augstskolā (Stokholma) izstrādātajam runas artikulācijas modelim APEX, kurš balstīts uz ilgiem un rūpīgiem skaņu izrunas un skaņu akustisko parametru pētījumiem. APEX modelī vieni no galvenajiem skaņu artikulāciju nosakošajiem parametriem ir mēles ķermeņa un mēles gala parametri.

APEX autori par mēles ķermeņa sākumstāvokli ir izvēlējušies tādu, kas atbilst velārai artikulācijai, no kuras mēles ķermenis var tikt virzīts uz priekšu līdz palatālai artikulācijai, vai atpakaļ un uz leju līdz faringālai artikulācijai. Tā kā mēles ķermeņa parametri APEX modelī nosaka mēles muguras veidoto artikulācijas vietu pie pasīvā runas orgāna, tad izvēlējāmies velāro mēles muguras vietu par sākumstāvokli „mēles muguras vietas“ jeb „mīkstuma“ dimensijā, apzīmējot to ar neitrālo 0 vērtību. Virzot mēles muguru uz cieto aukslēju pusi, pakāpeniski pieaug ar dzirdi uztveramais skaņas „gaišums“ un mīkstums, tāpēc pozitīvā skaitliskā vērtība palielinās virzībā no „velāra“ (0) uz „palatalizēta“ (1) un „palatāla“ (2). Virzot mēles muguru uz rīkles dobuma jeb faringa lejas daļas pusi, skaņa kļūst „tumšāka“, dobjāka, cietāka, tāpēc šajā virzienā negatīvā vērtība pieaug no „velāra“ (0) uz „uvulāra“ (-1) un „faringāla“ (-2). Dimensijā „ne mēles muguras vieta“ apvienojām artikulācijas vietas, ko APEX modelī nosaka mēles gala un lūpu parametri. Par sākumstāvokli pieņēmām atbrīvotu mēles gala stāvokli, kad mēles gals aktīvi nepiedalās skaņas izrunā, nosaucot šo artikulācijas vietu par „neitrāla“ un piešķirot tai vērtību 0. Pārvietojoties artikulācijas vietai uz priekšu no maksimāli priekšējas (palatālas) mēles muguras artikulācijas, šīs dimensijas vērtība pakāpeniski pieaug virzienā no „neitrāla“ (0) uz „alveolāra“ (1), „dentāla“ (2) un „labiāla“ (3). Lai šādā sistēmā apzīmētu laminmāli vai apikāli dentālas, interdentālas un labiodentālas skaņas, tām būtu jāpiešķir atbilstoši vērtības 1,25, 1,5, 2 un 2,5. Ja mūsu izstrādātajā fonēmu telpā kā atsevišķas fonēmas būtu jāiekļauj retrofleksie līdzskaņi, tad „ne mēles muguras vietas“ dimensija būtu jāpapildina ar stāvokli „retrofleksa“, piešķirot tai vērtību -1, jo šo līdzskaņu artikulācija notiek ar uz augšu un atpakaļ atlocīta mēles gala apakšējo malu.

Bez tam, lai izveidotu tādu telpu, kas pilnībā raksturotu visas latviešu izloksnēs sastopamās fonēmas, respektīvi, nebūtu divu dažādu fonēmu, kuru koordinātes telpā sakristu, tika izveidotas vēl piecas dimensijas: lūpiskuma (neitrāla-labializēta), trīcīguma (neitrāla-vibrants) un nebalsīguma (balsīga-nebalsīga), šķērseniskuma (mediāla-laterāla) un deguniskuma (neitrāla-nazāla), tādējādi iegūstot astoņu dimensiju telpu. Lūpiskuma dimensija ļauj nošķirt skaņas, kuru izrunā lūpas nepiedalās, no skaņām, kuras tiek izrunātas ar lūpu stiepumu vai noapaļojumu. Tā kā vairumā gadījumu lūpas skaņas izrunā ieņem pasīvu lomu, šāda izruna tika klasificēta kā „neitrāla“, piešķirot tai vērtību 0, bet „labializētai“ izrunai ar aktīvi noapaļotām vai izstieptām lūpām – vērtību 1. Trīcīguma dimensija tika ieviesta, lai no pārējiem līdzskaņiem varētu nošķirt citādi līdzīgos vibrantus. Ja līdzskanis tiek izrunāts bez runas orgāna vibrācijas, tā artikulācija tiek klasificēta kā

„neitrāla“ un tai piešķirta vērtība 0, bet, ja līdzskaņa izrunai nepieciešamas kāda runas orgāna vibrācijas, tas tiek klasificēts kā „vibrants“, piešķirot tam vērtību 1. Skaņu balsīgumu raksturoja dimensija „(ne)balsīgums“, kurā par neitrālām skaitliskā vairākuma dēļ tika atzītas balsīgas skaņas, piešķirot tām vērtību 0, bet nebalsīgām skaņām – vērtību 2, lai atstātu iespēju arī balsīguma starppakāpēm. Šķērseniskuma dimensija nepieciešama laterālo līdzskaņu nošķiršanai. Ja gaisa plūsma līdzskaņa artikulācijas laikā virzās runas orgānu veidotā rezonatora garenvirzienā, tad artikulācija tiek klasificēta kā „neitrāla“, piešķirot tai vērtību 0. Ja līdzskaņa artikulācijas rezultātā gaisa plūsmas ceļā ir radīts šķērslis, kuru apejot gaisa plūsma tiek virzīta uz sāniem, artikulācija tiek klasificēta kā „laterāla“ un tai piešķirta vērtība 1.

Deguniskuma dimensija ļauj nošķirt orālās skaņas no nazālajām. Ja skaņas artikulācija notiek caur muti, bloķējot gaisa plūsmai ceļu caur deguna dobumu, kā tas notiek lielākās skaņu daļas artikulācijas gadījumā, artikulācija tiek klasificēta kā „neitrāla“, piešķirot tai vērtību 0. Ja skaņas izrunas laikā gaisa plūsmai ceļš caur deguna dobumu ir atvērts, tad artikulācija tiek klasificēta kā nazāla, piešķirot tai vērtību 1. Nāseņu jeb nazālo līdzskaņu klasifikācija rosināja pārdomas par tās neatbilstību tradicionāli pieņemtajai, kad tos sauc par slēdzeņiem. Asociatīvi „slēdzenis“ saistās ar gaisa plūsmas pilnīgu pārtraukumu kaut vai uz īsu brīdi. Protams, nazālo līdzskaņu izrunas laikā gaisa plūsmai ceļš caur muti tiek bloķēts, taču gaisa plūsma un skaņa samērā brīvi izplūst caur deguna dobumu, iegūstot nazālajām skaņām raksturīgās papildrezonanses. Tieši šis brīvais gaisa plūdums un skanīgums ir pamats nazālo līdzskaņu ierindošanai skaneņu grupā. Ņemot vērā šīs nāseņu īpašības, autori atļāvās tos klasificēt pēc atvēruma nevis kā „slēdzenis“, bet gan kā „vairāk atvērts spraudzenis“.

Vadoties no latviešu izloksnēs satopamo fonēmu fonētiskajām īpašībām, tām piešķīrām nosacītas koordinātes, kuras, neraugoties uz savu nosacītību, ataino fonēmu savstarpējo tuvību pietiekami labi, respektīvi, atbilst tradicionālajiem, intuitīvajiem priekšstatiem. Papildus fonēmām šādas koordinātes tika piešķirtas arī izloksnēs sastopamo fonēmu variantiem jeb alofoniem.

Lai būtu iespējams aprēķināt vērtību ne tikai skaņas aizvietošanai ar kādu citu, bet arī tās dzēšanai vai pievienošanai, ir jādefinē arī t. s. neitrālais punkts jeb ε, šinī gadījumā to varētu saukt par tukšās skaņas koordinātēm. Izskatījām trīs iespējamos risinājumus: koordināšu sistēmas sākumpunktu (tam atbilstu līdzskanis /g/), izmantoto intervālu viduspunktu (šādai parametru kombinācijai atbilstu mistiska, neizrunājama skaņa), kā arī fizioloģiski pamatotu, īpaši definētu punktu. Vispamatotākais šķita trešais ceļš, tāpēc izšķīrāmies par labu tam. Tomēr arī fizioloģiski definēta punkta izvēle nebija tik viennozīmīga: piemēram, punkta, kas atbilst mierīgai elpošanai caur degunu koordinātes būtu (2, 0, 3, 0, 0, 0, 0, 1), bet mierīgai elpošanai caur muti - (2, 0, 0, 0, 4, 0, 0, 0). Salīdzinot šo punktu koordinātes ar fonēmu koordinātēm redzams, ka pirmais ir tuvāks līdzskaņiem, nekā patskaņiem, bet otrais – otrādi. Skaitliskā izteiksmē par to var pārliecināties, sarēķinot attālumus. Tā kā mūsu intuitīvajā uztverē patskaņi ir izrunājami ar mazāku piepūli, piekam līdzskaņi ir grūti izrunājami bez patskaņa pieskaņas, tad pamatotāka šķita esam tāda nulles punkta izvēle, kurš atrodas tuvāk patskaņiem. Tāpēc izšķīrāmies par labu elpošanai caur muti: $\varepsilon = (2, 0, 0, 0, 4, 0, 0, 0)$.

Programmēšanas valodā PERL izstrādājām programmu, kura pēc augstāk definētajām koordinātēm aprēķina Eiklīda attālumu starp visiem fonēmu pāriem un uzskatāmi izvada rezultātus tabulas veidā tīmekļa pārlūka logā. Programmā iespējams interaktīvi mainīt dimensiju svara koeficientus, respektīvi – mazināt vai palielināt dimensiju ietekmi uz attālumu. Eksperimentējot nonācām pie secinājuma, ka labāk intuitīvajam priekšstatam par attālumiem starp fonēmām atbilst tabula, kurai mīkstuma

dimensijas svara koeficients ir 0,5, pārējām atstājot pēc noklusējuma esošos koeficientus ar vērtību 1.

Rediģēšanas attālumu metode paralēliem vārdu komplektiem

Iepriekš aplūkojām n-grammu metodi, kura pielietojama patvaļīgi izvēlētiem tekstiem. Viens no virzieniem, kurā virzīties, būtu pāriet no patvaļīgiem tekstiem (korpusiem) uz paralēliem tekstiem (korpusiem) – līdz ar to pozitīvu rezultātu iegūšanai nepieciešamo datu apjoms būtiski samazinātos. Protams, šādas paralēlas informācijas savākšana ekspedīcijās vienalga prasītu lielus laika un darba ieguldījumus, tomēr izrādījās, ka latviešu izloksnēm paralēli vārdu komplekti ir savākti, tiesa, tie glabājas burtnīcās un tālab jāciparo. 1954. gadā Latvijas PSR ZA Valodas un literaturas instituta izdotajā „Latviešu valodas dialektoloģijas atlanta materialu vākšanas programā“ ir aprakstīta ne tikai fonētiskā pieraksta sistēmas specifikācija, bet arī dots jautājumu saraksts, atbildes uz kuriem jādabū katrai izloksnei. Šinī sarakstā iraid 670 jautājumi, 103 no kuriem sastādīti fonētisko, 160 – morfoloģisko, 107 – sintaktisko un 300 – leksisko atšķirību atainošanai. Kā lasāms grāmatas paskaidrojumos: „Saskaņā ar dialektoloģijas atlanta uzdevumu – dot sintetisku pārskatu par tādām valodas parādībām, kas diferencējas pa dialektiem un izloksnēm, – programā ir mēģināts ietvert tādas latviešu valodas fonetikas, morfoloģijas, sintakses un leksikas parādības, ko dažādās izloksnēs runā dažādi.“ Tā kā arī mūs valodās pamatā interesē atšķirīgais, nevis kopīgais, tad šādi materiāli pilnībā atbilst mūsu vajadzībām un ir izmantojami mūsu pētījumā.

Nākamais solis bija piemērotas salīdzināšanas metodes izvēle. Izrādījās, ka ar līdzīgiem datiem jau ir strādājuši gan īru, gan holandiešu kolēģi, tāpēc nolēmām izmēģināt viņu ieteikto metodi – definēt attālumu starp valodām kā summāro vai vidējo Levenšteina attālumu starp vienādas nozīmes vārdiem (to fonētiskajā pierakstā) un pēc tam iegūto rezultātu analīzei izmantot vidējā attāluma hierarhisko aglomeratīvo kategorizāciju.

Par Levenšteina attālumu sauc vienas teksta rindas pārrediģēšanas citā cenu, kur darbības ir simbola dzēšana, pievienošana vai aizvietošana, bet katras darbības cena ir vienāda ar 1. T.i., dotajā gadījumā cena sakrīt ar minimāli rediģēšanas darbību skaitu. Piemēram, lai no vārda „maģistrs“ iegūtu vārdu „doktors“, mums nāksies veikt sekojošas darbības: nodzēst „m“ un „a“, aizvietot „ģ“ ar „d“, „i“ ar „o“, „s“ ar „k“, kā arī attiecīgajā vietā ievietot „o“. Tātad Levenšteina attālums starp šiem vārdiem ir vienāds ar 6.

Mēs ievadījām skaitļotājā nelielu datu izlasi – 13 vārdus, kas raksturo fonētiskas, 10 – morfoloģiskas un 8 – leksiskas atšķirības, 13 izloksnēm no dažādām Latvijas vietām. Lai atvieglotu ievadi (tā būtu iespējama no jebkuras konsoles) un apstrādi (būtiska ir visu skaņzīmju apzīmēšana ar vienādu, fiksētu baitu skaitu), dati tika ievadīti fonētiskā pseidorakstībā, kurā katra fonēma tika apzīmēta ar 6 baitiem (1., 2. – pati fonēma, 3. – garums, 4. – intonācija, 5. – uzsvērtība, 6. – zilbiskums.

Programmēšanas valodā PERL mēs uzrakstījām programmu, kura rēķina vidējos Levenšteina attālumus starp izlokšņu pāriem (t.i., vidējos starp doto izlokšņu vārdu pāriem) un pēc tam veic rezultātu kategorizāciju.

Pēc vairākiem izmēģinājumiem mēs konstatējām, ka vidējā attāluma hierarhiskās aglomeratīvās kategorizācijas metode reizēm darbojas neatbilstoši (gadījumos, kad acīmredzama ir divu izlokšņu tuvība – vienai esot grupā, bet otrai – ārpus, pēc vidējā aritmētiskā tuvāka izrādās otra izloksne), un tāpēc saskaņā ar aplūkojamo objektu īpašībām labāk būtu izmantot mazākā attāluma hierarhiskās aglomeratīvās kategorizācijas metodi, proti, kad divas kategorijas tiek apvienotas vienā, attālumu starp jauno, apvienoto kategoriju un pārējām kategorijām mēs nosakām kā mazāko, nevis vidējo aritmētisko, attālumu starp jaunās un citas kategorijas elementiem.

