Liepājas Universitāte

INGA ZNOTIŅA

OTRĀS BALTU VALODAS APGUVĒJU KORPUSS: IZVEIDES METODOLOĢIJA UN LIETOJUMA IESPĒJAS

Promocijas darbs
filoloģijas doktora grāda iegūšanai
valodniecības zinātņu nozares lietišķās valodniecības apakšnozarē

Zinātniskā vadītāja
vadošā pētniece ILZE AUZIŅA,
Dr. philol.

Doktora studiju attīstība Liepājas Universitātē
Vienošanās Nr.2009/0127/1DP/1.1.2.1.2./09/IPIA/VIAA/018

Liepāja 2017

Saturs

Ievads

Viena no visstraujāk augošajām valodniecības apakšnozarēm ir korpusa lingvistika. Tās centrā ir korpuss jeb valodas paraugu kopums, kas ir atlasīts pēc noteiktiem kritērijiem un piemērots datorizētai apstrādei un analīzei. Tā kā datori sniedz iespēju apstrādāt lielu informācijas apjomu īsā laikā, korpusa lingvistika ļauj izvērsti analizēt valodu, izmantojot autentiskus tekstus: ātri veikt dažādus mērījumus un aprēķinus, noteikt dažādu tekstu veidu īpatnības, tās salīdzināt ar citos tekstos konstatētajām utt.

Attīstoties korpusa lingvistikai, veidojas jauni tās atzari, un par vienu no tādiem var uzskatīt valodas apguvēju korpusu veidošanu un izpēti. Šādi korpusi satur noteiktas valodas apguvēju producētus valodas paraugus mērķvalodā. Valodas apguvēju korpusi tiek arvien plašāk izmantoti, pētot valodu apguvi, kļūdas un dzimtās valodas ietekmi uz tām. Tomēr šādi korpusi ir radīti lielākoties tikai pasaulē izplatītākajām valodām: angļu, spāņu, ķīniešu u. c. Tādi ir izveidoti arī Latvijā un Lietuvā (Vinčela 2010a, Rutenberga 2012, Juknevičienė 2013a, Grigaliūnienė u. c. 2008 u. c.).

Maz izplatītos „dzimtās valodas : apgūstamās valodas" pāros valodas apguvēju korpusu ir maz, un, tādus veidojot, nepieciešams atrisināt dažādus metodoloģiskus jautājumus, piem., tekstu atlases, kļūdu anotēšanas un autortiesību jomā. Līdz ar to šī promocijas darba **iecere** ir izveidot publiski pieejamu otrās baltu valodas apguvēju korpusu. Ar otro baltu valodu baltistikā parasti saprot latviešu valodu, ja subjekta dzimtā valoda ir lietuviešu valoda, un lietuviešu valodu, ja subjekta dzimtā valoda ir latviešu[1] (Butkus 2008, 57). Tātad korpuss paredzēts, lai pētītu latviešu valodas kā otrās baltu valodas un lietuviešu valodas kā otrās baltu valodas apguves sākumposmu no lingvistiska skatpunkta.

Promocijas darba **novitāti** nosaka tas, ka, valodu tehnoloģiju pētniecībai Latvijā veiksmīgi attīstoties, tā joprojām ir visnotaļ fragmentēta un daļai potenciālo darba virzienu netiek pievērsts pietiekami daudz uzmanības (Skadiņa u. c. 2014, 232). To var attiecināt arī uz valodas apguvēju korpusiem. Šis ir pirmais publiski pieejamais baltu valodu apguvēju korpuss, kā arī pirmais publiski pieejamais valodas apguvēju korpuss Latvijā.

Darbs pie otrās baltu valodas apguvēju korpusa izveides ir netipisks vairākos aspektos. Viens no tiem – šis ir (nosacīti) divvirzienu korpuss: tajā ir iekļauti nevis vienas, bet divu valodu apguvēju radītie teksti, ar nosacījumu, ka abos valodu pāros iekļautās valodas ir vienas un tās pašas, tikai katra no tām vienā pārī ir apguvēja dzimtā valoda, bet otrā pārī – mērķvaloda.

[1] Lai gan pēdējā laikā zināmu popularitāti ir guvusi arī prūšu valoda, tā tomēr galvenokārt tiek uzskatīta par mirušu valodu (Blažek 2007, 100), un lielāka uzmanība tiek pievērsta dzīvo baltu valodu apguvei.

Nav zināms, ka citi pētnieki jebkur pasaulē kādā valodu pārī būtu strādājuši ar divvirzienu materiālu saturošu valodas apguvēju korpusu, tāpēc var pieņemt, ka šis ir, ja ne pirmais, tad vismaz viens no pirmajiem tāda veida darbiem. Korpusa lingvistikā kopumā gan tas nav gluži jaunums – pētnieki daudzviet, tai skaitā arī Baltijas valstīs, strādā ar *divvirzienu paralēlajiem korpusiem*, proti, tulkojumu korpusiem, kuros iekļauti tulkojumi no A valodas B valodā līdz ar tulkojumiem no B valodas A valodā (Johansson 2007, 302; Rimkutė u. c. 2013, 73). Tulkojumos, līdzīgi kā valodu apguvē, nozīmīga loma ir avotvalodas ietekmei uz mērķvalodu (Sīlis 2009, 29–35), nereti tiek runāts arī par interferenci (Zauberga 2001), un divvirzienu pētījumi var ne vien atklāt praksē nostiprinājušās ietekmes tendences, bet arī sniegt savu artavu divu valodu sistēmu kontrastīvā analīzē (piem., norādāmo vietniekvārdu salīdzinājumu latviešu un lietuviešu valodā tulkojumu aspektā, sk. Znotiņa 2012). Domājams, ka līdzīgu labumu varētu gūt arī no valodas apguvēju korpusiem ar atbilstošu (t. i., divvirzienu) materiālu.

Pētījuma **aktualitāti** pamato fakts, ka valodas apguvēju korpusi ir populāri visā pasaulē, un jaunu korpusu izveide tām valodām, kurām vēl tādu nav, ļauj papildināt valodas apguves pētījumu atziņas un veidot pilnīgāku ainu globālā kontekstā. Tas ir aktuāls arī lokālā aspektā: otrās baltu valodas apguvēju korpuss palīdz noskaidrot galvenās grūtības otrās baltu valodas apguvē, dodot ierosmi mācību materiālu pilnveidošanai, kā arī abu valodu salīdzinošiem pētījumiem.

Promocijas pētījuma **problēma** ietver dažādu teorētisku un praktisku, t. sk. juridisku, metodoloģisku un tehnisku jautājumu risinājuma nepieciešamību publiski pieejama otrās baltu valodas apguvēju korpusa izveides nolūkā.

Pētījuma ieceres un problēmas ietekmē promocijas darba **temats** ir „Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas", savukārt tā **priekšmets** – nepieciešamās darbības otrās baltu valodas apguvēju korpusa izveidei un metodes šāda korpusa materiāla izpētē.

Pētījuma **jautājums** ir – kā veidot un izmantot otrās baltu valodas apguvēju korpusu?

Promocijas darba **mērķis** ir divpusīgs. Tas ir – izveidot valodas apguvēju korpusu, balstoties uz otrās baltu valodas apguvēju (latviešu, kas apgūst lietuviešu valodu, un lietuviešu, kas apgūst latviešu valodu) patstāvīgi rakstītajiem tekstiem apgūstamajā valodā un izstrādājot šī mērķa sasniegšanai nepieciešamo metodoloģiju. No otras puses, tas ir arī raksturot publiski pieejama otrās baltu valodas korpusa izveidi, pamatojot būtiskākās izvēles, un noskaidrot šī korpusa lietojuma iespējas pētniecībā.

Lai sasniegtu darba mērķi, izvirzīti šādi **uzdevumi**:

1) raksturot valodas apguvēju korpusa jēdzienu, veidus un izveides soļus;

2) savākt korpusā iekļaujamos tekstus, nodrošinot autortiesību ievērošanu un personas datu aizsardzību;
3) marķēt un raksturot korpusā iekļaujamos tekstus;
4) nodrošināt tehnisko atbalstu korpusa veiktspējai un publiskai pieejamībai;
5) izveidot vai pielāgot esošas korpusa anotēšanas sistēmas;
6) anotēt korpusu;
7) aprakstīt korpusa būtiskāko iespēju izmantošanu.

Promocijas darba izstrādē ir izmantotas vairākas **metodes**. Valodas apguvēju korpusu vēstures un izplatības raksturojums galvenokārt ir balstīts zinātniskās literatūras deskriptīvajā analīzē. Savukārt valodas apguvēju korpusu būtība un lietojums skaidrots ar zinātniskās literatūras referatīvās analīzes palīdzību.

Raksturojot populārās metodes, kas tiek lietotas darbā ar valodas apguvēju korpusiem, izmantota metožu salīdzināmā analīze, kas palīdz izvēlēties, kuras no tām uzskatīt par primāri svarīgām otrās baltu valodas apguvēju korpusa gadījumā. Kad izvēle ir izdarīta, ar analītiski sintētisko metodi ir izveidots pats otrās baltu valodas apguvēju korpuss. Korpusam dots nosaukums „Esam", un šī nosaukuma pamatā ir savstarpēji līdzīgās darbības vārdu la. *būt* un lie. *būti* 'būt' pirmās personas daudzskaitļa formas: la. *esam*, lie. *esame* (lie. sarunvalodā arī saīsināta forma – *esam*). Vārda leksiskā un gramatiskā nozīme uzsver baltu valodu kopību, kā arī korpusā iekļauto tekstu autoru, korpusa veidotāju un zināmā mērā arī korpusa lietotāju piederību baltu kultūrtelpai.

Promocijas darba noslēgumā rādīts, kā jaunradīto korpusu var izmantot pētniecībā. Šajā nodaļā uzmanība īpaši tiek pievērsta divām metodēm: kontrastīvai starpvalodas analīzei un datorizētai kļūdu analīzei.

Par darba **teorētisko pamatu** uzskatāmas valodas apguvēju korpusu pētnieku publikācijas – Karinas Aijmeres (*Karin Aijmer*), Maikla Bārlova (*Michael Barlow*), Silvianes Greindžeres (*Sylviane Granger*), Ankes Līdelingas (*Anke Lüdeling*), Fanijas Menjē (*Fanny Meunier*), Utes Rēmeres (*Ute Römer*), Jukio Tono (*Yukio Tono*), kā arī Baltijas valstu pārstāvju Linas Bikelienes (*Lina Bikelienė*), Nidas Burneikaites (*Nida Burneikaitė*), Jones Grigaļūnienes (*Jonė Grigaliūnienė*), Ritas Juknevičienes (*Rita Juknevičienė*), Vinetas Rūtenbergas, Zigrīdas Vinčelas u. c. darbi; esošo valodas apguvēju korpusu – *ICLE, Falko, Lindsei* u. c. izstrādes informācija un tehniskais apraksts; literatūra par dažādu korpusu izveidi, marķēšanu un anotēšanu, sevišķi attiecībā uz baltu valodām – Duglasa Baibera (*Douglas Biber*), Silvijas Bernardini (*Silvia Bernardini*), Džefrija Līča (*Geoffrey Leech*), Tonija Makenerija (*Tony McEnery*), Aleksandra Rosena (*Alexandr Rosen*), Džona Sinklēra (*John Sinclair*), kā arī

Latvijas un Lietuvas pārstāvju Kristīnes Levānes-Petrovas, Rūtas Marcinkevičienes (*Rūta Marcinkevičienė*) Gunas Rābantes-Bušas, Erikas Rimkutes (*Erika Rimkutė*), u. c. darbi; baltu valodu apguves pētījumi un lingvodidaktikas terminoloģijas avoti – Inetas Dabašinskienes (*Ineta Dabašinskienė*), Ivetas Grīnbergas, Vitas Kalnbērziņas, Deivida Kristala (*David Crystal*), Ingas Laizānes, u. c. darbi; korpusa lingvistikas terminoloģijas avoti latviešu, lietuviešu un angļu valodā – Pola Beikera (*Paul Baker*), Normunda Grūzīša, Endrū Hārdija (*Andrew Hardie*), Anitas Helvigas, Andreja Spektora, Andrjus Utkas (*Andrius Utka*) u. c. darbi, kā arī spēkā esošie normatīvie akti, kas regulē autortiesību un personas datu aizsardzības jautājumus Latvijas Republikā un Lietuvas Republikā.

Ir vispārpieņemts uzskats, ka dalīt korpusus „derīgos korpusos" un „nederīgos korpusos" nav iespējams, jo „tas, kā korpuss tiek veidots, ir atkarīgs no tā, kāda veida korpuss tas ir un kā to ir iecerēts izmantot" (Hunston 2008, 155). Līdz ar to ne mazāk svarīgi par korpusa izveidi ir saprast, ko no tā var gaidīt, bet ko – ne, kādus secinājumus pēc tajā veikto pētījumu rezultātiem var izdarīt, bet kādus – ne. Tāpēc ikvienā pētījumā, kurā ir izmantots korpuss, ir būtiski ņemt vērā korpusa un tajā iekļautā materiāla veidu, īpatnības, apjomu u. c. faktorus. Promocijas darba **materiāls** un pētījuma **avoti** ir otrās baltu valodas apguvēju patstāvīgi rakstītie teksti apgūstamajā valodā. To autori tekstu tapšanas brīdī ir studenti augstākās izglītības iestādēs. Šos tekstus no studentiem ir saņēmuši un tālāk korpusa veidošanai izsnieguši viņu otrās baltu valodas pasniedzēji četrās universitātēs: Latvijas Universitātē, Liepājas Universitātē, Viļņas Universitātē un Vītauta Dižā universitātē.

Promocijas darba **struktūru** veido ievads, divas daļas ar nodaļām un apakšnodaļām, nobeigums un secinājumi, promocijas darba aizstāvēšanai izvirzītās tēzes un bibliogrāfija.

Darba pirmajā daļā ir trīs nodaļas. Pirmajā nodaļā raksturots valodas apguvēju korpusa termins, jēdziens un definīcija, kā arī sniegta informācija par valodas apguvēju korpusu veidiem. Otrajā nodaļā aprakstīta valodas apguvēju korpusu vēsture un izplatība, koncentrējoties galvenokārt uz Latviju un Lietuvu. Trešajā nodaļā skaidrots, ar kādām metodēm un procedūrām tiek pētīti valodas apguvēju korpusi un kāds ir to izmantojums valodas apguvē.

Darba otrajā daļā raksturota otrās baltu valodas apguvēju korpusa izveide. Pirmajā nodaļā runāts par avotu atlasi, ieskaitot ne vien tekstu ieguvi, bet arī personas datu aizsardzības un autortiesību jautājumus. Otrajā nodaļā aprakstīta korpusa tekstu marķēšana un anotēšana, savukārt trešajā nodaļā skaidrots, ar kādu programmatūru korpuss ir veidots un ar kādu to ir paredzēts izmantot. Ceturtajā nodaļā sniegts īss apraksts ar piemēriem par šī korpusa izmantošanu konkrētu metožu kontekstā.

Darbam ir pievienots izmantotās literatūras saraksts (342 vienības) un 6 pielikumi: atļauju paraugi, marķētu un anotētu tekstu paraugi u. c. Darba saturu palīdz atspoguļot 9 tabulas un 16 attēli.

Promocijas darbs ir **aprobēts** 22 referātos dažādās vietēja un starptautiska mēroga zinātniskās konferencēs un semināros Latvijā (Liepājā, Ventspilī, Rīgā, Daugavpilī, Jelgavā), Lietuvā (Viļņā, Kauņā), Igaunijā (Tartu) un Lielbritānijā (Šefīldā, Lenkasterā):

1. Inga Znotiņa. „Valodas apguvēju korpuss: lietuviešu un latviešu termins un definīcija." Liepājas Universitātes 18. starptautiskā zinātniskā konference *Vārds un tā pētīšanas aspekti*. Liepājā 2013. gada 28.–29. novembrī.

2. Inga Znotiņa. „Learner corpora in Latvia and Lithuania." Tartu Universitātes 8. starptautiskā zinātniskā konference *Native Language and Other Languages*. Tartu 2013. gada 28.–29. novembrī.

3. Inga Znotiņa. „Learner corpus research methods and requirements for the corpora used." Tartu Universitātes, Igauņu valodas institūta un Tallinas Universitātes organizētā starptautiskā konference *Mapping Methods: Approaches to Language Studies*. Tartu 2014. gada 8.–10. maijā.

4. Inga Znotiņa. „Valodas apguvēju korpusi Latvijā un Lietuvā." Latvijas Universitātes, Liepājas Universitātes un Ventspils Augstskolas valodniecības nozares doktorantu zinātniskais seminārs. Rīgā 2014. gada 16.–17. maijā.

5. Inga Znotiņa. „Pētniecības iespējas nemarķētā baltu valodu apguvēju korpusā." Liepājas Universitātes 19. starptautiskā zinātniskā konference *Vārds un tā pētīšanas aspekti*. Liepājā 2014. gada 27.–28. novembrī.

6. Inga Znotiņa. „Error-tagging a learner corpus of Baltic languages." Šefīldas Universitātes ikgadējā lingvistikas studentu konference *ShefLing PGC*. Šefīldā 2015. gada 5.–6. martā.

7. Inga Znotiņa. „Valodas apguvēju korpusa anotēšanas veidi." Ventspils Augstskolas un Liepājas Universitātes 3. starptautiskā zinātniskā konference *Via scientiarum*. Ventspilī 2015. gada 12.–13. martā.

8. Inga Znotiņa. „Semantiski tuvu leksēmu apguves izpēte neanotētā valodas apguvēju korpusā." Rīgas Stradiņa universitātes ikgadējā zinātniskā konference. Rīgā 2015. gada 26.–27. martā.

9. Inga Znotiņa. „Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība.” Liepājas Universitātes 20. starptautiskā zinātniskā konference *Vārds un tā pētīšanas aspekti*. Liepājā 2015. gada 3.–4. decembrī.

10. Inga Znotiņa, Daiva Puškorjute-Riduliene. „Kļūdu anotēšana otrās baltu valodas apguvēju korpusā.” Liepājas Universitātes 20. starptautiskā zinātniskā konference *Vārds un tā pētīšanas aspekti*. Liepājā 2015. gada 3.–4. decembrī.

11. Inga Znotiņa. „Besimokančiųjų tekstynų anotavimas Lietuvoje ir Latvijoje.” Vītauta Dižā universitātes Svešvalodu institūta un Lietuvas valodu pedagogu asociācijas organizētā starptautiskā zinātniskā konference *Darnioji daugiakalbystė: kalba, kultūra, visuomenė*. Kauņā 2015. gada 29.–30. maijā.

12. Inga Znotiņa. „Learner corpus *Esam*: a new corpus for researching Baltic interlanguage.” Lenkasteras Universitātes organizētā astotā starptautiskā zinātniskā konference *Corpus Linguistics*. Lenkasterā 2015. gada 21.–24. jūlijā.

13. Inga Znotiņa. „Lemmatization in a beginner learner corpus.” 3. Baltijas valstu studentu starptautiskā zinātniskā konference *Bridges in the Baltics*. Viļņā 2015. gada 2.–3. oktobrī.

14. Inga Znotiņa. „Kļūdu klasifikācija otrās baltu valodas apguvēju tekstos.” XII starptautiskais baltistu kongress. Viļņā 2015. gada 28.–31. oktobrī.

15. Inga Znotiņa, Daiva Puškorjute-Riduliene. „*Mans* un *savs* baltu valodās: vietniekvārdu kļūdu klasifikācijas problēmas”. Daugavpils Universitātes 26. starptautiskā zinātniskā konference *Zinātniskie lasījumi*. Daugavpilī 2016. gada 28.–29. janvārī.

16. Inga Znotiņa. „Publiski pieejama valodas apguvēju korpusa izveide: programmatūras meklējumi.” Latvijas Universitātes, Liepājas Universitātes un Ventspils Augstskolas valodniecības nozares doktorantu zinātniskais seminārs. Rīgā 2016. gada 20. maijā.

17. Inga Znotiņa. „Publiski pieejama valodas apguvēju korpusa izveide: programmatūras meklējumi.” Latvijas Universitātes, Liepājas Universitātes un Ventspils Augstskolas valodniecības nozares doktorantu zinātniskais seminārs. Rīgā 2016. gada 20. maijā.

18. Inga Znotiņa. „Vārdšķiru un pamatformu noteikšana otrās baltu valodas korpusā: problemātiskie gadījumi latviešu valodā.” Daugavpils Universitātes 27. starptautiskā zinātniskā konference *Zinātniskie lasījumi*. Daugavpilī 2017. gada 26.–27. janvārī.

19. Inga Znotiņa. „Otrās baltu valodas korpusa sintaktiska anotēšana.” Rīgas Stradiņa universitātes ikgadējā zinātniskā konference. Rīgā 2017. gada 6. –7. aprīlī

20. Inga Znotiņa, Inga Laizāne. „Otrās baltu valodas apguve Latvijas un Lietuvas augstākās izglītības iestādēs.” Ventspils Augstskolas un Liepājas Universitātes 3. starptautiskā zinātniskā konference *Via scientiarum*. Liepājā 2017. gada 7. –8. aprīlī.

21. Inga Znotiņa, Inga Laizāne. „Digitālie resursi baltu valodu apguvē." 14. starptautiskā zinātniskā konference *Valodu apguve: problēmas un perspektīva*. Liepājā 2017. gada 21. aprīlī.

22. Inga Znotiņa. „Computer-aided error analysis for researching Baltic interlanguage." 11th International Scientific Conference *Rural Environment, Education, Personality 2018*. Jelgavā 2017. gada 11.–12. maijā.

Atsevišķas pētījuma daļas un atziņas ir publicētas 6 zinātniskos rakstos un 11 konferenču tēžu un anotāciju krājumos.

Raksti:

1. Znotiņa, Inga. Valodas apguvēju korpuss: lietuviešu un latviešu termins un definīcija. *Vārds un tā pētīšanas aspekti* : rakstu krājums, 18 (2). Red. kolēģijas vadītāja Benita Laumane. Krājuma atb. red. Linda Lauze. Liepāja : LiePA, 2014. 265. –271. lpp.

2. Znotiņa, Inga. Learner corpus annotation in Latvia and Lithuania. *Sustainable Multilingualism,* No. 7. 2015, pp. 145–159.

3. Znotiņa, Inga. Pētniecības iespējas neanotētā baltu valodu apguvēju korpusā. *Vārds un tā pētīšanas aspekti* 19 (2), 2015, 208.–221. lpp.

4. Znotiņa, Inga. Otrās baltu valodas apguvēju korpusa morfoloģiska anotēšana. *Via Scientiarum* : starptautiskās jauno lingvistu konferences rakstu krājums. 3. laidiens. Sastādītājas I. Laizāne, I. Znotiņa. Ventspils, Liepāja : Ventspils Augstskola, Liepājas Universitāte, 2016, 148.–160. lpp.

5. Znotiņa, Inga. Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība. *Vārds un tā pētīšanas aspekti* 20 (2), 2016, 219.–227.lpp.

6. Znotiņa, Inga. Computer-aided error analysis for researching Baltic interlanguage. Rural Environment, Education, Personality. Proceedings of the 10th International Scientific Conference, 2017, pp. 238–244.

Tēzes un anotācijas:

1. Znotiņa, Inga. Valodas apguvēju korpusa anotēšanas veidi. *3. starptautiskā jauno lingvistu konference Via Scientiarum.* Konferences materiāli. Ventspils : Ventspils Augstskola, 2015.

2. Znotiņa, Inga. Learner corpora in Latvia and Lithuania [online]. *Native Language and Other Languages VIII.* Abstracts of the conference. Tartu : University of Tartu, 2013 [viewed

18 August 2014]. Available: http://emakeeljateisedkeeled.weebly.com/uploads/7/7/4/8/7748994/emakeel_ja_teised_keeled_viii_teesid_abstracts_201113.doc

3. Znotiņa, Inga. Learner corpus research methods and requirements for the corpora used [online]. *Mapping Methods: Approaches to Language Studies*. Abstracts of the conference. Tartu : University of Tartu, 2014 [viewed 18 August 2014]. Available: http://mappingmethods.eki.ee/images/docs/Znotina%20Inga%20 _abstract_MaMe_2014.pdf

4. Znotiņa, Inga. Error-tagging a learner corpus of Baltic languages [online]. *ShefLingPGC 2015*. Book of abstracts. Sheffield : The University of Sheffield, 2015 [viewed 16 March 2015]. Available: https://drive.google.com/file/d/0BxuHJWsEyyLLT2xxc1BhSFU4UVk/ view

5. Znotiņa, Inga. Semantiski tuvu leksēmu apguves izpēte neanotētā valodas apguvēju korpusā. *2015. gada zinātniskā konference.* Tēzes. Rīga : Rīgas Stradiņa universitāte, 2015, 428. lpp.

6. Znotiņa, Inga. Besimokančiųjų tekstynų anotavimas Lietuvoje ir Latvijoje. *Darnioji daugiakalbystė: kalba, kultūra, visuomenė*. Konferencijai priimtos santraukos. Kaunas : Vytauto Didžiojo universitetas, 2015. [viewed 23 October 2015]. Available: http://daugiakalbyste.vdu.lt/wp-content/uploads/docs/03/ abstracts/Zlotina.pdf

7. Znotiņa, Inga. Learner corpus *Esam*: a new corpus for researching Baltic interlanguage. *CL2015*. Abstract book. Lancaster : Lancaster University, 2015, pp. 447–448.

8. Znotiņa, Inga. Lemmatization in a beginner learner corpus. *Third Baltic Student Conference* Bridges in the Baltics. Abstracts. Vilnius : Vilnius University, 2015 [viewed 24 October 2015]. Available: http://www.keelekeskus.ut.ee/sites/default/files/maailmakeeled/parallel_session_abstracts_2015_bb.pdf

9. Znotiņa, Inga. Kļūdu klasifikācija otrās baltu valodas apguvēju tekstos. *XII starptautiskais baltistu kongress.* Referātu tēzes. Viļņa : Viļņas Universitāte, 2015 [skatīts 24.10.2015.], 280. lpp. Pieejams tiešsaistē: http://www.baltistukongresas.flf.vu.lt/failai/XII_Tarptautinio_baltistu_kongreso_tezes.pdf

10. Znotiņa, Inga. Valodas apguvēju korpusu izmantojums svešvalodas mācību procesā. *2016. gada zinātniskā konference.* Tēzes. Rīga : Rīgas Stradiņa universitāte, 2016, 326. lpp.

11. Znotiņa, Inga. Otrās baltu valodas korpusa sintaktiska anotēšana. *2017. gada zinātniskā konference.* Tēzes. Rīga : Rīgas Stradiņa universitāte, 2017, 326. lpp.

Promocijas darba izstrādi veicinājusi stažēšanās Kauņā Vītauta Dižā universitātes Datorlingvistikas centrā Lietuvā no 2013. gada 1. maija līdz 31. jūlijam. Stažēšanās laikā tika

pilnveidota promocijas darba teorētiskā bāze, kā arī risināti ar autortiesībām un personas datu aizsardzību saistītie jautājumi.

Promocijas darba izstrādē nepieciešamās zināšanas papildinātas Lenkasteras Universitātes Valodas korpusu datorizētas izpētes centra (*University Centre for Computer Corpus Research on Language, UCREL*) organizētajā korpuslingvistikas vasaras skolā Lenkasteras Universitātē, Apvienotajā Karalistē no 2014. gada 15. jūlija līdz 18. jūlijam.

1. Valodas apguvēju korpusi, to izveide un lietošana

Līdz ar datortehnoloģiju izplatības pieaugumu valodniecībā arvien plašāku popularitāti gūst korpusi. Latvijas Universitātes Latviešu valodas institūta izdotajā *Valodas pamatterminu skaidrojošajā vārdnīcā* norādīts, ka valodniecībā korpuss ir „datorizētai analīzei pieejams apjomīgs tekstu kopums" (VPSV 2007, 196), bet attiecīgo lietišķās valodniecības apakšnozari sauc par korpusa lingvistiku (VPSV 2007, 196). Korpusā ievietojamie teksti parasti tiek izvēlēti tā, lai pēc iespējas precīzāk atspoguļotu kādu komunikācijas veidu vai valodas paveidu (ELL 2005, 234). Tāpēc nereti tiek veidoti dažādi specializēti korpusi – korpusi, kuros ir iekļauti viena noteikta veida teksti (par to sīkāk Flowerdew 2004), pretstatā vispārīgajiem korpusiem, kuru mērķis ir pēc iespējas pilnīgāk atspoguļot valodu kopumā (McEnery u. c. 2006, 15). Par vienu no specializēto korpusu veidiem var uzskatīt valodas apguvēju korpusus (starp tādiem tos min, piem., Koester 2010).

1.1. Valodas apguvēju korpusa jēdziens un raksturojums

Izpratne par to, kas ir valodas apguvēju korpuss, dažādu pētnieku darbos visā pasaulē ir diezgan līdzīga, taču trūkst vienprātības par robežgadījumiem, tāpēc šajā nodaļā skaidroti izplatītākie viedokļi, kas daļēji arī rāda nozares attīstību Latvijā, Lietuvā un citviet pasaulē. Turklāt raksturotas tendences valodas apguvēju korpusu izveidē, kā arī šādu korpusu paveidi un klasifikācijas iespējas.

1.1.1. Valodas apguvēju korpuss: termins un definīcija

Visbiežāk sastopamā valodas apguvēju korpusa definīcija nosaka, ka tas ir „svešvalodas vai otrās valodas apguvēju producētu autentisku tekstu elektronisks kopums" (Granger 2003b, 538)[2]. Angļu valodā to parasti dēvē par *learner corpus*, atsevišķos gadījumos sinonīmiski lietojot arī vārdkopas *learner text corpus* un *test taker corpus* (Rutenberga 2012; López-Lago, Saiz de Lobado García 2011; Rūtenberga, Kalnbērziņa 2013) vai arī, veidojot nosaukumu nevis pēc tekstu autoriem, bet gan pēc to atspoguļotā valodas paveida, *interlanguage corpus* vai *L2 corpus* (Granger 2003a, 465). Vācu valodā parasti lieto terminu *das Lernerkorpus* (Lüdeling et al. 2008, Siemen u. c. 2006), bet krievu valodā – *учебный корпус текстов* (Камшилова 2013, 301; Мальцева 2011, 209).

[2] Šeit un turpmāk – mans tulkojums latviešu valodā – I. Z.

Minētajai definīcijai ir atrodami vairāki varianti. Piem., daļa pētnieku to formulē šādi: „Elektronisks tekstu korpuss, kuru autori ir personas, kas mācās svešvalodu." (Мальцева 2011, 209) Būtībā šī definīcija izsaka to pašu, ko iepriekš minētā. Līdzīgā veidā arī citos darbos valodas apguvēju korpusi dažkārt definēti nedaudz atšķirīgi, taču saglabājot to pašu definīcijas saturu: „Valodas apguvēju producētu tekstu vai valodas paraugu vākums." (LDLTAL 2010, 138); „Valodas apguvēju tipiski rakstveida snieguma vai atdevuma digitāls atainojums." (Barlow 2005, 335); „Principos balstīts apguvēju valodas vākums." (Lüdeling u. c. 2005, 1) u. c. Būtība definīcijā saglabājas nemainīga. Saskaņā ar to, valodas apguvēju korpusam svarīgi ir šādi parametri:

- tas ir elektronisks;
- to veido teksti (ar to saprotot gan mutvārdu, gan rakstveida valodas paraugus);
- tekstu autori ir noteiktas svešvalodas (dažkārt precizēts: svešvalodas vai otrās valodas) apguvēji.[3]

Termins šāda veida korpusu nosaukšanai dažādās valodās veidots pēc atšķirīgiem principiem.

Angļu valodas populārākais termins *learner corpus* ir veidots no divām daļām: *learner* 'skolēns; tas, kurš mācās' un *corpus*, ko latviešu valodnieki dēvējuši par *datorfondu, tekstu masīvu, elektronisko tekstu krājumu* (Spektors 2000), bet pēdējā laikā visbiežāk – par *korpusu* (VPSV 2007), ar šo vārdu veidojot arī jaunus terminus (Helviga 2012, 106).

Angļu valodā tiek runāts arī par *translational learner corpus* vai *learner parallel corpus*, kuru nozīme ir viena un tā pati – tas ir tulkošanas apguvēju radīto tulkojumu korpuss (Arhire 2011, tas kā valodas apguvēju korpusu paveids pieminēts arī Мальцева 2011, 209; Granger 2002, 10). Citu valodu terminiem šāds lietojums nav plaši izplatīts.

Par to, vai šāds korpuss arī ir uzskatāms par valodas apguvēju korpusu, domas var dalīties. Neapšaubāmi tam ir citāds raksturs nekā tiem valodas apguvēju korpusiem, kuros ir iekļautas, piem., esejas vai studiju noslēguma darbi, jo tulkojumu korpusā iekļautie teksti parāda nevis to autoru spēju veidot savu tekstu, bet gan spēju precīzi tulkot jau esošu tekstu. Ņemot vērā, ka tulkošana tiek uzskatīta par „komunikatīvu aktu specifiskā situācijas kontekstā" (Sīlis 2009, 29), var gan arī uzskatīt, ka šajā aspektā tā ir pielīdzināma dažādiem citiem komunikatīviem aktiem, kas izpaužas valodā, piem., stāstījums vai kaut kā mācīšana citam. Līdz ar to, ja stāstījuma teksti var veidot valodas apguvēju korpusu, tāpat tas būtu jāvar arī tulkojuma tekstiem. Turklāt pieredze liecina, ka, mācoties svešvalodu, teksti sākotnēji nereti

[3] Valodas apguvēju korpusam raksturīgie parametri sīkāk skaidroti šī darba 1.1.2. apakšnodaļā.

tiek nevis rakstīti uzreiz svešvalodā, bet gan sacerēti dzimtajā valodā un pēc tam iztulkoti. To var novērot, arī veidojot otrās baltu valodas apguvēju korpusu – daži tekstu autori pat kopā ar lietuvisko tekstu ir iesnieguši latvisko sākotnējo tekstu, no kura tas tulkots.

Šķiet, tulkojumu korpusus tomēr nevajadzētu ierindot starp valodas apguvēju korpusiem arī tad, ja tajos iekļautos tekstus radījuši tulkotāji, kas šo prasmi mācās (Bowker, Bennison 2003 u. c.). Tulkojumi gan atklāj tulkotāja mērķvalodas prasmi, tomēr tulkojuma mērķvaloda nereti ir tulkotāja dzimtā valoda (Castagnoli u. c. 2006), un šādā gadījumā runa ir drīzāk nevis par valodas apguvi, bet gan par ekvivalences meklēšanas prasmi divās vai vairākās valodās. Turklāt valodas apguvēju korpusā iekļauj svešvalodas, tātad nedzimtās valodas lietojuma paraugus.

Uzsverot nošķīrumu, dažkārt lieto terminu *learner language corpus*, ar vārdu *language* ‘valoda’ pretstatot to citiem apguvēju korpusiem, šajā gadījumā – *learner translation corpus* jeb „tulkošanas apguvēju korpusam” (Bernardini u. c. 2003, 7).

Vācu valodā tiek lietots termins *das Lernerkorpus* (Lüdeling et al. 2008), un tā nozīme ir tāda pati kā angļu valodā: *der Lerner* nozīmē ‘tas, kurš mācās’, bet *das Korpus* – ‘korpuss’. Līdz ar to arī vācu valodā var veidoties iepriekš aprakstītās neskaidrības ar tulkojumu korpusiem (piem., sk. tulkojumu korpusa *MELLANGE* dokumentācijā[4]).

Krievu valodā ir divi līdzīgi termini, kuru nozīmes ir jāšķir: *учебный корпус текстов* un *обучающий корпус*. *Учебный корпус текстов* ir korpuss, kas sastāv no kādas valodas apguvēju radītiem tekstiem, tātad tas, ko angļu valodā sauktu par *learner corpus* (Камшилова 2013, 301; Мальцева 2011, 209). Savukārt *обучающий корпус* ir korpuss, kas paredzēts mērķvalodas apguvei un satur dzimtās valodas tekstus (Савчук, Сичинава 2009, 317). *Учебный корпус* burtiski varētu tulkot kā *mācību korpuss*[5], savukārt *обучающий корпус* saprotams kā ‘korpuss, kas māca; korpuss, ar kuru mācās’. Angļu un vācu valodas terminiem nav ekvivalenta krievu valodā **корпус учеников*.

Lietuviešu un latviešu valodā pagaidām vēl ir maz publikāciju par šo tēmu, tāpēc nevienā baltu valodā nav nostabilizējies viens termins. Lielākoties termini darināti pēc angļu un vācu, nevis krievu valodas parauga, proti, tiek nosaukta korpusā iekļaujamo tekstu autoru grupa.

Latviešu valodā šie korpusi dēvēti gan par *valodas apguvēju korpusiem*, gan par *valodas apguvēja korpusiem*, gan par *studentu korpusiem*. Publikāciju, kurās tie latviešu valodā

[4] Skatīts 18.08.2015. Pieejams tiešsaistē: *http://mellange.eila.jussieu.fr/public_doc.de.shtml*
[5] Sal.: *учебное заведение* – mācību iestāde, *учебная программа* – mācību programma.

pieminēti, gan ir maz. K. Aijmere sniedz *valodas apguvēju korpusa* definīciju, kas būtībā sakrīt ar iepriekš minēto, proti, tas ir elektronisku tekstu krājums, kurā iekļauti tikai valodas apguvēju radīti teksti, atlasīti pēc īpašiem kritērijiem un ar papildu informāciju par to izcelsmi u. c. (Aijmer 2011). Šis termins ir lietots arī Latvijas valodu skolotāju asociācijas (LVASA) pētījumā par latviešu valodas apguvi mazākumtautību skolās (Kalnbērziņa u. c. 2011, 16 u. c.). Tomēr šeit vērojama nekonsekvence – korpusa nosaukums ir *Latviešu valodas apguvēja* [izcēlums mans – I. Z.] *korpuss*, savukārt visur citur vārds *apguvējs* lietots daudzskaitlī. Turklāt salīdzinājumam minēts arī *dzimtās valodas lietotāju korpuss*, tātad atkal ar daudzskaitļa formu. Līdz ar to var pieņemt, ka pētnieces tomēr devušas priekšroku daudzskaitļa formai terminā. Par *valodas apguvēju korpusu* runā arī Everita Andronova un V. Rūtenberga (Andronova 2009; Rūtenberga 2012). Tātad šķiet, ka šis ir vispopulārākais termins šī korpusu veida nosaukšanai. Termins *studentu korpuss* minēts LU Matemātikas un informātikas institūta sagatavotajā latviešu valodas korpusa koncepcijā kā sinonīms terminam *valodas apguvēju korpuss* (LVKK 2005, 17).

Minētie darbi nav vienīgās Latvijas pētnieku publikācijas šajā jomā, tomēr daļa no tām izdotas svešvalodās un līdz ar to nesniedz terminu latviešu valodā.

Lietuviešu terminos vērojama lielāka dažādība nekā latviešu valodnieku darbos. Lietoti nosaukumi *mokinių kalbos tekstynas*, *besimokančiųjų tekstynas*, *negimtakalbių X*[6] *kalbos studentų tekstynas* un *X kalbos kaip užsienio kalbos tekstynas*. Visu nosaukumu sastāvā ir vārds *tekstynas* ‘korpuss’.

R. Juknevičiene promocijas darbā lieto terminu *mokinių kalbos tekstynas* (Juknevičienė 2011, 7), ko varētu burtiski tulkot apmēram kā *skolēnu, mācēnu*[7] vai *mācekļu valodas korpuss*. Valodas apguvēju korpusa definīcija gan šeit nav sniegta. Terminu *besimokančiųjų tekstynas* – ‘to, kas mācās, korpuss’ lieto, piem., J. Grigaļūniene, L. Bikeliene un R. Juknevičiene angļu valodā publicēta raksta anotācijā lietuviešu valodā (Grigaliūnienė u. c. 2008, 62). Tā paša raksta nosaukumu tulkojot lietuviešu valodā, korpuss gan nosaukts citādi: *negimtakalbių anglų kalbos studentų tekstynas* – ‘angļu valodas kā nedzimtās[8] valodas

[6] Šeit un turpmāk – konkrētās valodas nosaukums ģenitīvā, piem.: *negimtakalbių anglų kalbos studentų tekstynas* – angļu valodas kā nedzimtās valodas studentu korpuss.

[7] Vārds *mācēns* ir piedāvāts kā atbilsme angļu valodas vārdam *learner*, taču, šķiet, nav plaši pieņemts kā termins (Skujiņa$_{-e}$).

[8] Lietuviešu valodā vārds *negimtakalbis* nozīmē ‘cilvēks, kuram konkrētā valoda nav dzimtā’. Šeit nav šķīruma starp otrās valodas un svešvalodas pratējiem/apguvējiem, tāpēc, lai neradītu pārpratumus, šajā darbā gadījumos, kad šķīrums starp otro valodu un svešvalodu nav vajadzīgs, lietots vārdu savienojums *nedzimtā valoda*. Termins *nedzimtā valoda* nav atrodams 2007. gadā iznākušajā *valodniecības terminu vārdnīcā* (VPSV 2007), taču kā krievu valodas termina *неродной язык* latviskā atbilsme ir norādīts sešdesmitajos gados izdotajā vārdnīcā (VTV 1963). Savukārt krievu valodnieki terminu *неродной язык* pielīdzina angļu valodas terminam *non-native*

studentu korpuss' (Grigaliūnienė u. c. 2008, 62). Lietuviešu valodnieču publikācijā minēta arī valodas apguvēju korpusa definīcija – tā pati, kas norādīta šīs apakšnodaļas sākumā. Agne Zujevaite (*Agnė Zujevaitė*) un Egle Žilinskaite (*Eglė Žilinskaitė*), šobrīd – Egle Žiliskaite-Šinkūniene (*Eglė Žilinskaitė-Šinkūnienė*), raksta par *latvių kalbos kaip užsienio kalbos tekstynas* – 'latviešu valodas kā svešvalodas korpusu' (Zujevaitė, Žilinskaitė 2012, 55).

Latviešu valodā kā visprecīzāk jēgu izsakošs tiek piedāvāts *valodas apguvēju korpuss* (vairāk sk. Znotiņa 2014, 269), tāpēc šajā darbā tiek lietots šis termins. Savukārt lietuviešu valodā, meklējot kompromisu starp precizitāti un valodas līdzekļu ekonomiju, ieteikts lietot terminu *besimokančiųjų tekstynas* (Znotiņa 2014, 269).

Runājot par valodas apguvēju korpusa jēdzienu, jāmin vēl kāds aspekts, kas dažkārt izraisa domstarpības. Kā norādīts šīs apakšnodaļas sākumā, izplatīta ir S. Greindžeres sniegtā valodas apguvēju korpusa definīcija, taču jāpiebilst, ka ne visi pētnieki par to ir vienisprātis. Piem., pastāv viedoklis, ka var būt ne vien otrās valodas vai svešvalodas, bet arī dzimtās valodas apguvēju korpuss. Viens tāds – igauņu valodas kā dzimtās valodas apguvēju korpuss (angļu valodā – *native Estonian learner text corpus*) *EMMA* – ir izveidots Tartu Universitātē (Sõrmus 2014), cits – *KoKo* – tapis vācu valodai kā dzimtajai valodai (Abel u. c. 2014b). Šeit, tāpat kā gadījumā ar tulkošanas apguvēju korpusiem, droši vien varētu raisīties diskusijas ne tik daudz par korpusa būtību vai lietderību, bet gan tā piederību valodas apguvēju korpusa grupai.

Korpusi, ko tagad sauc par valodas apguvēju korpusiem, vēsturiski radušies kā *nedzimtās valodas korpusi* (angļu valodā – *corpora of non-native language*), par primāru uzskatot tieši šo aspektu, nevis, piem., valodas kā tādas apguves procesu (Granger 2002, 2). Runājot par dzimtās valodas apguvi, parasti tiek minēti *bērnu* un *jauniešu valodas korpusi*. Ja tiek salīdzināti dzimtās valodas runātāju un tās pašas valodas kā otrās valodas vai svešvalodas apguvēju radītie teksti, tad parasti runa ir par, piem., *spāņu kā dzimtās valodas korpusu* un *spāņu valodas apguvēju korpusu*, pat ja korpusi ir veidoti maksimāli līdzīgi viens otram ar mērķi nodrošināt iespēju salīdzināt tajos iekļautos datus (Díez-Bedmar, Papp 2008; Orol González, Alonso Ramos 2013 u. c.). Diskusiju gan vēl raisa pētījumi, kuros tiek aplūkots, kā cilvēks apgūst jaunu savas dzimtās valodas paveidu – izloksni vai specializētu valodas paveidu, piem., populāri ir pētīt akadēmiskās valodas attīstību studentu darbos (Hardy, Römer 2013; Römer, O'Donnell 2011; Janulienė, Dziedravičius 2015; Šimčikaitė 2012). Dažkārt arī šādus

language, ar to saprotot valodu, kas konkrētajam subjektam nav dzimtā, taču bez obligāta šķīruma otrajā valodā un svešvalodā (Багироков, Блягоз 2012).

korpusus to veidotāji uzskata par valodas apguvēju korpusu paveidu (piem., Bohát u. c. 2015). Tomēr citi autori tos nošķir atsevišķi (piem., Zevakhina u. c. 2015), uzsverot, ka abu veidu korpusu pētnieki cits no cita var aizgūt metodes, tomēr dzimtās valodas paveida apguve būtiski atšķiras no jaunas valodas apguves. Ņemot vērā visu izklāstīto, šajā darbā atstāts spēkā šķīrums starp dzimtās valodas korpusiem un nedzimtās valodas apguvēju korpusiem.

Lai arī definīcijas pamats dažādu autoru formulējumā saglabājas tas pats, reizēm to mēdz paplašināt, pievienojot vēl citus, pēc definīcijas autora domām, svarīgus nosacījumus. Piem., arī S. Greindžere pati savu definīciju izvērš, balstīdama to Dž. Sinklēra, viena no korpusa lingvistikas pamatlicējiem, radītajā korpusa definīcijā. Viņas formulējums ir šāds: „Datorizēti valodas apguvēju korpusi ir autentisku otrās valodas/svešvalodas tekstuālu datu elektroniski vākumi, kas ir apkopoti, vadoties pēc izvērstiem kritērijiem, noteiktam otrās valodas/svešvalodas apguves vai otrās valodas/svešvalodas mācīšanas mērķim. Tie ir kodēti standartizētā viendabīgā veidā, un to izcelsme ir dokumentēta." (Granger 2002, 4)

Šajā definīcijā papildus jau minētajiem trim nosacījumiem – valodas apguvējiem, tekstiem un elektroniskam formātam – parādās vēl citas pazīmes, kas ļauj precizēt izpratni par valodas apguvēju korpusu. Pirmkārt, šeit ir minēta gan otrā valoda, gan trešā valoda jeb svešvaloda, kas īsākajās definīcijās, iespējams, tieši lakoniskas izteiksmes labad nav nošķirtas un mēdz tikt apvienotas zem tādiem apzīmējumiem kā *nedzimtā valoda* vai *svešvaloda*. Otrkārt, šeit uzsvērta nepieciešamība pēc skaidriem izvērstiem kritērijiem tekstu atlasē, ļaujot saprast, ka ne kurš katrs pēc nejaušības principa savāktu apguvēju valodas paraugu kopums ir uzskatāms par korpusu. Treškārt, norādīts, ka korpusu veido, paturot prātā noteiktus pētījuma mērķus. Šī prasība ir spēkā korpusa lingvistikā vispār (Reppen 2010, 31), ne tikai valodas apguvēju korpusos.

Kā jau iepriekš minēts, dažkārt valodas apguvēju korpusi tiek dēvēti par *starpvalodas korpusiem*, jo tajos iekļautos tekstus var uzskatīt par starpvalodas paraugiem. Starpvaloda (lietuviešu val. *tarpukalbė,* angļu val. *interlanguage*) tiek raksturota kā „individuāls, mainīgs valodas paveids vai lietojums, kam raksturīgas divu valodu pazīmes. Šāda valoda var veidoties svešvalodas apguves procesā, pārejā no dzimtās valodas uz apgūstamo." (VPSV 2007, 373). Lingvodidaktikā to definē, koncentrējoties tieši uz valodas apguves aspektu, un skaidro sīkāk: „Starpvaloda .. [ir] valoda, ko producē persona, kura mācās kādu mērķvalodu .. [un] kas atšķiras no dzimtās valodas lietotāju .. valodas. [Tā] ir nevis kļūdaina mērķvaloda, bet patstāvīga apguvēja radīta valodas sistēma, kura mācību procesā pastāvīgi mainās. [..] [To] ietekmē apguvēja dzimtā valoda, mērķvaloda, iedevums mācību procesā, mācību stratēģijas un komunikācijas stratēģijas." (LTŽ 2012, 200); „Mainīgs valodas paveids vai lietojums, kurā uz

apgūstamo valodu tiek pārnestas runas struktūras no dzimtās valodas vai citas zināmas valodas, veidojot jaunu runas struktūru ar divu vai vairāku valodu pazīmēm.” (LTSV 2011, 86)

Šajā pētījumā tiek runāts par otrās baltu valodas apguvi, proti, lietuviešu valodas apguvi latviešiem vai latviešu valodas apguvi lietuviešiem. Līdz ar to šeit tiek lietots termins *baltu starpvaloda*, ar to saprotot starpvalodu, kāda veidojas otrās baltu valodas apguves gadījumā.

Pētnieki dažkārt norāda, ka korpusam jābūt veidotam bez noteikta pētījuma mērķa, lai līdz ar to netiktu ne tīši, ne nejauši ietekmēti pētījuma rezultāti (Baker 2006, 11). Šāda pieeja ir skaidri motivēta, taču iespējas pēc tās vadīties ir ierobežotas, jo, piemērojot to nekritiski, tā apgrūtina specifisku gadījumu izpēti, kuriem nepieciešams īpaša veida tekstu korpuss. Vismaz pašreizējā korpusa lingvistikas attīstības posmā neviens korpuss nav pilnīgi visaptverošs un derīgs jebkādiem pētījuma mērķiem. Tāda izveidošana prasītu ļoti daudz resursu, un nav pārliecības, ka tas atmaksātos, tālab korpusa lingvistikā ierasta prakse ir veidot specializētus korpusus noteiktu tekstu veidu izpētei, piem., valodas apguvēju korpusus, kas satur noteiktu tekstu tipu – tādus tekstus, ko ir radījuši attiecīgās valodas apguvēji.

Iepriekšminētajam uzskatam skaidrojums ir rodams arī šaurākā izpratnē. Vadoties pēc šāda skaidrojuma, jebkādas darbības korpusa izveidē neatkarīgi no tā, kādus tekstus tas satur, būtu jāveic bez noteiktas pētījuma problēmas un/vai hipotēzes. Tas zināmā mērā attiecas arī uz korpusa anotēšanu (par to vairāk sk. 1.3.1. apakšnodaļā „Korpusa lingvistikas procedūras darbā ar valodas apguvēju korpusiem”). Ja ar korpusu tiek veiktas manipulācijas (pārstrukturēšana, marķēšana, anotēšana utt.) ar konkrētu pētījuma mērķi, tad šīs manipulācijas jau būtu uzskatāmas par konkrētā pētījuma daļu un to veikšana jāskaidro un jāpamato attiecīgā pētījuma kontekstā.

Otrās baltu valodas apguvēju korpuss tiek veidots bez noteiktas pētījuma problēmas, ar nolūku kļūt par pēc iespējas universālu bāzi baltu starpvalodas pētījumos.

1.1.2. Valodas apguvēju korpusa raksturīgie parametri

Kā iepriekš norādīts, valodas apguvēju korpusa definīcijas koncentrējas uz trim parametriem, kas ir būtiski, lai kādu materiālu vākumu atzītu par valodas apguvēju korpusu: tam jābūt elektroniskam, jāsastāv no tekstiem, un šo tekstu autoriem jābūt valodas apguvējiem. Lai labāk raksturotu valodas apguvēju korpusus, turpmāk sīkāk aprakstīts katrs no šiem parametriem.

1.1.2.1. Materiāla autori

S. Greindžere norāda, ka „valodas apguvēju korpusiem piemīt visas tās pašas īpašības, kas parasti tiek attiecinātas uz korpusiem .., vienīgā atšķirība ir tā, ka dati nāk no valodas apguvējiem" (Granger 2008a, 259). Tātad korpusā iekļaujamo valodas paraugu autori ir pirmā un galvenā valodas apguvēju korpusiem raksturīgā īpatnība.

Jau pats valodas apguvēju korpusa nosaukums rāda, ka tajā iekļauto tekstu autors ir kādas konkrētas valodas apguvējs. Vispārīgi runājot, tā ir jebkura persona, kas apgūst noteikto valodu, neatkarīgi no tā, ar kādām metodēm apguve tiek realizēta, kādā vidē tas notiek utt. Tomēr, veidojot korpusu, datu atlasē ar šādu pieeju nepietiek. Dažādi autori uzsver, ka materiāliem jābūt vāktiem pēc konkrētiem izvērstiem kritērijiem, jo tikai tā ir iespējams izvērtēt dažādus faktorus, lai iegūtu informāciju, uz kuru var paļauties (Meyer 2004, 30 u. c.).

Kritēriji, pēc kādiem tiek vākti un atlasīti teksti ievietošanai korpusā, var būt divējādi:

- tādi, kas raksturo attiecīgo valodas lietojuma situāciju vai uzdevumu;
- tādi, kas raksturo teksta autoru, tātad šajā gadījumā – valodas apguvēju (Granger 2008b, 338; par korpusa „Esam" tekstu atlases kritērijiem sk. 2.1. nodaļu „Avotu atlase").

Līdz ar tiešajiem korpusā iekļaujamajiem valodas paraugiem tiek vākta arī papildinformācija, kas ļauj labāk izprast šo paraugu producēšanas situāciju. Papildinformācijas vākšanai ir liela nozīme, jo, izmantojot šo papildinformāciju, var atdalīt samērā homogēnus apakškorpusus, lai analizētu attiecīgā faktora ietekmi uz valodas producēšanu un/vai valodas apguvi (Granger 2008b, 339; par apakškorpusu atdalīšanu sk. arī Hunston 2008, 154).

1.1.2.2. Teksti

Autentiskums. Viens no galvenajiem kritērijiem, ko mēdz izvirzīt dažādos korpusos iekļaujamiem valodas paraugiem, ir šo valodas paraugu autentiskums – tiem būtu jābūt iegūtiem īstā komunikācijas procesā, kura dalībnieki valodu lieto normāli (Meyer 2004, 30). Pētnieki uzsver, ka valodas apguvēju korpusos tiek apkopoti dati, kas ir tuvi to autoru dabiskajam valodas lietojumam, vienlaikus atzīstot, ka, saglabājot nepieciešamo kontroli pār korpusā iekļaujamo tekstu veidiem, pilnīgi dabisku tekstu iegūšana nav iespējama (Granger 2008b, 338).

Valodas apguvēju korpusos centieni pēc autentiskuma var izrādīties problemātiski arī tāpēc, ka valodas apguvēji reti lieto mērķvalodu ikdienas vajadzībām. S. Greindžere no valodas apguvējiem iegūstamos datus raksturo drīzāk kā skalu, kurā visautentiskākie būtu tādi valodas

paraugi, kuros apguvējam nav prasīts lietot konkrētus valodas līdzekļus (piem., gramatiskas konstrukcijas) vai paust konkrētu domu (piem., tulkot vai atstāstīt lasītu tekstu). Līdz ar to šādi teksti arī būtu vispiemērotākie iekļaušanai valodas apguvēju korpusos (Granger 2008a, 261). Tomēr tiek pieļauta arī tādu valodas apguvēju korpusu veidošana, kuros ir iekļauti mazāk autentiski teksti, kā piemēru minot tulkošanas apguvēju korpusus (par tulkošanas korpusu piederību valodas apguvēju korpusiem sk. arī iepriekš – 1.1.1. apakšnodaļā „Valodas apguvēju korpuss: termins un definīcija"). Šādus korpusus piedāvāts uzskatīt par perifēriem valodas apguvēju korpusiem (Granger 2008a, 261).

Runājot par autentiskumu, ir būtisks vēl kāds faktors. Mācoties jaunu valodu, apguvēji neizbēgami pieļauj dažādas kļūdas, un tās atspoguļojas arī apguvēju producētajos tekstos. Pat tad, ja šajos tekstos nav nekā tāda, kas būtu uzskatāms par valodas kļūdu, dzimtās valodas runātājam tie var šķist savādi, neierasti. Valodas apguvēju korpusos šie teksti tiek iekļauti oriģināli, t. i., nelaboti, jo tieši šīs īpatnības ir to galvenais izpētes objekts. Strādājot ar valodas apguvēju korpusiem, tāpat jāatceras, ka, pastāvot atšķirībai starp valodas prasmi un sniegumu (Ellis 1994, 12–13), valodas apguvēju korpuss sniedz iespēju pētīt sniegumu, taču par prasmi ļauj tikai izteikt pieņēmumus, balstītus attiecīgajos snieguma pētījumos.

Saistījums. Līdzīgi kā citi korpusi, arī valodas apguvēju korpusi sniedz iespēju pētīt plašākas valodas vienības par teikumu, kas agrāk bijis uzmanības centrā kā svešvalodas apguves pētniecības pamatvienība (Granger 2008b, 338). Tas nozīmē, ka tiek vākti veseli teksti, nevis atsevišķi to fragmenti, noteikta tipa teikumi vai tml.

Pastāv dažādas izpratnes par teksta jēdzienu, un tā definējums parasti „atbilst konkrētam pētniecības nolūkam, proti, kurš(-i) teksta aspekts(-i) ir šī pētījuma priekšmets" (Laiveniece 2010, 84). Korpusa lingvistikā par tekstu plašā nozīmē var dēvēt „datni, kas sastāv no mašīnlasāmiem datiem", jo korpusā var būt arī, piem., videoierakstu datnes (McEnery, Hardie 2012, 2). Tomēr šajā promocijas darbā tiek izmantota definīcija no *Valodniecības pamatterminu skaidrojošās vārdnīcas*: teksts tiek izprasts kā „mutvārdos izteikts vai rakstveidā fiksēts loģiski strukturēts, funkcionāli vienots jēdzieniski saistītu izteikumu kopums vai atsevišķs izteikums" (VPSV 2007, 392). Šo definīciju var attiecināt uz vairumu gadījumu korpusa lingvistikā vispār, lai arī tā, kā jau norādīts, nav pietiekama, lai raksturotu dažus atšķirīgus korpusos iekļaujamu datu veidus. Tieši tekstu (nevis, piem., teikumu kopumu) atlasīšana ir svarīga tāpēc, lai izprastu, vai un kā valodas apguvēju sniegumu ietekmē faktori, kas var nebūt nosakāmi viena teikuma robežās: domas plūdums, teksta uzbūve utt.

Jāuzsver – kā tas redzams arī definīcijā – šajā gadījumā ar tekstiem nav noteikti jāsaprot tikai rakstveidā producētas valodas paraugi. Pasaulē tiek veidots arī gana daudz valodas apguvēju mutvārdu valodas korpusu, lai arī joprojām dominē rakstveida valodas korpusi (Granger 2008b, 340). Tas ir tādēļ, ka mutvārdu valodas paraugus savākt ir sarežģītāk – nepieciešams audio vai video ieraksts, un runāto nepieciešams transkribēt, lai to varētu apstrādāt ar korpusa lingvistikas rīkiem.