Eksperimenta rezultāti izrādījās gana labi un atbilda priekšstatam par attiecīgo izlokšņu savstarpējo tuvību. Tomēr mēs nolēmām padarīt mūsu eksperimentu sarežģītāku, pārejot no Levenšteina attāluma uz Vāgnera-Fišera attālumu, proti, konstantas fonēmas aizvietošanas cenas vietā ieviešot cenu, kura atkarīga no attiecīgo fonēmu tuvības. (Piemēram, šaurā *e* aizvietošanas ar plato cenai acīmredzot jābūt mazākai, nekā aizvietošanai ar *o*, bet tai, savukārt – mazākai, nekā aizvietošanai ar *b*.)

Tā mēs nonācām līdz problēmai, par kuras risinājumu – fonēmu telpas izveidi – jau pastāstījām iepriekšējā nodaļā. Šī eksperimenta veikšanas laikā gan fonēmu telpas izstrāde bija tikai procesā, tāpēc to veicām ar tās „darba variantu“ – nevis astoņdimensiju, bet sešdimensiju (balsīgums, mīkstums, nemuguras vieta, lūpstiepums, atvērums un trīcīgums).

Arī nulles punktu mēs pagaidām izvēlējāmies nevis fonoloģiski pamatotu, bet tīri matemātiski – kā mūsu izmantoto ašu intervālu viduspunktu.

Tāpat mēs papildinājām katras skaņzīmes aprakstu ar vēl vairākiem parametriem jeb dimensijām, kas raksturoja ne tik daudz pašu fonēmu, cik tās pielietojumu: skaņas garumu, intonāciju (uzdevām bināru attālumu, jo intonācijas attiecīgajās valodās ir piecas, tomēr nav skaidrs, kuras – tuvākās, kuras – tālākas, un cik lielā mērā), uzsvērtību un zilbiskumu.

Skriptu Vāgnera-Fišera attāluma rēķināšanai veidojām uz Levenšteina attāluma skripta bāzes.

Vāgnera-Fišera attāluma izmantošana nenesa būtiskas izmaiņas, salīdzinot ar Levenšteina attālumu. Tomēr ir šādas tādas nianses, kuras intuitīvi šķiet pamatotas esam. Tā, piemēram, ar Vāgnera-Fišera attālumu divas dažādu guberņu sēliskās izloksnes (Līvānu – Latgalē un Rites – Kurzemē) veido pāri gan pēc fonētiskajiem, gan pēc morfoloģiskajiem jautājumiem, bet pēc leksiskajiem Līvānu izloksne izrādās nedaudz tuvāka citām Latgales izloksnēm (kas izskatās diezgan loģiski, jo ir leksikas slānis, piemēram, katoliskais, kurš raksturīgs tieši Latgalei). Levenšteina attāluma gadījumā šī smalkā nianse izpaliek, sēliskās izloksnes nonāk pārī visos trijos gadījumos.

Pirmajā brīdī negaidīta varētu šķist Atašienes un Ziemera morfoloģiskā tuvība, kura parādās tikai pēc Vāgnera-Fišera attāluma, jo Atašiene atrodas Latgalē un tanī runā dziļā sēliskā izloksnē, bet Ziemeris atrodas Vidzemē un tanī runā dziļi latgaliskā izloksnē, pie tam arī ģeogrāfiski šīs vietas ir ļoti tālu viena no otras. Tomēr, ja padomājam, tad varam atrast izskaidrojumu: abas izloksnes ilgstoši atrodas lielā viduslatviešu izlokšņu iespaidā, un tieši morfoloģija no tā ir ietekmējusies visvairāk.

Ir arī tik spilgti jautājumu grupu nošķiršanās piemēri, kuri parādās gan pēc Vāgnera-Fišera, gan Levenšteina attāluma, piemēram, Ģeros runā Vidzemes lībiskajā izloksnē, tāpēc nepārsteidz rezultāts, ka fonētiski tā izrādās tuvāka Kurzemes lībiskajām izloksnēm, bet leksiski – saviem kaimiņiem, Kocēnu Vidzemes vidus izloksnei.

Protams, jau pats morfoloģisko un leksisko jautājumu kategorizēšanas rezultātu sakrišanas fakts liecina par mazāku ar Levenšteina attālumu iegūto rezultātu niansētību. Tāpēc varam secināt, ka šāda veida mērījumos ir vērts izmantot Vāgnera-Fišera attālumu gadījumos, kad mūs interesē valodas nianses, bet Levenšteina attāluma izmantošana attaisnojas tad, kad vajadzīga rupja, bet ātrdarbīga kategorizācija.

Šo metodi publicējām 2006. gadā. Daudzus gadus vēlāk, 2014. gadā, Helsinku universitātes līdzstrādnieki publicēja līdzīga virziena rakstu. Tanī par ieejas datiem tiek izmantoti izcilā krievu valodnieka Sergeja Starostina izveidotās daudzvalodu etimoloģiskās datubāzes *Starling* dati, šinī gadījumā faktiski tiek izgūtas paralēlas daudzvalodu vārdnīcas fonētiskajā pierakstā, tātad tas pats, kas mūsu eksperimentā. Novērtēšanai tiek izmantots normalizētais saspiešanas attālums, kas ir jauns un interesants risinājums un turpmāk būtu izmēģināms arī mūsu darbos.

Fonēmu atpazinēju metode nedalītām fonogrammām

Līdz šim mēs strādājām nodalīti – vai nu ar fonētisko transkripciju, vai pa taisno ar runas ierakstiem. Šinī nodaļā aplūkosim metodi, kura šos pēc savas būtības tik saistītos datu veidus sasaista arī praksē.

Spontānas runas ierakstiem pielietojamu fonēmu atpazinēju parādīšanās un attīstība uzvedināja mūs uz domu par kombinētu metodi – sākumā ar fonēmu atpazinēju no fonogrammām tiek radītas attiecīgas transkripcijas, bet pēc tam tām tiek pielietotas tuvības pakāpes noteikšanas metodes, kuras jau pārbaudītas eksperimentos ar manuāli veidotām fonētiskajām transkripcijām.

Pirmajā acu uzmetienā varētu šķist, ka uzstādījums ir nepareizs pēc būtības, jo pagaidām nav zināms par universāla (t.i., valodneatkarīga) fonēmu atpazinēja eksistenci – visi mums zināmie atpazinēji ir apmācīti (resp., priekš tiem radīti modeļi) uz konkrētām valodām un, strikti veroties, priekš šīm valodām arī ir paredzēti. Bet sarunās ar kolēģiem, kuri piedalījās atpazinēja PhnRec izstrādē, uzzinājām, ka intereses pēc viņi ir mēģinājuši atpazīt arī citu valodu skaņu ierakstus, un rezultāti izrādījušies lietojami, kaut, protams, sliktāki, nekā modeļa valodai. Par zināmu apstiprinājumu tam kalpoja arī izstrādātāju publicētie rezultāti par fonēmu atpazinēja pielietošanu valodu identificēšanā, jo tas arī, tāpat kā mūsu gadījumā, ir fonotaktikas metodes izmantojošs risinājums (kaut arī viņi apmācībā izmanto visu potenciāli identificējamo valodu datus). Tomēr arī iespēja ar vienvalodīga fonēmu atpazinēja izmantošanu ir aprakstīta un tai pat atrasta interpretācija: tā modelē situāciju, kad vienvalodīgs cilvēks dzird dažādas svešvalodas.

Tāpēc mēs atļāvāmies izvirzīt hipotēzi, ka izmantojot fonēmu atpazinēju, kura modelis apmācīts kādai valodai, pārim citu valodu, kļūdas var būtiski neietekmēt fonotaktiskos raksturlielumus, kurus mēs izmantojam tuvības pakāpes noteikšanai.

Mūsu rīcībā bija mūsu pašu savāktie piecu Latvijas izlokšņu spontānas runas ieraksti. Visi ieraksti tika vākti saskaņā ar mūsu uzdotajiem automatizētai fonogrammu analīzei paredzētas informācijas vākšanas principiem, proti, visi ieraksti bija vienveidīgi, ierakstīti ar viena veida aparatūru (tika izmantots dinamisks vienvirziena mikrofons, kurš bija fiksēts pie informanta galvas) samazinātā ārēju trokšņu ietekmē. Visi ieraksti tika manuāli iztīrīti (informācijas, ne skaņas kvalitātes nozīmē): tika izgrieztas visas citas skaņas un balsis, atstājot tikai informanta tiešo runu. Ieraksta kvalitāte 44,1 kHz / 16 biti. Atkarībā no konkrētā fonēmu atpazinēja modeļa prasībām (šinī gadījumā – apmācības ierakstu tehniskās kvalitātes), fonogrammām tika izpildīta diskretizācijas frekvences pazemināšana vai nu līdz 8 kHz, vai 16 kHz.

Fonēmu atpazinējs *PhnRec* izstrādāts Brinnes Tehniskās universitātes Informācijas tehnoloģiju fakultātes Runas apstrādes grupā. Tā galvenā īpatnība ir ilga laika konteksta (līdz vairākiem simtiem milisekunžu) izmantošana. Runas raksturlielumu izgūšana balstās laika konteksta sadalīšanā, klasifikatora lomu pilda mākslīgie neironu tīkli, bet rindu dekodēšana tiek veikta ar Viterbi algoritma palīdzību. Programmas pakotnē iekļauti jau apmācīti čehu, angļu, krievu un ungāru valodas modeļi. Lielāku salīdzināšanas iespēju labad nospriedām izmēģināt tos visus.

Programmu pakotnes *Sphinx 4* sastāvā ir iekļauta programma *pocketsphinx,* kura darbojas tai skaitā arī kā fonēmu atpazinējs. Diemžēl, mums neizdevās atrast publikāciju, kura aprakstītu Sphinx 4 izmantoto fonēmu atpazīšanas algoritmu, bet pašiem pētīt programmas kodu nebija laika. No kopējās dokumentācijas izdevās saprast, ka programma valodas modeļa apmācības laikā rada slēpto Markova modeli katrai fonēmai (resp., liels laika konteksts netiek ņemts vērā), bet dekodēšana tiek veikta ne tikai ar „klasiskā“ Viterbi algoritma, bet arī ar Buš-derbi algoritma palīdzību.

Pēc noklusējuma pakotnē iekļauts tikai angļu valodas modelis. Ir pieejami ārēju izstrādātāju radīti citu valodu modeļi, tomēr, lai izslēgtu kļūdas, kas varētu rasties nepietiekami kvalitatīvas modeļu izstrādes rezultātā, mēs nolēmām izmantot tikai tādus modeļus, ko izstrādājuši paša atpazinēja veidotāji.

Pirmīt mēs aprakstījām n-grammu metodi un tās pielietojumu fonētiskajām transkripcijām, kuras veidotas, manuāli atšifrējot skaņas ierakstus. Tā kā rezultāti bija labi, tad šo metodi varam pielietot arī mūsu aturomātiski radītajām transkripcijām. Atgādināsim, ka pierādījām, ka metodes ietvaros definētais attālums ir metrika, tātad, faktiski arī šīs kombinētās metodes palīdzību iegūtais attālums būs metrika.

Eksperiments tika veikts ar skriptu palīdzību, ko uzrakstījām programmēšanas valodā PERL. Sākumā mēs pielietojām visus mūsu rīcībā esošos atpazinējus un modeļus (proti, kopā 5: *PhnRec* priekš čehu, krievu, angļu un ungāru, kā arī *Sphinx* priekš angļu valodas) visām mūsu rīcībā esošajām attiecīgo izlokšņu fonogrammām. Pēc tam mēs iegūtās fonēmu datnes nokonvertējām divbaitīgā fonotekstā, kā arī izveidojām katrai izloksnei savu datni, tanī apvienojot visus attiecīgās izloksnes informantus. Un pēc tam ar mūsu modificētās un izlabotās PERL-programmas *TextCat* palīdzību, izrēķinājām attālumus starp jauniegūtajiem fonotekstiem. Rezultāti bija samērā labi: izloksnes, kas tuvākas intuitīvā izpratnē, ir tuvākas arī pēc mūsu attālumiem.

Lielākas uzskatāmības labad nolēmām sakategorizēt izloksnes, izmantojot mazākā attāluma hierarhiskās aglomeratīvās kategorizācijas metodi. Rezultātā mēs ieguvām divu veidu kokus. Abi atsevišķā zarā izdalīja Ziemeļlatgales izloksnes (kas patiešām bija sagaidāms, ņemot vērā to lielo līdzību). Atsevišķi tika izdalīta arī vienīgā Kurzemes – tāmnieku – izloksne (te nu noteikti savādāk nedrīkstēja būt). Atšķirība parādījās ar viduslatgales izloksnēm: pirmais koks (pēc krievu un čehu modeļu rezultātiem) tās izdalīja kopīgā zarā, bet otrais – nē. No intuitīva un lingvoanalītiska redzes punkta pirmais variants izskatās pareizāks, kaut arī otrajā ir saskatāma zināma jēga. Tā kā krievu un čehu ir slāvu valodas, bet pēc ģeneoloģiskās valodu klasifikācijas vistuvākās baltu valodām ir tieši slāvu valodas, tad gribi-negribi uzprasās secinājums par kvalitatīvākiem rezultātiem radniecīgāku modeļa apmācības valodu gadījumā.

Pirmkārt, mēs pierādījām, ka mūsu pamathipotēze apstiprinās: fonēmu atpazinēji ir pielietojami runas fonokorpusiem ar automātiskas valodu tuvības pakāpes novērtēšanas nolūkā.

Otrkārt, mēs pārliecinājāmies, ka modeļa apmācības valoda ietekmē rezultātus, tomēr ne tādā mērā, lai pielietošanu izslēgtu. Proti, principā mēs varam izmantot jebkādu atpazinēju ar modeli jebkādai valodai, bet priekš eksperimenta tīrības būtu vēlams, lai radniecības pakāpe starp modeļa apmācības valodu un dažādām eksperimenta valodām nebūtu pārlieku atšķirīga.

Treškārt, fonēmu atpazinēja arķitektūra uz rezultātiem būtisku iespaidu neatstāj. (Kaut paši atpazīšanas rezultāti *PhnRec* gadījumā bija „lasāmāki“.)

Un, ceturtkārt, mēs faktiski esam uzdevuši metriku valodu telpā, pie tam – aprēķināmu automātiski, bez būtiskas datu priekšsagatavošanas, t.i., bez manuāla jeb t.s. „melnā“ darba, kas izskatās ļoti perspektīvi.

Slēpto Markova modeļu metode nedalītām fonogrammām

Par SMM objektu izvēlējāmies nedalītu runātāja fonogrammu, un būvējām modeļus uz pietiekama skaita attiecīgās valodas informantu ierakstiem.

Skaidrs, ka šāda metode ir pielietojama jebkurai ļaužu valodai, tai skaitā arī tādai, kurai nav rakstu formas. Tomēr, tā kā latviešu – gan baltiešu, gan latgaliešu, izloksnes mums bija pieejamākas, tad sākotnēji eksperimentu veicām ar tām.

2008. gada rudenī mēs izbraucām vairākās ekspedīcijās Latgalē un Kurzemē. To rezultātā tika savākts 4 latgaliešu un 1 kurzemnieku izloksnes materiāls: 30 informanti, kuri runāja Viļakas (Vileks), 23 – Baltinavas (Baļtinovys), 29 – Rudzātu, 14 – Aulejas (Aulejis) un 17 – Dundagas (Dundags) izloksnē. Visi informanti stāstīja savu dzīves gājumu: vecākiem, vecvecākiem, brāļiem, māsām, ģimeni, skolu, darbu, kāzām, bērniem, saimniecību u.tml. Daļu savāktā materiāla izmantojām mūsu eksperimenta veikšanai.

Faktiski, iedomātās metodes pārbaudīšanai tika veikti vairāki eksperimenti. Tie visi tika realizēti, izmantojot programmu pakotni HTK, t.i., nebija nepieciešamības pašiem programmēt attiecīgos algoritmus un pat pētīt to realizācija dotajā pakotnē, jo tā ir pārbaudīta un atzīta pasaules runas pētnieku vidū. Eksperimentu būtība: vairāku slēpto Markova modeļu radīšana uz atbilstošiem fonogrammu komplektiem un mēģinājumi atpazīt citu, apmācībā neizmantotu fonogrammu veidu. Protams, datu apstrādes un darbību atumatizēšanas nolūkos tika izstrādāti attiecīgi skripti.

Pirmo eksperimentu veicām ar tā paša cilvēka lasītu vienāda satura tekstu trijās valodās – baltiešu (jeb latviešu), latgaliešu un krievu. Katrā valodā tika ielasītas četras fonogrammas: trīs – vidējā tempā un viena – paātrinātā; fonogrammu garums – no 1 līdz 2 minūtēm. Katrai valodai uz vidējā tempa fonogrammām tika izveidoti slēptie Markova modeļi. Tālāk ar utilītas HVite (tā ir Viterbi algoritma realizācija pakotnē HTK) palīdzību tika meklēts tuvākais modelis priekš ierakstiem paātrinātā tempā. Pie neliela Gausa maisījuma komponenšu (tā saukto „modu“) skaita rezultāti bija neapmierinoši, bet sākot ar četrām un augstāk darbojās nevainojami – ātrā tempa ieraksta valoda tika noteikta nekļūdīgi.

Šī eksperimenta pozitīvie rezultāti mudināja mūs veikt otru eksperimentu, šoreiz jau uz īstiem, mūs interesējošiem datiem.

No mūsu savāktajām izloksnēm mēs izvēlējāmies divas – Rudzātu un Viļakas (Vileks), proti, abas latgaliešu, bet no pretējiem Latgales galiem: Ziemeļaustrumiem un Dienvidrietumiem. Tādējādi aplūkojamās valodas bija ļoti tuvas (kas, protams, pastiprina rezultāta nozīmi pozitīva iznākuma gadījumā), bet tai pat laikā varējām rēķināties ar to, ka atšķirības nebūs izsmērējušās runātāju sociālo kontaku ceļā. No katras izloksnes mēs nejaušā kārtā izvēlējāmies astoņas sieviešu dzimuma informantes, kuras tāpat nejaušā veidā sadalījām divās apakškopās: pa piecām – modeļu izveidei, bet pa trim – pārbaudei. Rezultāti izrādījās tādi pat, kā iepriekšējā eksperimentā: neliela modu skaita gadījumā valoda tika noteikta kļūdaini (pie tam dažādos veidos, bez izsekojamām konsekvencēm), bet sākot ar četrām – nekļūdīgi.

Tādējādi varam secināt, ka mūsu hipotēze par iespēju apmācīt Markova modeļus uz nedalītām fonogrammām, lai raksturotu valodu kā tādu, ir izrādījusies pareiza. Ja reiz tā darbojas atpazīšanas uzdevumos, t.i., pēc šādiem modeļiem tiek pareizi noteikta citu fonogrammu valoda, tad tai vajadzētu darboties arī attāluma starp valodām raksturošanas uzdevumos, respektīvi, attālums starp attiecīgajiem slēptajiem Markova modeļiem raksturos valodu tuvību vai tālību.

Tāpēc nospriedām izveidot SMM visām mūsu rīcībā esošajām izloksnēm un definēt dažāda veida metrikas to telpā.

Sākotnēji nolēmām izmēģināt laimi ar pazīstamākās metrikas – Eiklīda metrikas palīdzību. Vēl atlika izvēlēties, tieši kuri sadalījumu raksturojošie dati būs metrikas telpas dimensijas. Likumsakarīga šķita sadalījuma vidējās vērtības vektoru izmantošana.

Vispirms veicām attālumu aprēķinus augšminētajai latviešu/latgaliešu/krievu ielasītajai runai. Aprēķinājām Eiklīda metriku, normētu Eiklīda metriku (normētu pēc pirmā, otrā un abiem argumentiem) un Žordāna metriku.

Korektu attālumu gadījumā būtu sagaidāms, ka attālumi starp runas paraugiem vienas valodas ietvaros ir mazāki, latviešu un latgaliešu valodām – vidēji, krievu un latgaliešu – lielāki, bet krievu un latviešu – paši lielākie. Tomēr visām metrikām, redzams, ka attālumi ir ļoti līdzīgi, pie tam tie „lēkā", intuitīvi tuvākām valodām bieži vien ir lielāki nekā tālākām un otrādi.

Intereses pēc veicām šo eksperimentu arī mūsu savāktajiem izlokšņu runas paraugiem.

Diemžēl HTK pakotnē ietilpstošajai programmai HERest, kura veic SMM komplekta parametru pārrēķinu, izmantojot Bauma-Velša algoritmu, acīmredzot, piemīt kāda kļūme – pie lielāka skaita ievades datņu tā vienubrīd izvada kļūdas paziņojumu, ka tuvinājumu nav iespējams aprēķināt: *WARNING [-7324] StepBack: File [ceļš līdz datnei] - bad data or over pruning*. Šādai problēmai vajadzētu rasties, ja skaņas ieraksts ir tehniski nekvalitatīvs vai tam piemīt kāda cita vaina. Tomēr interesanti, ka pie lielāka datņu kopskaita šī kļūda parādās tādām datnēm, kurām pie mazāka kopskaita neparādās – tātad nav atkarīga no datnes ieraksta kvalitātes, bet no kaut kā cita. Tas ļauj secināt, ka tā ir programmas vaina, un vienīgais veids, kā no tās izvairīties, ir to apiet tai pielāgojoties. Tā kā vienkārši atmest daļu datņu nevēlējāmies, tad nospriedām vīru un sievu balsis izdalīt atsevišķās grupās – katrā grupā nu bija mazāk datņu un HERest pārstāja lamāties. Tādējādi eksperiments tapa lielāks un, iespējams, arī interesantāks, tomēr tam ir arī viens trūkums – mēs tā rezultātus nevarēsim tieši salīdzināt ar citu metožu rezultātiem.

Arī šinī gadījumā visas metrikas „dancoja" – tās pašas izloksnes puišu runa izrādījās tālāka nekā citu izlokšņu sievu runa, tālas izloksnes brīžam izrādījās tuvas, bet tuvas – tālas.

Pēc profesora A. Lorenca ieteikuma nolēmām izmēģināt šīs pašas metrikas, bet jau vidējo vērtību dalījumam ar dispersiju, resp., jo vērtības mainīgākas, jo mazāks to svars – tās mazāk ietekmē attāluma vērtību.

Kā redzam, nevienā gadījumā, proti, nevienai datu kopai un nevienai metrikai, šis ievedums nav padarījis rezultātus sakarīgus.

Tāpēc mūsu secinājums ir negatīvs, respektīvi, ka šādā veidā attālumu definēt nevaram un ir jāmeklē citi ceļi, kā to paveikt.

Izplatītākais SMM līdzības novērtējums ir Kuļbaka-Leiblera diverģence, kuru autori pieteikuši savā 1951. gada publikācijā.

Tā ir logaritmiskās starpības starp diviem varbūtību sadalījumiem matemātiskā cerība pēc pirmā sadalījuma. Līdz ar to, likumsakarīgi, ka tā nav simetriska, tātad neatbilst vienai no metrikas aksiomām un nav metrika. Šo problēmu bieži vien risina, par metriku definējot vidējo aritmētisko no diverģences vērtībām vienā un otrā virzienā.

Kuļbaka-Leiblera diverģenci aprēķinājām ar nedaudz modificēta Brinnes Tehniskajā universitātē uzrakstīta Pitona skripta palīdzību.

Pirmajā acu uzmetienā zināms sakarīgums rezultātos ir saskatāms (piem., tas, ka Dundaga izskatās patāla no pārējām mēlēm, vai tas, ka Baltinava un Viļaka – savstarpēji vistuvākās), kaut, protams, simetrijas trūkums un sievu un vīru balsu atsevišķums jauc galvu un neļauj rezultātus sakarīgi izanalizēt. Tāpēc nolēmām tos novienkāršot, pirmkārt, tabulu simetrizējot uz vidējajām aritmētiskajām vērtībām un, otrkārt, apvienojot sievu un vīru balsis, arī paņemot vidējo aritmētisko vērtību.

Šāda rīcība visas vērtības stipri satuvināja, kas apliecina, ka lielais vērtību diapazons bija atkarīgs nevis no mēlēm, bet no kādiem citiem apstākļiem. Tas, protams, nav labi. Bet arī šādas, līdzīgas vērtības varētu kaut ko atspoguļot – tāpēc tās papētīsim.

Aulejas attālumi izskatās gaužām labi – Dundaga pati tālākā, Rudzāti – paši tuvākie, Baltinava tuvāka par Viļaku.

Baltinavas rezultāti arī būtu labi (Viļaka ļoti tuva, Rudzāti tālāki), ja ne Dundagas tuvākums par Auleju.

Rudzāti izskatās švakāk – Baltinava tuvāka par Auleju, Dundaga tuvāka par Viļaku.

Savukārt, Viļaka izskatās ļoti labi – Baltinava pati tuvākā, tad Rudzāti, tad Auleja, un Dundaga pati tālākā.

Kopumā ņemot, metode ir izmantojama. Tomēr tā ir tehniski sarežģīta un rezultāti, lai arī lietojami, tomēr nav spīdoši. Tāpēc droši vien reālā pielietošanā jāizvēlas cita metode.

i-vektoru metode nedalītām fonogrammām

i-vektori ir samērā jauns atpazīšanas uzdevumu risinājuma veids, kas nu jau tiek pielietots arī cita veida objektu atpazīšanai, tomēr sākotnēji doma par tiem radās tieši runas atpazīšanas metodes meklējumos.

Pirmā plašāk zināmā publikācija, kurā šī jaunā doma tika pausta (runātāja noteikšanas sakarā), bija 2009. gadā, tanī gan i-vektoru nosaukums vēl neparādās, bet to telpa tiek saukta par pilnīgas mainības pazīmju telpu. 2010. gada sākumā i-vektoru nosaukums parādās kā blakusnosaukums, bet tā paša gada otrajā pusē jau tiek lietots pilnā sparā – jau aprakstot valodu noteikšanas izaicinājumu.

i-vektoru metodes pamatā ir izteikumu Gausa maisījuma modeļu attēlošana ar apslēptu mazdimensiju mainīgo un šī izteikuma attēlojuma izmantošana pazīmju vektora lomā valodu klasifikātorā.

Mēdz būt dažādi i-vektori, atkarībā no tā, kādu lingvostatistisko informāciju tie sevī satur, piemēram, akustisku, prosodisku, fonotaktisku, gan uz kāda veida datiem tie būvēti – nepārtrauktiem vai diskrētiem, gan no tā, kam tie paredzēti – vai runātāja atpazīšanai (SID), vai valodas atpazīšanai (LID), vai kādiem citiem uzdevumiem. Tātad, faktiski, mēs varētu runāt pat par veselu metožu kopumu, tomēr ieslīgšana šādos smalkumos nav mūsu disertācijas mērķis.

Mūsu eksperimentā tika izmantots izlokšņu runas ierakstu kopums, kurš aprakstīts iepriekšējā nodaļā.

Tā kā mums bija pieejami Brinnes Tehniskās universitātes izstrādātie skripti i-vektoru rēķināšanai, tad mēs, protams, izmantojām tos. 2015. gadā BTU Runas grupa nāca klajā ar priekšlikumu veidot kopīgu balss biometrijas standartu – „Voice Biometry Standart“ jeb saīsināti VBS, jo dažādie tobrīd izmantotie tehniskie standarti neļāva veikt operatīvu datu sniegšanu un apmaiņu. Standarta pakotnē arī iekļauti pitona skripti i-vektoru sarēķināšanai (to darbināšanai ievadei nepieciešami tikai runas ieraksti, tomēr labāki rezultāti sasniedzami, ja jau ir ārēji noteikti balss aktivitātes intervāli, t.s., VAD jeb „Voice Activity Detection“, jo iebūvētais aktivitātes noteicējs ir ļoti primitīvs). Skaidrs, ka biometrijas standarts ir paredzēts runātāja identificēšanas uzdevumiem, proti, orientēts uz runātāja, resp., konkrēta cilvēka runas (t.sk. balss) īpatnībām, tātad i-vektori, kurus ģenerē šī pakotne ir tā saucamie SID (jeb „Speaker IDentification“) i-vektori. Mūsu uzdevumam tie teorētiski ir mazāk piemēroti, tomēr nolēmām tos izmēģināt, jo atvērts standarts un skriptu publiska pieejamība ir būtiski faktori tehnoloģijas izvēlē.

Tāpat kā iepriekšējā nodaļā aprakstītajā gadījumā ar SMM, i-vektorus rēķinājām pilnām informantu fonogrammām, tādējādi sagaidot, ka tie raksturos valodu kopumā. Visu izlokšņu pāriem aprēķinājām kosīnus novērtējumu starp iegūtajiem i-vektoriem.