Tekstu definēt ir svarīgi ne vien korpusa apraksta veidošanas nolūkā, bet arī tāpēc, ka definējums ļauj izkristalizēt kritērijus, pēc kādiem izvēlēties, kurus datus iekļaut korpusā, bet kurus – ne. Piem., korpusam „Esam" savāktajos materiālos ir atrodams arī tāds darbs, kurā tā autors gluži kā gramatikas uzdevumā sastādījis astoņus savstarpēji nesaistītus teikumus lietuviešu valodā pēc viena un tā paša modeļa:

> *Aš pirkau pusė kilogramą vištienos.*
> *Aš pirkau dvi pamidorius ir dvi agurkius.*
> *Aš pirkau dvi apelsinus ir kilogramą braškes.*
> *Aš pirkau vieną pakelę multivitaminų suliu ir buteli vanduo.*
> *Aš pirkau vieną kepalą džiuvesių.*
> *Aš pirkau vieną pakelę varškes ir vieną pakelę rugpienio.*
> *Aš pirkau pusė kilogramą salotos.*
> *Aš pirkau kilogramą upetakes.*[9]

Salīdzinājumam – cita autora darbs lietuviešu valodā (arī astoņu teikumu garumā) par to pašu tēmu:

> *Prieš aš eidama į parduotuvė, aš padarau sąrašą. Paprastai sąrašas nėra ilgas, jei einu apsipirkti pries paprastų pietų arba vakarienių. Tada aš perku daržovę: agurkus, bulves, česnākus, kopūstus, morkas, pomidorus, špinatus, vaisius ir uogas: abrikosus, ananasus, mango, apelsinus, bananus, citrinas, datules, abuolius, persikus, vynuoges. Aš galvoju, kad šie produktai labai sveikas, todėl perku tuos daug ir neretai. Svarbi valgymo sudėtinė yra gėrimas, todėl aš perku irgi limonadą, irgi kakavą, irgi sulą, irgi vandenį, irgi arbatą. Taip pat sviestą, sūrį, varškę, bet pieno produkto aš neperku dažnai. Aš perku saldumynus: ledą, pyragaičius, šokoladus, jei man skoniui saldumynai. Iš mėsos produktų aš perku tik vištieną.*[10]

[9] 'Es pirku puskilogramu vistas gaļas. Es pirku divus tomātus un divus gurķus. Es pirku divus apelsīnus un kilogramu zemeņu. Es pirku vienu paciņu multivitamīnu sulas un pudeli ūdens. Es pirku vienu klaipu [sic!] rīvmaizes. Es pirku vienu paciņu biezpiena un vienu paciņu rūgušpiena. Es pirku puskilogramu salātu. Es pirku kilogramu foreles.'

[10] 'Pirms es eju uz veikalu, es izveidoju sarakstu. Parasti saraksts nav garš, ja es eju iepirkties parastām pusdienām vai vakariņām. Tad es pērku dārzeņus: gurķus, kartupeļus, ķiplokus, kāpostus, burkānus, tomātus, spinātus; augļus un ogas: aprikozes, ananasus, mango, apelsīnus, banānus, citronus, dateles, ābolus, persikus, vīnogas. Es domāju, ka šie produkti ir ļoti veselīgi, tāpēc pērku tos daudz un bieži. Svarīga ēšanas daļa ir dzēriens, tāpēc es pērku arī limonādi, arī kakao, arī sulu, arī ūdeni, arī tēju. Tāpat arī sviestu, sieru, biezpienu, bet piena produktus es nepērku bieži. Es pērku saldumus: saldējumu, pīrādziņus, šokolādes, ja man garšo saldumi. No gaļas produktiem es pērku tikai vistas gaļu.'

Redzams, ka pirmais piemērs neatbilst teksta definīcijai – teikumi nav savstarpēji jēdzieniski saistīti, un arī funkcionālu veselumu neveido. Gluži pretējs ir otrais piemērs, kurš, lai arī var tikt kritizēts par uzbūvi, teksta definīcijai tomēr atbilst. Līdz ar to korpusā ir iekļauts otrais šeit citētais darbs, taču pirmais – ne.

Apjoms. Runājot par tekstiem, korpusa lingvistikā ir svarīga ne vien to kvalitāte, bet arī kvantitāte, jo liela daļa no korpusa lingvistikā izmantotajām procedūrām savā būtībā ir kvantitatīvas (par to vairāk sk. 1.3.1. apakšnodaļā „Korpusa lingvistikas procedūras darbā ar valodas apguvēju korpusiem"). Par korpusiem vispārīgi runājot, to apjoms parasti „ir ļoti liels – miljons vai vairāk vārdlietojumu" (VPSV 2007, 196). Par to, kas būtu uzskatāms par lielu korpusu, gan var rasties domstarpības, jo arī atšķirības starp dažādiem korpusiem ir lielas. Vislielākie korpusi parasti ir vispārīgi, nevis specializēti. Piem., lietuviešu valodas balansētajā korpusā ir vairāk nekā 140 miljonu vārdu (DLKT$_{-e}$), arī dažādu citu korpusu apjoms sniedzas līdz simtiem miljonu vārdu (Leech 2014, 10), un lielākie korpusi pasaulē jau tiek skaitīti miljardos vārdu, piem., *Collins* korpusā ir 4,5 miljardi vārdu (Collins$_{-e}$). Korpusa lingvistikā tomēr par šo jautājumu ir plašas diskusijas. Tas, vai korpusa apjoms ir pietiekams, parasti tiek sasaistīts ar prasību pēc reprezentativitātes, proti, korpusam jābūt tik lielam, lai tas būtu pietiekami reprezentatīvs secinājumiem par valodas daļu, ko tas pārstāv (Leech 2014, 11). Līdz ar to konkrētus skaitļus pētnieki min negribīgi.

Realitātē korpusa apjomu nosaka ne tikai idealizēti priekšstati par to, cik datu būtu jāsavāc, lai tas būtu reprezentatīvs, bet arī tas, cik daudz datu ir iespējams savākt un apstrādāt. Piem., tā kā mutvārdu korpusa sagatavošana ir sarežģītāka (pirms jebkādas apstrādes tie vēl arī jātranskribē), šādi korpusi mēdz būt mazāki.

Valodas apguvēju korpusu datu ieguve ir specifiska, tāpēc arī to apjoms parasti nav tik liels kā, piem., vispārīgajiem valodas korpusiem, taču arī šajā jomā jau ir radīti gana lieli tekstu krājumi (sevišķi – angļu valodā).

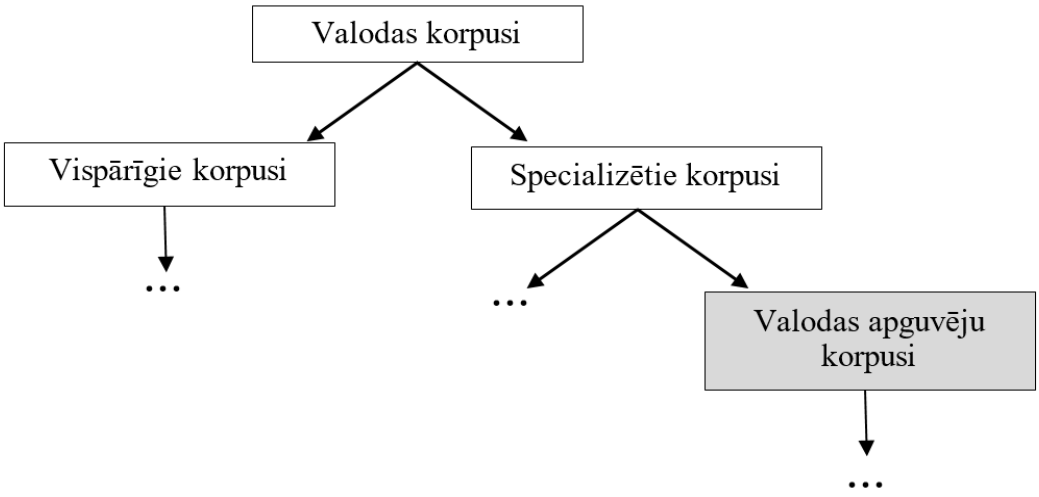
Liela daļa līdz šim tapušo valodas apguvēju korpusu sastāv no apmēram 50 tūkstošiem līdz 150 tūkstošiem vārdlietojumu (Мальцева 2011, 211). Par samērā lielu jau var uzskatīt valodas apguvēju korpusu, kurā ietilpst apmēram 250 tūkstoši vārdlietojumu vai vairāk, piem., angļu valodas apguvēju korpuss *NOSE*, kurā ir nedaudz vairāk par 300 000 vārdu (Díaz-Negrillo 2012, 43; par apjomu sk. arī 1.1.3. apakšnodaļā „Valodas apguvēju korpusu veidi").

1.1.2.3. Mašīnlasāmība

Neatkarīgi no tā, kā teksti ir iegūti, tie jāpadara mašīnlasāmi, tātad jāpārveido elektronisku simbolu virknē, kuru dators ir spējīgs nolasīt un apstrādāt. Ja tas nav izdarīts, tad datu vākumu ar korpusa lingvistikas metodēm pētīt nevar. Mutvārdu runa jātranskribē. Daži pētnieki par piemērotāku runas korpusu izveidē uzskata ortogrāfisko transkripciju (Rābante-Buša 2012, 126), citkārt tiek izvēlēta fonētiskā transkripcija (Auziņa u. c. 2015). Savukārt rakstveida teksti, ja tie jau ir elektroniski, jāsaglabā piemērota formāta teksta datnēs, bet, ja nav elektroniski, jāpārraksta. Tā tiek iegūts tekstu kopums bez jebkāda veida papildinformācijas norādēm.

Ja korpusa datiem ir nolemts pievienot papildinformāciju, tad to dara, tekstus *marķējot* un *anotējot*. Termini *marķēšana* un *anotēšana* Latvijā dažkārt tiek lietoti sinonīmiski, taču šajā darbā tie tiek šķirti, vadoties pēc angliski runājošajās zemēs nostiprinājušās tradīcijas: par *marķēšanu* uzskatāma metadatu pievienošana tekstam (piem., teksta valoda, autora dzimtā valoda utt.), turpretim *anotēšana* savā ziņā ir teksta interpretēšana – tā ir lingvistiskas analīzes rezultātu (piem., vārda sastāva, palīgteikumu veida, valodas apguvēju korpusa gadījumā – kļūdu tipa) pievienošana korpusā iekļautajam tekstam (par šo šķīrumu sk. McEnery, Hardie 2012, 29). Tātad korpusa *marķējums* sniedz objektīvu informāciju, kas ir iegūta līdz ar tekstiem, savukārt *anotējums* – samērā subjektīvu dziļāku pētnieka skatījumu uz teksta lingvistiskajām vai saturiskajām īpatnībām (vairāk par anotēšanu un marķēšanu sk. 1.3.1. apakšnodaļā „Korpusa lingvistikas procedūras darbā ar valodas apguvēju korpusiem").

1. attēls. Valodas apguvēju korpusu vieta valodas korpusu klasifikācijā



1.1.3. Valodas apguvēju korpusu veidi

Valodas apguvēju korpusi ir valodas korpusu veids (sk. 1. attēlā 24. lpp.). Taču, lai arī salīdzinājumā ar visu valodas korpusu kopumu šī daļa ir viendabīga, tā nav klasifikācijas sīkākā iespējamā vienība. Gluži otrādi – tāpat kā dažādos citos korpusos, ir iespējamas dažādas variācijas, kuras atspoguļo arī sazarotas klasifikācijas iespējas.

Mēģinājumi valodas apguvēju korpusus klasificēt ir bijuši dažādi. Nav vienas vispārpieņemtas klasifikācijas, kuru nozarē varētu uzskatīt par etalonu. Ja pētnieki šim jautājumam vispār pievēršas, tad parasti norāda, ka valodas apguvēju korpusus var iedalīt pēc dažādiem parametriem, un nosauc dažus no tiem, dažkārt precizējot korpusu tipus, kādi var veidoties, izvēloties attiecīgo parametru par klasifikācijas pamatu. Tā kā šie parametri bieži vien ir tie paši, kurus pētnieki ņem par pamatu, veidojot korpusa koncepciju, korpusu klasifikācija lielā mērā sasaucas ar korpusu izveidi un lēmumiem, kas izveides gaitā ir jāpieņem.

20. gadsimta 90. gadu sākumā tika radīts korpusu izveides pamatstandarts, kura mērķis ir panākt, lai taptu arvien vairāk „augstas kvalitātes savstarpēji saderīgu korpusu dažādām valodām, dažādiem mērķiem, dažādās vietās, kā arī izmantojot dažādu programmnodrošinājumu un tehnoloģijas" (Atkins 1992, 1), un kurā minēti dažādi kritēriji, kurus ieteicams ņemt vērā dažādu korpusu izstrādē. Uz to atsaucas arī S. Greindžere, norādīdama, ka ir īpaši svarīgi noteikt skaidrus kritērijus valodas apguvēju korpusam, jo valodas apguvēju dati jau savā būtībā ir ļoti dažādi (Granger 1997, 176). Šie kritēriji ir nepieciešami, lai precizētu, kādi tieši dati tiek iekļauti korpusā, taču līdz ar to šie paši kritēriji var noderēt arī korpusu veidu noteikšanā. S. Greindžere no Sjū Atkinsas (*Sue Atkins*) minētajiem nosauc šādus galvenos kritērijus īpaši valodas apguvēju korpusa izveidei:

- valodas produkcijas veids (rakstveida vai mutvārdu);
- teksta tips (eseja, vēstule, saruna utt.);
- teksta funkcija (stāstījums, skaidrojums, viedoklis utt.);
- teksta tehniskums (netehnisks, tehnisks, daļēji tehnisks)[11];
- teksta autora dzimtā valoda (Granger 1997, 177).

Citur S. Greindžere īpaši valodas apguvēju korpusiem kā klasifikācijas parametrus nosauc mērķvalodu, dzimto valodu, valodas producēšanas veidu (mutvārdu vai rakstveida),

[11] Šis kritērijs skaidrots kā balstīts autora un lasītāja speciālajās un/vai tehniskajās zināšanās par noteiktu tēmu. Netehnisks teksts būtu tāds, kurā ne no autora, ne no lasītāja šādas zināšanas netiek gaidītas; daļēji tehnisks – tāds, kura autoram šādas zināšanas ir, taču lasītājam tādu nav; savukārt tehnisks teksts pēc šī kritērija ir tāds, kura autoram ir šādas zināšanas un tās ir nepieciešamas arī teksta lasītājam (Atkins 1992, 17).

tekstu žanru, tekstu ieguves perioda ilgumu (vai no viena un tā paša autora iegūti teksti tikai noteiktā brīdī, vai arī atkārtoti ilgstošā laika posmā) un korpusa pedagoģisko lietojumu (Granger 2008a, 262). Jarmo Jantunena nosauktie parametri ietver žanru skaitu, tēmu skaitu, valodu skaitu un citus parametrus, un pētnieks norāda, ka viņa minētie klasifikācijas veidi ir balstīti vairāku citu autoru darbos, kuri nav noteikti veltīti tieši valodas apguvēju korpusiem (Jantunen 2011, 90). Ir klasifikācijas iespējas, kas valodas apguvēju korpusiem ir kopīgas ar citiem korpusu veidiem. Tālāk raksturoti dažādās publikācijās minētie valodas apguvēju korpusu klasifikācijas veidi atbilstoši parametram, pēc kāda klasifikācija tiek veikta.

Klasifikāciju veidus var skatīt kā piederīgus atsevišķām lielākām grupām, piem., J. Jantunens kā tādas min vispārīgās korpusa klasifikācijas iespējas un tieši valodas apguvēju tekstiem specifiskās (Jantunen 2011, 90). Šajā darbā iedalījums ir citāds. 1.1.2. apakšnodaļā „Valodas apguvēju korpusa raksturīgie parametri" jau ir raksturotas būtiskākās valodas apguvēju korpusa pazīmes atbilstoši trim parametriem, proti, *teksti*, *autori* un *tehniskā lasāmība*. Šie parametri kā nozīmīgākās valodas apguvēju korpusa raksturojuma zonas ļauj arī apkopot informāciju par korpusu klasifikācijas veidiem. Korpusa veidošanā var izdalīt trīs galvenās dimensijas:

1) korpusā iekļaujamie teksti un to īpašības;
2) korpusā iekļaujamo tekstu tapšanas apstākļi, autori un ieguve;
3) korpusā iekļaujamo tekstu apstrāde un tehniskais noformējums, korpusu sagatavojot pētniecības darbam.

Šim sarakstam var pievienot arī ceturto dimensiju, kas gan vairs tik lielā mērā neattiecas uz korpusa izveidošanu, proti – 4) korpusa lietošana, tās iespējas un līdzšinējais darbs ar attiecīgajiem korpusiem. Par šīs dimensijas nošķiršanas lietderību, korpusus klasificējot, var rasties diskusijas, jo korpusa lietošanas iespējas veidojas korpusa tapšanas gaitā, tātad par to varētu runāt, balstoties iepriekšējās trijās dimensijās. Tomēr korpusa lietojumu nosaka ne vien tas, kā tas ir sagatavots, bet arī pētnieku intereses un vajadzības, kā arī zinātniskās domas attīstība[12], tāpēc uzskatāms, ka lietojuma dimensijas nošķīrums palīdz atklāt papildu klasifikācijas iespējas.

Valodas apguvēju korpusu klasifikācijas iespējas šeit ir grupētas atbilstoši minētajām dimensijām.

[12] Stigs Johansons (*Stig Johansson*) šajā sakarā uzsver: „Mums nevajadzētu jautāt: šeit man ir daudz materiāla, ko es ar to varu iesākt? Bet: man ir šāds pētījuma jautājums. Kā lai es atrodu atbilstošus pierādījumus? .. Izaicinājums ir izmantot korpusus un uzdot jautājumus." (Johansson 2011, 121)

1. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamajiem tekstiem un to īpašībām.

- **Mērķvaloda** var būt jebkura valoda, kuru kāds apgūst, un līdz ar to pēc šī kritērija nav iespējams veidot pabeigtu klasifikācijas grupu sarakstu. Turklāt vienā valodā var būt dažādas variācijas atkarībā no apgūstamā valodas paveida. Tas var būt reģionāls paveids (piem., britu angļu valoda vai austrāliešu angļu valoda; latviešu literārā valoda vai augšzemnieku dialekts), bet var būt arī profesionāls, sociāls vai kāds cits paveids (piem., akadēmiskā valoda, celtnieku slengs, juridisko tekstu valoda). Tomēr parasti uzmanības centrā ir vispārlietojamās valodas apguve, ieskaitot arī spēju lietot dažādus valodas stilus dažādās sabiedrības grupās. Ņemot vērā lielo angļu valodas dominanci nozarē, dažkārt iedalījums pat ir nevis pēc dažādām valodām vai valodu grupām, bet gan binārs: angļu un ne-angļu valodas apguvēju korpusos (Granger 2008a, 262).
- Pēc **valodu skaita** korpusi mēdz tikt iedalīti vienvalodīgos, divvalodīgos (paralēlos) un daudzvalodīgos korpusos, un J. Jantunens šī klasifikācijas principa piemērojamību valodas apguvēju korpusiem tālāk neskaidro, vien norāda, ka viņa raksturojamais somu valodas apguvēju korpuss ir vienvalodīgs, proti, tajā atrodami tikai somu valodā rakstīti teksti (Jantunen 2011, 90). Tā kā valodas apguvēju korpusos tiek iekļauti teksti apgūstamajā valodā jeb mērķvalodā, tad šajā gadījumā korpusa valodu skaits būtu tajā pārstāvēto mērķvalodu skaits – parasti valodas apguvēju korpusos tiek iekļauti dati ar vienu mērķvalodu, taču ir arī tādi, kuros mērķvalodas ir vairākas – piem., Francijā tapis divvalodīgs franču un angļu valodas apguvēju mutvārdu korpuss *COREIL* (Delais-Roussarie, Yoo 2011), savukārt, Vācijas un Čehijas augstākās izglītības iestādēm sadarbojoties, izveidots korpuss *MERLIN*, kurā ir gan čehu, gan vācu, gan arī itāliešu valodas apguvēju rakstīti teksti attiecīgajā mērķvalodā (vairāk sk. Abel u. c. 2014a).
- **Valodas producēšanas veids** – mutvārdu vai rakstveida (Granger 2008a, 262, Jantunen 2011, 90).
- **Tekstu tips.** J. Jantunens klasifikācijā pēc šī parametra piedāvā izdalīt divas valodas apguvēju korpusu grupas: vienā no tām būtu korpusi, kuros ir iekļauti

tikai viena žanra teksti, savukārt otrā – tādi, kuros ir dažādu žanru teksti (Jantunen 2011, 90). Viņš skaidro, ka dažādu žanru esamība korpusā ļauj labāk noteikt īpatnības, kas ir raksturīgas valodai kopumā, nevis tikai noteiktam žanram (Jantunen 2011, 90).

- **Tekstu tematika.** Šo parametru izvirzīdams, J. Jantunens nosauc divus iespējamos korpusa veidus: netematisks korpuss, kurā ir dažādu tēmu teksti, un tematisks korpuss jeb korpuss, kurā ir vienai tēmai, piem., medicīnai piederīgi teksti.
- **Tekstu oriģinalitāte** – tā varētu nosaukt parametru, atbilstoši kuram korpusus iedala tulkotu tekstu un oriģinālu tekstu korpusos. Valodas apguvēju korpusu gadījumā šis ir problemātisks jautājums. Pat tad, ja apguvējam ir bijis tiešs uzdevums radīt jaunu tekstu mērķvalodā, nevis to tulkot no citas zināmas valodas, tulkošana ir viena no valodas apguves stratēģijām (Jantunen 2011, 91). Ja arī valodas apguvējs tekstu sākotnēji neuzraksta citā valodā, teikumi prātā, sevišķi valodas apguves sākumposmā, nereti rodas citā apguvējam zināmā valodā un pēc tam tiek apzināti tulkoti mērķvalodā. Tāpēc, lai arī iedalījums tulkotos un netulkotos tekstos noteiktos gadījumos var būt noderīgs, jāpatur prātā tā zināmā nosacītība.
- **Specializācija.** Pēc šī parametra korpusus var iedalīt specializētos vai vispārīgos (sabalansētos, reprezentatīvos) korpusos. Kā jau minēts, valodas apguvēju korpusi kopumā ir viens no specializēto korpusu veidiem, taču pastāv arī nedaudz atšķirīga izpratne, ka arī valodas apguvēju korpusi var būt vai nu reprezentatīvi visai noteiktas valodas apguvēju producētajai valodai, vai arī specializēti kāda tās paveida pētīšanai (Jantunen 2011, 90).

2. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamo tekstu tapšanas apstākļiem, autoriem un ieguvi.

- Lai arī par **dzimtās valodas** lomu valodas apguvē speciālistu vidū vienprātības nav (par to sīkāk sk. Jordens 2003), dažādi pētnieki atzīst, ka dzimtā valoda ir svarīgs valodas apguvi ietekmējošs faktors (Gass, Selinker 1983). Tāpēc valodas apguvēju korpusu veidotājiem ir īpaši svarīgi norādīt, kāda(-as) ir korpusā iekļaujamo tekstu autoru dzimtā(-ās) valoda(-as). Tas redzams, piem., starptautiskajā valodas apguvēju korpusu

sarakstā, kurā dzimtā valoda ir viena no nedaudzajām informācijas vienībām, kas ir norādītas visiem sarakstā esošajiem korpusiem (Dumont, Granger 2017).

- **Dzimto valodu skaits** ļauj valodas apguvēju korpusus iedalīt tādos, kuros ir iekļauti tikai vienas dzimtās valodas runātāju rakstīti teksti, un tādos, kuros esošo tekstu autori pārstāv vairākas dzimtās valodas (Jantunen 2011, 92).
- **Valodas apguvēju vecums:** tā kā valodas apguve pieaugušajiem un bērniem notiek dažādi (RELTL 2004, 21–23, Dubovičienė, Gulbinskienė 2014, 138), arī pieaugušo valodas apguvēju valodas paraugi var būtiski atšķirties no datiem, kas iegūti no bērniem, kas apgūst to pašu valodu. Līdz ar to var būt pieaugušo valodas apguvēju korpusi, bērnu valodas apguvēju korpusi un jaukti valodas apguvēju korpusi. Var dalīt arī vēl sīkāk, piem., nošķirot pirmskolas un sākumskolas vecuma bērnus no pusaudžiem u. tml. Izpratne par to, kas ir pieaugušais, bet kas – bērns, gan var atšķirties, piemēram, dažkārt augstskolas studentus, kādi ir korpusa „Esam" informanti, par pieaugušiem valodas apguvējiem (angļu val. *adult learners*) neuzskata (Dubovičienė, Gulbinskienė 2014, 139). Šādā gadījumā lielāka nozīme var būt nevis precīzam vecumam, bet gan valodas apguves iemesliem, motivācijai, uzkrātajai pieredzei un citiem faktoriem.
- **Valodas prasmju līmenis**. Tā kā starpvalodas īpatnības lielā mērā ir saistītas ar to, cik augstā līmenī apguvējs prot mērķvalodu, valodas apguvēju korpusos ir svarīgi noteikt mērķvalodas prasmes līmeni. Tas tiek darīts pēc dažādām skalām. Viena no izplatītākajām balstās Eiropas Savienības valodu prasmju sešu pamatlīmeņu dalījumā – no A1 līdz C2 (EKP 2006), taču pastāv arī citas skalas (ACTFL 2012, ILR-e u. c.). Klasificēt var arī pēc iegūtā vērtējuma eksāmenā (Vinčela 2014).
- **Valodas prasmju līmeņu skaits**: pēc šī parametra iedala korpusus, kuros ir tikai vienam noteiktam valodas prasmju līmenim atbilstoši teksti, un tādos, kuros ir pārstāvēti vairāki prasmju līmeņi (Jantunen 2011, 92).
- **Pēc valodu apguves secības** valodu apguvē var nošķirt dzimto valodu un nedzimto valodu, kas tālāk var tikt iedalīta vēl sīkāk – otrajā valodā un svešvalodā (sīkāk sk. Laizāne 2014a). Līdz ar to attiecīgi var klasificēt arī valodas apguvēju korpusus, nošķirot otrās valodas apguvēju korpusus un

svešvalodas apguvēju korpusus (Jantunen 2011, 92). Daļa pētnieku gan uzskata, ka korpuss, kas atklāj mērķvalodas kā otrās valodas runātāju sniegumu, nebūtu saucams par valodas apguvēju korpusu. Vienlaikus gan tiek atzīts, ka tās ir „vienas monētas divas puses", kas nav stingri nošķiramas, un ka tām būtu jāapvieno spēki, lai atklātu nedzimtās valodas īpatnības (Granger 2008a, 260). Šim šķīrumam pamatā ir galvenokārt izpētes fokuss: ja svešvalodas apguvēju korpusa gadījumā tas galvenokārt ir centrēts uz atšķirībām starp apguvēja sniegumu un mērķvalodas normām, tad otrās valodas gadījumā drīzāk tiktu pētīts un raksturots, kā otrās valodas runātāji sazinās, nevis uz kļūdām, kas neietekmē komunikāciju (Granger 2008a, 260). Tomēr šķiet, ka šādā aspektā piemērotāk būtu orientēties nevis uz to, vai apguvējam attiecīgā valoda ir otrā valoda vai svešvaloda, bet gan uz to, vai minētās valodas apguve turpinās vai arī valodas lietotājs izmanto esošās prasmes, īpaši necenzdamies apgūt jaunas – proti, šādā gadījumā jānošķir kādas valodas kā nedzimtās valodas korpuss no tās pašas valodas apguvēju korpusa. Visbeidzot jāpiebilst, ka arī par dzimtās valodas apguvēju korpusiem nav vienprātības, vai tie būtu uzskatāmi par piederīgiem pie valodas apguvēju korpusiem (sīkāk par to sk. iepriekš – 1.1.1. apakšnodaļā „Valodas apguvēju korpuss: termins un definīcija").

- **Tekstu tapšanas secīgums** ļauj valodas apguvēju korpusus iedalīt sinhroniskajos un diahroniskajos jeb tādos, kuros iekļautie teksti ir tapuši (nosacīti) vienlaikus, un tādos, kuri apkopo dažādos laikos tapušu materiālu (Jantunen 2011, 91).
- Pēc **tekstu rakstīšanas veida** iedalot, rakstveida valodas apguvēju korpusi var būt tādi, kuros iekļautie teksti ir rakstīti ar roku un pēc tam digitalizēti, vai arī tādi, kuru teksti ir tapuši uzreiz datorrakstā (Jantunen 2011, 91).

3. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamo tekstu apstrādi un tehnisko noformējumu, korpusu sagatavojot pētniecības darbam.

- **Tekstu nedalāmība** ir parametrs, pēc kura veidotā klasifikācijā tiek šķirti pilnu tekstu korpusi un tekstu fragmentu korpusi.
- **Kopējais apjoms** ļauj valodas apguvēju korpusus iedalīt lielos un mazos (Granger 2008a, 262), taču šādā dalījumā ir grūti ievērot konsekvenci. Pat ja

pētnieki norāda kādu piemēru katrai no grupām (piem., liels korpuss – miljoniem vārdlietojumu; mazs korpuss – 40 000 vārdlietojumu – sk. Granger 2008a, 262), parasti nav norādes par to, kur atrastos grupu robežas. Turklāt, laikam ejot un pieaugot korpusu apjomam, šīs robežas var būtiski mainīties.