Rezultāti bija gaužām labi. Te varbūt jāpaskaidro, ka kosīnus novērtējums ir kosīnus vērtības, un ar arkkosīnu no tām var iegūt leņķus. Leņķus, savukārt, ir vieglāk iztēloties: iedomājamies plaknē uzzīmētu nulles līniju un uz tās punktu jeb centru; tad no šī centra ejošā stara un nulles stara (nulles līnijas labā stara) veidotais leņķis raksturo attiecīgo attālumu starp divām izloksnēm – jo leņķis mazāks, jo mēles tuvākas.

Tā Baltinavas un Viļakas izloksnes izrādās vistuvākais pāris (47°). Arī starp Dienvid- un Rietumlatgales izloksnēm – Rudzātiem un Auleju – attālums mazāks, nekā starp tām un Ziemeļlatgales izloksnēm. Dundaga ir vistālākā Aulejai, Baltinavai un Viļakai. Vienīgais novērtējums, kurš šķiet esam izteikti nepareizs, ir attālums starp Rudzātu un Dundagas izloksnēm – tam noteikti nebija jābūt mazākam, pie tam tik ļoti, par attālumu starp Rudzātu un pārējām trim latgaliešu izloksnēm.

Intereses pēc nolēmām izmēģināt Eiklīda un citas vektoru metrikas (piem. Žordāna) i-vektoriem. Nospriedām izmēģināt arī pilsētas kvartāla metriku. Lai šāda veida metrikas varētu pielietot, katrai valodai no tās informantu i-vektoriem aprēķinājām vidējo aritmētisko i-vektoru.

Pati sliktākā no šīm metrikām (kaut arī ne galīgi slikta) mūsu gadījumā izrādījās Žordāna: pēc tās gan Viļakai, gan Rudzātiem Auleja izrādās krietni tālāka par Dundagu.

L_1 un Eiklīda (gan nenormēta, gan normēta, jo normēšana lietas būtību nemaina) metrika izskatās vienlīdz labi un pie tam labāk arī par kosīnus novērtējumu: Viļaka ar Baltinavu ir pašas tuvākās, Dundaga visām Latgales izloksnēm – pati tālākā. Vienīgais jautājums, kas rodas, ir: kāpēc Aulejai Baltinava izrādās tuvāka par Rudzātiem? Tā it kā nevajadzētu būt. Tā varētu būt gan metrikas vaina, gan datu nepilnība, gan arī tomēr objektīvs novērtējums, kas ņēmis vērā kādas izlokšņu nianses, kuras teorētiskā salīdzināšanā parasti atstāj bez vērības. Lai uz šo jautājumu atbildētu, nepieciešami papildus eksperimenti ar lielākiem datu apjomiem un lielāku izlokšņu skaitu.

Stažēšanās laikā Brinnes Tehniskajā universitātē tās Runas apstrādes pētnieciskās grupas līdzstrādniekam Oldriham Plotam un iedevām pašu savāktos izlokšņu runas ierakstus, jo viņš tos palūdza saviem eksperimentiem, un pēc kāda laika saņēmām rezultātus.

Pirms eksperimenta veikšanas katras izloksnes dati ar nejauša procesa palīdzību tika sadalīti divās daļās: lielākā apmācības daļā un mazākā pārbaudes daļā. Tad, izmantojot katru daļu atsevišķi, tika aprēķināti i-vektori. Tad uz apmācības daļas i-vektoriem tika apmācīts Gausa lineārais klasifikātors, savukārt uz pārbaudes daļas i-vektoriem tas tika pielietots. Pēc procentuālā sadalījumu, cik liela daļa pārbaudes datu tika iedalīta pareizi un cik liela – atdota citām izloksnēm, rezultāti bija gaužām labi: Dundaga kā visatšķirīgākā tiek atpazīta vislabāk; Rudzāti arī tīri labi, tas ka „atdod“ kādu daļu citām Latgales izloksnēm, ir likumsakarīgi; Baltinava un Viļaka, ņemot vērā to savstarpējo tuvību, arī uzrāda samērā labus rezultātus, pie tam starpības lielāko daļu „atdod“ viena otrai; vienīgais, kas pārsteidz, ir Aulejas salīdzinoši sliktie rezultāti „par labu“ Rudzātiem.

Ņemot vērā labos brinnenieku rezultātus, nolēmām izmēģināt savus SID i-vektoriem veiktos eksperimentus viņu LID i-vektoriem. Pēc mūsu lūguma viņi mums laipnu roku šos i-vektorus iedeva. Sagaidījām, ka rezultāti būs līdzīgi SID i-vektoriem, tomēr kaut nedaudz, bet labāki, jo LID i-vektori taču tiek veidoti, orientējoties līdzīgāka uzdevuma veikšanai.

Mums par lielu pārsteigumu kosīnus novērtējums sanāca pilnīgi bezjēdzīgs. Šobrīd vēl strādājam pie cēloņu noteikšanas.

Tomēr nolēmām izmēģināt arī pārējos SID i-vektoriem izmantotos attālumus. Eiklīda un L_1 metrikai rezultāti bija līdzīgi, kā SID gadījumā. Tomēr interesanti, ka Žordāna metrika, kura SID i-vektoriem nebija tik laba, LID vektoru gadījumā uzvedās krietni labāk – tādu problēmu kā SID gadījumā tai nebija un, varētu teikt, ka tā LID gadījumā bija gandrīz līdzvērtīga Eiklīda un L_1 metrikai.

Esam pārliecinājušies, ka i-vektori pietiekami labi raksturo valodas un tāpēc ir izmantojami valodu atšķirīguma skaitliskai novērtēšanai. Pie tam to izmantošana ir ērtāka, nekā slēpto Markova modeļu izmantošana: gan datu parocības, gan populāru metriku izmantošanas ziņā. Saskaņā ar mūsu eksperimentu rezultātiem, ieteicams izmantot Eiklīda vai L_1 metriku.

Gausa maisījuma modeļu metode nedalītām fonogrammām

Vēl viens runas modelēšanas veids ir Gausa maisījuma modeļu (GMM) izmantošana.

Gausa maisījums ir galīga skaita Gausa jeb normālo sadalījumu kopums, precīzāk – (ar skalāriem svara koeficientiem) nosvērta summa. Šādu modeli apraksta trīs lielumi – matemātiskās cerības vektors, kovariācijas matrica un svaru koeficientu vektors. Tā kā neatkarīgu normālu sadalījumu summa ir normāls sadalījums, tad arī Gausa maisījums ir normāls sadalījums.

GMM tiek plaši izmantoti datu klasificēšanā (piem., ekonomikā, demogrāfijā, ekoloģijā un citur), ja pastāv pamats domāt, ka katra no šīm klasēm atbilst normālam sadalījumam. Līdz ar to GMM tiek izmantoti arī dažādu veidu (gan attēlu, gan skaņas, gan citu objektu) atpazīšanas uzdevumos, jo tie pēc savas būtības ir objekta visvarbūtīgākās piederības kādai no dotajām klasēm noskaidrošana.

Tā kā mūs interesē runas modelēšana, tad strādājām ar runas signālu spektrālās informācijas pazīmju sadalījumiem. Pazīmju vektori tiek veidoti no frekvenču Furjē kosīnus pārveidojuma koeficientiem jeb t.s. MFCC (no angļu *Mel Frequency Cepstral Coefficients*), katrā kadrā izmantojot 14 koeficientus.

Modeļi tiek veidoti ar sagaidāmās vērtības maksimizēšanas algoritma jeb EM palīdzību, bet divējādi – kā pilnībā neatkarīgi, atsevišķi apmācīti modeļi, katrs no kuriem tiek veidots tikai uz vienas attiecīgās izloksnes runas datiem, kā arī kā no kopīga, uz visu izlokšņu datu daļu veidota t.s. universālā fona modeļa jeb UBM ar MAP-pielāgošanu atvasināti modeļi.

Arī šīs nodaļas eksperimentos tika izmantots iepriekš aprakstītais izlokšņu runas ierakstu kopums. Un arī šinī gadījumā izmantojām jau ar citām metodēm pielietoto paņēmienu – apmācīt modeļus uz pilnām informantu fonogrammām. To darījām gan uz pilniem datiem (kuri apjoma ziņā dažādām izloksnēm atšķiras), gan uz vienāda apjoma datiem, izloksnēm, kurām datu apjoms lielāks, visas fonogrammas proporcionāli saīsinot (nogriežot to beigu daļas).

Eksperimenti tika veikti, izmantojot pakotnes *MatLab* sniegtās iespējas. Skriptus izstrādājām paši, par pamatu ņemot Meksikas Nacionālās autonomās universitātes Inženieru fakultātes Balss laboratorijas doktoranta Hosē Benito Trangola sniegto informāciju par viņa eksperimentiem runas atpazīšanas nozarē un tajos izmantotajām bibliotēkām, funkcijām un parametriem.

Kā **pirmo** veicām eksperimentu, veidojot Gausa maisījuma modeļus tieši, bez pielāgošanas, respektīvi, katrs modelis tika veidots tikai no attiecīgās izloksnes fonogrammām un pilnībā neatkarīgi no citu izlokšņu modeļiem.

Šinī un turpmākajos eksperimentos aprēķinājām Eiklīda, L_1 (jeb pilsētas kvartālu), Žordāna (jeb Čebišova) metrikas un Kuļbaka-Leiblera diverģenci visām trim modeļu sastāvdaļām – gan vidējo vērtību vektoriem, gan kovariācijas matricām, gan svaru vektoriem.

Kopumā ņemot, Eiklīda metrikas vērtības atbilst intuitīvam priekšstatam par attiecīgo izlokšņu tuvību: Dundaga ir pati tālākā visām pārējām izloksnēm, Viļaka un Baltinava ir savstarpēji vistuvākās u.t.t. Nedaudz jocīgi ir tas, ka Rudzāti un Auleja ir savstarpēji tālākas nekā gan Rudzāti, gan Auleja ar Viļaku un Baltinavu. Tomēr tam var būt zināms izskaidrojums, jo Aulejas izloksne ir diezgan specifiska, savukārt gan Rudzātu, gan Ziemeļlatgales izloksnes satur atsevišķas līdzīgas formas (piemēram, *tuo*, nevis *tai* u.tml.).

Savukārt Kuļbaka-Leiblera diverģences vērtības izskatās pilnīgi bezjēdzīgas, jo satur negatīvas vērtības. Pilsētas kvartālu metrika šinī gadījumā uzvedas tāpat, kā Eiklīda metrika. Žordāna metrika izskatās gaužām švaki: Dundaga nav pati tālākā visām četrām latgaliskajām izloksnēm, Viļaka un Baltinava nav pašas tuvākās.

Tagad no vidējo vērtību vektoriem pāriesim pie **kovariācijas matricām**. Eiklīda metrika uzvedas tāpat kā vidējo vektoru gadījumā: viss ir labi, tikai nelielus jautājumus izsauc Rudzātu un Aulejas attiecības uz Ziemeļlatgales izlokšņu fona.

Kuļbaka-Leiblera diverģence izskatās daudz maz jēdzīgi, tomēr tai piemīt nelielas neatbilstības, piemēram, Aulejai Viļaka ir tālāka nekā Dundaga. Tā kā otrā virzienā šādas

problēmas nav, tad uzreiz nāk prātā iespēja diverģenci simetrizēt. Tomēr simetrizēšana šo problēmu tikai samazina, nevis pilnībā atrisina.

Pilsētas kvartālu metrika izskatās samērā labi, tomēr nedaudz sliktāk nekā Eiklīda metrika, jo, piemēram, attālums Auleja-Viļaka ir līdzīgāks attālumam Auleja-Dundaga nekā attālumam Auleja-Baltinava. Mums par pārsteigumu kovariācijas matricām laba izrādās arī Žordāna metrika, kaut vidējo vērtību vektoriem tā bija pilnīgi bezjēdzīga.

Eiklīda metrika **svaru vektoriem** arī rada daudzmaz sakarīgu iespaidu, tomēr tā satur arī būtiskas kļūdas: Aulejai Dundaga ir tuvāka par Viļaku un Baltinavu; Viļakai Rudzāti ir tuvāki par Baltinavu. Kuļbaka-Leiblera diverģencei uzrādās tie paši trūkumi, kas Eiklīda metrikai, pie tam ir skaidrs, ka simetrizēšana līdzēt nevar, jo tie ir abos virzienos.

Arī pilsētas kvartālu metrika uzrāda tās pašas neatbilstības, kādas redzējām Eiklīda metrikas un Kuļbaka-Leiblera diverģences gadījumā. Žordāna metrikas rezultāti atšķiras: tā neuzrāda pirmo problēmu, tomēr otrā diemžēl uzrādās stiprākā izpausmē – priekš Viļakas Baltinava izrādās tālāka ne tikai par Rudzātiem, bet pat par visām pārējām izloksnēm.

Mahalanobja attālums neuzrāda nekādas jēgas pazīmes: priekš Aulejas Viļakai nav jābūt pašai tālākajai, pie tam tik ļoti lielā mērā – vairākas reizes, priekš Baltinavas Dundagai nav jābūt pašai tuvākajai un arī tik ļoti lielā mērā – starpība starp attālumiem Dundaga-Baltinava un Dundaga-Rudzāti ir gandrīz 200(!) reizes u.t.t. Bez tam tas nav simetrisks, tātad nav metrika. Arī t.s. īpašais Eiklīda attālums ir pilnībā bezjēdzīgs: lai arī proporcijas nav tik kliedzošas, tomēr pēc būtības tā rezultāti ir līdzīgi Mahalanobja attālumam.

Otrajā eksperimentā nolēmām veikt tādus pašus aprēķinus kā pirmajā, bet ar MAP pielāgošanas palīdzību veidotiem modeļiem. Tālab izveidojām kopīgu fona modeli, kurš tika būvēts uz 30% no katras izloksnes katra runātāja ierakstiem. Tad ar atlikušo 70% attiecīgās izloksnes datu palīdzību šis fona modelis tika adaptēts un izveidots attiecīgās izloksnes modelis – tā visām izloksnēm. Respektīvi, datu sagatavošanas procesā visas datnes tika sadalītas divās – 30% garā sākuma daļā un 70% garā beigu daļā.

Teorētiski pielāgošanu var veikt ne tikai pēc visiem trim, bet arī pēc diviem vai viena no GMM parametriem, visbiežāk – vidējās vērtības. Tomēr tādā gadījumā rezultāti nav interesanti – attālumiem, kuri balstās vidējo vērtību atšķirībās, tie sakrīt ar tādu modeļu, kuri adaptēti pēc visiem trim parmetriem, attālumiem, savukārt, attālumi, kuri balstīti kovariācijas vai svaru atšķirībās, ir nulle. Tāpēc izvēlējāmies eksperimentēt ar modeļiem, kuri tiek adaptēti pēc visiem trijiem parametriem.

Interesanti, ka Eiklīda attālums **vidējo vērtību vektoriem** ar pielāgošanu veidotiem modeļiem ir sliktāks, nekā tas bija tieši veidotiem modeļiem: Baltinavai un Viļakai Dundaga izrādās „tuvāka“ par Auleju. Ar Kuļbaka-Leiblera diverģenci kaut kas nav lāga – tā atgriež negatīvas vērtības.

Pilsētas kvartālu metrikai šinī gadījumā iraid tās pašas problēmas, kas Eiklīda metrikai. Kā jau ierasts, Žordāna metrika uzvedas sāvādāk un šoreiz, jāteic, arī labāk: Dundaga ir pati tālākā visām latgaliskajām izloksnēm. Tomēr neliela problēmiņa neizpaliek: Viļakai Rudzāti rādās tuvāki, nekā Baltinava.

Eiklīda attālumam starp **kovariācijas matricām** piemīt tās pašas problēmas, kādas tam bija starp vidējo vērtību vektoriem, tikai šoreiz tās saskatāmas stiprākā formā. Arī Kuļbaka-Leiblera diverģencei šoreiz piemīt tās pašas vainas. Un vērtības ir gandrīz simetriskas, tāpēc simetrizēšana šoreiz neglābs.

Pilsētas kvartālu metrikai šoreiz piemīt ne tikai visi tie paši trūkumi, kas Eiklīda metrikai, bet pat par vienu vairāk: priekš Viļakas Rudzāti rādās tuvāki par Baltinavu. Žordāna metrikai pēc būtības piemīt tā pati vaina, kas Eiklīda un pārējām metrikām, tomēr tā uzrādās citiem izlokšņu komplektiem: Dundaga izrādās tuvāka gan Aulejai par Viļaku, gan Viļakai par Auleju, savukārt, Rudzātiem par Dundagu tālākas ir gan Auleja, gan Viļaka.

Tagad pāriesim pie **svaru vektoriem**. Eiklīda metrika uzrāda negaidīti labus rezultātus: Dundaga ir tālākā priekš visām pārējām izloksnēm, bet starp Viļaku un Baltinavu ir pats mazākais attālums. Interesanti, ka, pretēji 1. eksperimentā novērotajam, Eiklīda metrika svaru

vektoriem uzrāda labākus rezultātus nekā konvariācijas matricām. Pie tam tās rezultāti vairāk līdzinās kovariācijas matricu rezultātiem bez pielāgošanas. Loģisku izskaidrojumu tam neradām, jo pielāgošanas procesā svaru vektoros netiek akumulēti pielāgošanas datu kovariācijas matricu dati, kas šādu efektu varētu radīt.

Kuļbaka-Leiblera diverģencei piemīt nelielas problēmas: Baltinavai un Viļakai Rudzāti izrādās nedaudz tālāki par Dundagu. Tā kā otrā virzienā viss ir kārtībā, tad pamēģināsim to labot ar simetrizēšanas palīdzību. Simetrizēšana tik tiešām šoreiz palīdz, un pēc tās Kuļbaka-Leiblera (jau) attālums izskatās pat ļoti labi.

Pilsētas kvartālu uzvedas līdzīgi, kā Eiklīda, tomēr nedaudz sliktāk: Baltinavai Dundaga izrādās nedaudz tuvāka par Rudzātiem. Žordāna metrika uzvedas savādāk: ar Dundagu viss ir kārtībā, tā tik tiešām ir tālākā priekš visām izloksnēm, tomēr pāris Baltinava-Viļaka nav labs: šis attālums izrādās krietni lielāks, nekā attālumi no Viļakas līdz Aulejai un Rudzātiem.

Mahalanobja attālums arī šoreiz ir pilnīgi bezsakarīgs: attālumi „lēkā“ vairrākkārtīgi un bez kāda saprotama izskaidrojuma. Arī īpašais Eiklīda attālums ir tikpat bezsakarīgs.

Izlokšņu runas materiāla daudzums nekad nebūs pilnīgi vienāds – tas ir atkarīgs no apstākļiem, veiksmes u.c. neparedzamiem vākšanās laikā klātesošiem faktoriem. Protams, arī mūsu savāktais runas daudzums pa izloksnēm atšķiras. Tas ir normāli, un tas nav nekas slikts, tomēr šīs daudzuma atšķirības var ietekmēt eksperimentu rezultātus – modeļi, kuri veidoti uz būtiski lielāka runas daudzuma, var atšķirties no citiem tāpēc vien, ka runas daudzums lielāks, tādējādi tās – valodiskajās īpašībās balstītās – atšķirības, kuras interesē, var nonivelēties un rezultāti izkropļoties.

Tādēļ nolēmām veikt eksperimentus, kuros runas daudzums tiek nolīdzsvarots, resp., izloksnēm ar lielāku runas daudzumu katrai informanta fonogrammai tiek atstāta tikai tās sākuma daļa tā, lai kopējais izloksnes runas daudzums sanāktu apmēram tāds pats, kā pašai „mazākajai“ izloksnei. Mūsu hipotēze bija, ka rezultātiem vajadzētu būt labākiem, nekā nenolīdzsvarotiem runas datiem.

Šis – **trešais** – eksperiments ir tāds pats kā 1. eksperiments, tikai pilno fonogrammu vietā tiek izmantotas „nolīdzsvarotās“.

Tāpat kā iepriekšējos eksperimentos, sāksim ar kovariācijas matricām. Eiklīda metrikas rezultāti nolīdzsvarotā gadījumā būtiski neatšķiras no pirmajā eksperimentā iegūtajiem rezultātiem nenolīdzsvarotiem runas datiem, proti, tie ir labi: Dundaga ir pati tālākā visām izloksnēm, Viļaka un Baltinava ir pats tuvākais izlokšņu pāris. Tāpat kā bez līdzsvarošanas, Kuļbaka-Leiblera diverģence satur arī negatīvas vērtības.

Pilsētas kvartālu metrikai parādās problēmas: Dundaga izrādās tuvāka par Rudzātiem gan Aulejai, gan Baltinavai, gan Viļakai. Bet bez līdzsvarošanas L_1 metrikai problēmu nebija. Tātad izskatās, ka Eiklīda metrika ir stabilāka pret līdzsvarošanu, nekā L_1 metrika. Žordāna metrika tieši veidotu modeļu vidējo vērtību vektoriem izskatās tikpat bezjēdzīga kā bez līdzsvarošanas: Dundaga ir tuvāka Viļakai nekā visas(!) pārējās latgaliskās izloksnes, no kā izriet arī tas, ka Baltinava ar Viļaku nav tuvākais pāris, arī Aulejai Dundaga izrādās tuvāka nekā Viļaka.

Eiklīda metrika **kovariācijas matricām** uzvedas līdzīgi kā bez līdzsvarošanas, kā arī līdzīgi kā vidējo vērtību vektoriem, proti, neatbilstības nav novērojamas. Arī Kuļbaka-Leiblera diverģence uzvedas līdzīgi kā bez līdzsvarošanas; tiesa, problēmas ir citās vietās: Aulejai un Baltinavai Dundaga izrādās tuvāka par Rudzātiem. Pēc simetrizēšanas Aulejas problēma pazūd, bet Baltinavas – paliek.

Pilsētas kvartāla metrika pret līdzsvarošanu izrādās jūtīgāka – tai parādās nopietnas vainas, kaut pirms līdzsvarošanas tādu praktiski nebija: Baltinavai Auleja ir tuvāka par Viļaku un Dundaga – tuvāka par Rudzātiem, savukārt Aulejai Dundaga izrādās tuvāka gan par Rudzātiem, gan par Viļaku. Arī Žordāna metrikai, kurai pirms līdzsvarošanas problēmu nevaidēja, tagad tās uzrodas, pie tam pulka: Aulejai Dundaga tuvāka par Rudzātiem, Baltinavai Viļaka tālāka par Auleju, bet Viļakai Baltinava tālāka par Rudzātiem.

Svaru vektoru gadījumā Eiklīda metrikai piemīt tās pašas vainas, kas nelīdzsvarotā gadījumā, tomēr tiesa, kas tiesa, tās ir mazāk izteiktas, proti, starpība ir samazinājusies: Aulejai Dundaga tuvāka par Viļaku, Viļakai Rudzāti tuvāki par Baltinavu.

Arī Kuļbaka-Leiblera diverģencei uzrādās tādas pat problēmas kā bez nolīdzsvarošanas – daļa tanīs pat, bet daļa – citās vietās: vienā virzienā Aulejai Dundaga ir tuvāka par Baltinavu un Viļaku, Baltinavai – Viļaka tālāka par Auleju un pat Dundagu, Viļakai – Baltinava tālāka par Rudzātiem, otrā virzienā problēmas ir līdzīgas, resp., abpusējas. Tāpēc ir skaidrs, ka simetrizēšana šinī gadījumā palīdzēt nevar un to nav vērts veikt.

Pilsētas kvartālu metrika šinī gadījuma satur vienu problēmu mazāk nekā bija bez līdzsvarošanas. Palikusi tikai Aulejas neatbilstība, kurai Dundaga izrādās tuvāka par Baltinavu un Rudzātiem. Atšķirībā no iepriekšējās, Žordāna metrika uzvedas sliktāk nekā bez līdzsvarošanas: it kā problēmas ir tādas pašas, bet to skaits ir dubultojies – Dundagai dikti tuva izrādās ne tikai Viļaka, bet arī Rudzāti.

Mahalanobja attālums mūs nepārsteidz – tas aizvien ir pilnīgi bezjēdzīgs. Un tieši tāpat arī īpašais Eiklīda attālums.

Savukārt **ceturtais** eksperiments ir tāds pats kā 2. eksperiments, tikai pilno fonogrammu vietā tiek izmantotas „nolīdzsvarotās“.

Eiklīda attālums **vidējo vērtību vektoriem** uzvedas tāpat kā bez līdzsvarošanas, respektīvi, sliktāk, nekā bez pielāgošanas: Viļakai un Baltinavai Dundaga ir tuvāka par Auleju. Kuļbaka-Leiblera diverģence nav lāga – tā satur negatīvas vērtības.

Pilsētas kvartālu metrikai piemīt tāda pati vaina kā bez līdzsvarošanas un kā Eiklīda metrikai, tomēr mazākā mērā. Žordāna metrika izskatās puslīdz pieklājīgi, bet tomēr tai ir viena, savādāka vaina: Dundaga ir stipri tuvāka Viļakai nekā Rudzāti.

Eiklīda metrika **kovariācijas matricām** uzvedas tāpat kā gadījumā bez līdzsvarošanas, proti, tai ir tāda pat problēma, kā vidējo vērtību gadījumā, bet izteiktākā formā. Kuļbaka-Leiblera diverģencei iraid tādas pat vainas kā bez nolīdzsvarošanas. Un tā ir gandrīz simetriska, tāpēc skaidrs, ka simetrizēšana nelīdzēs.

Arī pilsētas kvartālu attālums uzvedas tāpat kā bez nolīdzināšanas, proti, tam ir tāda pat problēma kā Eiklīda attālumam, kā arī papildus vēl viena: Viļakai Rudzāti ir stipri tuvāki par Baltinavu. Žordāna metrikai arī uzrādās tās pat vainas, kas bez līdzsvarošanas, tomēr tās daļēji ir citām izloksnēm un mazākā mērā.

Savukārt Eiklīda metrika **svaru vektoriem** nolīdzsvarošanas rezultātā ir sabojājusies: Baltinavai tagadiņ Dundaga tuvāka par Rudzātiem, bet Viļakai – Rudzāti tuvāki par Baltinavu. Kuļbaka-Leiblera diverģencei arī it kā ir tikai viena bēdiņa, gluži tāpat kā bez līdzsvarošanas. Tomēr atšķirība slēpjas tanī, ka šoreiz šī bēdiņa ir simetriska, tātad ar simetrizēšanu nav izlabojama, un tādēļ jāsecina, ka nolīdzsvarotā gadījumā šī diverģence uzvedas švakāk.

Pilsētas kvartālu metrika uzvedas tāpat kā bez līdzsvarošanas, proti, ir līdzīga Eiklīda metrikai, tomēr sliktāka. Žordāna metrika uzvedas sliktāk nekā bez līdzsvarošanas: vienas problēmas vietā tai to ir vesels lērums: Aulejai un Baltinavai Dundaga tuvāka par Rudzātiem, arī Rudzātiem Dundaga tuvāka par Auleju, savukārt Baltinavai Auleja tuvāka par Viļaku.

Mahalanobja attālums arī šoreiz izskatās pilnīgi bezjēdzīgs, kas top skaidrs kaut vai redzot, par cik kārtām attālumi savstarpēji atšķiras. Arī t.s. īpašais Eiklīda attālums nav būtiski labāks.

Varam **secināt**, ka kopumā metode ir izmantojama, jo tās ietvaros tiek aprēķināti tādi dati un iraid tādas metrikas, kas dod labus rezultātus.

MAP pielāgošana jeb adaptācija, kopumā ņemot, nāk par sliktu. To tā arī varējām iedomāties, jo, faktiski, pielāgošana ir datu sagatavošanas procesa atvieglināšana uz modeļu atšķirīguma rēķina, tiesa, pietiekami nelielā mērā, lai atpazīšanas procesi darbotos. Bet, kas ir gana labi runas atpazīšanas uzdevumos, var jau nebūt labi attālumu aprēķināšanā.

Hipotēze par rezultātu uzlabošanos datu apjoma nolīdzsvarošanas rezultātā ir izrādījusies aplama – rezultāti ne tikai nav uzlabojušies, bet pat ir nedaudz pasliktinājušies. Iespējams tādēļ, ka tas vien, ka atsevišķām izloksnēm datu bija vairāk, ļāva šos modeļus izveidot kvalitatīvākus, tādējādi arī spilgtāk iezīmējot objektīvās izlokšņu atšķirības. No kā jāsecina, ka varam cerēt uz

labākiem rezultātiem runas datu apjoma palielināšanas gadījumā.