- **Pēc anotējuma** valodas apguvēju korpusus var iedalīt vairākos līmeņos: vispirms tie būtu neanotēti un anotēti korpusi (Jantunen 2011, 92), tālāk var dalīt sīkāk pēc anotēto parādību tipiem, piem., sintaktiski, leksiski u. c. veidos anotētos, kā arī vēl sīkāk – pēc konkrētu parādību anotējuma, piem., korpusi, kuros ir anotēti teikumu veidi, korpusi, kuros ir anotēti teikuma locekļi utt.
- **Izmantotā programmatūra**. Valodas apguvēju korpusu izpētē var lietot dažādas korpusa pārlūkprogrammas un rīkus, pat ja tie nav paredzēti konkrēti valodas apguvēju korpusiem, jo atšķirība starp valodas apguvēju korpusiem un citu veidu korpusiem ir galvenokārt materiālu (tekstu) saturā un programmatūras sniegto rezultātu interpretācijā, nevis tekstu tehniskajā raksturojumā un/vai apstrādes pamatprincipos. Kā norāda Marine Maļceva (Мальцева 2011, 211), bieži tiek lietotas tādas programmas kā *Mike Smith's Wordsmith, Paul Nation's VocabProfile, Cobb T. The Compleat Lexical Tutor, Web Concordancer, Web Frequency Indexer* u. c.
- **Valodu pāru virzieni** – visbiežāk valodas apguvēju korpusos virziens ir viens un mērķvaloda ir viena. Taču, ja mērķvalodas ir vairākas, korpuss var būt vienvirziena, divvirzienu vai jaukts. Vienvirziena korpuss ir tāds, kurā neviena no pārstāvētajām dzimtajām valodām nav arī mērķvaloda. Divvirzienu korpuss ir tāds, kurā visas pārstāvētās dzimtās valodas ir arī mērķvalodas. Savukārt jauktā valodas apguvēju korpusā vismaz viena no pārstāvētajām dzimtajām valodām ir arī starp mērķvalodām, taču ir pārstāvēta arī vismaz viena dzimtā valoda, kuras starp mērķvalodām nav. Jāatzīst gan, ka šis iedalījums pašlaik vēl nav sevišķi aktuāls un varētu kļūt nozīmīgāks nākotnē, jo pašlaik divvirzienu un jauktu valodas apguvēju korpusu vēl tikpat kā nav.

4. Klasifikācija pēc parametriem, kas ir saistīti ar korpusa lietošanu, tās iespējām un līdzšinējo darbu ar attiecīgajiem korpusiem.

- **Korpusa pedagoģiskais lietojums** var būt tūlītējs (angļu val. *corpus for immediate pedagogical use*) vai pastarpināts (angļu val. *corpus for delayed pedagogical use*), attiecīgi ļaujot iedalīt arī korpusus. Tūlītēja pedagoģiskā lietojuma korpusi ir tādi, kuros valodas apguvēji, kuru producētie teksti korpusā ir iekļauti, paši strādā ar šo korpusu (Ragan 2001, 211). Pastarpināta pedagoģiskā lietojuma korpusi tiek izmantoti pētījumos, mācību materiālu sagatavošanā u. tml., taču ne tiešā pedagoģiskajā darbā ar tiem pašiem apguvējiem, kuri ir korpusā iekļauto tekstu autori. Vairums pašreiz pastāvošo valodas apguvēju korpusu pēc šī klasifikācijas principa joprojām ir pastarpināta pedagoģiskā lietojuma korpusi (Granger 2009, 20).
- **Korpusa komerciāls vai akadēmisks lietojums** nosaka korpusu dalījumu komerciālos korpusos, kurus galvenokārt veido uzņēmumi ar mērķi attīstīt savus produktus, lai gūtu peļņu, un akadēmiskos korpusos, kuri tiek veidoti zinātniskos nolūkos (Granger 2008a, 261).
- Lai arī par to, šķiet, atsevišķi runāts daudz netiek, viens no svarīgākajiem parametriem valodas apguvēju korpusam ir arī **pieejamība**, kas līdz ar pārstāvētajām valodām, tekstu veidu un korpusa apjomu ir viena no valodas apguvēju korpusu sarakstā norādītajām informācijas vienībām (Dumont, Granger 2017). Pēc pieejamības valodas apguvēju korpusus var iedalīt publiski pieejamos, ierobežotas piekļuves (piem., kādai universitātes struktūrvienībai pieejamos) un individuālai lietošanai paredzētos. Šo parametru var uzskatīt arī par korpusa sagatavošanas trešajai dimensijai piederīgu, taču tas var samērā viegli mainīties pēc korpusa izveides, neko nemainot pārējā korpusā, tāpēc šajā gadījumā ir piešķirts korpusa lietojuma dimensijai.

Bez šaubām, šie nav vienīgie iespējamie valodas apguvēju korpusu klasifikācijas veidi. Būtībā katrs faktors, kas ļauj veidot atsevišķus apakškorpusus, var kalpot arī kā parametrs korpusu klasifikācijā.

Nav nepieciešams katru jaunu valodas apguvēju korpusu obligāti piemērot katram no šiem klasifikācijas principiem. Klasifikācijai ir divas funkcijas: no vienas puses, tā parāda

valodas apguvēju korpusu daudzveidību un iespējamās variācijas, kā to mēģināts atspoguļot arī šeit. No otras puses, noskaidrojot konkrēta korpusa vietu klasifikācijā, kļūst skaidrāka tajā iekļauto datu būtība un to sniegtās pētniecības iespējas. S. Greindžere atgādina, ka korpusā iekļauto datu veids nosaka, kādus secinājumus var izdarīt no rezultātiem, kas iegūti, pētot šo korpusu: ja korpusā ir tikai viena veida teksti, pēc tā nevar spriest par apguvēju valodu kopumā, taču var izteikt diezgan noteiktus secinājumus par attiecīgo apguvēju valodas paveidu (Granger 1997, 177). Tāpēc šīs darba 2. daļā „Otrās baltu valodas apguvēju korpuss" vietām ir īsumā raksturota korpusa „Esam" piekritība tai vai citai korpusu grupai pēc šeit minētajiem klasifikācijas parametriem.

1.2. Vēsture un izplatība

Pirmie valodas apguvēju korpusi ir parādījušies salīdzinoši nesen, un to pētniecība ir veiksmīgi iekļāvusies starp vairākām nozarēm, it sevišķi lingvodidaktiku un korpusa lingvistiku. Šis starpdisciplinārais pētniecības virziens tiek raksturots kā „ļoti dinamisks" (Granger 2008a), un ne vien tādās valstīs kā Apvienotajā Karalistē, Vācijā, Beļģijā, bet arī Baltijas valstīs ir vērojamas nozīmīgas iestrādnes. Šajā nodaļā tās īsi raksturotas pētniecības virziena attīstības kontekstā. Dažādu jau pastāvošu korpusu konkrētas īpašības (piem., anotējuma veidi) minētas arī citās promocijas darba nodaļās, kurās nepieciešams aprakstīt pētnieku līdzšinējo pieredzi noteiktā aspektā (sk., piem., aprakstu par valodas apguvēju korpusu anotēšanu Latvijā un Lietuvā 2.2.3. apakšnodaļā „Anotēšanas veidu izvēle").

1.2.1. Valodas apguvēju korpusi pasaules kontekstā

Korpusa lingvistikas, kādu to pazīstam tagad, saknes meklējamas 20. gadsimta vidū – 50. un 60. gados (ELL 2005, 207). Valodas apguvēju korpusu izstrāde un pētniecība aizsākās vēlāk – 20. gadsimta 80. gadu beigās un 90. gados, sākotnēji koncentrējoties galvenokārt uz angļu valodas apguvēju radītajiem tekstiem (Granger 2003b, 538; 2004, 123; 2008, 337). Par šo korpusu priekštečiem uzskata kļūdu kartotēkas, kurās pētāmā materiāla apjoms gan bija salīdzinoši diezgan neliels – tas reti pārsniedza 2000 vārdlietojumu, un informantu skaits parasti bija ne vairāk kā desmit (Мальцева 2011, 209). Tomēr materiāla apjoms nebūt nav vienīgā valodas apguvēju korpusu priekšrocība, salīdzinot ar kļūdu kartotēkām, jo tie ar attiecīgu korpusa lingvistikas rīku palīdzību ļauj ne vien analizēt tekstos atrodamās kļūdas, bet arī raksturot tajos lietoto leksiku, gramatiskās formas u. c. (Камшилова 2009).

Par pirmajiem valodas apguvēju korpusiem uzskata Starptautisko angļu valodas apguvēju korpusu (*International Corpus of Learner English*), Honkongas Zinātnes un tehnoloģiju universitātes valodas apguvēju korpusu (*Hong Kong University of Science and Technology learner corpus*) un *Longman* valodas apguvēju korpusu (*Longman Learners' Corpus*), savukārt pirmās publikācijas šajā jomā iznākušas 1993. gadā (Granger 2015). S. Greindžeres īsajā publikācijā raksturots viņas veidotais angļu valodas apguvēju korpuss (Granger 1993), savukārt Honkongas valodnieki izmanto korpusa lingvistikas metodes, lai pētītu Honkongas studentu rakstītos tekstus angļu valodā, kas visiem korpusā iekļauto tekstu autoriem ir svešvaloda (Milton, Tsang 1993). Šajā pašā gadā tapusi arī par pirmo valodas apguvēju korpusā balstīto mācību līdzekli uzskatītā *Longman Language Activator* – vārdnīca, kurā raksturota angļu valodas lietošana kontekstā (LLA 1993). Šajā vārdnīcā viss materiāls ir iegūts korpusu izpētes rezultātā, un viens no šiem korpusiem ir jau iepriekš minētais *Longman* angļu valodas apguvēju korpuss, kurā gūts ieskats grūtībās, ar kādām sastopas angļu valodas apguvēji dažādās valstīs visā pasaulē, kā arī pētīts, kuras valodas struktūras apguvējiem sevišķas problēmas nerada (LLA 1996, F8).

Pirmā grāmata, kuras pētījumu centrā ir valodas apguvēju korpusi (tā apgalvots šīs grāmatas anotācijā), ir *Learner English on Computer*, kas ir izdota 1998. gadā (Granger 1998a; sk. arī Callies, Paquot 2015, 160) un atkārtoti 2013. gadā (Granger 2013). Šī grāmata paredzēta plašam interesentu lokam, un tā skata valodas apguvēju korpusu izveides un pētniecības pamatjautājumus, t. sk. saistību ar citām nozarēm, metodes, kā arī iespējas praktiski izmantot pētījumu rezultātus, veidojot un pilnveidojot mācību grāmatas, vārdnīcas u. c. (Granger 1998a).

Laika gaitā ir izveidoti ne tikai angļu, bet arī citu valodu apguvēju korpusi, un to pētniecība mūsdienās ir būtiski paplašinājusies. Valodas apguvēju korpusi ir izveidoti, piem., franču, zviedru, norvēģu, holandiešu, spāņu, kā arī vācu valodā (Granger 2008a, 262), un to klāsts arvien paplašinās. Pašlaik valodas apguvēju korpusi vispopulārākie ir Āzijā un Eiropā (Мальцева 2011, 209), taču ir sastopami arī citur. Pētnieki atzīst, ka lielā daļā gadījumu izaicinājums ir nodrošināt šo korpusu pieejamību plašai akadēmiskās sabiedrības daļai (Granger 2008a, 262).

Tādi valodas apguvēju korpusi, kas ir pieejami plašam pētnieku lokam un kuros iekļauts liels materiāla apjoms, pārsvarā arī gūst vislielāko popularitāti. Visbiežāk tie ir angļu valodas apguvēju korpusi, piem., *ICLE* (*International Corpus of Learner English*, ICLE 2015), kas 20. gadsimta 90. gados tapis Beļģijā sadarbībā ar vairāk nekā desmit citu valstu universitātēm (Granger 2003b), jāmin arī Mičiganas (ASV) Akadēmiskās angļu valodas runas

korpuss (*Michigan Corpus of Academic Spoken English*, MiCASE 2007) un Somijā, Tamperē izveidotais *ELFA* (*English as Lingua Franca in Academic Settings*) korpuss (ELFA 2014).

Eiropā un citur pasaulē ir tapuši arī dažādu citu, t. sk. mazāk izplatītu, valodu apguvēju korpusi, daļā gadījumu arī ar noteiktu dzimtās valodas–apgūstamās valodas pāri. Piem., Čehijā ir tapis ne vien čehu valodas apguvēju korpuss, kurā iekļauto tekstu autoriem ir dažādas dzimtās valodas, bet arī atsevišķs korpuss, kurā iekļauti tikai romu izcelsmes skolēnu darbi (Hana u. c. 2012). Arī krievu valodnieku publikācijās atrodami gan vispārīgāki skaidrojumi, kas ir valodas apguvēju korpuss un kādas iespējas tas paver (Мальцева 2011), gan arī informācija par jau tapušiem korpusiem un pētījumiem tajos (piem., Smolovskaya u. c. 2015).

Viena no galvenajām autoritātēm valodas apguvēju korpusu pētniecībā Eiropā ir S. Greindžere no Luvēnas Katoļu universitātes, kura kopā ar kolēģiem ir izveidojusi vairākus valodas apguvēju korpusus (ICLE 2015; LINDSEI-$_{e}$), ir viena no Valodas korpusu asociācijas (*Learner Corpus Association*, LCA) dibinātājām (LCA-$_{e}$) un papildus atsevišķu parādību izpētei korpusos dažādās publikācijās sniedz arī pārskatu par valodas apguvēju korpusiem, nozares attīstību un tendencēm (Granger 2002; 2008a; 2015 u. c.). Savukārt Āzijā ar valodas korpusu izpēti aktīvi nodarbojas J. Tono no Tokijas Ārvalstu studiju universitātes, kurš tāpat ir raksturojis dažādus valodas apguvēju korpusu izveides un lietojuma aspektus (Tono 2002; 2003 u. c.) un kura interešu centrā galvenokārt ir japāņu valodas runātāju angļu valodas apguves jautājumi (Tono u. c. 2014). Vēl vērtīgu ieguldījumu nozares attīstībā devuši Markuss Kaliss (*Marcus Callies*), Silvija De Koka (*Sylvie De Cock*), A. Līdelinga, F. Menjē (*Fanny Meunier*), Džons Miltons (*John Milton*), Nadja Neselhaufa (*Nadja Nesselhauf*), Magalī Pako (*Magali Paquot*), U. Rēmere, Odrija Robersone (*Audrey Roberson*) Gaetanele Žilkēna (*Gaëtanelle Gilquin*), un daudzi citi pētnieki.

Kā jau minēts, valodas apguvēju korpusos var apkopot gan otrās valodas, gan svešvalodas apguvēju producētās valodas paraugus. S. Greindžere atzīst – lai arī starp pirmajiem valodas apguvēju korpusiem bijis arī otrās valodas apguvēju korpuss, vairums valodas apguvēju korpusu veidotāju līdz šim tomēr koncentrējušies uz svešvalodas apguvēju valodu (Granger 2008b, 338).

Valodas apguvēju korpusi to aizsākumos tika raksturoti kā „revolūcija lietišķajā valodniecībā" (Granger 1994), „jauns pētījumu novirziens, jauns veids, kā domāt par apguvēju valodu, kas liek izvērtēt dažas no mūsu visdziļāk sakņotajām idejām par apguvēju valodu" (Granger 2004, 123). To pētniecības rezultāts ir jaunas atziņas dažādās jomās, piem., angļu valodas kā otrās valodas frāžu izpētē, savukārt kļūdu anotēšana valodas apguvēju korpusos ir likusi pārskatīt līdzšinējos priekšstatus par kļūdām, to veidiem un dažādību valodu apguvē

(Granger 2015). Jaunā pieeja ir ļāvusi valodas apguves pētījumos ērti izmantot rakstveida valodas paraugus pretstatā pirms tam dominējušajai tendencei pētīt mutvārdu valodas paraugus, kā arī introspektīvus vērojumus (Lessard 1999, 302). Lai gan pētnieki atzīst, ka joprojām nav daudz praktisku mācību materiālu, kas būtu sagatavoti, izmantojot valodas apguvēju korpusus, arī šādam lietojumam tomēr ir vērā ņemams potenciāls (Granger 2015). Korpusu pētnieki arī meklē iespējas tos izmantot nepastarpināti valodas apguves procesā (Tono u. c. 2014).

Valodas apguvēju korpusu pētnieku loka straujo paplašināšanos mūsdienās apliecina arī Valodas apguvēju korpusu asociācijas darbība. Tā ir reģistrēta Beļģijā, un tajā ir vairāk nekā simts biedru no visas pasaules. Organizācija ir dibināta 2013. gadā, un tās mērķis ir atbalstīt jaunu valodas apguvēju korpusu tapšanu dažādām valodām, jaunu rīku un metožu izstrādi. Turklāt tā tiecas veicināt valodas apguvēju korpusu plašāku lietošanu dzimtās valodas un svešvalodas apguves pētniecībā, kā arī valodniecībā kopumā. Organizācija arī uztur un papildina jau esošo valodas apguvēju korpusu sarakstu un plašu bibliogrāfiju par šajā jomā tapušajiem pētījumiem[13].

To, ka valodas apguvēju korpusi ir guvuši popularitāti samērā plašā lingvodidaktikas speciālistu lokā, netieši apliecina arī fakts, ka par šādiem korpusiem tiek runāts dažādās šai jomai veltītās vārdnīcās, enciklopēdijās un rokasgrāmatās (LDLTAL 2010; ELE 2008; Barlow 2005 u. c.). Latviešu valodā iznākušajā *Lingvodidaktikas terminu skaidrojošajā vārdnīcā* gan tie, šķiet, netiek pieminēti, kaut arī tiek runāts par konkordanču lietošanu, „analizējot valodas apguvēju rakstudarbus" (LTSV 2011, 48). Līdzīgi kā citi korpusa lingvistikas virzieni, arī šis vēl nav kļuvis par pašsaprotamu pētniecības metodi līdzās citām, tradicionālākām metodēm. Līdz ar to, lai arī interese par valodas apguvēju korpusiem palielinās, reizēm tiek norādīts, ka tie šķiet nevajadzīgi nodalīti atsevišķi no citiem lingvistikas pētījumu virzieniem, radot mākslīgu šķirumu (Meunier 2006, 111). Arī pētnieki, kas strādā ar valodas apguvēju korpusiem, atzīst, ka šis pētījumu virziens ir kļuvis par nozīmīgu korpusa lingvistikas daļu, tomēr tas tomēr vēl nav guvis plašu atpazīstamību lietišķajā valodniecībā kopumā, un tā devums redzamāks kļūs nākotnē (Callies, Paquot 2015, 161).

Valodas apguvēju korpusu pētniecībā un attīstībā plašāku ieskatu dažādos laika posmos ir devusi S. Greindžere (Granger 2002, 2004), Florense Mailza (*Florence Myles;* Myles 2005), N. Neselhaufa (Nesselhauf 2004), Norma Praveca (*Norma Pravec;* Pravec 2002),

[13] Vairāk informācijas par Valodas apguvēju korpusu asociāciju un tās darbību atrodams organizācijas interneta mājaslapā *http://www.learnercorpusassociation.org/*.

J. Tono (Tono 2003) u. c. pētnieki. Vienlaikus atzīts, ka „redzot tempu, kādā visā pasaulē tiek veidoti valodas apguvēju korpusi, ikviens mēģinājums inventarizēt valodas apguvēju korpusus ātri novecos" (Granger 2008a, 261), tāpēc esošo korpusu saraksti šādās publikācijās parasti netiek sniegti, aprobežojoties vien ar nedaudzu piemēru minēšanu.

1.2.2. Valodas apguvēju korpusi latviešu un lietuviešu valodniecībā

Latviešu un lietuviešu valodniecībā valodas apguvēju korpusi vēl netiek plaši izmantoti, taču pēdējos gados situācija sāk uzlaboties. Lietuvā pētnieki nereti pievēršas jau esošo angļu valodas apguvēju korpusu *ICLE* un *LINDSEI* papildināšanai un iegūtā materiāla izpētei, savukārt Latvijas valodnieki vairāk veido individuāli pielāgotus valodas apguvēju korpusus savu konkrēto pētījumu vajadzībām. Tā kā tendences atšķiras un sadarbība nav īpaši cieša un izvērsta, darbs ar valodas apguvēju korpusiem minētajās kaimiņvalstīs sīkāk raksturots atsevišķi. Par anotētajiem valodas apguvēju korpusiem papildu informācija ir atrodama arī 2.2.2. apakšnodaļā „Anotēšanas veidu izvēle".

Lietuviešu un latviešu korpusa lingvistikā strādājošie valodnieki savā starpā ir sadarbojušies, piem., lietuviešu-latviešu-lietuviešu paralēlā korpusa *LiLa* izveidē (Levāne-Petrova 2012a; Utka u. c. 2012), taču valodas apguvēju korpusu pētniecībā līdz šim aktīva sadarbība nav notikusi.

Iespējams, vienīgais līdz šim tapušais valodas apguvēju korpuss, kurā pārstāvēta gan lietuviešu, gan latviešu valoda, ir *Igauņu starpvalodas korpuss* (*Estonian Interlanguage Corpus*), kuru ir izveidojuši Tallinas Universitātes pētnieki (Eslon 2014, 438). Šeit abas baltu valodas ir pārstāvētas kā informantu dzimtās valodas, savukārt korpusa mērķvaloda ir igauņu valoda (EIC_{-e}). Diemžēl aktīvs darbs ar šī korpusa lietuviešu un/vai latviešu daļu pašlaik nenotiek. Kā jau minēts 1.1.1. apakšnodaļā „Valodas apguvēju korpuss: termins un definīcija", Tartu Universitātē ir izveidots arī igauņu kā dzimtās valodas apguvēju korpuss *EMMA*. Par šāda veida korpusu lietošanu kopā ar nedzimtās valodas apguvēju korpusiem plašāk runāts šī darba 1.3.2. apakšnodaļā „Valodas apguvēju korpusu izpētes metodes".

Igauņu starpvalodas korpuss nav vienīgais gadījums, kurā attiecīgās nozares pētnieki ir interesējušies gan par latviešu, gan lietuviešu valodu. Viļņas Universitātē ir bijis mēģinājums radīt valodas apguvēju korpusu tieši lietuviešu-latviešu valodu pārim (vairāk par to – tālāk).

1.2.2.1. Valodas apguvēju korpusi Latvijā

Viens no pirmajiem nozīmīgākajiem līdz šim latviešu valodniecībā izstrādātajiem darbiem par valodas apguvēju korpusiem ir Z. Vinčelas Latvijas Universitātē izstrādātais

promocijas darbs „Studentu elektroniskais diskurss kā lietišķās valodniecības pētījuma rezultāts" (Vinčela 2010a). Pētniece izveidojusi pati savu angļu valodas apguvēju korpusu *STUDTEXREG* (Vinčela 2013b, 217; 2014, 124). To veido angļu filoloģijas bakalaura studiju programmas studentu rakstīti dažādu žanru elektroniski teksti, un korpusa sastādītāja ir anotējusi tajā iekļautos datus morfoloģiski – pēc vārdšķirām. Tos viņa tālāk analizējusi statistiski (Vinčela 2010b, 346–347).

Z. Vinčela ir viena no aktīvākajām valodas apguvēju korpusu pētniecēm Latvijā. Viņa izstrādājusi vairākas publikācijas, tajās turpinādama darbu ar minēto korpusu arī pēc promocijas darba tapšanas. Līdzīgi kā citi pētnieki, kas veido korpusus galvenokārt tikai savām pētniecības vajadzībām, arī Z. Vinčela savos darbos mazāk pievēršas korpusa uzbūves jautājumiem vispārējā nozares attīstības kontekstā un vairāk pamato tā piemērotību konkrētajiem izvirzīto pētījuma jautājumu analīzei. Šis korpuss tiek izmantots plašā izpētes spektrā: autores interešu lokā ir gan vietniekvārdu lietojums (Vinčela 2010b, 2013a), gan adverbiālās frāzes (Vinčela 2013b), gan personu vietniekvārdu lietojums sastatījumā ar angļu valodas kā dzimtās valodas korpusu un latviešu valodas kā dzimtās valodas korpusu (Vinčela 2013a), gan lingvistiskais variatīvums (Vinčela 2011a, 2011c), gan reģistra īpatnības un grūtības, ar kādām studenti sastopas tādos valodas lietojuma aspektos kā ekspresivitāte, kauzalitāte un teikumu saskaņošana (Vinčela 2011b) vai pieturzīmju lietojums saliktos teikumos (Vinčela 2016). Dažos darbos pievērsts vairāk uzmanības metodoloģiskiem jautājumiem (Vinčela 2014).

Ir vēl viens šīs jomas projekts, kurā Z. Vinčela piedalījusies – Latvijas Valodu skolotāju asociācijas (LVASA) paspārnē V. Kalnbērziņa, Ilze Lokmane, Tatjana Kunda, Z. Vinčela un Kristīne Baiža izveidojušas *Latviešu valodas apguvēja runas un rakstu korpusu* (Kalnbērziņa u. c. 2011). Tajā iekļautas eksāmena mutvārdu uzdevumu atbildes un rakstveida esejas, kuras pamatskolas beidzēji no mazākumtautību skolām rakstījuši 9. klases latviešu valodas un literatūras eksāmenā 2009. gadā. Pētījuma jautājums ir gana plašs: „Kāda ir pamatskolas beidzēju latviešu valodas apguves kvalitāte mazākumtautību skolās, kādi vārdi un sintakse raksturīga latviešu valodas apguvējam, kāds ir valodas plūdums, kāds ir paužu lietojums, cik labi skolēni spēj uztvert dzirdēto un lasīto tekstu, cik veiksmīgi viņi var izteikt savu viedokli sarunā un rakstu valodā?" (Lokmane u. c. 2009, 20) Tāpēc ar savākto producētās

valodas materiālu veikts diezgan vispusīgs darbs: iegūtos valodas paraugus transkribējot, marķējot un anotējot[14], izveidots gan runas, gan rakstu korpuss (Kalnbērziņa u. c. 2011, 1).

Mutvārdu tekstos, neizmantojot korpusa lingvistikas rīkus, analizētas tādas parādības kā paužu skaits un garums vai nepiemērotu izteikumu iespraušana runā, lai aizpildītu laiku, kas A līmeņa apguvēju valodā ir mazāk raksturīga nekā skolēniem ar vājākām prasmēm. Vēl uzmanība pievērsta leksikai (Kalnbērziņa u. c. 2011, 14–15). Savukārt daļa no tekstiem (vairāk nekā 8700 vārdu) transkribēti un iekļauti korpusā tā mutvārdu daļā. Tālāk analizēts vārdu skaits teikumā, vārdu garums utt. (Kalnbērziņa u. c. 2011, 17)

Rakstu korpuss veidots no esejām, tās transkribējot, kārtojot pēc tematikas, anotējot un analizējot morfoloģiski un sintaktiski. Korpusa „Esam" kontekstā īpaši vērtīgs ir anotēšanas kategoriju (Lokmane u. c. 2009, 91–92) un problēmgadījumu apraksts (Kalnbērziņa u. c. 2011, 18–19) (par šo anotējumu sīkāk sk. 2.2. nodaļā „Tekstu marķēšana un anotēšana").

Pētījums galvenokārt ir orientēts uz eksāmena validāciju un pielīdzināšanu *Eiropas kopīgajās pamatnostādnēs valodu apguvei* (EKP 2006) šķirtajiem līmeņiem. Tas sniedz arī ieteikumus turpmākajam darbam, t. sk., izmantojot jaunizveidoto korpusu:

> *Sintaksē būtu nepieciešams apzināt saikļa vārdu dažādību gan sakārtojumā, gan pakārtojumā, biežāk lietotos palīgteikumu tipus, detalizēti pētīt jaukto salikto teikumu struktūru un semantiku. Morfoloģijā pētāms ne tikai vārdšķiru lietojuma biežums, bet arī leksēmu dažadība* [sic!] *katrā vārdšķirā (piemēram, vietniekvārdu grupas, īpašības vārdu semantiskās grupas). Uz korpusa bāzes iespējams pētīt tipiskās formveidošanas kļūdas katrā vārdšķirā.* [..] *uz izpētīta korpusa bāzes var veidot uzdevumus; no korpusa izvēlētus kļūdainus teikumus, vārdu savienojumus .. vai vārdu formas .. analizēt klasē.* (Kalnbērziņa u. c. 2011, 20)

Publikāciju par latviešu valodas apguvēju korpusu ir maz (Kalnbērziņa u. c. 2011, Kalnbērziņa 2015), un pārējās autores, izņemot Z. Vinčelu, savā pētnieciskajā darbā vairāk pievēršas citām tēmām. Diemžēl ar šo korpusu daudz plašāk vairs nav strādāts, taču korpusā iekļautais materiāls ir saglabāts, tāpēc darbu var turpināt; autores norāda, ka būtu īpaši vēlams strādāt divos virzienos: (1) papildināt esošo korpusu un tā anotējumu un (2) izveidot arī atbilstošu pieaugušo valodas apguvēju korpusu, „lai radītu drošu pamatu latviešu valodas apguves pārbaudes un mācību materiālu izveidei" (Kalnbērziņa u. c. 2011, 1). LVASA

[14] Pētnieces pašas runā tikai par marķēšanu, lietojot šo terminu abās nozīmēs – terminu *anotēšana* un *marķēšana* šķīrums šajā darbā skaidrots 1.1.2. nodaļā „Valodas apguvēju korpusa raksturīgie parametri".

pētījuma mērķi gan ir sasniegti, tāpēc, lai turpinātu apkopotā materiāla pētniecību, būtu nepieciešama jauna iniciatīva.

Latvijā valodas apguvēju korpusus izmanto arī V. Rūtenberga. Viņa, līdzīgi Z. Vinčelai, izveidojusi savu valodas apguvēju korpusu, kurā balstīto promocijas darbu „Sintaktiskās kriteriālās pazīmes angļu un franču valodas rakstveida snieguma vērtēšanā" aizstāvējusi Latvijas Universitātē (Rūtenberga 2014). V. Rūtenberga galvenokārt interesējas par sintaktisko konstrukciju lietojumu kā rādītāju par vispārējo valodas prasmju līmeni. Viņa ir apkopojusi vidusskolas beidzēju angļu un franču valodas eksāmenu esejas un materiālus sintaktiski anotējusi. Ir arī norādīts, kādam valodas prasmes līmenim pēc eksāmena vērtējumiem katrs no tekstiem atbilst, tādējādi radot iespēju pētīt sakarības starp valodas prasmes līmeni un sintakses lietojumu (Rūtenberga 2012; Rūtenberga, Kalnbērziņa 2013).

Arī Indra Karapetjana promocijas darbā „Bakalaura darbu valoda kā lingvofunkcionālās kompetences attīstības rezultāts" (Karapetjana 2007) strādā ar korpusu, ko var uzskatīt par valodas apguvēju korpusu – viņa ar korpusa lingvistikas rīkiem pēta piecdesmit bakalaura darbus angļu valodā, kurus rakstījuši Latvijas studenti. Šeit uzmanība pievērsta tekstu struktūrai – pētniece adaptē žanra analīzes metodoloģiju un korpusā manuāli anotē bakalaura darbos atkārtojošos retoriskos elementus, kas pēc tam tiek aplūkoti gan kvantitatīvā, gan kvalitatīvā griezumā (Karapetjana 2007, 149). Valodas apguvēju korpusa jēdziens šajā promocijas darbā, šķiet, tā arī neparādās.