Ja aplūkojam konkrētas metrikas un datus, kuriem tās aprēķinātas, tad laba un stabila pret nolīdzsvarošanu ir Eiklīda metrika vidējo vērtību vektoriem un kovariācijas matricām. Izmantojami ir abi varianti, bet iespējams, ka uz šo divu datu veidu pamata var izveidot kādu jaunu eiklīdveidīgu metriku, kas ņemtu vērā tos abus.

Līdz ar to galvenais šī nodaļas secinājums ir: Gausa maisījuma modeļi ir izmantojami valodu tuvības pakāpes novērtēšanai, pie tam to var veikt ar vienkāršu un visiem pieejamu – Eiklīda – metriku.

Rezultātu pārbaude ar ekspertu novērtējuma metodi

Visās iepriekšējās nodaļās jaunizstrādāto metožu novērtēšanai izmantojām savas zināšanas, kuras iegūtas no valodnieciski-analītiskiem avotiem un dzīvas komunikācijas ar izlokšņu nesējiem. Šāds novērtējums radīja mūsos pārliecību, ka metodes darbojas un ir izmantojamas, tomēr formāli, lai arī objektīvos avotos balstīts, tas var tikt uzskatīts par subjektīvu, tāpēc nepieciešams veikt arī formāli nevainojumu papildus novērtējums, vēlams – skaitlisku.

Ir tādas skaitļotājvalodniecības jomas, kurās jau izstrādāti t.s. „zelta standarti“. Skaidrs, ka Latvijas izlokšņu runas ierakstu analīzes starp tām nav. Arī pašiem izstrādāt šādu standartu būtu neiespējami gan no darbietilpības, gan no atbildības uzņemšanās puses raugoties. Tāpēc rezultātu pārbaudē bija jāiet cits – nosacīta salīdzinošā novērtējuma ceļš.

Pieejamākā un īstenojamākā šinī gadījumā mums šķita ekspertu novērtējuma metode: gan tādēļ, ka citi – atzīti – skaitliskie novērtējumi nav izstrādāti, gan tādēļ, ka Latvijas izlokšņu tuvības novērtēšanai vajadzīgie speciālisti pieejami tepat uz vietas, proti, Latvijā.

Literatūrā norādīts, ka par ekspertiem jāizvēlas labākie attiecīgās jomas speciālisti, pie tam padsmit esot optimāls ekspertu skaits. Tā kā eksperimenti tika veikti uz latviešu izlokšņu materiāla, tad šāda veida manuālu novērtējumu spēja veikt ļoti šaurs cilvēku loks, kuri Latvijā nodarbojas vai ir nodarbojušies ar dialektoloģiju. Apzinājām visus potenciāli kompetentos ļaudis un gandrīz visiem palūdzām veikt mūsu sagatavoto datu novērtējumu. Tātad varam secināt, ka esam veikuši optimālu ekspertu izvēli – gan kvalitatīvo, gan kvantitatīvo rādītāju izpratnē.

Visas mūsu izstrādātās metodes ir izmantojamas attālumu starp valodām aprēķināšanai, un šie skaitliskie attālumi arī ir metožu izejas dati. Tomēr šāda veida rezultāti ir pārāk grūti novērtējami: lai salīdzinātu tos savā starpā, tie būtu kaut kādā (mums šobrīd nezināmā) veidā jānormē, jo attālumi ir nosacīti, tātad katrai metodei absolūtos skaitļos savādāki, bet, lai salīdzinātu ar ekspertu vērtējumu, ekspertiem būtu jāuzņemas šādu attālumu skaitliska nosaukšana, kas diez vai būtu iespējams, it jo sevišķi tāpēc, ka šīs nozares speciālistu lielākā daļa ir humanitāri tendēti ļaudis.

Arī paši gandrīz katrā nodaļā metodes sekmīgumu noteicām pēc hierarhiskās kategorizēšanas shēmām, kuras bija daudz uzskatāmākas par skaitlisko attālumu tabulām. Acīmredzot, veicot ekspertu aptauju, tiem arī jālūdz sastādīt šāda veida hierarhiskās kategorizēšanas kokus, pēc attiecīgā eksperta profesionālajām zināšanām un izjūtas katrā nākamajā solī izvēloties tuvāko pāri.

Vispirms nolēmām aplūkot zināmās ekspertu novērtējuma metodes (ENM) un izvērtēt to pielietojamību mūsu datiem. Dažādos avotos to skaits svārstās no 4 līdz 7, tomēr tās visas ir orientētas uz elementu sakārtošanu uz vienas ass, salīdzināšanu pa pāriem, ranžēšanu vai novērtēšanu kvantitatīvās vienībās; tātad mūsu hierarhiskās kategorizēšanas rezultāts – binārie koki, pie tam ar birkām jeb etiķetēm uz visām lapām, bet bez birkām mezglos, neatbilst nevienai no dotajām metodēm.

Tā kā mūsu eksperimenti tika veikti uz 3 dažādiem datu komplektiem – 4 (neparalēliem spontānās runas pierakstiem fonētiskajā transkripcijā), 5 (spontānās runas fonogrammām) un 13 (paralēlām vārdnīcām) izloksnēm, aptaujas anketā iekļāvām 3 attiecīgus uzdevumus: hierarhiski sakategorizēt katru no komplektiem. Labākai izpratnei pievienojām arī papildus lapu ar ilustratīvu piemēru.

Tā kā salīdzināmos datus jau izvēlējāmies – tie ir hierarhiskās kategorizēšanas rezultātā radītie binārie koki, tad turpmākajam darbam jāpiemeklē metrikas, kuras definētas šādu objektu telpā. Jāteic, ka attālumu aprēķināšana starp šāda veida objektiem nav pārāk izplatīts uzdevums, tomēr ir bijuši atsevišķi mēģinājumi to darīt, līdz ar to pasaules kontekstā ir atrodami arī attiecīgo metriku piedāvājumi.

Visplašāk izplatītais ir t.s. rotāciju attālums – tas tiek definēts kā mazākais nepieciešamo rotāciju skaits, lai no viena koka iegūtu otru. Ir atrodami arī izstrādāti aprēķina algoritmu piedāvājumi. Un, lai arī pamatā tiek strādāts ar kokiem bez birkām, tiek apgalvots, ka pāreja uz birkainiem kokiem lietas būtību nemainot. Tomēr tas, kā tiek definētas pašas šīs rotācijas, nonāk

pretrunā ar mūsu ar binārajiem kokiem aprakstīto objektu būtību: sanāk, ka, piemēram, ar vienu rotācijas darbību var iegūt (mūsu izpratnē) principiāli atšķirīgu koku, kas neatbilst tik tuvam (rotāciju izpratnē) attālumam.

Atradām vēl vienu, mazāk zināmu – AKM attālumu. Tanī mūs neapmierināja jau pati pieeja, ka zemākos zaros atšķirībām ir mazāka vērtība, jo mūsu gadījumā visas izvēles ir no svara, neatkarīgi no tā, kad tās veiktas.

Tādējādi meklējumu rezultātā nonācām pie secinājuma, ka neviena no atrastajām metrikām neatbilst mūsu uzdevumam. Tālab jāstrādā no otras puses – jāmēģina pašiem definēt metrika, kura skaitliski raksturotu mūsu objektu savstarpējo attiecību būtību.

Kas tad ir tas atšķirīgais, kas var piemist mūsu vērtējumiem – gan automatizētiem, gan manuāliem? Principā tās ir izvēles, kuras katrā solī tiek veiktas, izvēloties tuvāko pāri. Tātad, ja dots ir *n* valodu komplekts, tad radītos kokus varam aprakstīt arī kā *n*-elementīgas kopas, kuras sastāv no jaunizveidotajiem (jeb izvēlētajiem) pāriem, kas jau nākamajā solī top par izvēles elementiem.

Par **hierarhisko izvēļu attālumu** sauksim starpības starp visu hierarhiskās kategorizācijas netriviālo izvēļu elementu kopu kopām apjoma dalījumu ar netriviālo izvēļu skaitu. Hierarhisko izvēļu attālums ir metrika.

Lai ekspertu novērtējums būtu izmantojams, jāpārliecinās, ka to kompetences līmenis ir pietiekoši augsts. Ir vairāki paņēmieni, kā to darīt, tomēr mums vienīgais reālais pieejamais variants bija kompetences novērtēšana, balstoties pašos aptaujas datos – šo vērtējumu mēdz saukt arī par vienprātības pakāpes noteikšanu, kuru pie tam var arī uzlabot. Protams, šeit var rasties riski, ka lielākā daļa ir nekompetenti, un tieši kompetentie tiek izbrāķēti, tomēr mēs, izvēlēdamies labākos profesionāļus, šos riskus jau samazinājām līdz minimumam. Protams, tās visas ir aprakstītas standarta novērtējumu gadījumiem, bet ne hierarhiskajai kategorizēšanai. Tāpēc bija jāmēģina izdomāt līdzīgu ekpertu kompetences/vienprātības novērtējumu, kas būtu piemērots mūsu datiem.

Aprēķināsim hierarhisko izvēļu attālumu savstarpēji starp visiem ekspertiem. Katram ekspertam izrēķināsim vidējo aritmētisko attālumu līdz pārējiem ekspertiem. Šo vidējo attālumu sarēķināsim katram valodu komplektam, kuram kategorizēšanas veikta, mūsu gadījumā tādu ir trīs. Tad izrēķināsim vidējo aritmētisko visiem komplektiem. Tos, kuru vidējais attālums būs lielāks par pusi (0,5), izslēgsim un ar mazāku ekspertu skaitu atkārtosim visus soļus no sākuma. Kad visu atlikušo ekspertu vidējie attālumi būs mazāki par 0,5, procesu pārtrauksim.

Tā kā pirmajam komplektam (4 izlokšņu) visi vērtējumi sakrita, tad tiem neveidojām atsevišķu tabulu, jo visi attālumi ir vienādi ar nullēm. Aprēķinot vidējās vērtības pa komplektiem, šo nulli ņēmām vērā kā vienu no trim saskaitāmajiem. No rezultātiem tapa redzams, ka viena eksperta vidējais attālums ir lielāks par 0,5, tātad to kā neatbilstošu vienprātības kritērijiem slēdzam ārā un nākamajā solī pārrēķinām tabulu bez tā.

Nākamajā solī ekspertu vienprātība ir augstāka, nekā bija pirmajos aprēķinos, tātad ārpus mūsu noteiktajām robežām esošā eksperta izslēgšana ir devusi pozitīvu pienesumu. Vairāk nav neviena eksperta, kura vidējais attālums būtu lielāks par 0,5, tātad ekspertu vētīšana ir sekmīgi noslēgusies, un mēs varam pāriet pie galvenā uzdevuma – metožu vērtēšanas.

Tālāk aprēķinājām attālumus starp visām metodēm un visiem ekspertiem – bez un ar kritērijiem neatbilstošā eksperta izslēgšanu (lai redzētu, kā tā ietekmē rezultātus). Metožu izmantojamības robeža intuitīvi jānosaka pašiem. Izdarījām to divos veidos – relatīvā (robeža ir ekspertu vidējais attālums dotajam komplektam, resp., metodes kļūda nedrīkst pārsniegt ekspertu kļūdu) un absolūtā (robeža ir mūsu noteikts absolūts skaitlis, kura izvēle balstīta intuīcijā un pieredzē). No rezultātiem redzams, ka ir pietiekami daudz metožu, kuru rezultāti pēc ekspertu novērtējuma metodes pielietošanas uzskatāmi par labiem.

Arī kļūdas novērtējumam gadījumā, kad mērījumi ir bināri koki, nav gatavas receptes. Pēc būtības par metodes kļūdu var tikt uzskatīta ekspertu vērtējuma vidējā vērtība – jo tā lielāka, jo metode sliktāka, proti, kļūdaināka. Tā kā šāda pieeja empīriski atbilst, kā arī neprasa papildus darba ieguldījumus, tad pie tās arī pieturējāmies.

Pirmkārt, ar ekspertu metodi esam formāli pierādījuši mūsu automatizēto metožu dzīvotspēju un novērtējuši katras metodes kļūdu. Vienai tās nebija vispār, citām bija niecīga, dažām – maza (šīs metodes noteikti ir izmantojamas), dažām – vidēja (iespējams, izmantojamas), pārējām – pārlieku liela (neizmantojamas).

Otrkārt, jāsecina, ka ekspertu metode hierarhiskajā kategorizēšanā ir izmantojama salīdzinoši nelieliem (līdz 10?) elementu (valodu) komplektiem – pie 13 elementiem cilvēkam jau ir grūti orientēties un lēmumu spontānums sāk prevalēt pār izsvērtību.

Treškārt, ja automatizētais novērtējums tiek veikts pēc noteiktām pazīmēm (piem., fonētiskām, morfoloģiskām, leksiskām vai sintaksiskām), tad arī ekspertu novērtējums jāveic atsevišķi pēc tām pašām pazīmēm. Pretējā gadījumā, veicot vienu kopīgu novērtējumu, kļūda var jūtami (un pat kritiski) pieaugt.

Pēcvārdi

Kā pārliecinājāmies disertācijas izstrādes ietvaros veikto eksperimentu laikā, mūsu galvenā hipotēze apstiprinās: gan runas fonētiskā transkripcija, gan runas ieraksti kā ievades dati ir pietiekami, lai no tiem ar statistiskām metodēm iegūtu skaitlisku valodu tuvības pakāpes novērtējumu.

Darba mērķis ir sasniegts – izstrādājām sešas jaunas metodes, kas atbilst mūsu definētajām prasībām (skat. priekšvārdus). Arī visi četri darba uzdevumi ir izpildīti.

Tā kā fonētiski transkribētas runas apjoms ir ierobežots, pie tam transkribēšanas tradīcijas ir dažādas un to novienādošana prasa diezgan lielus darba ieguldījumus, tad perspektīvāka, protams, šķiet, darbošanās ar runas ierakstiem. Tanī pat laikā fonētiskajai transkripcijai paredzētās metodes ir uzskatāmākas un elegantākas, tāpēc nekādā gadījumā nebūtu atstājamas novārtā. Bez tam, kā mēs parādījām nodaļā par fonēmu atpazinēju metodi, iezīmējas arī iespēja šīs it kā pilnīgi dažādās pieejas sapludināt.

Jaunākā un vieglāk pielietojamākā metode runas ierakstiem ir i-vektoru metode. Veidojot kompleksu sistēmu valodu tuvības pakāpes noteikšanai, šī metode droši vien būtu jāņem par pamatu, tomēr tā jāpapildina arī ar citām metodēm. Šādu sistēmu varētu izmantot mērķa sasniegšanai, kas mūs pamudināja sākt strādāt pie šīs tēmas: ierobežot sabiedriskas un politiskas spekulācijas par jautājumu, kur beidzas dialekts un sākas valoda.