Savukārt Nataļja Cigankova promocijas darbā „Lingvistiskā variācija elektroniskajā akadēmiskajā diskursā" (Cigankova 2009) apraksta darbu ar korpusu, kurā ir iekļauti akadēmiska diskursa teksti angļu valodā. Šo tekstu autori pārstāv plašu dzimto valodu klāstu – kopā 42 valodas, un ne vairāk par 0,3 % autoru dzimtā valoda ir angļu valoda (Cigankova 2009, 81). Korpusā ir iekļauti gan rakstveida, gan mutvārdu valodas paraugi, turklāt īpaši uzsvērts: „Tika pieņemts, ka dalībnieku angļu valodas prasme bija pielīdzināma dzimtās valodas prasmei, jo pētījuma laikā tie pasniedza akadēmisko rakstīšanu angļu valodā universitātēs." (Cigankova 2009, 81) Tātad par valodas apguvēju korpusu šo var uzskatīt tikai nosacīti.

Šie abi piemēri uzskatāmi ilustrē atšķirību starp valodas apguvēju korpusiem un nedzimtās valodas korpusiem (sk. 1.1.2. apakšnodaļā „Valodas apguvēju korpusa raksturīgās īpašības"). Arī N. Cigankovas un Z. Vinčelas kopīgie raksti ir par variatīvuma izpēti ar korpusa lingvistikas metodēm dažādos korpusos, nevis tieši valodas apguvēju korpusos (Cigankova, Vinčela 2012; 2013).

N. Cigankovas un V. Kalnbērziņas vadībā tapuši arī vairāki studentu pētījumi, kuros izmantoti angļu valodas apguvēju korpusi (Bērziņa 2013; Mihailova 2015; Līduma 2013;

Ločmele 2015; Dzērve 2013). Tajos iekļauti augstskolas studentu, vidusskolēnu un pamatskolas beidzēju teksti, un šķiet, ka 9. klases nobeiguma eksāmenu darbi pārstāv zemāko valodas prasmju līmeni, kāds Latvijā ir pētīts, izmantojot valodas apguvēju korpusus. Studiju gala darbos angļu valodas apguvēju valodas paraugi ir pētīti arī bez korpusa lingvistikas metožu iesaistīšanas (Lokastova 2007; Savenkova 2011). Lai arī minētie studiju galadarbi pagaidām plašākos pētījumos nav izvērtušies, to esamība rāda, ka Latvijā ir gan interese, gan iestrādnes valodas apguvēju korpusu pētniecībā.

Neviens no Latvijas pētnieku izmantotajiem valodas apguvēju korpusiem nav publiski pieejams.

1.2.2.2. Valodas apguvēju korpusi Lietuvā

Īpaši šī darba kontekstā nozīmīgs ir fakts, ka latviešu-lietuviešu valodu pāris valodas apguvēju korpusu kontekstā valodniekus ir ieinteresējis jau iepriekš. A. Zujevaite E. Žilinskaites vadībā Viļņas Universitātē digitalizēja šajā universitātē latviešu valodu mācījušos studentu rakstītos tekstus, kas pēc tam (datu zuduma dēļ digitalizējot no jauna) tika iekļauti šī promocijas darba 2. daļā raksturotajā otrās baltu valodas apguvēju korpusā „Esam" (sīkāk sk. apakšnodaļā nr. 2.1.1 *Tekstu ieguve un atlases kritēriji*). Tekstu digitalizācija tika veikta ar mērķi izveidot valodas apguvēju korpusu, un, lai arī tobrīd tas netika izdarīts, jau darba gaitā radušies novērojumi apkopoti nelielā publikācijā. Tajā stāstīts par biežāk sastopamajām četrām kļūdu grupām:

- pareizrakstības kļūdas – it īpaši diakritisko zīmju trūkums vai pārdaudzums, patskaņu un divskaņu šķīrums, īpašvārdu atveide;
- morfoloģiskās kļūdas – nepareiza locījuma lietojums, kļūdas piederības vietniekvārdu saskaņojumā un piederības konstrukcijā *man ir*;
- leksikas kļūdas – lietuviešu valodas vārda lietojums gadījumā, ja nav zināms latviskais; jaundarinājumi;
- semantikas kļūdas – personu vietniekvārdu lietojums norādāmo vietniekvārdu vietā, arī citu lietuviešu valodas vārdiem līdzīgu latviešu valodas vārdu neiederīgs lietojums pēc lietuviešu valodas parauga (Zujevaitė, Žilinskaitė 2012).

Lai arī šī publikācija nav apjomīga un turpmāki pētījumi ar tajā izmantoto materiālu līdz šim nav veikti, tas ir vērtīgs ieskats latviešu valodas apguvē un ir noderīgs arī otrās baltu

valodas korpusa „Esam" izstrādē (vairāk sk. 2.2. nodaļā „Tekstu apstrāde, marķēšana un anotēšana").

Vairums citu valodas apguvēju korpusu pētnieku Lietuvā ir pievērsušies angļu valodai. Augsti vērtējot iespēju iekļaut jaunos pētījumus jau esošā plašākā kontekstā, daļa no esošajiem valodas apguvēju korpusiem ir veidoti pēc citviet pastāvošu korpusu parauga, un ievērojot to vadlīnijas, lai savākto materiālu varētu iekļaut tajos kā apakškorpusus.

Viens no būtiskākajiem paveiktajiem darbiem ir lietuviešu materiāla savākšana Starptautiskajam angļu valodas apguvēju korpusam (ICLE – *International Corpus of Learner English*). Tas sastāv no angļu filoloģijas bakalaura studiju trešā un ceturtā kursa studentu rakstītajiem tekstiem Viļņas Universitātē un Vītauta Dižā universitātē (plašāk sk. Grigaliūnienė u. c. 2008). Šajā materiālā balstītas publikācijas izdevušas četras autores: L. Bikeliene, R. Juknevičiene, J. Grigaļūniene un Ale Šimčikaite (*Alė Šimčikaitė*).

L. Bikeliene *ICLE* korpusa lietuviešu daļu salīdzinājusi ar citām, piem., zviedru un ungāru daļām, kā arī *LOCNESS* angļu valodas kā dzimtās valodas korpusu, plašā pētījumā aplūkodama dažādu saistītājvārdu grupu lietojumu un to raksturodama vairākās publikācijās (Bikelienė 2008a, 2008b, 2009a, 2009b, 2010), kā arī doktora disertācijā (Bikelienė 2012). Arī pēc zinātniskā grāda iegūšanas viņa turpinājusi angļu valodas saistītājvārdu lietojuma kontrastīvos pētījumus, izmantojot sastatījumam vēl vairākus citus angļu valodas kā dzimtās valodas korpusus (Bikelienė 2013), un pievērsusies arī īpašības vārdu lietojuma salīdzinājumam (Bikelienė 2016).

A. Šimčikaite pētījusi sarunvalodas diskursa marķierus angļu valodas apguvēju akadēmiskajos tekstos, un arī šeit, līdzīgi kā L. Bikelienes darbos, *ICLE* korpusa lietuviešu daļas dati salīdzināti ar angļu valodas kā dzimtās valodas korpusu *LOCNESS*, ar kontrastīvās analīzes palīdzību pārliecinoties – lai gan šādi diskursa marķieri ir sastopami arī dzimtās valodas runātāju tekstos, valodas apguvēji tos tomēr lietojuši biežāk (Šimčikaitė 2012).

R. Juknevičiene savukārt abos minētajos korpusos pēta kolokācijas ar bieži sastopamiem darbības vārdiem: *have, do, make, take* un *give*, arī šeit konstatēdama atšķirības starp lietojumu dzimtajā valodā un apgūstamajā valodā (Juknevičienė 2008).

Redzams, ka lietuviešu materiāli korpusā *ICLE* nereti tiek salīdzināti ar dzimtās valodas runātāju radītiem līdzīgiem tekstiem vai arī tā paša korpusa citu apakškorpusu materiālu. Taču salīdzinājums ir veikts arī ar valodas apguvēju mutvārdu valodu. Arī šādi dati ir pieejami – lietuviešu materiāls ir savākts arī Luvēnas Starptautiskajai angļu mutvārdu starpvalodas datubāzei (*Louvain International Database of Spoken English Interlanguage*; *LINDSEI*). Tā nosaukums ir *LINDSEI-LITH*, un tajā ir apkopoti Viļņas Universitātē angļu

valodu studējošo jauniešu valodas paraugi (Grigaliūnienė, Juknevičienė 2011, 13). J. Grigaļūniene un R. Juknevičiene šo korpusu salīdzina ar *ICLE* lietuviešu materiālu, pētot angļu valodas darbības vārdu formas ar izskaņu *-ing*, lai demonstrētu kontrastīvās pieejas iespējas (Grigaliūnienė, Juknevičienė 2012). Nedaudz vēlākā pētījumā šī pati korpusu kombinācija izmantota arī formulas rakstura izteikumu analīzei (Grigaliūnienė, Juknevičienė 2013). J. Grigaļūniene korpusus *ICLE* un *LINDSEI* un to lietuviešu daļas piemin arī savā mācību grāmatā par korpusa lingvistiku, kurā vesela nodaļa ir veltīta valodas apguvēju korpusiem un to pētniecībai (Grigaliūnienė 2013a, 56–67).

Arī *LINDSEI* korpusa lietuviskās daļas materiāls ir ticis salīdzināts ar dzimtās valodas runātāju korpusa materiālu (šajā gadījumā tas ir *LOCNEC – the Louvain Corpus of Native English Conversation*). J. Grigaļūniene ir skaidrojusi vārda *right* lietojumu tajos, konstatēdama atšķirības starp dzimtās valodas un apgūstamās valodas lietojumu gan kvalitatīvā, gan kvantitatīvā aspektā (Grigaliūnienė 2013c).

R. Juknevičiene promocijas darba izstrādes gaitā ir izveidojusi arī pati savu angļu valodas apguvēju korpusu. Tas ir līdzīgs *ICLE* korpusa materiāliem, taču R. Juknevičienes korpusā ietilpst pirmā, nevis trešā un ceturtā kursa angļu filoloģijas studentu rakstītie darbi. Šī korpusa nosaukums ir *AFK1,* un korpuss nav paredzēts publiskai pieejai. Autore to izmanto, lai pētītu tematam specifisku leksiku (Juknevičienė 2007) un leksiskos savienojumus salīdzinājumā ar *ICLE* korpusā iekļautajiem datiem un dzimtās valodas runātāju radītajiem tekstiem (Juknevičienė 2009).

Lietuvas Nacionālajā eksaminācijas centrā ir veidots arī vidusskolas beidzēju angļu valodas eksāmenu rakstu darbu korpuss (jeb *NEC* korpuss), uz kura pamata ir runāts gan par valodas prasmes līmeņa noteikšanu, gan arī par negaidītu veidu, kā nepieciešamība kārtot eksāmenu ietekmē mācību procesu: atklāts neparasti biežs noteiktu frāžu lietojums, kuras angļu valodā parasti tik bieži sastopamas nav. Pēc tam izrādījies, ka šīs frāzes ir iekļautas eksāmenu labotāju vadlīnijās un, pēc labošanas informācijai nonākot skolās, tās apgūtas pārlieku centīgi (Juknevičienė 2013a, 56). Šī korpusa izpētē R. Juknevičienei pievienojusies arī Inesa Šeškauskiene (*Inesa Šeškauskienė*) kopīgos pētījumos par valodas prasmes līmeņa izpausmēm un pazīmēm skolēnu eksāmenos rakstītajos tekstos (Juknevičienė, Šeškauskienė 2014a, 2014b).

Līdzīgi kā Latvijā, arī Lietuvā daļa zinātnieku korpusus veido paši savām izpētes vajadzībām. Šādi ir rīkojušās I. Šeškauskiene un N. Burneikaite, kas ir apkopojušas korpusā Viļņas Universitātē tapušos valodniecības nozares bakalaura un maģistra darbus angļu valodā. I. Šeškauskienes korpuss sastāv no bakalaura darbiem, un līdz šim tajā aplūkoti piesardzīga

formulējuma līdzekļi (Šeškauskienė 2008; vairāk sk. 2.2.2. apakšnodaļā „Anotēšanas veidu izvēle"). N. Burneikaites sastādītajā korpusā savukārt ir iekļauti maģistra darbi angļu valodā: vienā apakškorpusā ir angļu kā dzimtās valodas runātāju darbi, savukārt otrā – angļu valodas apguvēju darbi, kas tapuši Lietuvas universitātēs. Šajā korpusā pētniece analizē dažādus metadiskursa elementus (Burneikaitė 2007, 2008, 2009b, 2009c) u. c. izteiksmes līdzekļus (Burneikaitė 2006, 2009a, 2012, 2013), tai skaitā, piem., jautājumus (Burneikaitė 2011b) un pavēles (Burneikaitė 2011a). Arī pirms šī korpusa izveides N. Burneikaite ir interesējusies par valodas apguvēju tekstiem, kā to parāda jau 2003. gadā iznākusi publikācija par informācijas struktūru valodas apguvēju tekstos (Burneikaitė, Zabiliūtė 2003). Šajā darbā, šķiet, vēl nav runas par korpusu – drīzāk gan par pētījumam savāktu materiālu kopu.

Minētās valodnieces plaši sadarbojas savā starpā, gan veicot kopīgus pētījumus, gan izmantojot cita citas savākto materiālu. Tā L. Bikeliene vairākos pētījumos izmantojusi šādus korpusus:

- R. Juknevičienes izveidoto *AFK1* korpusu;
- korpusa *ICLE* divus lietuviešu apakškorpusus – attiecīgi Viļņas Universitātes un Vītauta Dižā universitātes studentu rakstu darbus;
- akadēmiskās angļu valodas apguvēju korpusa (*Corpus of Academic Learner English, CALE*) divus lietuviešu apakškorpusus, kuri attiecīgi satur studentu rakstītus kopsavilkumus un pētījumus.

Minētajos korpusos aplūkotas britu un amerikāņu angļu valodas pareizrakstības sistēmiskās atšķirības, piem., izskaņu *-ence* un *-ense* lietojums (Bikelienė 2015a) un norādes uz personu studentu akadēmiskajā prozā (Bikelienė 2015b).

Arī R. Juknevičiene līdzīgā veidā apkopojusi vairāku korpusu materiālu. Korpusos *NEC* un *LICLE* pētītas dažādas bieži atkārtojošās vārdu virknes (Juknevičienė 2013b), savukārt *ICLE* korpusa lietuviešu daļā (pielīdzināta B2-C1 valodas prasmes līmenim), *AFK1* korpusā (B2 līmenis) un *NEC* korpusā (vidusskolēnu rakstītie teksti angļu valodā, B1 līmenis) pētīta frazeoloģiskās kompetences attīstība (Juknevičienė 2014) un sadarbībā ar I. Šeškauskieni analizēts prievārdu *in* un *on* lietojums (Šeškauskienė, Juknevičienė 2015).

Jomas attīstība var mudināt interesentu pulkam pievienoties arī citus pētniekus. Iespējams, tieši tas, ka Lietuvā līdz šim jau ir bijis samērā plašs angļu valodas apguvēju korpusu pētījumu klāsts, kalpojis par papildu ierosmi Aušrai Janulienei (*Aušra Janulienė*) un Justinam Dziedravičum (*Justinas Dziedravičius*), aplūkojot apstākļa vārdu lietojumu saikļavārdu funkcijā studentu rakstītajās akadēmiskajās esejās angļu valodā, daļu pētījuma veikt ar korpusa

lingvistikas metodēm. Viņi tieslietu studentu darbus pēta gan ar kvalitatīvām, gan ar kvantitatīvām metodēm, ieskaitot noteiktu leksēmu biežuma noteikšanu ar *AntConc* programmu (Janulienė, Dziedravičius 2015, 81).

Neraugoties uz to, ka Lietuvā plaši tiek izmantoti vieni un tie paši angļu valodas apguvēju korpusi, kuri turklāt ir savstarpēji salīdzināmi, pētījumiem pašlaik ir samērā individuāls raksturs – pagaidām nav iznācis neviens plašāka apjoma darbs par angļu valodas apguvēju valodu Lietuvā dažādos aspektos.

Vēl jāpiemin Margita Brekla (*Margit Breckle*), kura ir viena no *AleSKo* vācu valodas apguvēju korpusa veidotājām (Dumont, Granger 2017) un pētniecēm (Zinsmeister, Breckle 2012). No 2009. gada līdz 2012. gadam viņa strādāja Viļņas Pedagoģiskajā universitātē (kopš 2011. gada pārdēvēta par Lietuvas Izglītības zinātņu universitāti, sk. Petryla 2015) un šobrīd iesāktos darbus turpina Vāsas Universitātē Somijā (Breckle 2015). Jāteic gan, ka viņas darbā ar valodas apguvēju korpusiem vispār nav materiāla, kas būtu saistīts ar latviešu vai lietuviešu valodu vai kas balstītos Baltijas valstīs iegūtos datos, un arī viņas kolēģi Lietuvā, kā šķiet, valodas apguvēju korpusus pētniecības darbā neizmanto. Tāpat arī kolēģi no citām Lietuvas universitātēm viņas darbus Lietuvas kontekstā neaplūko. Līdz ar to, lai arī M. Brekla ir veikusi vērtīgus pētījumus (piem., par diskursa anotēšanu valodas apguvēju korpusā, kā arī par īpatnībām vācu valodas apguvēju rakstītajos tekstos salīdzinājumā ar vācu valodas kā dzimtās valodas runātāju tekstiem, attiecīgi sk. Zinsmeister, Breckle 2010; Breckle, Zinsmeister 2010; 2012 u. c.), tiem nav tiešas ietekmes uz valodas apguvēju korpusu pētniecības virziena attīstību Baltijas valstīs.

1.3. Darbs ar valodas apguvēju korpusiem

Valodas apguvēju korpusi visā pasaulē tiek izmantoti gan kā izziņas avots valodu apguves un nedzimtās valodas pētniecībā, gan arī kā palīgrīks valodu apguvē, valodu skolotāju izglītošanā utt. Tāpat tie var palīdzēt datorlingvistiem, kas veido pareizrakstības pārbaudes rīkus datorprogrammām, taču jāteic, ka šajā virzienā vairāk tiek strādāts ar dzimtās valodas korpusiem.

Dažādās publikācijās uzsvērts (Reppen 2010, 31), ka, korpusu veidojot, ir nepieciešams noteiktā mērā paredzēt tā lietojumu. Valodas apguvēju korpuss no tehniskā viedokļa ir līdzīgs dažādu citu veidu korpusiem, un galvenās procedūras to lietojumā maz atšķiras. Korpusā iekļauto datu raksturs tomēr ietekmē to, uz kādiem jautājumiem šīs procedūras palīdz rast atbildes, un to, kādus secinājumus no iegūtajiem rezultātiem var izdarīt.

Valodas apguvēju korpusu izmantojumam ir divi galvenie virzieni. Viens no tiem ir pētniecības darbs – valodas apguvēju korpusos interese galvenokārt ir vērsta uz valodas apguves un attīstības jautājumiem (LDLTAL 2010, 138), proti, starpvalodas attīstību, valodas pārnesi, kā arī kļūdām valodas apguvē (LDLTAL 2010, 327) un faktoriem, kas ietekmē valodas apguvēju sniegumu (Granger 2008a, 259). Savukārt otrā virzienā – t. s. pedagoģiskajā virzienā – uzmanības centrā ir pedagoģisko rīku un metožu izstrādāšana un pielāgošana valodas apguvēju vajadzībām (Granger 2008a, 259). Pie šī virziena var pieskaitīt arī pašu korpusu izmantošanu valodas mācību procesā. Valodas apguves procesā gan, protams, var būt noderīgi ne tikai valodas apguvēju korpusi, bet arī citu veidu korpusi (Koo 2006).

Šajā nodaļā raksturots darbs ar valodas apguvēju korpusiem un galvenās metodes to izmantojumā. Vispirms raksturotas galvenās korpusa lingvistikas procedūras un to lietojuma īpatnības valodas apguvēju korpusa gadījumā. Pēc tam skaidrots, kādos pētījumos un ar kādām metodēm valodas apguvēju korpusi lielākoties tiek lietoti. Visbeidzot aprakstīts arī valodas apguvēju korpusu lietojums valodas apguvē un valodas apguvei paredzētu materiālu izstrādē.

Korpusa lingvistikā kopumā par galvenajām tekstu izpētes metodēm vai procedūrām (dažādos avotos lietoti dažādi termini) parasti uzskata biežuma sarakstu sastādīšanu, konkordanču rindu atlasi un kolokāciju analīzi (Barnbrook 2008, 24; McEnery, Hardie 2012, 1 u. c.). To lietojums ir plaši izplatīts darbā ar dažādu tipu korpusiem, ieskaitot specializētos korpusus – arī valodas apguvēju korpusus. Pašas par sevi šīs procedūras ir „ārkārtīgi vienkāršas, taču normālā gadījumā pat vistiešākajam to lietojumam pamatā ir sarežģīti lēmumi" (Barnbrook 2008, 24). Līdz ar to ir svarīgi ne tikai raksturot procedūras, bet arī skaidrot lēmumus, uz kuriem balstās to lietojums.

1.3.1. Korpusa lingvistikas procedūras darbā ar valodas apguvēju korpusiem

Viens no svarīgākajiem zinātniskās pētniecības elementiem ir metode. Kā norāda Vitālijs Koduhovs:

> *Par zinātniskās pētniecības metodi sauc kādas parādības vai parādību grupas pētīšanas paņēmienu un likumu kopumu. Zinātniskās pētniecības metode ir tai pašā laikā pētīšanas aspekts, kam pakļauti metodes rīcībā esošie pētīšanas paņēmieni, to izmantošanas metodika (tehnika, procedūra), kā arī iegūto rezultātu aprakstīšanas veidi.* (Koduhovs 1987, 217)

Turpat tālāk gan V. Koduhovs norāda arī, ka terminam „metode" iespējamas dažādas izpratnes, pats minēdams vairākas, t. sk. procedūru (Koduhovs 1987, 217).

Arī korpusa lingvistikā jautājums, ko uzskatīt par metodi (un ko – par metodoloģiju), netiek skaidrots viennozīmīgi. Rīki un paņēmieni joprojām attīstās, un tiem nav skaidru robežu, lai arī daži no tiem, piem., konkordanču rindu atlase, jau ieguvuši stabilu vietu kā korpusa lingvistikas pamatprocedūras (McEnery, Hardie 2012, 1). Šajā darbā tiek šķirti procedūras un metodes jēdzieni atbilstoši iepriekš citētajai definīcijai: termins *procedūra* tiek lietots, runājot par konkrētām darbībām, piem., konkordanču rindu atlasi, biežuma sarakstu sastādīšanu, kolokāciju noteikšanu u. c., savukārt *metode* tiek izprasta kā plašāks jēdziens, kurā ietilpst gan procedūras (viena vai vairākas), gan to kombinējums un dažādi lietojuma aspekti. Kā metodes piemēru šādā izpratnē var nosaukt kontrastīvo starpvalodas analīzi (vairāk par to sk. 1.3.2. apakšnodaļā „Valodas apguvēju korpusu izpētes metodes").

Līdz ar to šajā apakšnodaļā runāts par populārākajām korpusa lingvistikas procedūrām kvalitatīvā un kvantitatīvā izpētē un to lietojuma īpatnībām valodas apguvēju korpusos, cenšoties nevis aptvert visu korpusa lingvistikas procedūru klāstu, bet gan izcelt un raksturot galvenās.

Zinātniskās pētniecības metodes var iedalīt pēc dažādiem principiem, un viens no izplatītākajiem ir pētniecības metožu dalījums kvalitatīvajās un kvantitatīvajās metodēs. Tam pamatā ir uzskats, ka ikviens pētījums atbilst vienai no divām pretējām kategorijām:

1. pētījumiem, kas individuāli aplūko katru fenomenu ar maksimālu iedziļināšanos, jeb kvalitatīviem pētījumiem

   vai

2. pētījumiem, kas fenomenus skata grupās un cenšas vispārināt iegūtos rezultātus, jeb kvantitatīviem pētījumiem.

Atkarībā no tā, cik plaši tiek izprasts metodes jēdziens, var būt arī metodes, kurās kvalitatīva un kvantitatīva pieeja tiek kombinēta (sk. tālāk), taču arī šādā gadījumā šis šķīrums palīdz izvēlēties atbilstošas procedūras.

Lai arī pētniecības metožu iedalījums kvalitatīvajās un kvantitatīvajās metodēs ir ierastāks sociālajās zinātnēs, tas ir attiecināms arī uz valodniecību, ieskaitot korpusa lingvistiku. Strādājot ar valodas korpusiem, šis šķīrums nosaka veicamās procedūras un rīkus, ar kādiem apkopotais materiāls tiek apstrādāts.

1.3.1.1. Kvalitatīvās un kvantitatīvās metodes korpusa lingvistikā

Kvalitatīvo un kvantitatīvo metožu šķīruma pamatprincips ir redzams jau šo metožu grupu nosaukumos – proti, kvantitatīvās pieejas gadījumā lielāka nozīme ir noteikta veida objektu skaitam, apjomam, resp., „kvantitātei", mēģinot noskaidrot sakarības un tendences, savukārt kvalitatīvā pieeja pievērš vairāk uzmanības aplūkojamā objekta vai nedaudzu objektu

īpašībām, individuālajam raksturojumam, resp., „kvalitātei", cenšoties skaidrot arī konstatēto parādību cēloņus. Šī atšķirība izpaužas vairākos aspektos, kas uzskatāmības labad parādīti 1. tabulā „Kvalitatīvo un kvantitatīvo pētniecības metožu vispārīgs salīdzinājums".

1. tabula. Kvalitatīvo un kvantitatīvo pētniecības metožu vispārīgs salīdzinājums

Kvantitatīvās metodes	**Kvalitatīvās metodes**
Uzskata, ka pastāv viena kopīga realitāte visiem cilvēkiem (Benz, Newman 2008, 2): objektivitāte	Uzskata, ka realitāte ir sociāla parādība, var būt vairākas vienlīdz īstas realitātes (Benz, Newman 2008, 2): subjektivitāte
Dati tiek pētīti skaitliskā izpausmē (Dörnyei 2007, 19)	Pētāmie dati var būt dažādos formātos, ieskaitot tekstu (Dörnyei 2007, 19)
Pētījums sākas ar teoriju vai hipotēzi, kas tiek pārbaudīta statistiski (Benz, Newman 2008, 3)	Vispusīgi raksturo un skaidro pētāmo fenomenu; šo skaidrojumu var uzskatīt par teoriju (Benz, Newman 2008, 3)
Orientētas uz lēmuma pieņemšanu (Benz, Newman 2008, 8)	Orientētas uz risinājumu meklēšanu (Benz, Newman 2008, 8)

Valodniecībā kopumā šis šķīrums nav sevišķi ierasts un lielākoties pamatliteratūrā netiek pieminēts vai arī tiek pieminēts tikai garāmejot. Piem., *Valodniecības pamatterminu skaidrojošajā vārdnīcā* (VPSV 2007) šādu šķirkļu nav – tas gan ir sagaidāms arī tādēļ, ka šie nav valodniecībai specifiski termini. Arī V. Koduhovs šo pētniecības metožu klasifikācijas principu piemin tikai garāmejot (Koduhovs 1987, 219, 248), taču neskaidro atšķirības starp kvalitatīvajām un kvantitatīvajām metodēm.

Tomēr arī šīs nozares, resp., valodniecības, pētījumos dalījums kvalitatīvajās un kvantitatīvajās metodēs ir izmantots. Samērā bieži tas novērojams korpusa lingvistikas pētījumos (piem., Mair 1991, Doorslaer 1995, Durian 2002 u. c.).

Korpusa lingvistika tiek skaidrota kā „valodniecības apakšnozare, kas analīzei izmanto elektronisku uzkrātu tekstu kopumu" (VPSV 2007, 196), taču jāatzīst, ka tā nereti tiek dēvēta nevis par valodniecības apakšnozari, bet gan drīzāk par metodoloģiju (Andronova, Andronovs 2011, 41; Aston 2011, 2). Tam par pamatu ir atzinums, ka tā „nav valodniecības nozare tādā pašā izpratnē kā sintakse, semantika, sociolingvistika utt. Visas šīs disciplīnas koncentrējas uz kāda valodas lietojuma aspekta aprakstīšanu/skaidrošanu." (McEnery, Wilson 2001, 2) Korpusa lingvistika savukārt sniedz iespēju pētīt dažādus valodas lietojuma aspektus (ieskaitot, piem., jau minētajām valodniecības apakšnozarēm – sintaksei, semantikai,

sociolingvistikai – būtiskos), izmantojot noteiktus rīkus. Šos rīkus var izmantot gan kvalitatīvos, gan kvantitatīvos pētījumos. Sīkāks ieskats korpusa lingvistikā un tās attīstībā Latvijā sniegts vairākās publikācijās (Andronova, Andronovs 2011; Grūzītis 2012; Skadiņa, Vasiļjevs 2013; Skadiņa u. c. 2014 u. c.).

Kristians Mērs (Mair 1991, 68) sniedz apskatu par būtiskākajām kvalitatīvās un kvantitatīvās analīzes īpatnībām korpusa lingvistikā salīdzinošas tabulas veidā (sk. 2. tabulu „Manuāla un automātiska korpusu analīze").

2. tabula. Manuāla un automātiska korpusu analīze (tulkots no Mair 1991, 68)

	Manuāla analīze	**Automātiska analīze**
Datu ieguves veids	Analizē cilvēks (analīze atkarīga no konteksta)	(Marķēta/anotēta) korpusa automātiska apstrāde
Visvairāk piemērots	Fenomeniem, kas parādās reti	Fenomeniem, kas parādās bieži
Uzmanības centrā	Izņēmumi, īpatnējas gramatiskas formas, neatbilstības tendencēm, jauktas un hibrīda tipa konstrukcijas	Formāli nosakāmas (anotējamas) galvenās gramatiskās kategorijas
Gramatikas izpratne	Saprātīga eklektika (t. i., kategorijas nosaka jaukti, pēc strukturāliem, semantiskiem un funkcionāliem kritērijiem, ja nepieciešams, papildinot ar funkcionālu vai diskursā balstītu pieeju)	Kāds autonomas sintakses paveids (t. i., gramatiskās struktūras tiek definētas kā formāli algoritmi)
Precizitātes pakāpe	Augsta	Zema
Sagaidāmais rezultāts	Dati tiek aplūkoti saiknē ar kontekstu; primitīva statistika – reti tiek iets tālāk par konkrēto gadījumu skaita norādīšanu	Neskaidriem vispārinājumiem tiek sniegts stingrs statistisks pamats, noformulēts kā atbilde uz primitīviem jautājumiem

Tabulas autors gan koncentrējies galvenokārt uz gramatikas izpēti, taču tabulā norādītās īpatnības ar nelieliem pielāgojumiem var attiecināt arī uz citu valodas līmeņu pētniecību.

Neatkarīgi no metodes korpusa lingvistikā pastāv divas atšķirīgas pieejas: *korpusā balstīta* (angļu val. *corpus-based*) un *korpusa vadīta*[15] (angļu val. *corpus-driven*)[16] pieeja. Par *korpusā balstītu* pieeju runā tad, ja pētījuma jautājums vai hipotēze radusies nesaistīti ar korpusa izpētes darbu, un korpuss tiek izmantots, tikai lai atrastu pierādījumus – t. i., konkrētus piemērus, kas hipotēzi apstiprina vai apgāž. Šajā gadījumā analīzē nav nepieciešams uzrādīt visus noteiktā vārda (vārdformas, gramatiskās kategorijas utt.) lietojumus, kādi ir atrasti korpusā – var tos aplūkot fragmentāri un izmantot tikai īpaši interesantos un vērtīgos piemērus, lai ilustrētu pētāmo fenomenu (Marcinkevičienė 2000, 18). *Korpusa vadītas* pieejas gadījumā hipotēze vai pētījuma jautājums izriet no agrāk veiktas korpusa analīzes. Strādājot pēc šī principa, tiek veikta izsmeļoša korpusa datu analīze, aplūkojot pilnīgi visus tajā atrodamos pētāmās parādības piemērus (Marcinkevičienė 2000, 18).

1.3.1.1.1. Korpusa lingvistikas procedūru lietojums kvantitatīvās metodēs

Kā atzīst Sarma Kļaviņa, statistiskās metodes valodniecībā samērā plaši tika lietotas jau 20. gadsimta 20.–30. gados, un atsevišķi pētnieki tām pievērsušies vēl agrāk – 19. gadsimta vidū, taču pastāvīgu vietu valodniecībā tās ieņem kopš 20. gadsimta 50. gadiem (Kļaviņa 1980, 5). S. Kļaviņa piebilst arī, ka skaitļotāji (mūsdienās saukti par datoriem) īpaši atvieglo tādu pētniecību, kurā liela nozīme ir aprēķiniem (Kļaviņa 1980, 7). Līdz ar to datorlingvistikas, t. sk. korpusa lingvistikas viena no būtiskākajām priekšrocībām ir iespēja veikt plašus statistiskus aprēķinus, ja dati ir mašīnlasāmi.

Daži pētnieki, runājot par kvantitatīvo un kvalitatīvo metožu sadalījumu korpusa lingvistikā, tās dēvē par *manuālu* un *automātisku korpusa analīzi* (Mair 1991, 68). Ņemot vērā valodas korpusu specifiku, šāds formulējums nav uzskatāms par pilnīgi nepareizu, jo, kā skaidrots tālāk šajā nodaļā, visbiežāk statistisko aprēķinu algoritmi ir iebūvēti korpusu apstrādes rīkos, un pētniekam tikai precīzi jānorāda, kādu elementu attiecības nepieciešams noteikt. Savukārt korpusa datu dziļa kvalitatīva analīze automātiski nav veicama – datorlingvistikas rīki tajā var palīdzēt, taču kvalitatīvā pētniecībā kā salīdzinoši subjektīvā

[15] Šis termins latviešu valodniecībā nav nostabilizējies, jo par šo pieeju šķīrumu latviski nav daudz runāts. Līdz ar to angliskā termina *corpus-driven* latviskās atbilsmes meklējumi joprojām ir problemātisks jautājums, kuram būtu nepieciešams pievērsties atsevišķi.

[16] Tims Džonss (*Tim Johns*) ieviesa jēdzienu *data driven learning – datu vadīta mācīšanās* (Boulton 2011, 23), taču Elena Tonini-Bonelli (*Elena Tognini-Bonelli*) ieteica šķīrumu starp abām pieejām (Tognini-Bonelli 2001).

pieejā nozīme ir ne tikai stingri noteiktām datu kopām, bet arī pētnieka papildu zināšanām, skatījumam, viedoklim un dažkārt pat intuīcijai – dators to nodrošināt nevar. Tomēr nekritiski pārdēvēt par manuālām un automātiskām šīs metožu grupas nevar, jo arī korpusa datu kvantitatīvu analīzi ir iespējams veikt manuāli, ja tas nepieciešams (piem., ja nepieciešami specifiski aprēķini, kādu automātiska veikšana nav paredzēta ar pieejamajiem programmrīkiem). Turklāt arī datu anotēšana, kas principā ir kvalitatīva rakstura procedūra, var tikt veikta, ja ne pilnīgi automātiski, tad noteikti pusautomātiski (sk. tālāk).

Korpusa lingvistikā „līdz ar salīdzinoši vienkāršām aprakstošām statistiskām metodēm tiek izmantoti arī statistiskie testi, kolokāciju analīzes, daudzdimensiju statistiskās analīzes metodes un mašīnmācīšanās paņēmieni" (Lüdeling 2006, 28). Tātad procedūru klāsts ir gana plašs.

Viena no visbiežāk lietotajām procedūrām kvantitatīvajos pētījumos korpusa lingvistikā ir dažādu veidu biežuma sarakstu (angļu val. *frequency list*) sastādīšana. Šādi saraksti sniedz informāciju, cik reižu kurš vārds ir atrodams korpusā – tas ir absolūtais biežums. Taču, tā kā nereti ir nepieciešams salīdzināt dažādus korpusus, kuros iekļauto tekstu kopējais apjoms nav vienāds, bieži tiek veikta rezultātu apstrāde, iegūstot relatīvo biežumu – piemēru skaitu uz noteiktu apjomu, piem., uz 1000 vārdiem (Biber u. c. 2006, 263).

Morfēmu, vārdu un frāžu biežuma dati (angļu val. *frequency counts*) ir visvienkāršākie kvantitatīvie dati, ko var iegūt no neanotēta korpusa (Granger 1997, 177). Tie nereti tiek izmantoti salīdzinājumā ar to pašu morfēmu, vārdu un frāžu biežuma datiem dzimtās valodas runātāju radītos atbilstošos tekstos, norādot uz neatbilstoši biežu vai retu kāda valodas elementa lietojumu valodas apguvēju tekstos. Šāds salīdzinājums nereti sniedz ierosmi arī tālākai analīzei (Granger 1997, 177).

Korpusā iespējams statistiski noteikt kolokācijas (angļu val. *collocation*) jeb vārdu[17] kombinācijas, kuras noteiktā kontekstā parādās īpaši bieži (Baker u. c. 2006, 36)[18]. Tas nozīmē, ka, meklējot kādam vārdam kolokācijas, tiek atrasti tādi vārdi, kas attiecībā pret to piemēru skaitu korpusā statistiski biežāk ir atrodami meklētā vārda tiešā tuvumā (Marcinkevičienė 2010, 144).

[17] Vārds šajā gadījumā ir jebkura korpusā iekļautā materiāla daļa, ko pēc formāliem kritērijiem (piem., atstarpju esamība abās pusēs) kā vārdu atpazīst programmatūra. Tas nozīmē, ka vārds var būt arī, piem., grafiska simbolu virkne @#$%^, ja korpusā iekļautajos tekstos tāda tiek lietota necenzētas leksikas aizstāšanai.

[18] Šis termins var tikt saprasts nedaudz atšķirīgi atkarībā no valodniecības apakšnozares. Minētā definīcija atbilst korpuslingvistu izpratnei (sk. arī Baker u. c. 2006, 36-37), savukārt leksikoloģijā par kolokāciju tiek uzskatīta tāda vārdkopa, kurā starp vārdiem pastāv idiomātiska semantiska saite (Bussmann 1996, 200).

Vēl var meklēt arī koligācijas[19] (angļi val. *colligation*). Koligācija ir īpašs kolokācijas paveids – nevis „vārds ar vārdu[20]" tipa bieži sastopamās kombinācijas, bet gan „vārds ar vārdšķiru", „vārdšķira ar vārdšķiru" vai tml. tipa savienojumi. Tātad, ja par kolokācijām galvenokārt runā leksikas līmenī, tad gramatikas līmenī tiek meklētas koligācijas (Baker u. c. 2006, 36). Šī darbība ir iespējama tikai tad, ja vārdšķiras vai cita interesējošā gramatiskā informācija korpusā ir kodēta, resp., ja korpusa dati ir atbilstoši anotēti. Tātad ir nepieciešama korpusa datu kvalitatīva analīze, lai tos sagatavotu kvantitatīvai analīzei ar koligāciju meklēšanas paņēmienu.

Latviešu valodniecībā kolokācijas, ieskaitot koligācijas, līdz šim nav guvušas ļoti plašu ievērību, taču citur pasaulē, ieskaitot Lietuvu (Marcinkevičienė 2010), tiek pētītas daudz. Tas attiecas arī uz valodas apguvēju korpusiem. Piem., viena no atzītākajām valodas apguvēju korpusu pētniecēm S. Greindžere pētījusi kolokācijas angļu valodas apguvēju korpusā jau 1998. gadā, salīdzinādama tās ar dzimtās valodas runātāju tekstos atklātajām (Granger 1998b).

Strādājot ar kvantitatīvajām metodēm, būtiski pievērst uzmanību tam, ka datiem jābūt sistēmai saprotamiem, t. i., mašīnlasāmiem. Tā kā dators visus aprēķinus veic automātiski, pētniekam ir jāpārliecinās, ka tas spēj atpazīt nepieciešamās datu vienības (resp., vārdus, vārdformas u. c.) un atšķirt tās no citām. Atsevišķos gadījumos pētījuma jautājums var būt tik specifisks, ka nekāda papildu sagatavošana korpusā iekļautajiem tekstiem nav vajadzīga, tomēr visbiežāk pirms statistisko metožu izmantošanas nepieciešams korpusu marķēt vai anotēt. Tā ir kvalitatīva rakstura darbība, tāpēc par to vairāk runāts tālāk.

Kvantitatīviem pētījumiem var noderēt arī konkordanču rindu atlase, meklējot korpusa datos kādu noteiktu simbolu virkni vai marķētu/anotētu teksta elementu. Tā ir galvenokārt kvalitatīvajai pieejai noderīga metode, tāpēc ir sīkāk aprakstīta nākamajā nodaļā. Vienīgie kvantitatīva rakstura dati, ko no konkordanču rindām var iegūt, ir konkordanču rindu skaits – tas ir vienāds ar korpusā atrodamo meklētā elementa piemēru skaitu. Reizumis tiek veikti pētījumi, kas prasa šo skaitu salīdzināt ar, piem., cita līdzīga elementa piemēru skaitu tajā pašā korpusā vai arī ar tā paša elementa piemēru skaitu citā salīdzināmā korpusā. To var darīt, un tā jau būtu kvantitatīva pieeja, taču tā nav raksturīga tieši korpusa lingvistikai, tāpēc šeit netiek sīkāk aplūkota.

[19] Šis termins, šķiet, datorlingvistikas pētījumos latviešu valodā līdz šim nav lietots.
[20] *Vārds* šeit lietots iepriekš skaidrotajā nozīmē – kā jebkura formālajiem vārda kritērijiem atbilstoša simbolu virkne.

1.3.1.1.2. Korpusa lingvistikas procedūru lietojums kvalitatīvās metodēs

Kvalitatīvās pētniecības metodes korpusa lingvistikā nav būtiski atšķirīgas no kvalitatīvajām metodēm valodniecībā vispār. Galvenā atšķirība starp kvalitatīvo metožu lietojumu korpusa pētniecībā un šo metožu lietojumu cita materiāla (ne korpusa) izpētē ir pētāmo gadījumu atlasē. Korpusu analīzes rīki ļauj meklēt noteiktus elementus, kā arī skatīt materiālu konkordanču rindās (sk. tālāk), kuras var sakārtot alfabētiskā secībā pēc atrastajām simbolu virknēm vai tām blakus (pa labi vai pa kreisi) esošajiem simboliem.

Kā jau tika pieminēts, korpusa datu kvalitatīva pētniecība dažkārt tiek dēvēta arī par *manuālu korpusa analīzi* pretstatā automātiskai, akcentējot faktu, ka korpusu pētniecības rīki automātiski nespēj veikt kvalitatīvu analīzi, bet gan tikai sniegt pētniekam viņa meklētos datus viņam ērtā formā. Tālākais darbs jāveic manuāli.

Līdz ar to nevar nosaukt konkrētus kvalitatīvās pētniecības paņēmienus, kas attiektos tieši uz korpusa lingvistiku, jo to izvēli, lai arī ierobežo, tomēr nenosaka tehniskās iespējas – korpuss kalpo tikai kā datu avots.

Viena no visbiežāk pieminētajām un lietotajām procedūrām korpusa lingvistikā ir konkordanču rindu atlase (angļu val. *concordancing*). Konkordance ir „alfabētisks kādas grāmatas vai noteikta valodas korpusa vārdu un to lietojuma apkaimes rādītājs, kas ļauj fiksēt un analizēt noteiktu vārdu lietojumu un to biežumu tekstā" (LTSV 2011, 48). Tā kā datorprogramma nekādu papildu analīzi neveic, šī procedūra ir piemērota tieši kvalitatīvas analīzes veikšanai, lai arī, kā jau iepriekš minēts, noteiktā mērā var tikt izmantota arī kvantitatīvos pētījumos.

Pie kvalitatīvajām metodēm noteikti jāpiemin arī marķēšana (angļu val. *markup*) un anotēšana (angļu val. *annotation*)[21]. Tās nav gluži pētniecības metodes, taču ļoti bieži ietilpst korpusa sagatavošanas darbos, un zinātniskās pētniecības metodē ietilpst arī datu sagatavošana. Turklāt gan marķēšana, gan anotēšana ir specifiska tieši korpusa lingvistikai. Neanotēti korpusi ir tādi, kuriem nav pievienota nekāda papildu lingvistiska informācija. Angļu valodā to mēdz saukt arī par *raw corpus* (angļu val. 'neapstrādāts, jēls korpuss'; Granger 1997, 177).

Datus var marķēt un anotēt pēc dažādiem principiem. Marķēšana pie kvalitatīvajām metodēm iederas mazāk, jo tajā neietilpst datu analīze, bet gan tikai savāktā materiāla īpašību kodēšana mašīnlasāmā formā. Savukārt anotēšana, kā jau minēts, prasa iepriekšēju lingvistisku analīzi par katru anotējamo vienību (morfēmu, vārdu, teikumu u.tml.), tātad ietver kvalitatīvu izpēti.

[21] Par šo terminu šķīrumu sk. 1.1.2. nodaļā „Valodas apguvēju korpusa raksturīgās īpašības"

Pastāv dažādi anotējuma veidi. Izplatīta prakse ir vārdšķiru vai teikuma veidu anotēšana, norādot katrai attiecīgajai valodas vienībai atbilstošo raksturojumu. Populārs ir pamatformu anotējums, katrai vārdformai norādot pamatformu. Nereti ir sastopami arī tādi valodas apguvēju korpusi, kuros ir anotētas kļūdas, t. i., korpusā iekļautajos tekstos sastopamās neatbilstības attiecībā pret mērķvalodas normu (vairāk par kļūdas jēdzienu sk. 2.2.7. apakšnodaļā „Kļūdu anotēšana otrās baltu valodas apguvēju korpusā"). Tomēr šādam anotējumam ir arī trūkumi. Ne vienmēr ir viegli noteikt, ko tieši teksta autors ir gribējis uzrakstīt, tādēļ pastāv dažādas interpretācijas iespējas. Tāpēc valodas apguvēju korpusiem tiek piedāvāts arī cits anotējums – uz otrās valodas apguvi orientēta anotēšana (angļu val. *SLA-oriented tagging*), kas radīta ar nolūku pētīt starpvalodu kā fenomenu, kuram ir savas iekšējās likumības, kas ne vienmēr sakrīt ar mērķvalodas likumībām (Rastelli 2009). Tomēr šis anotēšanas veids nav plaši izplatīts.

Korpusa lingvistikā ļoti izplatīta ir sintaktiskā anotēšana, morfoloģiskā anotēšana, semantiskā anotēšana, bet it īpaši darbā ar valodas apguvēju korpusiem (tomēr ne tikai – sk., piem., Deksne, Skadiņa 2014) – kļūdu anotēšana korpusā iekļautajos tekstos. Šeit minētie anotēšanas veidi, protams, nav pilnīgs visu iespējamo anotēšanas veidu saraksts. Te nosaukti tikai tie anotēšanas veidi, kas ir īpaši nozīmīgi valodas apguvēju korpusu kontekstā.

Korpusa anotēšana tehniski notiek, izmantojot tagus[22] (angļu val. – *tag*). Anotējot korpusa datus, tiek izmantota tagu sistēma atbilstoši veiktajai analīzei. Piem., S. Greindžere ir sastādījusi sistēmu kļūdu anotēšanai valodas apguvēju korpusos, kurā katrai kļūdai tiek piešķirts kods atkarībā no līmeņa (fonētiskais, morfoloģiskais, sintaktiskais utt.) un kļūdas tipa (piem., morfoloģiskajā līmenī – dzimte, skaitlis, persona utt.), lai atvieglotu noteikta kļūdu veida atrašanu korpusā (Granger 2003a).

Abas šīs darbības – anotēšana un marķēšana – bieži tiek veiktas, lai atvieglotu korpusa datu statistisko analīzi vai sniegtu iespēju dažādot meklēšanas vaicājumus konkordanču rindu atlasē. It īpaši anotēšana papildus parāda korpusa datos problemātiskus jautājumus lingvistiskajā analīzē un rosina tos risināt (tas nepieciešams, lai korpusa anotēšanu pabeigtu un iegūtu konsekventu datu kopumu). Piem., anotējot vārdšķiras, jālemj, kā anotēt tādus pāru saikļus kā *gan-gan*, kas tekstā neatrodas blakus. Otrās baltu valodas apguvēju korpusā katra šāda saikļa daļa anotēta kā atsevišķs saiklis *gan*, taču par šo jautājumu var būt nepieciešamas arī plašākas diskusijas.

[22] Latvijas Zinātņu akadēmijas Terminoloģijas komisijas Informācijas tehnoloģijas, telekomunikācijas un elektronikas terminoloģijas apakškomisija šim jēdzienam apstiprinājusi divus terminus: *tags* un *birka* (Akadterm$_{-e}$), taču šķiet, ka *tags* datorlingvistikā tiek lietots biežāk, tāpēc tas lietots arī šajā darbā.

Kā jau iepriekš minēts, anotēšana ne vienmēr ir manuāls process. Ir izveidoti dažādi rīki (tai skaitā latviešu valodai), kas anotē korpusa tekstus automātiski pēc noteiktiem algoritmiem (automātiski morfoloģiski anotēts ir, piem., *Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss* – sk. Levāne-Petrova 2011; 2012b u. c.). Lietojot šos rīkus, gan jāņem vērā iespējamās kļūdas anotācijās, jo nereti šie algoritmi tomēr neaptver visus iespējamos gadījumus – traucē homoformas. Tādā gadījumā, ja pētnieks šādas anotēšanas rezultātu pārbauda un, kur tas ir nepieciešams, arī labo, šis anotēšanas process kopumā varētu tikt uzskatīts drīzāk par pusautomātisku.

Apguvēju tekstos mēdz būt īpaši daudz dažādu noviržu no sagaidāmajiem un algoritmos ietvertajiem likumiem, un automātiskās anotēšanas rīku pielāgošana var izrādīties nepietiekami efektīvs uzdevums, ņemot vērā ieguldāmos resursus (par to un iespējamajiem risinājumiem vairāk sk. Rastelli 2009). Tāpēc valodas apguvēju korpusu anotēšana parasti notiek manuāli. Piem., V. Rūtenberga savos pētījumos strādā ar pašas izveidotiem salīdzināmiem franču un angļu valodas apguvēju korpusiem (Rūtenberga, Kalnbērziņa 2013; Rūtenberga 2014 u. c.). Tos viņa ir manuāli sintaktiski anotējusi, piebilstot, ka „tas ir ārkārtīgi sīkumains un laikietilpīgs uzdevums, it sevišķi zemākajos valodas apguves līmeņos" (Rūtenberga, Kalnbērziņa 2013, 124).

1.3.1.1.3. Jaukta tipa pieeja

Kā jau iepriekš norādīts, katrai no pieejām – gan kvalitatīvajai, gan kvantitatīvajai – ir gan priekšrocības, gan trūkumi, tādēļ pētnieki nereti izvēlas jaukta tipa pieeju, izmantojot elementus gan no kvalitatīvajām, gan no kvantitatīvajām metodēm. Turklāt nereti tiek pausts uzskats, ka tām būtu jābūt nevis divām pretnostatītām kategorijām, kas nekādā ziņā nesaskaras viena ar otru, bet gan drīzāk rīkiem, ko var izmantot vienlaikus, laukiem, kas daļēji pārklājas, vai vienas skalas diviem pretpoliem (jaukta tipa metožu tipoloģija piedāvāta plašā klāstā dažādu publikāciju, piem., Benz, Newman 2008; Leech, Onwuegbuzie 2009 u. c.). Arī korpusa lingvistikā nav obligāti jāizvēlas tikai viena vai otra pieeja, ja pētniekam šķiet, ka noderīgas varētu būt abas.

Pieeja ir jaukta tipa, ja gan kvalitatīvajām, gan kvantitatīvajām metodēm raksturīgās pazīmes parādās vienā vai vairākos no šiem pētījuma posmiem:

- pētījuma mērķis;
- datu un darbību veids;
- analīzes veids;

- secinājumu veids (Leech, Onwuegbuzie 2009, 267).

Tas nozīmē, ka visdažādākās iepriekš aprakstīto un arī citu kvalitatīvo un kvantitatīvo metožu kombinācijas ir uzskatāmas par jaukta tipa pieejas piemēriem. Jāatzīst, ka gan kolokācijas, gan koligācijas, kas atrastas ar statistisku aprēķinu palīdzību, var pēc tam analizēt arī kvalitatīvi, atlasot konkordanču rindas. Arī biežuma sarakstos izvēlētus īpaši interesējošus elementus konkordanču rindās var pētīt kvalitatīvi.

No otras puses, ne katrā kvalitatīvas pētniecības gadījumā iespējams arī kvantitatīvs darbs. Tas ir lielā mērā atkarīgs gan no pētījuma jautājuma, gan no korpusa uzbūves un anotāciju veida. Tomēr arī šāda secība ir bieži sastopama pētījumos. Tā kā pirms kvantitatīvo metožu izmantošanas nereti nepieciešams, lai korpuss būtu atbilstoši anotēts, tad, ja korpusa sagatavošana ir viens no tā paša pētījuma uzdevumiem, tā noteikti iekļauj vismaz kaut kāda līmeņa kvalitatīvu analīzi. Tāds ir, piem., Z. Vinčelas pētījums par apstākļa vārdiem studentu rakstītajos tekstos angļu valodā (Vinčela 2013b): viņa sāk ar konkordanču rindu atlasi un manuālas kvalitatīvas analīzes rezultātā iegūst materiālu – interesējošos apstākļa vārdus –, kuram pēc tam tiek aprēķināts katras vienības biežums un salīdzināts ar citu pētnieku atklāto (Vinčela 2013b, 217). Savukārt promocijas darbā (Vinčela 2010a) pētniece sāk ar korpusa tekstu strukturālu marķēšanu, kvalitatīvu analīzi un korpusa anotēšanu, tad izgūst no tā konkordanču rindas ar interesējošajiem elementiem un to skaitu salīdzina statistiski (Vinčela 2010a, 86). Biežums gan abos pētījumos aprēķināts nevis ar specifiski korpusa lingvistikai raksturīgām procedūrām, bet gan manuāli vai ar vispārīgas statistikas rīkiem veicot aprēķinus pēc skaitliskiem datiem, kas iegūti ar korpusa lingvistikas metožu palīdzību, resp., sastādot biežuma sarakstus un atlasot konkordanču rindas.

Runājot par marķēšanu un anotēšanu – ja pētījuma uzdevumos ietilpst gan marķēšana/anotēšana, gan arī iegūtā marķētā/anotētā korpusa izpēte ar kvantitatīvu pieeju, tad tā jau ir jaukta tipa pieeja pētniecībai. Savukārt, ja pētījumā tiek izmantots jau iepriekš marķēts/anotēts korpuss, tad konkrētais pētījums tāpēc vien vēl nav uzskatāms par jaukta tipa pieejas paraugu.

Neanotētu korpusu var apstrādāt arī ar citu veidu īpašām programmām, iegūstot papildu lingvistisku informāciju, kas nav iegūstama ar ierastajām korpusa lingvistikas metodēm. Taču tās mēdz būt valodspecifiskas. Piem., Maikla Brenta (*Michael Brent*) izveidotā programma, kas veido darbības vārdu sarakstus no neanotētiem valodas apguvēju korpusiem un grupē šos darbības vārdus pēc lietojuma veida, ir tapusi tieši angļu valodai, un, lai to izmantotu citu valodu materiāla pētīšanai, būtu jāpārskata valodnieciskie likumi, ar kuriem saskaņā ir sastādīts programmas algoritms (Brent 1991, 210). Turklāt M. Brents norāda arī, ka

programmatūras darbu valodas apguvēju korpusā apgrūtina nepieciešamība atpazīt dažādas kļūdas (Brent 1991, 213).

1.3.2. Valodas apguvēju korpusu izpētes metodes

Valodas apguvēju korpusi ir attīstījušies nesen, taču tajos ir balstīts jau diezgan plašs pētījumu loks – piem., tiešsaistes valodas apguvēju korpusu bibliogrāfijā[23] 2008. gadā bija apmēram 300 vienību (Granger 2008a, 266), savukārt 2013. gada nogalē to bija jau apmēram 1100 – turklāt ne visas šajā jomā tapušās publikācijas tūlīt pēc iznākšanas nonāk bibliogrāfijas veidotāju redzeslokā.

Dž. Līčs min dažus tipiskus jautājumus, kuru izpētē noderīgs ir valodas apguvēju korpuss:

- Kādus valodas elementus mērķvalodas apguvēji lieto daudz mazāk vai daudz vairāk nekā dzimtās valodas lietotāji?
- Cik lielā mērā mērķvalodas lietojumu ietekmē dzimtā valoda?
- Kurās jomās bieži tiek izmantotas t. s. „izvairīšanās stratēģijas" (angļu val. *avoidance strategies*), apguvējam nespējot lietot pilnu valodas līdzekļu klāstu?
- Kurās jomās ir vieglāk, kurās – grūtāk sasniegt līmeni, kurā lietojums ir pielīdzināms dzimtās valodas lietojumam?
- Kādās jomās apguvējiem no noteikta apvidus ir visgrūtāk sasniegt lietojuma līmeni, kas ir pielīdzināms dzimtās valodas lietojumam?
- Kādās jomās pārmērīgs vai nepietiekams kāda valodas elementa lietojums, kā arī kļūdas ir vairāk raksturīgas vienas dzimtās valodas pārstāvjiem pretstatā citu dzimto valodu pārstāvjiem? (Leech 1998, xiv–xv)

Valodas apguvēju korpusus un to lietojumu mēdz aplūkot no valodas apguves un valodas mācīšanas skatpunkta (Granger 2002, 2). Valodas apguves pētnieki cenšas izprast mehānismu, kā cilvēks apgūst valodu, un raksturot starpvalodu kā patstāvīgu valodas paveidu, kamēr valodas mācīšanas skatpunkts koncentrējas uz valodas mācīšanas un mācīšanās efektivitātes paaugstināšanu. Tādējādi šie divi aspekti viens otru papildina. Pētījumi valodas apguvēju korpusos ne vien ļauj aprakstīt noteiktas apguvēju valodas īpatnības, bet arī sniedz iedvesmu un pamatojumu mācību metodikas attīstīšanai un pielāgošanai (Мальцева 2011, 209). Tādējādi tie sniedz ieguldījumu ne vien valodniecības, bet arī pedagoģijas nozarē (par to sk. arī 1.3.3. apakšnodaļā „Valodas apguvēju korpuss kā palīglīdzeklis valodas apguvē").

[23] Pieejama tiešsaistē: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html

S. Greindžere (Granger 2007, 52) norāda, ka valodas apguvēju korpusu izpētē parasti tiek izmantota viena no divām metodēm:

- kontrastīvā starpvalodas analīze (angļu val. *Contrastive Interlanguage Analysis*)

vai

- datorizētā kļūdu analīze (angļu val. *Computer-Aided Error Analysis*).

1.3.2.1. Kontrastīva starpvalodas analīze

Kontrastīvā starpvalodas analīze koncentrējas uz salīdzinājumu starp valodas apguvēju datiem un dzimtās valodas datiem vai arī starp divu dažādu valodas apguvēju grupu datiem (Granger 2008a, 267) gan kvalitatīvā, gan kvantitatīvā aspektā (Granger 2007, 52). Starpvalodas jēdziens skaidrots jau iepriekš, 1.1.1. apakšnodaļā „Valodas apguvēju korpuss: termins un definīcija", taču jāpiemin vēl divi termini – sakarā ar valodas apguvēju korpusiem bieži tiek runāts par *valodas pārneses* un *interferences* parādībām.