Šobrīd strādājam pie lielāka apjoma materiālu ieguves un sagatavošanas (kas ir liels un apjomīgs darbs). Vēlamies savas metodes pielietot lielākam valodu skaitam, lai kategorizācijas rezultāti atainotu veselas valodu saimes (vai vairāku saimju) visu valodu savstarpējo tuvību – tad rezultātu salīdzināšana ar analītisko valodu iedalījumu būs daudz interesantāka, kā arī sniegs lielākas iespējas novērtēt metožu atbilstību analītiskiem valodu iedalījumiem.

Stažēšanās ietvaros Leišu valodas institūtā (Viļņā) sagatavojām visu leišu izlokšņu skaņas ierakstu korpusu, kuram mēģināsim savas metodes pielietot. Tāpat stažēšanās laikā Mehiko Nacionālajā autonomajā universitātē uzsākām darbu pie spāņu valodas izlokšņu runas korpusa un indiāņu valodu runas korpusa izveides, kuriem arī būtu iespējams pielietot šīs disertācijas ietvaros izstrādātās metodes.

Perspektīvā mēs labprāt redzētu uz savu atsevišķo metožu pamata veidotu virsmetodi jeb supermetodi – tanī būtu apvienotas visas pieejamās atsevišķās metodes un tā jebkuram valodu pārim aprēķinātu attālumu starp tām. Programmatūra būtu ar ērtu saskarni, noslīpēta, plaši lietojama. Aprēķinātie attālumi būtu starptautiski atzīti un tiktu ņemti vērā gan zinātniskos, gan sabiedriskos atzinumos. Visām pasaules valodām būtu sagatavoti pietiekami ieejas dati virsmetodes izmantošanai. Tas, protams, nav viena cilvēka, bet gan vairāku institūtu darbs. Bet mēs savā darbā esam padarījuši galveno – to, ko viens cilvēks var izdarīt: pavēruši šīs durvis un pierādījuši, ka šādu tehnoloģiju var izveidot un pielietot.

Vēres

Vēres veidojām hronoloģiskā secībā – tā, kā, rakstot disertāciju, avoti pie mums „atnāca". Šī secība atšķiras no disertācijas secības, jo tās nodaļas tika rakstītas jauktā secībā.

Literatūras atspoguļošanā centāmies būt maksimāli godīgi – norādījām visus avotus, kurus izziņas procesā izmantojām, tai skaitā tīmekļa vietnes un tai skaitā arī tīmekļa enciklopēdijas. Uzskatām, ka godīgums un atklātums statāms pirmajā vietā arī zinātnē un, ja reiz mēs kādu avotu izmantojam, tad tas ir norādāms. Uz potenciāli nievājošiem komentāriem par tai skaitā Vikipēdijas izmantošanu paskaidrojam, ka arī Vikipēdija ir lietojams un pat noderīgs avots, protams, pie nosacījuma, ka ar to māk pareizi apieties.

1. *Никольский В.К., Яковлев Н.Ф.* Как возникла человеческая речь. Москва: Государственное издательство культурно-просветительной литературы, 1949. 64 стр. с ил.
http://genling.ru/books/item/f00/s00/z0000029/index.shtml
Piekļūts 2015. gada 1. aprīlī.

2. *Ferdinand de Saussure.* Cours de linguistique générale. Paris: Payot, coll. «Grande bibliothèque Payot », 1995 (1re éd. 1916).
http://fr.wikisource.org/wiki/Page:Saussure_-_Cours_de_linguistique_g%C3%A9n%C3%A9rale,_%C3%A9d._Bally_et_Sechehaye,_1971.djvu/32
Piekļūts 2015. gada 1. aprīlī.

3. *Иванов В.В.* Моногенеза теория // Лингвистический энциклопедический словарь. - М., 1990. - С. 308-309.
http://www.philology.ru/linguistics1/ivanov-90c.htm
Piekļūts 2015. gada 1. aprīlī.

4. *Кочеткова В.И.* Палеоневрология. М.: Изд-во Моск. ун-та, 1973. С. 188-215.
http://www.ido.rudn.ru/psychology/anthropology/ch4_2.html
Piekļūts 2015. gada 1. aprīlī.

5. *Кареев Н.И.* О «новом взгляде» г. Шапиро на современную систему сравнительного языкознания (Возражение) // Филологические записки. Воронеж, 1874.
http://vrn-id.ru/filzaps741.htm
Piekļūts 2015. gada 1. aprīlī.

6. *Топоров В.Н.* Сравнительно-историческое языкознание // Лингвистический энциклопедический словарь. - М., 1990. - С. 486-490.
http://www.philology.ru/linguistics1/toporov-90.htm
Piekļūts 2015. gada 1. aprīlī.

7. *Иванов В.В.* Генеалогическая классификация языков // Лингвистический энциклопедический словарь. - М., 1990. - С. 93-98.
http://tapemark.narod.ru/les/093d.html
Piekļūts 2015. gada 1. aprīlī.

8. *Яхонтов С.Е.* Оценка степени близости родственных языков // Теоретические основы классификации языков мира. - М., 1980. - С. 148-157

9. *Swadesh M.* Perspectives and problems of Amerindian comparative linguistics // Word, 1954, № 10, pp. 306-332.

10. *Дьячок М.Т., Шаповал В.В.* Генеалогическая классификация языков. - Новосибирск, 2002. – 32 с.
http://www.philology.ru/linguistics1/dyachok-shapoval-02.htm
Piekļūts 2015. gada 1. aprīlī.

11. *Реформатский А.А.* Введение в языкознание. М.: ГУПИ МП РСФСР, 1960. 431 стр.

12. מ. ריַיזגיַיוו. דער ייִוואָ און די פּראָבלעמען פֿון אונדזער צייַט // Yivo Bleter. – New York, 1945. Vol. XXV, No.3, p. 3-18.

13. *Зиндер Л.Р.* Общая фонетика. М.: Высшая школа, 1979. – 312 с.

14. Latviešu valodas dialektoloģijas atlanta materialu vākšanas programa. Rīga: Latvijas PSR ZA izdevniecība, 1954.

15. *Rudzīte M.* Latviešu dialektoloģija. Rīga: Latvijas valsts izdevniecība, 1964.

16. *Tambovtsev Y.* Phonological Similarity Between Basque and Other World Languages Based on the Frequency of Occurrence of Certain Typological Consonantal Features // The Prague Bulletin of Mathematical Linguistics 79-80, pp. 121-126, 2003.

17. *Canvar W.B., Trenkle J.M.* N-Gram-Based Text Categorization // In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 1994.

18. Latviešu izlokšņu teksti. Sast. Marta Rudzīte. Rīga: P. Stučkas Latvijas Valsts universitāte, 1963.

19. Augšzemnieku dialekta teksti. Latgaliskās izloksnes. Sast. N. Jokubauska. Rīga: Zinātne, 1983.

20. Latviešu izlokšņu teksti. Sast. Benita Laumane. Liepāja: Liepājas Pedagoģiskā akadēmija, 2000.

21. *Ball M.J.* Teaching Vowels in Practical Phonetics: The Auditory or Articulatory Route?
http://www.phon.ucl.ac.uk/home/johnm/ball.htm
Piekļūts 2007. gada 26. oktobrī.

22. Corporate Author International Phonetic Association. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 1999.

23. *Kessler B.* Computational dialectology in Irish Gaelic // Proceedings of the European ACL. Dublin, 1995. Pp. 60-66.

24. *Markus D., Grigorjevs J.* Fonētikas pētīšanas un vizualizēšanas metodes, II grām. sēr.: Fonētikas pētīšanas un vizualizēšanas metodes. Rīga: Rasa ABC, 2004.

25. *University College London, Dept. of Phonetics and Linguistics.* Cardinal Vowels by Daniel Jones. London: UCL Press, 1996.

26. *Пиотровский Р.Г.* Еще раз о дифференциальных признаках фонемы // Вопросы языкознания, № 6, М.: РАН, 1960, стр. 24-38.

27. *Nerbonne J., Heeringa W., van den Hout E., van der Kooi P., Otten S., van de Vis S.W.* Phonetic Distance between Dutch Dialects // CLIN VI, Papers from the sixth CLIN meeting. Antwerp: University of Antwerp, Center for Dutch Language and Speech. Pp. 185-202.

28. От аналоговой записи – к цифре.
http://its-journalist.ru/Articles/ ot_analogovoj_zapisi_k_cifre.html
Piekļūts 2012. gada 5. martā.

29. Звукозапись, цифровая или аналоговая?
http://www.midi.ru/forumd.php?id=181648
Piekļūts 2012. gada 5. martā.

30. *Дубровский Д.Ю.* Чем цифровая запись лучше аналоговой?
http://demorecord.ru/analogsound.html
Piekļūts 2012. gada 5. martā.

31. *Музыченко Е.В.* Принципы цифрового звука. 1998-1999.
http://www.websound.ru/articles/theory/digsnd.htm
Piekļūts 2012. gada 5. martā.

32. Audio file format.
http://en.wikipedia.org/wiki/Audio_file_format
Piekļūts 2012. gada 5. martā.

33. Сжатие без потерь.
http://ru.wikipedia.org/wiki/Сжатие_без_потерь
Piekļūts 2012. gada 5. martā.

34. Сжатие данных с потерями.
http://ru.wikipedia.org/wiki/Сжатие_данных_с_потерями
Piekļūts 2012. gada 5. martā.

35. A-Law Compressed Sound Format.
http://www.digitalpreservation.gov/formats/fdd/fdd000038.shtml
Piekļūts 2012. gada 5. martā.

36. *Salvi G.* Mining Speech Sounds. Stockholm: KTH, 2006. Pp. 18-19.

37. *Melin H.* Automatic speaker verification on site and by telephone: methods, applications and assessment. Stockholm: KTH, 2006. Pp. 103-104.

38. Микрофон.
http://ru.wikipedia.org/wiki/Микрофон
Piekļūts 2012. gada 5. martā.

39. Электретный микрофон.
http://ru.wikipedia.org/wiki/Электретный_микрофон
Piekļūts 2012. gada 5. martā.

40. Характеристики микрофонов.
http://ingibit.rigalink.lv/info/c2/mikro01.html
Piekļūts 2012. gada 5. martā.

41. Сравнение конденсаторных и динамических микрофонов.
http://www.microphone.ru/articles/paragraph_1.html
Piekļūts 2012. gada 5. martā.

42. Динамические, конденсаторные микрофоны и фантомное питание.
http://midi.ucoz.ru/publ/1-1-0-16
Piekļūts 2012. gada 5. martā.

43. Конденсаторный микрофон.
http://ru.wikipedia.org/wiki/Конденсаторный_микрофон
Piekļūts 2012. gada 5. martā.

44. *Костоломов В.* Ленточные микрофоны. 2000.
http://www.oktava-mics.net/shop/a-2/lentochnye_mikrophony.html
Piekļūts 2012. gada 5. martā.

45. Угольный микрофон.
http://ru.wikipedia.org/wiki/Угольный_микрофон
Piekļūts 2012. gada 5. martā.

46. *Rabiner L.* A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of the IEEE, vol. 77, No 2, February 1989, p. 262-286.

47. *Кушнир Д.А.* Алгоритм формирования структуры эталона для пословного дикторонезависимого распознавания команд ограниченного словаря // Штучный інтелект, № 3'2006. Київ, 2006.

48. Dainuskapis. Sast. Kr. Barons.

49. *Haugland Tokheim Å.E.* iVector Based Language Recognition. Trondheim: NTNU, 2012.

50. *Li H., Ma B., Lee K.A.* Spoken Language Recognition: From Fundamentals to Practice // Proceedings of the IEEE, Vol. 101, No. 5, May 2013.

51. *Plchot O., Diez M., Soufifar M., Burget L.* PLLR Features in Language Recognition System for RATS // Interspeech. Singapore, 2014. Pp. 3047-3051.

52. *Tebelskis J.* Speech Recognition using Neural Networks. Pittsburgh: Carnegie Mellon University, 1995.

53. *Liu Z., Huang Q.* A new distance measure for probability distribution function of mixture type // IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. Vol. 1, pp. 616-619.

54. *Young S., Evermann G., Gales M., Hain Th., Liu X., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland Ph.* The HTK Book (for HTK Version 3.4). Cambridge: Cambridge University Engineering Department, 2009.

55. *Johnson D., Sinanovic S.* Symmetrizing the Kullback-Leibler Distance. Computer and Information Technology Institute, Department of Electrical and Computer Engineering, Rice University, Houston, 2001.

56. Метрика // Математическая энциклопедия. – М.: Советская энциклопедия, 1982. – Т. 3.

57. *Howard D., Angus J.* Acoustics and psychoacoustics. Oxford: Focal Press, 2009.

58. *Kullback S., Leibler R.* On information and sufficiency // Annals of Mathematical Statistics. Vol. 22, No. 1, Mar 1951, pp. 79-86.

59. *Dehak N., Dehak R., Kenny P., Brummer N., Ouellet P., Dumouchel P.*. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification // Proceedings Interspeech. Brigthon, UK, 2009.

60. *Dehak N., Kenny P.J., Dehak R., Dumouchel P., Ouellet P.* Front-End Factor Analysis for Speaker Verification // IEEEE Transactions On Audio, Speech, And Language Processing. Piscataway: IEEE Press, 2011. Vol. 19, no. 4, pp. 788-798.

61. *Schwarz P.* Phoneme Recognition based on Long Temporal Context, PhD Thesis. Brno: Vysoké učení technické v Brně, 2009.

62. *Schwarz P., Matejka P., Cernocky J., Chytil P.* Phonotactic Language Identification using High Quality Phoneme Recognition // Proceedings Eurospeech, 2005.

63. Phoneme Recognition (caveat emptor) // CMU Sphinx.
http://cmusphinx.sourceforge.net/wiki/phonemerecognition
Piekļūts 2016. gada 16. februārī.

64. Phoneme recognizer based on long temporal context // BUT Speech@FIT.
http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context
Piekļūts 2016. gada 16. februārī.

65. *Lamere P., Kwok P., Walker W., Gouvea E., Singh R., Raj B., Wolf P.* Design of the CMU Sphinx-4 decoder // Proceedings of the 8th European Conference on Speech Communication and Technology, Geneve, Switzerland, 2003, pp. 1181–1184.

66. *Soufifar M.* Subspace Modeling of Discrete Features for Language Recognition. Trondheim: NTNU, 2014.

67. *Ghosh S., Vijay Girish K.V., Sreenivas T.V.* Relationship between Indian Languages Using Long Distance Bigram Language Models // Proceedings of ICON-2011: 9'th International Conference on Natural Language Processing. Chennai: Macmillan Publishers, 2011.