Valodas pārnese ir „dzimtās valodas ietekme uz mērķvalodu tās apguves procesā" (LTSV 2011, 62). Tai radniecīgs jēdziens ir interference. Šis termins tiek skaidrots kā „atkāpes no dzimtās valodas normām kontaktvalodas ietekmē, kā arī dzimtās valodas negatīva ietekme uz mērķvalodu tās apguves procesā" (LTSV 2011, 40). Interference dažkārt tiek definēta kā pārneses nevēlams, negatīvs paveids (LDLTAL 2011, 322–323). Kontrastīvā starpvalodas analīze gan nenosaka viena vai otra jēdziena lietojumu – tā neietver viena vai cita dzimtās valodas ietekmes veida pārākuma noteikšanu. Ja tas tiek darīts, tad tā ir attiecīgā pētnieka izvēle, kura ir atsevišķi pamatojama. Tomēr, ar šo metodi pētot valodas apguvēju korpusu, var gūt pamatojumu noteiktu parādību uzskatīšanai par dzimtās valodas ietekmētām.

Kontrastīvajai starpvalodas analīzei ir divi paveidi:

- starpvalodas jeb apguvēju valodas un dzimtās valodas kontrastīva analīze;
- divu vai vairāku starpvalodu kontrastīva analīze (Granger 2004, 128).

Starpvalodas un dzimtās valodas kontrastīva analīze „ļauj atklāt lietojuma modeļus, kas atšķir valodas apguvēju datus no dzimtās valodas datiem. Tos var iedalīt divās kategorijās: kvalitatīvās atšķirības (nepiemērots lietojums) un kvantitatīvās atšķirības (pārmērīgs vai nepietiekams lietojums)." (Granger 2003b, 541) Šādai analīzei gan ir arī trūkumi. Kā vienu no tādiem pētnieki min faktu, ka dzimtās valodas runātāja teksti tiek uzskatīti par mērķi valodas apguvējam (Granger 2004, 133). Tāpēc pēdējā laikā tiek vairāk runāts par dažādiem mērķvalodas paveidiem, kurus var izmantot „dzimtās valodas" vietā šādā analīzē (Granger 2015). Turklāt visbiežāk šādi pētījumi tiek veikti augsta valodas prasmes līmeņa

valodas apguvēju korpusos – šādos gadījumos ir lietderīgi uzzināt, kādos aspektos viņu producētās valodas paraugi tomēr vēl atšķiras no dzimtās valodas paraugiem (Granger 2004, 133).

Divu vai vairāku starpvalodu kontrastīva analīze savukārt „ir nepieciešama, lai noteiktu, vai atklātās atšķirības ir radušās valodas apguves procesa vai pārneses ietekmē" (Granger 2003b, 541). Tā parāda atšķirības starp dažādu dzimto valodu pārstāvju starpvalodām, esot vienai un tai pašai mērķvalodai.

Lielākā daļa pētījumu, kuros izmantota šī metode, tiek veikta neanotētos valodas apguvēju korpusos vai arī tādos korpusos, kuros ir anotētas vārdšķiras, ļaujot salīdzināt noteiktu vārdšķiru lietojumu (Granger 2004, 132).

1.3.2.2. Datorizēta kļūdu analīze

Lai arī lingvodidaktikas pētījumos kļūdu analīze tikusi kritizēta kā neskaidrs, nezinātnisks un neuzticams valodas apguves pētniecības veids (par to vairāk sk. James 2013, 15–18), dažādi pētnieki tomēr norāda, ka kļūdas ir neatņemama starpvalodas daļa (Granger 2003a, 466; LTŽ 2012, 92) un ka to analīze var būt noderīgs rīks valodas apguves izziņā (Ellis 1994, 20; Vanhaegendoren 2002, 35). Šim nolūkam tiek izmantoti arī korpusa lingvistikas rīki, it īpaši valodas apguvēju korpusi, kuros ir anotētas kļūdas. Pētnieki secinājuši, ka, atbilstoši lietojot šādus korpusus, samazinās tādas kļūdu analīzei raksturīgās problēmas kā kļūdu kategoriju neskaidrs sadalījums un datu neviendabīgums (Dagneaux u. c. 1998, 164).

Datorizētā kļūdu analīze no parastās kļūdu analīzes atšķiras ar to, ka sākas ar kļūdu anotēšanu korpusā – anotēt var visas atrodamās kļūdas vai arī tikai noteiktu kļūdu veidu –, un pēc tam šo anotējumu var izmantot, lai ar datorizētiem rīkiem dažādos griezumos kvantitatīvi vai kvalitatīvi aplūkotu anotētos kļūdu veidus (Granger 2003b, 542). Pētnieki gan norāda, ka šī metode ir ļoti darbietilpīga un līdz ar to nav tik plaši izmantota kā kontrastīvā starpvalodas analīze (Granger 2004, 133).

Uzsvērts, ka „pilnībā anotēts valodas apguvēju korpuss sniedz iespēju raksturot noteiktu valodas apguvēju populāciju pēc lielāko kļūdu kategoriju proporcijām" (Dagneaux u. c. 1998, 169) un ļauj arī salīdzināt ne vien kļūdu daudzumu, bet arī proporcijas starp dažādām populācijām (Dagneaux u. c. 1998, 169).

Minētās metodes var lietot arī kopā, lai labāk izprastu kāda valodas elementa lietojumu. Piem., S. Greindžere un Stefānija Taisone (*Stephanie Tyson*) veica kvantitatīvu kontrastīvu starpvalodas analīzi un konstatēja, ka angļu valodas apguvēji saistītājvārdus kopumā lieto tikpat daudz kā dzimtās valodas lietotāji. Taču kvalitatīva kontrastīvā analīze un

datorizētā kļūdu analīze atklāja, ka noteikti saistītājvārdi lietoti būtiski retāk, citi – būtiski biežāk, un daļa lietoti neiederīgi (Granger, Tyson 1996). Līdzīgi vienā pētījumā izmantojot gan datorizēto kļūdu analīzi, gan kontrastīvo starpvalodas analīzi, atklāts, ka angļu valodas apguvēji, kuru dzimtā valoda ir franču valoda, nepieļauj sevišķi daudz kļūdu frāzēs, kas satur darbības vārdu *make* – toties lieto šādas frāzes daudz retāk nekā angļu valodas kā dzimtās valodas lietotāji (Gilquin 2007). Šāds kombinēts metožu lietojums ļauj neizdarīt pārsteidzīgus secinājumus par noteiktas valodas parādības lietojuma prasmi.

Visbeidzot jāpiebilst arī, ka valodas apguves datus var izmantot arī kādas valodas gramatikas aprakstīšanā vai teorētiskajā valodniecībā (par to sīkāk sk. Jordens 2003), taču šis izpētes lauks netiek uzskatīts par primāru darbā ar valodas apguvēju korpusiem un publikācijās parasti netiek pieminēts.

1.3.3. Valodas apguvēju korpuss kā palīglīdzeklis valodas apguvē

Valodas apguvēju korpuss var būt noderīgs ne tikai pētniekiem, bet arī pašiem valodas apguvējiem. Jau 20. gs. beigās valodnieki aicināja valodas mācīšanās un mācīšanas procesā izmantot dažādus datorizētus rīkus, ieskaitot korpusus (vairāk sk. Bankava, Vinčela 1999; Grigaliūnienė 2013b), taču valodas apguvēju korpusi tiem pievienojās vēlāk. Mūsdienās to popularitāte arī šajā jomā strauji aug, un zinātnieki savās publikācijās aplūko veidus, kā šādus korpusus izmantot svešvalodas mācību procesā (Altay, Tilfarlioğlu 2012 u. c.).

Nereti šīs iespējas gan paliek novārtā, jo informācija par valodas apguvēju korpusiem praktizējošo pedagogu uzmanības lokā nonāk ierobežotā apjomā, galvenokārt – koncentrējoties uz angļu valodu. Mazāk populāru valodu mācīšanā lielākoties valodas apguvēju korpusus nodarbībās izmanto tie pedagogi, kas nodarbojas arī ar šādu korpusu pētniecību un/vai veidošanu.

Vispārīgu pārskatu par valodas apguvēju korpusu lietojumu lingvodidaktikā sniedz F. Menjē (Meunier 2007) un S. Greindžere (Granger 2008b). Viņu veiktie apraksti nedaudz atšķiras, un pilnīgākas ainas gūšanai ir derīgi aplūkot abu autoru sniegtos raksturojumus.

F. Menjē korpusu lietojumu pedagoģijā skata trīs plašos darbības laukos:

- mācību programmu izstrāde;
- mācību līdzekļu un rīku izstrāde;
- gramatikas mācīšana nodarbībās (Meunier 2007, 26).

Mācību programmu izstrāde

Mācību programmu izstrādē ir lietderīgi izmantot gan dzimtās valodas, gan valodas apguvēju korpusus (vairāk sk. Meunier 2000). Dzimtās valodas korpusi palīdz noteikt biežāk lietotos valodas elementus, taču, nosakot, kas būtu mācāms, ir jāņem vērā ne tikai attiecīgo valodas elementu lietošanas biežums, bet arī to apguves grūtības pakāpe un konsekvence mācību procesā. Valodas apguvēju korpusi palīdz noteikt formas, kas attiecīgās valodas apguvējiem sagādā vairāk grūtību, līdz ar to ļaujot secināt, kas būtu mācāms agrāk, kas – vēlāk un kam būtu jāpievērš vairāk uzmanības. Jāatzīst gan, ka mācību programmu izstrādē korpusi pašlaik vēl tikpat kā netiek lietoti (Meunier 2007, 27).

Mācību līdzekļu un rīku izstrāde

Mācību līdzekļu un rīku izstrādē korpusa lingvistikas lietojums ir daudz plašāk izplatīts, un ir sākts izmantot arī valodas apguvēju korpusus. F. Menjē gan šajā grupā piemin galvenokārt vārdnīcas, turklāt koncentrējoties lielākoties uz dzimtās valodas korpusu lietojumu to izveidē. Mācību grāmatas tiek aplūkotas kā daļa no gramatikas mācīšanas nodarbībās, kaut arī par šādu dalījumu varētu diskutēt, ņemot vērā arī to, ka autore pati tomēr nošķir mācību grāmatu izstrādi no mācību metodoloģijas (Meunier 2007, 32).

Kā viens no veidiem, kā mācību līdzekļu izstrādē var izmantot valodas apguvēju producētos datus, ir minēts kļūdu labošanas uzdevums, kurā tiek sniegti piemēri no korpusa ar autentiskām kļūdām, kuras apguvējam jālabo, attīstot spēju atpazīt problemātiskos gadījumus. Būtiski plašākas lietojuma iespējas gan netiek minētas, skaidrojot, ka lielākoties arī tādas mācību grāmatas, kuru sastādīšanā ir izmantoti korpusu dati, pēc uzbūves parasti neatšķiras no tradicionālajām (Meunier 2007, 32–33).

Gramatikas mācīšana nodarbībās

Gramatikas mācīšanu nodarbībās, ar to saprotot mācīšanas procesā izmantojamās metodes, korpusa lingvistika pēdējā laikā ir ietekmējusi īpaši plaši, attīstoties jaunam novirzienam, kas ir pazīstams kā *datos balstīta mācīšanās* (angļu val. *data-driven learning*) un paredz korpusa datu izpēti mācību laikā, ļaujot apguvējiem pašiem pētīt mērķvalodā bieži atkārtojošās parādības, lai tās raksturotu vai atrastu atbildes attiecīgi sastādītiem uzdevumiem (Meunier 2007, 33). Tas gan parasti tiek veikts ar dzimtās valodas korpusu datiem, lai nodrošinātu valodas apguvējiem pēc iespējas vairāk saskarsmes ar autentisku mērķvalodu, tomēr tos var izmantot arī kombinācijā ar valodas apguvēju datiem. Piem., tos salīdzinot, apguvējs var konstatēt atšķirības, un šāda apzināšanās var palīdzēt apguvējam tuvināt savu valodas lietojumu dzimtās valodas lietojumam (vairāk sk. Granger, Tribble 1998).

Jāpiebilst, ka apguvēju valodas paraugiem nebūtu jāaizstāj dzimtās valodas paraugi – katram no šiem valodas paraugu veidiem ir sava funkcija: vieni sniedz informāciju par to, kā mērķvalodā tiek veidoti izteikumi, savukārt otri ļauj saprast iespējamās neatbilstības, un, „pamanot atšķirību starp savām un mērķvalodas formām, apguvēji arī labāk spēj paātrināt mācīšanos" (Joyce, Burns 1999, 48, citēts no Meunier 2007, 35).

F. Menjē nosauc šādas priekšrocības datos balstītai valodas apguvei, izmantojot gan dzimtās valodas, gan valodas apguvēju korpusu:

- apguvējs kļūst par pētnieku – atklājuma prieks, papildu motivācija;
- ir pieejamas gan kļūdas un novirzes, gan pareiza, valīda informācija;
- šādas darbības veicina sadarbību un diskusijas;
- apguvējs labāk izprot valodas kā sistēmas dabu (Meunier 2007, 37).

Šādai pieejai gan ir arī trūkumi – tā prasa daudz laika un labu sagatavošanos gan uzdevumu izveidē, gan tehniskā aprīkojuma nodrošināšanā. Turklāt jāņem vērā, ka ir studenti, kuri pašapziņas trūkuma vai citu iemeslu dēļ nelabprāt strādā induktīvi un dod priekšroku tādam mācību procesam, kurā valodas likumības izskaidro pasniedzējs, nevis tās ir jāsecina pašam. Līdz ar to korpusa lingvistika var sniegt vērtīgu papildinājumu esošajām mācību darbībām, taču nevajadzētu censties ar to aizstāt visas līdzšinējās mācību metodes (Meunier 2007, 40).

S. Greindžere pedagoģiskās darbības laukus saistībā ar valodas apguvēju korpusiem iedala citādi – nevis pēc iespējamā valodas apguvēju korpusu izmantojuma, bet gan pēc tā, kā tie līdz šim galvenokārt jau tikuši izmantoti lingvodidaktikā. Viņas nosauc šādus pedagoģiskās prakses virzienus:

- pedagoģiskā leksikogrāfija;
- mācību materiālu izstrāde;
- valodas prasmes vērtēšana (Granger 2008b, 347–349).

Pedagoģiskā leksikogrāfija

Mācību vārdnīcu izstrāde, šķiet, visagrāk sāka izmantot valodas apguvēju korpusu sniegtās priekšrocības. Īpašas valodas apguvējiem paredzētas vārdnīcas (piem., LDOCE 2003, CALD 2003) papildus citai leksikogrāfiskajai informācijai sniedz arī piezīmes par bieži sastopamām kļūdām, lai palīdzētu apguvējiem tās atpazīt un no tām izvairīties (De Cock, Granger 2005). S. Greindžere norāda, ka būtu vērtīgi līdzīgā veidā bieži sastopamas kļūdas norādīt arī valodas apguvējiem paredzētās vārdnīcās (Granger 2008b, 348). Lai arī, piem., vācu valodas kļūdu leksikonā (Heringer 2001) nav norādīts, vai tas sastādīts, izmantojot korpusu,

vai bez tā, valodas apguvēju korpuss šādu vārdnīcu sastādīšanu ievērojami atvieglo, jo, pareizi sagatavots, var gan sniegt datus par kļūdu biežumu, gan arī ļaut atrast noteiktu kļūdu tipu dažādus piemērus, neveidojot atsevišķu kartotēku.

Mācību materiālu izstrāde

Mācību materiāli līdz šim galvenokārt tikuši izstrādāti pašu materiālu autoru un to audzēkņu lietošanai, nevis komerciālai izplatīšanai. S. Greidžere nosauc īpašības, kas tiem visbiežāk ir kopīgas:

1) šo materiālu darbības princips balstās valodas apguvēju korpusu tiešā, nevis pastarpinātā pedagoģiskā izmantojumā;
2) šie materiāli visbiežāk ir paredzēti valodas apguvējiem ar noteiktu dzimto valodu;
3) šie materiāli kalpo skaidri noteikta mērķa sasniegšanai;
4) šie materiāli ir elektroniski (Granger 2008b, 348).

Abi iepriekšējie S. Greindžeres nošķirtie darbības lauki pēc F. Menjē dalījuma ietilpst mācību līdzekļu un rīku izstrādes laukā. Savukārt trešais – **valodas prasmes vērtēšana** – nevienam no F. Menjē nosauktajiem nebūtu pieskaitāms. Valodas apguvēju korpusu analīze var atklāt tendences valodas lietojumā, kas ir īpaši raksturīgas kādam noteiktam valodas prasmes līmenim – piem., angļu un franču valodas apguvēju tekstos teikumu veidu sadalījums atšķiras atkarībā no attiecīgās valodas prasmes līmeņa (Rūtenberga 2014). Šāda veida informāciju tālāk var izmantot, veidojot sistēmas, kas automātiski nosaka valodas apguvēja prasmes līmeni, analizējot noteiktas vienības tajā (Granger 2008b, 349).

Lielākā daļa valodas apguvēju korpusu pedagoģiskā izmantojuma veidu ir saistīti ar izpēti, gatavojoties nodarbībām un izstrādājot tām paredzētos materiālus, tāpēc pasniedzējam, kurš pats šajā procesā nav iesaistīts un galvenokārt nodarbojas tikai ar tiešo mācīšanu, valodas apguvēju korpusi tiešā veidā nav noteikti nepieciešami. Tomēr arī šie pasniedzēji gūst labumu no šo korpusu izpētes, izmantojot citu valodnieku izstrādātos materiālus, turklāt tieši šādiem pasniedzējiem nereti ir plaša pieeja valodas apguvēju producētajam materiālam, kas, nodots pētnieku rokās, veicinātu šādu materiālu izstrādi. Līdz ar to arī gadījumos, kad pasniedzējs apzināti izvēlas pats valodu apguvēju korpusus neizmantot, sadarbība ar pētniekiem, kas to dara, var sniegt vērtīgu ieguldījumu nozares attīstībā.

Lai arī valodas apguvēju korpusu izmantojums var sniegt būtiskas priekšrocības valodas apguves procesā, nav sagaidāms, ka tas ievērojami mainītu līdzšinējo mācību praksi kopumā. Pētnieki norāda, ka svarīgākais ir nevis panākt, lai korpusus vai tajos balstītus materiālus lietotu visi, bet gan lai pedagogi apzinātos savas iespējas un tās atbilstoši izmantotu.

„Beigu beigās tas, kas patiešām sniedz rezultātu, ir nevis metode, bet gan pasniedzēja spēja pielāgoties un izvēlēties labāko pieeju konkrētajai apguvēju grupai, konkrētajā laikā, konkrētajā kontekstā.” (Meunier 2007, 40)

2. Otrās baltu valodas apguvēju korpuss

Valodas apguvēju korpusa noderīgumu jau 2012. gadā ieskicējušas A. Zujevaite un E. Žilinskaite-Šinkūniene, veidodamas latviešu valodas apguvēju korpusu (par to vairāk sk. iepriekš – 1.2.2. apakšnodaļā). Kā galveno šāda korpusa priekšrocību viņas norāda iespēju sistematizēt kļūdas un noteikt, kuras no tām ir radušās dzimtās valodas ietekmē, bet kuras – citu iemeslu dēļ (Zujevaitė, Žilinskaitė 2012, 55).

Runājot par valodas apguvēju korpusu, kurā vienas un tās pašas valodas mainās vietām kā dzimtā valoda un mērķvaloda, šajā darbā tiek lietots termins *divvirzienu valodas apguvēju korpuss*, lietuviešu valodā – *dvikryptis besimokančiųjų tekstynas,* angļu valodā – *bidirectional learner corpus.*

Jāņem vērā arī tas, kādam valodu pārim korpuss tiek veidots. Kā jau minēts, ierasti vismaz viena no valodām pārī ir angļu, vācu, spāņu vai cita ļoti izplatīta valoda. Tam iemesls ir ne vien fakts, ka par izplatītākajām valodām interesējas vairāk pētnieku, bet arī tas, ka šādos valodu pāros ir vairāk arī valodas apguvēju. Līdz ar to praktiski ir vieglāk savākt lielu apjomu datu, lai izveidotu korpusu un varētu uzticami strādāt ne vien ar kvalitatīvajām, bet arī ar kvantitatīvajām metodēm. Lietuviešu un latviešu valodu pārī apguvēju ir ievērojami mazāk, nekā, piem., latviešu vai lietuviešu, kas apgūst angļu vai krievu valodu (par populārākajām svešvalodām Latvijā sk. VL 2012, Lietuvā – MGUKM 2013), līdz ar to arī nevar gaidīt, ka korpusa apjoms sasniegtu pasaulē lielāko valodas apguvēju korpusu apjomu.

Tā kā šādi praktiski apsvērumi līdz ar pētnieka paša interesi nereti nosaka attiecīgajā korpusā iekļaujamos valodu pārus, parasti netiek uzsvērta abu pāra valodu radniecība. Korpuss „Esam" balstās citādā pieejā. Tā kā abas valodas ir tuvi radniecīgas, praksē novērots, ka liela nozīme ir ne vien interferencei, bet arī pozitīvajai pārnesei[24]. Valodas apguvēju korpuss ļauj noskaidrot, kādas ir baltu starpvalodas īpatnības.

Kā norādījusi Sjūzena Hanstone (*Susan Hunston*), „visi korpusi ir kompromiss starp vēlamo, proti, to, ko korpusa veidotājs ir iecerējis, un iespējamo. Korpusa izveidei ir daudz praktisku ierobežojumu, no kuriem svarīgākie ir: programmatūras noteiktie ierobežojumi, autortiesības un ētikas jautājumi, kā arī tekstu pieejamība." (Hunston 2008, 156–157) Otrās baltu valodas apguvēju korpuss „Esam" šajā ziņā nav izņēmums, tāpēc šajā promocijas darba daļā daudz runāts ne vien par to, kā būtu *vēlams* rīkoties, bet arī (un it īpaši) par to, kādi šķēršļi

[24] Dzimtās valodas ietekme uz apgūstamo mērķvalodu var būt negatīva vai pozitīva. Ja tā ir negatīva, to dēvē par interferenci (LTSV 2011, 40); savukārt, ja ietekme ir pozitīva, to sauc par pozitīvo pārnesi (LTSV 2011, 62–63)

parādījušies un kā ieceres saskaņotas ar iespējām gadījumos, kad korpusa izveidi ietekmē ierobežojoši faktori.

2.1. Avotu atlase

Kā norāda Rods Eliss (*Rod Ellis*) un Gerijs Barkhuizens (*Gary Barkhuizen*), datus korpusam var iegūt ar vienu no trim galvenajām metodēm: tiešā vērojuma, kontrolēta vērojuma vai eksperimentālo metodi. Izvēloties tiešo vērojumu, jāvāc valodas paraugi, kurus valodas apguvēji producē paši pēc savas iniciatīvas: mērķvalodā tapušas vēstules, sarunas utt. Šādā gadījumā ļoti zems ir kontroles līmenis, kādu pētnieks iegūst pār datiem – valodas parauga autors pats nosaka gan tēmu, gan stilu, gan izpausmes veidu (mutvārdu vai rakstveida, rokraksts vai datorraksts utt.), gan arī to, kad un vai vispār valodas paraugu producēt. Nedaudz vairāk kontroles piedāvā kontrolēts vērojums – pētnieks panāk, ka apguvējs rada tekstu vai mutvārdu valodas paraugu (tātad – dod uzdevumu), taču stingri nenosaka, kādam šim paraugam jābūt, ļaujot apguvējam koncentrēties uz saziņas situāciju un nododamo informāciju. Savukārt eksperimentālā metode paredz ciešu kontroli pār iespējami daudziem faktoriem, nosakot, kam valodas paraugā jābūt. Šādā gadījumā galvenokārt tiek pētīta forma, kādā noteikta, iepriekš konkretizēta informācija tiek pausta (Ellis, Barkhuizen 2005, 23–24). Dažādi pētnieki ir vienisprātis: jo dabiskāks ir valodas paraugs, jo vērtīgāks tas ir pētījumiem, tāpēc ka tas vairāk atspoguļo patieso valodas prasmi; tomēr vienlaikus tiek atzīts arī, ka grūtības dabisku paraugu ieguvē, kā arī vēlme kontrolēt dažādos faktorus, kas var ietekmēt valodas producēšanu, līdz šim ir mudinājušas pētniekus vairāk pievērsties kontrolētam vērojumam un eksperimentālajai metodei korpusu datu ieguvē (Granger 2008b, 337; Ellis, Barkhuizen 2005, 24).

Korpusam „Esam" teksti ir iegūti ar kontrolēta vērojuma palīdzību. Valodas apguvējiem tika dots uzdevums uzrakstīt tekstu, kā arī aptuvenais ieteicamais teksta apjoms, taču teksta saturu noteikuši autori paši. Turpmākajās apakšnodaļās sīkāk raksturota tekstu ieguves procedūra, kā arī ar to saistīto tiesisko jautājumu risinājums.

2.1.1. Tekstu ieguve un atlases kritēriji

Teksus korpusam iesnieguši pasniedzēji, kas Lietuvas un Latvijas augstskolās māca otro baltu valodu. Daļa tekstu iesniegti fiziski (papīra formā), daļa – elektroniski. Lai tekstus saņemtu un vienotos ar tekstu autoriem par to izmantošanu, izmantoti dažādi elektroniskās saziņas veidi, t. sk. e-pasts un iekšējās saziņas sistēma sociālajos tīklos *Draugiem.lv* un

Facebook.com. Ar daļu no tekstu autoriem korpusa veidotāja tikusies un tekstu izmantošanas nosacījumus pārrunājusi klātienē bez elektronisku saziņas rīku iesaistes.

Tālāk šajā apakšnodaļā raksturoti korpusa tekstu atlases kritēriji.

2.1.1.1. Mērķvaloda, avotvaloda, valodas prasmes līmenis

Korpuss „Esam" sastāv no lietuviešu un latviešu valodas apguvēju rakstītiem tekstiem apgūstamajā valodā jeb mērķvalodā. Korpusā iekļautie teksti ir Lietuvas un Latvijas augstskolu studentu rakstītie sacerējumi, kas tikuši uzdoti kā patstāvīgi pildāmi mājasdarbi otrās baltu valodas kursā iesācējiem. Tātad tekstu autori ir:

- Lietuvas augstskolās studējošie lietuvieši, kuri iesācēja līmenī mācās latviešu valodu;
- Latvijas augstskolās studējošie latvieši, kuri iesācēja līmenī mācās lietuviešu valodu.

Visi korpusā iekļautie teksti ir tapuši, to autoriem otro baltu valodu mācoties pirmo vai otro semestri bez priekšzināšanām, kuras gan var būt neapzinātas (LTSV 2011, 67). Daļa no autoriem ir dzirdējuši mērķvalodu arī iepriekš, taču ne pietiekami, lai šo pieredzi jau varētu uzskatīt par apzinātām valodas priekšzināšanām. Nelielu neapzināto priekšzināšanu esamība atsevišķos gadījumos gan ir iespējama, tāpēc nevar droši apgalvot, ka iegūtie dati ir pilnīgi vienadabīgi.

Valodas apguves ātrums un sekmes individuāliem apguvējiem atšķiras, tāpēc nevar uzskatīt, ka visi šajā korpusā iekļautie teksti noteikti atbilst vienam un tam pašam valodas prasmju līmenim saskaņā ar Eiropas Savienības vadlīnijām (EKP 2006). Par kritēriju tekstu atlasei ir izvēlēts valodas apguves semestris (pirmais vai otrais), nevis valodas prasmju līmenis. Jo ilgāks laiks ir pagājis kopš teksta tapšanas, jo mazāka ir iespēja noskaidrot, kādam līmenim tajā brīdī teksta autors atbildis. Šobrīd šo līmeni var mēģināt noteikt vairs tikai pēc pašiem tekstiem, tātad tas vairs nav zināms ārējs faktors, kuru varētu marķēt kā papildinformāciju.

Katram tekstam ir norādīts, kurš attiecīgās valodas apguves semestris tas ir šī teksta autoram. Vadoties pēc attiecīgo augstskolu noteiktā otrās baltu valodas kursu satura un apjoma, pirmajā semestrī būtu apgūstamas A1 līmeņa prasmes, otrajā semestrī tām parasti būtu jāpaaugstinās līdz A2 līmenim, trešajā un ceturtajā – līdz B1, varbūt līdz B2 līmenim. Tā kā korpusā līdz šim ir iekļauti tikai pirmajā un otrajā valodas apguves semestrī tapuši teksti, var pieņemt, ka kopumā visos tajos valodas prasmju līmenis ir (plašā izpratnē) A līmenī. Tomēr vienlaikus jāatzīst arī, ka semestru skaits nav uzticams valodas prasmes līmeņa indikators. Ļoti

nozīmīgs ir stundu skaits semestrī – ja tas ir liels, tad viena semestra laikā var sasniegt arī A2 līmeni. Izmantojot otrās baltu valodas apguvēju korpusu, jāņem vērā, ka attiecīgajās augstskolās situācija gadu gaitā ir mainījusies, un ne vienmēr ir iespējams skaidri noteikt, cik stundu kurš autors ir mācījies otro baltu valodu, turklāt dažādi teksti tiek rakstīti dažādos laikos – semestra sākumā, vidū vai beigās atkarībā no konkrētā pasniedzēja noteiktajām prasībām.

Balstoties klasifikācijā, saskaņā ar kuru apgūstamās valodas tiek iedalītas pirmajā jeb dzimtajā valodā, otrajā valodā un trešajā valodā jeb svešvalodā (LTSV 2011, 28, 59, 65, 88; par to sk. arī Laizāne 2014a), visiem korpusa tekstu autoriem otrā baltu valoda ir svešvaloda.