68. *Zha Sh., Peng X., Cao H., Zhuang X., Natarajan P., Natarajan P.* Text Classification via iVector Based Feature Representation // 2014 11th IAPR International Workshop on Document Analysis Systems. IEEE, 2014.

69. *Dehak N., Torres-Carrasquillo P.A., Reynolds D., Dehak R.* Language Recognition via Ivectors and Dimensionality Reduction // Proceedings of Interspeech 2011. Florence: International Speech Communication Association, 2011.

70. *Glembek O., Burget L., Matejka P.* Voice Biometry Standard – Draft. Brno: Speech@FIT, 2015.

71. *Nouri J., Yangarber R.* Measuring Language Closeness by Modeling Regularity // Proceedings of the EMNLP'2014 Workshop: Language Technology for Closely Related Languages and Language Variants. Doha: 2014.

72. Вавилонская Башня: Проект этимологической базы данных. *http://starling.rinet.ru/* Piekļūts 2016. gada 7. martā.

73. *Cilibrasi R., Vitányi P.M.B.* Clustering by compression // IEEE Transactions on Information Theory. Toronto: IEEE, 2005.

74. *Miao Y., Zhang H., Metze F.* Towards speaker adaptive training of deep neural network accoustic models // Proceedings of 15th Annual Conference International Speech Community Associtation. Singapore: Interspeech, 2014.

75. *Yao K., Yu D., Seide F., Su H., Deng L., Gong Y.* Adaptation of context-dependent deep neural networks for automatic speech recognition // Proceedings of the Spoken Language Technology Workshop. Miami: SLT, 2012.

76. *Saon G., Soltau H., Nahamoo D., Picheny M.* Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors // Proceedings of Automatic Speech Recognition and Understanding (ASRU) Workshop. Olomouc: IEEE, 2013.

77. *Trumpa E.* Latviešu ģeolingvistikas etīdes. R.: Zinātne, 2012.

78. *Садыхов Р.Х., Ракуш В.В.* Модели гауссовых смесей для верификации диктора по произвольной речи // Доклады БГУИР, № 4/2003. Минск: Белорусский государственный университет информатики и радиоэлектроники, 2003.

79. *Dobkeviča M.* Varbūtību teorijas un matemātiskās statistikas elementi. Daugavpils: RTU DF, 2004.

80. *Davis S., Mermelstein P.* Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, Vol. 28, No. 4, pp. 357-366.

81. *Reynolds D.* Universal Background Models. MIT Lincoln Laboratory. *https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics_UBM.pdf* Piekļūts 2016. gada 21. septembrī.

82. *Reynolds D., Quatieri T., Dunn R.* Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing 10, pp. 19-41, 2000.

83. *Smyth P.* The EM Algorithm for Gaussian Mixtures. // Probabilistic Learning: Theory and Algorithms. Irvine: University of California. *http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf* Piekļūts 2016. gada 21. septembrī.

84. *Reynolds D.* Gaussian Mixture Models. MIT Lincoln Laboratory. *https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf* Piekļūts 2016. gada 21. septembrī.

85. *Chang W., Cathcart Ch., Hall D., Garrett A.* Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis // Language, 2015, Vol. 91, No. 1, pp. 194-244.

86. *Hennig W.* Grundzüge einer Theorie der phylogenetischen. Berlin: Deutscher Zentralverlag, 1950.

87. *Hennig W.* Phylogenetic Systematics // Annual Review of Entomology, 1965, Vol. 10, pp. 97-116.

88. *Булатова Л.Н., Касаткин Л.Л., Строганова Т.Ю.* О русских народных говорах. М.: Просвещение, 1975.

89. *Zinkevičius Z.* Lietuvių kalbos tarmės. Kaunas: Šviesa, 1968.

90. *Chapman W.H., Olsen E., Lowe I., Andersson G.* Introduction to Practical Phonetics. Horsleys Green: Summer Institute of Linguistics, 1989.

91. *Кочерган М.П.* Вступ до мовознавства: Підручник для студентів філологічних спеціальностей вищих закладів освіти. - К.: Видавничий центр Академія, 2001.

92. *Scupin R.* Cultural Anthropology: A Global Perspective. Boston: Pearson, 2012.

93. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. Том 163, вып. 4, стр. 845-848. М.: Наука, 1965.

94. *Винцюк Т.К.* Распознавание слов устной речи методами динамического программирования // Кибернетика. Вып. 1, стр. 81-88. Киев: Наукова думка, 1968.

95. *Wagner R.A., Fischer M.J.* The string to string correction problem // Journal of the ACM. New York: ACM, 1974. Vol. 21, no. 1, pp. 168–173.

96. *Kaufman L., Rousseeuw P. J.* Finding Groups in Data – An Introduction to Cluster Analysis. New Jersey: Wiley, 1990.

97. *Van Noord G.* TextCat.
http://www.let.rug.nl/~vannoord/TextCat/
Piekļūts 2004. gada 10. oktobrī.

98. *Han J., Kamber M., Pei J.* Data Mining: Concepts and Techniques. 3rd Edition. Waltham: Morgan Kaufmann / Elsevier, 2012.

99. *Drgas Sz., Dąbrowski A.* Generalized cosine similarity in I-vector based automatic speaker recognition systems // Signal Processing: Algorithms, Architectures, Arrangements, and Applications. Poznan: IEEE, 2013. Pp. 73-77.

100. *Bai Zh., Zhang X.-L., Chen J.* Cosine Metric Learning for Speaker Verification in the i-Vector Space // Interspeech 2018. Hyderabad: 2018. Pp. 1126-1130.

101. *Берзиньш А.У.* Сравнение балтийских языков методом n-грамм // Труды международной коференции «Корпусная лингвистика - 2004». СПб.: Издательство С.-Петербургского университета, 2004.

102. *Berzinch A.A.* La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes // Identification des langues et des

variétés dialectales par les humains et par les machines. Paris: École National Supérieure des Télécommunications, 2004.

103. *Берзинь А.У.* Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, 2006.

104. *Bērziņš A.A., Grigorjevs J.* Latviešu izloksnēs sastopamo fonēmu telpa // Linguistica Lettica XVIII, R.: Latviešu valodas institūts, 2008.

105. *Берзинь А.* Возможности применения статистических методов распознавания речи для определения близости языков // Прикладна лінгвістика та лінгвістичні технології, Megaling-2009. Київ: «Довіра», 2009.

106. *ბერზინი ა.* ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის / Принципы сбора информации для автоматизированного анализа фонограмм // ქართული ენა და თანამედროვე ტექნოლოგიები - 2011. თბილისი: „მერიდიანი“, 2011.

107. *Берзинь А.У.* Применение распознавателей фонем для автоматического определения уровня близости языков // Труды международной конференции «Диалог 2016». М., 2016.

108. *Bērziņš A.A.* Usage of HMM-based Speech Recognition Methods for Automated Determination of a Similarity Level between Languages // AINL Proceedings. «Springer», 2019.

109. *Berziñch A.A., Chavarría Amezcua M.A.* El espacio de los alófonos del español. Iesniegts publicēšanai «Lengua y Habla», 2020.

110. *Павлинов И.Я.* Введение в современную филогенетику (кладогенетический аспект). М.: изд-во КМК, 2005.

111. *Šimko J., Suni A., Hiovain K., Vainio M.* Comparing Languages Using Hierarchical Prosodic Analysis // Proceedings Interspeech 2017. Stockholm: 2017. Pp. 1213-1217.

112. Ekonomikas skaidrojošā vārdnīca. *Sast. aut. kol. R. Grēviņas vadībā.* R.: Zinātne, 2000.

113. Valodniecības pamatterminu skaidrojošā vārdnīca. *Atb. red. V. Skujiņa.* R.: LU Latviešu valodas institūts, 2007.

114. LVS ISO 5127:2005. Informācija un dokumentācija. Vārdnīca. Informācijas zinātnes termini (bibliotēkas, arhīvi un muzeji). R.: 2005.

115. Естественный язык.
http://ru.wikipedia.org/wiki/Естественный_язык
Piekļūts 2019. gada 14. septembrī.

116. *Gay K.M.* Recent Advances and Issues in Computers. Phoenix, Arizona: Oryx Press, 2000.

117. Demogrāfija 2018, statistisko datu krājums. R.: Centrālā statistikas pārvalde, 2018.

118. *Mehl M.R., Vazire S., Ramírez-Esparza N., Slatcher R.B., Pennebaker J.W.* Are Women Really More Talkative Than Men? // Science, No. 317 (5832). Washington: American Association for the Advancement of Science, 2007.

119. *Liberman M.* Sex-Linked Lexical Budgets // Language Log. 2006/2007.
http://itre.cis.upenn.edu/~myl/languagelog/archives/003420.html
Piekļūts 2019. gada 15. septembrī.

120. *Čmejrková S.* The (Re)Presentation Of The Author In Czech And Slovak Scientific Texts // Jezik in slovstvo, Vol. 52 , Issue 3–4. Ljubljana: Zveza društev Slavistično društvo Slovenije, 2007.

121. *Жеребило Т.В.* Словарь лингвистических терминов: Изд. 5-е, испр. и дополн. — Назрань: Изд-во «Пилигрим», 2010.

122. *Sarkar A., Matrouf D., Bousquet P.-M., Bonastre J.-F.* Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification // Interspeech Proceedings. Portland, 2010.

123. *Громова Н.М., Громова Н.И.* Основы экономического прогнозирования. Учебное пособие. Старая Русса: Старорусский политехнический колледж, 2007.

124. *Markovičs Z.* Ekspertu novērtējuma metodes. R.: RTU izdevniecība, 2009.

125. Экспертное оценивание.
http://ru.wikipedia.org/wiki/Экспертное_оценивание
Piekļūts 2019. gada 31. oktobrī.

126. *Бешелев С.Д., Гурвич Ф.Г.* Экспертные оценки. М.: «Наука», 1973.

127. *Sleator D.D., Tarjan R.E., Thurston W.P.* Rotation Distance, Triangulations, and Hyperbolic Geometry // Journal Of The American Mathematical Society. Volume 1. Number 3. July 1988.

128. *Duda J.* Practical estimation of rotation distance and induced partial order for binary trees. Cornell University, 2016.

129. *Chen Y.J., Chang J.M., Wang Y.L.* An efficient algorithm for estimating rotation distance between two binary trees // International Journal of Computer Mathematics.
Vol. 82, No. 9, September 2005, Taylor & Francis.

130. *Dehornoy P.* On the rotation distance between binary trees // Advances in Mathematics. No. 223. Elsewier: 2010.

131. *Caspersen K.M., Madsen M.B., Eriksen A.B., Thiesson B.* A Hierarchical Tree Distance Measure for Classification // Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Porto: “SCITEPRESS”, 2017.

132. *Полегенько А.Ф., Князский О.В.* Оценка относительной компетентности экспертов в экспертной группе с использованием матриц парных сравнений // Озброєння та військова техніка. № 3. Київ: Центр. НДІ озброєння та військ. техніки ЗС України, 2014.

133. *Бурков Е.А.* Определение компетентности экспертов на основе поставленных ими оценок // Известия СПбГЭТУ «ЛЭТИ». № 4. Санкт-Петербург, 2009.

134. *Берзинь А.У.* Применение i-векторов для автоматизированного определения уровня близости языков // Труды Института системного программирования РАН. Том 31, № 5. М.: ИСП РАН, 2019.

135. *Bērziņš A.A.* Automated Comparison of Natural Languages – Software and Datasets of the Dissertation (Thesis). "Zenodo", 2019.
http://doi.org/10.5281/zenodo.3527981
Piekļūts 2019. gada 4. novembrī.

136. Preliminary recommendations on Corpus Typology. EAGLES – Expert Advisory Group on Language Engineering Standards Guidelines, 1996.
http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html
Piekļūts 2019. gada 5. novembrī.

137. *Maia B.* What are comparable corpora? // Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives. Lancaster: Saarland University, 2003.

138. Comparable Corpora // MT Research Survey Wiki. University of Edinburgh.
http://www.statmt.org/survey/Topic/ComparableCorpora
Piekļūts 2019. gada 5. novembrī.

139. Similarity (State of the art) // ACL Wiki for Computational Linguistics. The Association for Computational Linguistics.
https://aclweb.org/aclwiki/Similarity_(State_of_the_art)
Piekļūts 2019. gada 6. novembrī.

140. *Nissani M.* Fruits, Salads, and Smoothies: A Working Definition of Interdisciplinarity // The Journal of Educational Thought (JET) / Revue de la Pensée Éducative. Vol. 29, No. 2. Calgary: Werklund School of Education, University of Calgary, 1995.

141. EURAB report makes recommendations to promote interdisciplinary research // CORDIS, EU research results. European Commission, 2004.
https://cordis.europa.eu/article/rcn/21983/en
Piekļūts 2019. gada 6. novembrī.

142. Par prioritārajiem virzieniem zinātnē 2018.–2021. gadā. Kopsavilkums. Izglītības un zinātnes ministrijas Augstākās izglītības, zinātnes un inovāciju departaments, 2017.

143. *Тимофеев К.А.* Религиозная лексика русского языка как выражение христианского мировоззрения: учебное пособие. Новосибирск, 2001.

144. Muzeoloģijas terminu vārdnīca. R.: Latvijas Muzeju asociācija, 1997.

145. *Balodis M.* Kurzemes cietoksnis // Austrālijas latvietis. Nr. 2855. Haknija, 25.VII.2007.

146. *Gārša A.* Minoritātes Latvijā vēsturiskā skatījumā // Brīvā Latvija. Nr. 22. 11.VI.2011.

147. Solomon Kullback.
https://en.wikipedia.org/wiki/Solomon_Kullback
Piekļūts 2020. gada 8. februārī.

148. *Даль В.И.* Толковый словарь живаго великорускаго языка. Томъ второй. И – О. С.-Петербургъ/Москва: Изданіе книгопродавца-типографа М. О. Вольфа, 1881.

149. Moisejs Kuļbaks.
https://timenote.info/lv/Moisejs-Kulbaks
Piekļūts 2020. gada 8. februārī.

150. *Passricha V., Aggarwal R.K.* Convolutional Neural Networks for Raw Speech Recognition // From Natural to Artificial Intelligence: Algorithms and Applications. IntechOpen, 2018.

151. *Dauphin Y.N., Fan A., Auli M., Grangier D.* Language Modeling with Gated Convolutional Networks // Proceedings of the 34th International Conference on Machine Learning. Volume 70, August 2017.