Otrās baltu valodas apguvēju korpusam nav izveidots salīdzināms dzimtās valodas korpuss. Daudzi pētnieki nodarbojas galvenokārt ar kontrastīviem pētījumiem, kuros šāds salīdzināms korpuss būtu nepieciešams. Tomēr šajā gadījumā centieni šādu korpusu veidot būtu samērā neauglīgi. Pašā sākotnējā valodas apguves posmā radītos tekstus īsti nevar salīdzināt ar dzimtās valodas runātāju tekstiem, jo tēmu un konstrukciju izvēli lielā mērā nosaka līdz šim mācīto tēmu saturs, ko tieši ietekmē skolotājs un izmantotais mācību līdzeklis, nevis intralingvālas parādības kā noteiktu konstrukciju sarežģītība u. c. Piem., pirmo īso domrakstu „Mana ģimene" nevar salīdzināt ar dzimtās valodas runātāja līdzīga temata darbu, jo dzimtās valodas runātājs var izvēlēties sniegt pilnīgi citu informāciju, ko pieļauj viņa zināšanas, kamēr apguvējam šis teksts ir viņa pirmo, stingri ierobežoto zināšanu treniņa laukums. Savukārt, ja dzimtās valodas runātājam tiks dots uzdevums turēties stingri tajos rāmjos, kādus valodas apguves process ir uzlicis valodas apguvējam, zudīs dabiskums dzimtās valodas runātāja tekstos. Līdz ar to agrīnā valodas apguves posmā dzimtās valodas runātāju un valodas apguvēju producēto tekstu kontrastīva kvantitatīva analīze nebūtu jāuzskata par primāru, bet kontrastīvas kvalitatīvas analīzes gadījumā jāņem vērā ļoti daudz blakusfaktoru, ieskaitot mācību materiālu saturu un to, cik tālu apguvējs to ir izskatījis līdz brīdim, kad ir rakstījis analizējamo tekstu.

2.1.1.2. Pieejamība

Korpusa saturu būtiski ietekmē arī tas, kādi teksti ir bijuši pieejami korpusa izveides laikā. Šobrīd korpusā ir iekļauti dati no Latvijas Universitātes (Rīga, Latvija), Liepājas Universitātes (Liepāja, Latvija), Vītauta Dižā universitātes (Kauņa, Lietuva) un Viļņas Universitātes (Viļņa, Lietuva). Pēc vienošanās ar pasniedzējiem, tekstu vākšana turpinās, un iesaistīto universitāšu loks var arī paplašināties.

Otrās baltu valodas mācīšana skolās Lietuvā un Latvijā nav populāra, jo bieži par lietderīgākām tiek uzskatītas citu, izplatītāku valodu zināšanas (Puškoriutė-Ridulienė 2011). Tāpēc korpusā „Esam" iekļauti dažādos gados tapuši teksti, lai paplašinātu autoru loku un palielinātu korpusa apjomu. Šobrīd korpusā ir teksti, kas ir tapuši laika posmā no 2007. gada līdz 2014. gadam Latvijas Universitātē (LU), no 2013. gada līdz 2014. gadam Liepājas Universitātē (LiepU), no 2011. gada līdz 2012. gadam Vītauta Dižā universitātē (VDU) un no 2007. gada līdz 2012. gadam Viļņas Universitātē (VU).

Par otrās baltu valodas apguvi augstākās izglītības iestādēs nav pieejama detalizēta informācija. Par latviešu valodas apguvi ārpus Latvijas informāciju vāc Latviešu valodas aģentūra, taču tās izveidotajā latviešu valodas apguves kartē ir iekļautas tikai tās iestādes, ar kurām aģentūrai ir nodibināts pastāvīgs kontakts un sadarbība. Tas neizslēdz iespēju, ka latviešu valoda Lietuvā tiek mācīta arī citur. Līdz ar to šeit norādīta informācija, kas ir iegūta, tieši sazinoties ar augstākās izglītības iestādēm Latvijā un Lietuvā.

Lietuviešu valoda Latvijā iesācējiem no 2007. gada līdz 2014. gadam[25] tika mācīta šādās augstākās izglītības iestādēs:

- Latvijas Universitātē – pasniedzēji Edmunds Trumpa (*Edmundas Trumpa*), Jolanta Nagle;
- Liepājas Universitātē – pasniedzēja Ieva Ozola;
- Daugavpils Universitātē – pasniedzēja Vilma Šaudiņa;
- Rēzeknes Tehnoloģiju akadēmijā (iepriekšējais nosaukums – Rēzeknes Augstskola) – pasniedzēja Antra Kļavinska.

Latviešu valoda iesācējiem Lietuvā šajā pašā laika posmā ir tikusi mācīta šādās augstākās izglītības iestādēs[26]:

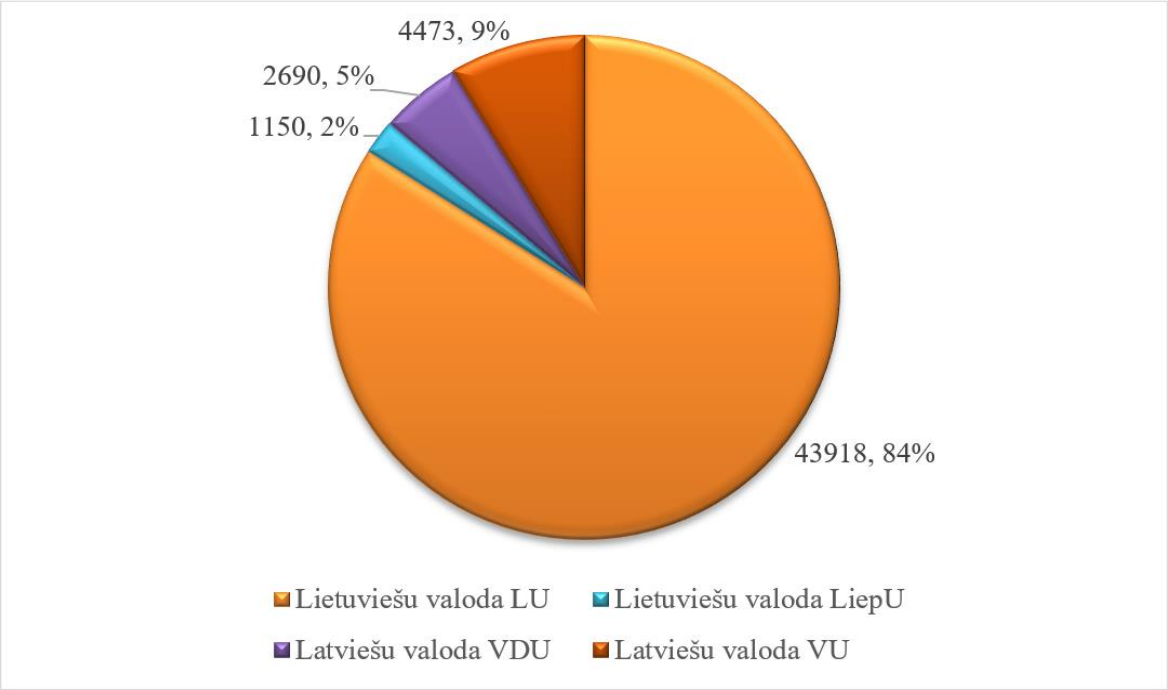
- Vītauta Dižā Universitātē – pasniedzēji Alvīds Butkus (*Alvydas Butkus*), Kristina Vaisvalavičiene (*Kristina Vaisvalavičienė*), Daiva Puškorjute-Riduliene (*Daiva Puškoriutė-Ridulienė*), Violeta Butkiene (*Violeta Butkienė*);
- Klaipēdas Universitātē – pasniedzēja Daļa Kiseļūnaite (*Dalia Kiseliūnaitė*);
- Šauļu Universitātē – pasniedzēja Regina Kvašīte (*Regina Kvašytė*);
- Viļņas Universitātē – pasniedzēji Egle Žilinskaite, Vītauts Rinkēvičs (*Vytautas Rinkevičius*), Agne Navickaite-Klišauskiene (*Agnė Navickaitė-Klišauskienė*), Ēvalds Švageris (*Evaldas Švageris*), Ērika Sausverde;

[25] Šāds laika posms izvēlēts, jo no šiem gadiem ir iegūti teksti korpusam vismaz no vienas augstākās izglītības iestādes.
[26] Datus savā topošajā promocijas darbā apkopojusi un analizējusi Inga Laizāne.

- Viļņas Pedagoģiskajā universitātē – pasniedzēja Žaneta Markevičiene (*Žaneta Markevičienė*).

Korpusā šobrīd ir 257 teksti no 83 autoriem. Kopējais vārdu skaits šobrīd ir apmēram 45 000 korpusa lietuviskajā daļā un apmēram 7 000 korpusa latviskajā daļā. Sīkāka informācija par vārdlietojumu skaitu un sadalījumu redzama 2. attēlā.

2. attēls. Pašreizējais korpusa apjoms pēc vārdlietojumu skaita



Kā redzams, šobrīd korpuss gan saturiski, gan valodu ziņā nav balansēts – lietuviešu valodā tekstu apjoms ir daudz lielāks nekā latviešu valodā. Latvijas Universitātes lietuviešu valodas pasniedzējs E. Trumpa vairākus gadus ir konsekventi vācis studentu rakstu darbus, bet citi pasniedzēji iniciatīvai pievienojušies vēlāk. Turklāt Latvijas Universitātē otro baltu valodu studējošo skaits ir daudz lielāks nekā, piem., Liepājas Universitātē. Jāņem vērā arī tas, ka Latvijas Universitātē iegūtie teksti, kā redzams 5. attēlā (sk. 73. lpp.), ir ievērojami garāki nekā Liepājas Universitātē un Viļņas Universitātē iegūtie (Vītauta Dižā universitātes studentu teksti gan ir vēl garāki, taču to ir daudz mazāk).

Nākotnē iecerēts panākt, lai korpuss valodas ziņā būtu pēc iespējas sabalansēts, tā pilnveidojot divvirzienu valodas apguvēju korpusu. Joprojām tiek apzināti jau savākto tekstu autori, lai iegūtu atļaujas tekstu ievietošanai korpusā, un, kā jau minēts, tiek vākti arī jauni

teksti. Primāri gan ir centieni padarīt pieejamus pēc iespējas vairāk atbilstošu materiālu, nevis saskaņot apjomu lietuviešu un latviešu valodā rakstītajiem tekstiem.

Nākotnē korpusā iespējams arī iekļaut tekstus, kurus rakstījuši apmaiņas studenti no Lietuvas Latvijā, mācīdamies latviešu valodu, un no Latvijas Lietuvā, mācīdamies lietuviešu valodu. Atsevišķos gadījumos var rasties neskaidrības, vai šādā gadījumā latviešu valoda ir uzskatāma par svešvalodu (kā tas ir ar pārējiem korpusā iekļautajiem tekstiem) vai otro valodu, taču, kā skaidro I. Laizāne, valodas vide nav vienīgais aspekts, kas nosaka šo šķīrumu, liela nozīme ir arī uzturēšanās laikam attiecīgajā valstī, zināšanām par attiecīgo valsti utt. (Laizāne 2014a, 143–144; sk. arī Laizāne 2015). Līdz ar to arī šādu studentu rakstītos tekstus varētu pielīdzināt un skatīt kopā ar pašreiz korpusā iekļautajiem.

2.1.1.3. Tekstu formāts

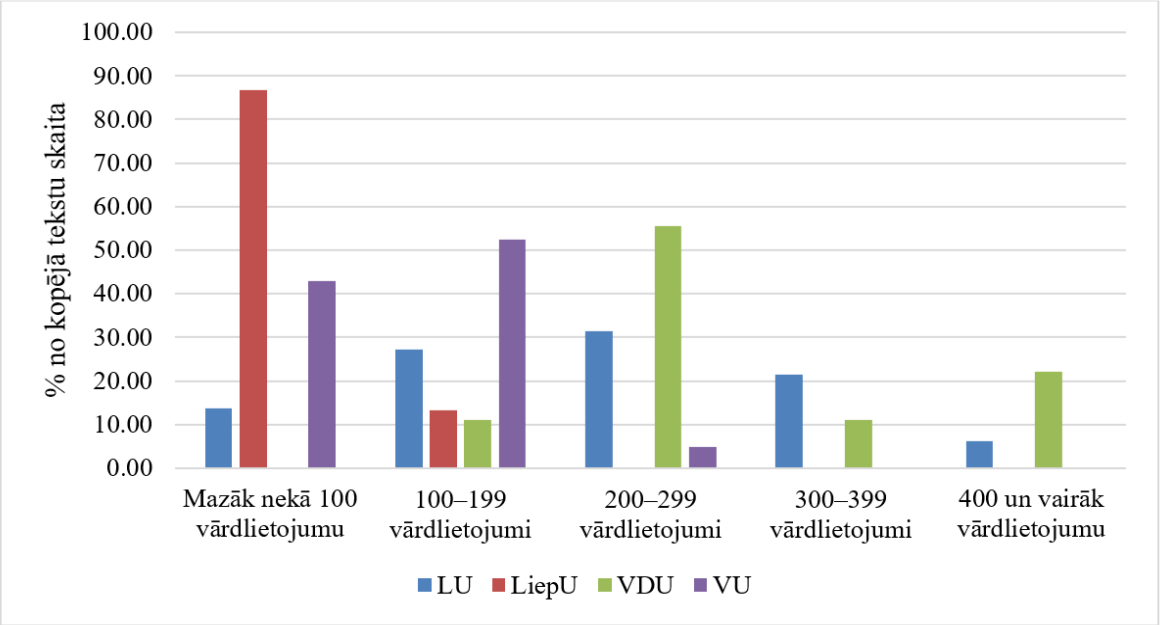
Paši pasniedzēji lemj par to, vai tekstiem būtu jābūt iesniegtiem rokrakstā vai datorrakstā, uz papīra vai elektroniskā formātā. Gadījumos, kad pasniedzējs šajā ziņā norādījumus nesniedz, studenti izvēlas sev ērtāko veidu. Līdz ar to iegūtais materiāls šajā ziņā nav viendabīgs. Ja teksts ir iesniegts rokrakstā vai izdrukātā veidā, to, manuāli pārrakstot, digitalizē promocijas darba autore[27], saglabājot visas rakstības īpatnības, izņemot dubultas atstarpes starp vārdiem. Arī gadījumos, kad teksts iesniegts rokrakstā, tas tiek digitalizēts, visu tekstu manuāli pārrakstot datorrakstā (vairāk sk. 2.2.1. apakšnodaļā „Digitalizēšana"). Tekstos, kas ir iegūti līdz šim, nav bijis gadījumu, kad rokraksts būtu tik grūti salasāms, lai digitalizēšanas laikā rastos neskaidrības.

2.1.1.4. Apjoms

Atšķirīgu prasību (tās nosaka paši pasniedzēji) dēļ atšķiras dažādu universitāšu studentu rakstīto tekstu garums. Garākie teksti sasniedz 500 vārdlietojumu apjomu, savukārt īsāko garums ir, sākot ar 40 vārdlietojumiem.

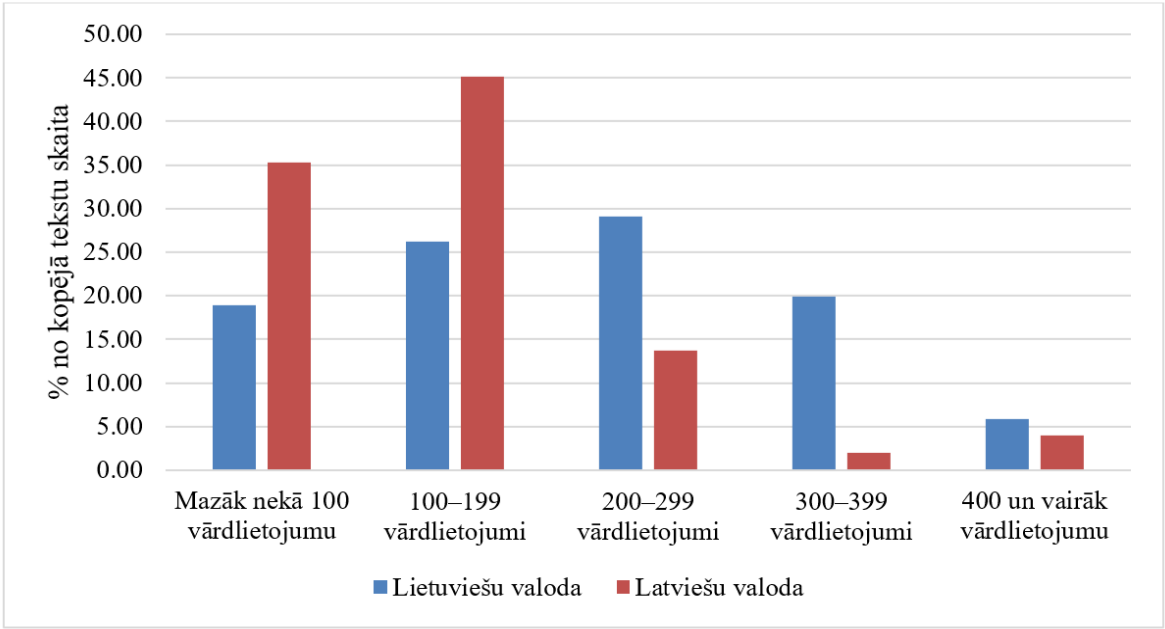
[27] Daļu no tekstiem pēc tiem pašiem principiem digitalizēja Evija Zubova.

3. attēls. Tekstu garuma sadalījums pa augstskolām

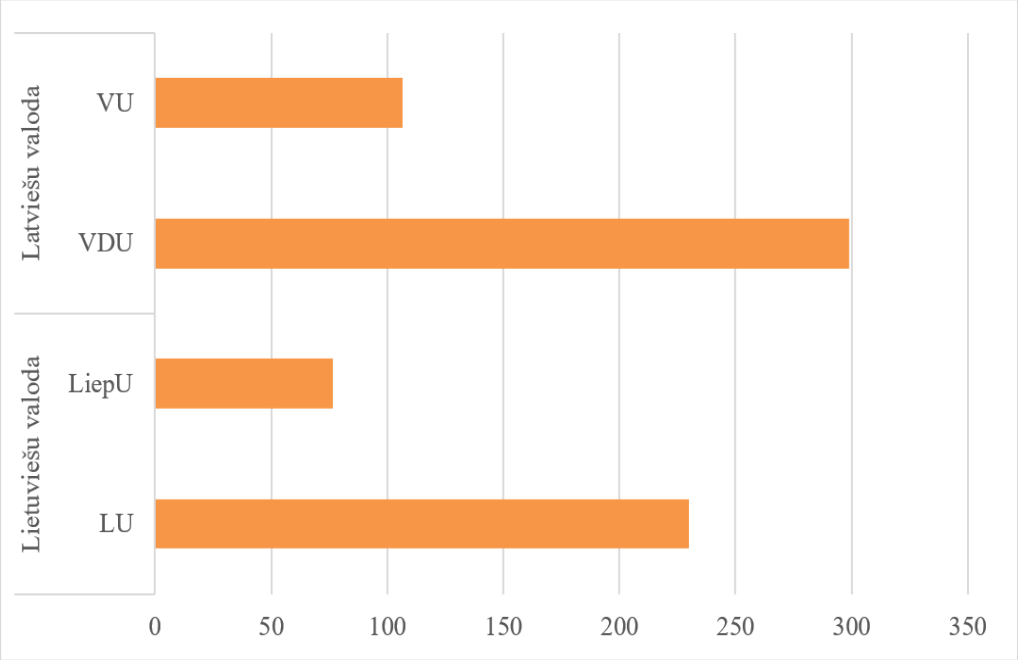


3. attēlā redzams salīdzinājums pēc universitātēm par tiem sacerējumiem, kuru iekļaušanai korpusā ir iegūta atļauja no autoriem, savukārt 4. attēlā šī pati informācija redzama salīdzinājumā pēc valodām. 5. attēlā (sk. 73. lpp.) redzams vidējais aritmētiskais tekstu garums pēc universitātes un pēc valodas.

4. attēls. Tekstu garuma sadalījums pa valodām



5. attēls. Vidējais aritmētiskais tekstu garums dažādās universitātēs tapušos tekstos



Attēlotajos datos nav iekļauta informācija par diviem LU studentes sacerējumiem, kuri ir tapuši padziļinātu lituānistikas studiju gaitā un ir ievērojami garāki par pārējiem (attiecīgi 1819 un 2394 vārdlietojumi). Šie divi sacerējumi korpusā nav iekļauti.

Redzams, ka Liepājas Universitātes studentu rakstītajiem tekstiem ir tendence būt daudz īsākiem nekā, piem., Vītauta Dižā universitātes studentu rakstītajiem. Apkopojot informāciju par tekstiem pēc to valodas, atšķirības mazliet izlīdzinās, taču dati nav pilnīgi viendabīgi arī šādā griezumā. Tātad šādā aspektā korpuss nebūtu uzskatāms par sabalansētu. Tomēr vienlaikus var apšaubīt, vai tekstu garums ir īpaši nozīmīgs faktors korpusā iekļaujamo datu atlasē. Visgarākais korpusā iekļautais sacerējums ir 500 vārdu garš, tātad nesasniedz pat vienu procentu no kopējā korpusa apjoma.

2.1.1.5. Autentiskums

Korpusā iekļautie teksti ir rakstīti patstāvīgi, brīvi izmantojot studentiem pieejamos palīglīdzekļus, bez laika ierobežojumiem (izņemot termiņu sacerējuma iesniegšanai). Tekstu rakstīšanas uzdevums sākotnēji dots tikai pedagoģiskos nolūkos, nevis korpusa tekstu vākšanai, un arī pēc tam, kad tika pieņemts lēmums veidot korpusu, uzdevuma nosacījumi nav mainījušies. Studentiem uzdots mājasdarbs līdz noteiktam datumam uzrakstīt tekstu par noteiktu tematu. Daļā gadījumu studenti paši šīs tēmas varēja modificēt (piem., sašaurināt un rakstīt nevis par savu ģimeni un draugiem, bet gan tikai par ģimeni) vai pilnībā mainīt (ja bija kāda cita tēma, par kuru attiecīgais students vēlējās rakstīt un jutās spējīgs to darīt). Šādā

gadījumā studentiem par to iepriekš bija individuāli jāvienojas ar pasniedzēju. Dažādu pasniedzēju uzdotās tēmas daļēji atšķīrās.

Daļā sacerējumu pasniedzēji, kas tos apkopoja, ir konstatējuši fragmentus, kas kvalificējami kā plaģiāts (kopēts teksts no interneta enciklopēdijām u. c.). Šīs daļas korpusā netika iekļautas. Ja tekstā ir gan plaģiāts, gan arī oriģināls teksts un šīs daļas ir skaidri nošķiramas (piem., vairākos darbos atrasta viena rindkopa no interneta resursa, taču pārējais teksts katrā darbā ir rakstīts individuāli), tad tā oriģinālā daļa ir iekļauta korpusa materiālā, bet plaģiāts ir dzēsts.

Daļa tekstu ir saglabājušies tikai izlabotā formā, t. i., tikai pasniedzēja rediģēti, bez iespējas atjaunot sākotnējo (nerediģēto) tekstu. Šādi teksti korpusā nav iekļauti. Arī tekstu nosaukumi korpusā nav iekļauti. Lai arī daļa autoru tos ir mēģinājuši formulēt patstāvīgi, lielākoties vērojama spēcīga pasniedzēju sniegtā formulējuma ietekme, tāpēc nolemts, ka tie nav uzskatāmi par valodas apguvēju patstāvīgi izveidotiem un var maldināt korpusa lietotājus. Tomēr nosaukumi ir saglabāti pieejami kā papildinformācija, lai vajadzības gadījumā varētu pārliecināties par teksta tematu un pēc šī kritērija tekstus atlasīt.

2.1.2. Korpusa „Esam" vieta valodas apguvēju korpusu klasifikācijā

Atbilstoši 1.1.3. nodaļā nosauktajiem valodas apguvēju korpusu klasifikācijas parametriem, korpusa „Esam" raksturojums ir šāds:

1. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamajiem tekstiem un to īpašībām.

- Korpusa **mērķvaloda** ir otrā baltu valoda. Korpusā ir divi apakškorpusi, no kuriem vienā mērķvaloda ir latviešu valoda un tekstu autoru dzimtā valoda ir lietuviešu valoda, bet otrā mērķvaloda ir lietuviešu valoda un tekstu autoru dzimtā valoda ir latviešu valoda.
- **Valodas producēšanas veids** – rakstveida. Pagaidām nav iecerēts šajā korpusā iekļaut arī mutvārdu tekstus, jo tas prasītu cita veida apstrādi, kā arī padarītu datu ieguvi sarežģītāku. Pētnieki gan norāda, ka valodas apguvēju mutvārdu valodas paraugu ieguvē ir zināmas priekšrocības, jo mutvārdu pārbaudījumi nereti tiek ierakstīti, tomēr atzīst arī, ka šādi iegūtais materiāls atspoguļo apguvēju sniegumu pārbaudījuma apstākļos, bet ne kopumā (Hunston 2008, 159).

- **Tekstu tips** – apraksts, dažos gadījumos robežojoties ar eseju. Daži autori ir iesnieguši arī atdzejas mēģinājumus, taču to ir nedaudz, un tie būtiski atšķiras no pārējiem tekstiem, tāpēc korpusā nav iekļauti. Nākotnē nav paredzēts šādus darbus iekļaut korpusā, vismaz ne tā anotētajā daļā, jo tekstiem, kuros ir daudz mākslinieciskās izteiksmes līdzekļu, ir apgrūtināta gan kļūdu, gan vārdšķiru utt. anotēšana.
- Pēc **tekstu tematikas** šis būtu uzskatāms par netematisku korpusu. Tam nav noteikta vienas vai nedaudzu nozaru tematika, un arī katra konkrēta teksta temati būtiski atšķiras savā starpā.
- Tekstu **oriģinalitāte** – korpusā ir oriģināli teksti, nevis tulkojumi. Ir iespējams, ka daļa no tekstu autoriem tekstus sākotnēji uzrakstījuši dzimtajā valodā un pēc tam iztulkojuši, taču arī šādā gadījumā autors ir tas pats un teksts jau sākotnēji bijis paredzēts tulkošanai mērķvalodā.

2. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamo tekstu tapšanas apstākļiem, autoriem un ieguvi.

- **Teksta autora dzimtā valoda** – viena no baltu valodām. Kā jau norādīts, lietuviešu valodas apguvējiem tā ir latviešu valoda, bet latviešu valodas apguvējiem – lietuviešu valoda. Šis apgalvojums gan nav simprocentīgi precīzs visos gadījumos – dažiem autoriem attiecīgā baltu valoda ir viena no vairākām dzimtajām valodām. Nākotnē varētu tikt izveidoti arī salīdzināmi korpusi, kuros būtu iekļauti arī citas dzimtās valodas pārstāvju teksti. Šādu korpusu izveide palīdzētu salīdzināt baltu starpvalodu ar to, kāda starpvaloda veidojas, ja apguvēja dzimtā valoda nepieder baltu valodu grupai. Visiem līdz šim iekļauto tekstu autoriem dzimtā valoda ir tā pati valoda, kas ir tikusi izmantota kā starpniekvaloda otrās baltu valodas apguves procesā.
- **Valodas apguvēju vecums** – korpusā iekļauto tekstu autori visi ir pieaugušie, tie ir augstskolu studenti, tātad viņiem visiem ir vismaz vidējā izglītība. Lielākā daļa no autoriem ir filoloģijas studenti, jo otrās baltu valodas apguve parasti ietilpst atbilstošās bakalaura studiju programmās, taču ir arī citu nozaru pārstāvji. Vecums lielākajai daļai ir līdzīgs – ap divdesmit gadiem, taču ir arī nedaudzi studenti, kas ir vecāki par kursabiedriem.

- **Valodas prasmju līmenis –** visi autori ir iesācēji bez priekšzināšanām un bez ikdienas saskarsmes ar otro baltu valodu ārpus studijām. Visi autori otro baltu valodu mācās pirmo vai otro semestri, sasniedzot A1 vai A2 līmeni.
- Pēc **valodu apguves secības** visiem autoriem otrā baltu valoda ir svešvaloda. Neliela daļa autoru ir darījusi zināmas savas citu svešvalodu prasmes, taču lielākā daļa autoru nespēja vai nevēlējās precizēt, kādas vēl svešvalodas viņi prata tekstu tapšanas laikā, tāpēc tika nolemts šo informāciju nepievienot. Lai gan būtu interesanti mēģināt noteikt arī šo valodu ietekmi, dažādu valodu iespaids var būt tik sarežģīts, ka, veidojot atsevišķus apakškorpusus (piem., apakškorpuss, kurā ir studenti, kuriem ir franču valodas zināšanas), šie apakškorpusi varētu izrādīties pārāk mazi, lai, tajos balstoties, varētu izdarīt drošticamus secinājumus. Tomēr ir skaidrs, ka nevienam no autoriem šī nav pirmā apgūstamā svešvaloda, jo gan Latvijā, gan Lietuvā ir noteikts, ka, lai iegūtu vidējo izglītību, ir jāmācās vismaz viena svešvaloda (MK 281; VUPA 2011), un nevienam no studentiem tā nav bijusi otrā baltu valoda – to apliecina fakts, ka otrās baltu valodas apguve bez priekšzināšanām ir uzsākta studiju laikā, t. i., pēc vidējās izglītības iegūšanas.
- Pēc **tekstu tapšanas secīguma** korpuss ir nosacīti sinhronisks. Tajā iekļautie teksti ir tapuši 8 gadu laikā no 2007. gada līdz 2014. gadam. Daļai no tekstiem nav zināms precīzs tapšanas gads, taču ir zināms, ka tas ir bijis minētajā laika posmā. Līdz ar to tekstiem obligāti pievienojamā informācija nesatur rakstīšanas gadu. Tā kā mācību metodes un apstākļi minētajā laika posmā nav būtiski mainījušies, var uzskatīt, ka konkrētam teksta tapšanas gadam nav nozīmes.
- Pēc **tekstu rakstīšanas veida** korpuss „Esam" ir neviendabīgs. Tajā ir iekļauti gan teksti, kas tapuši datorrakstā, gan arī tādi, kas ir rakstīti ar roku. Ar roku rakstītie teksti korpusa vajadzībām ir digitalizēti, un digitalizācijas process ir aprakstīts 2.2.1. apakšnodaļā „Digitalizēšana".

3. Klasifikācija pēc parametriem, kas ir saistīti ar korpusā iekļaujamo tekstu apstrādi un tehnisko noformējumu, korpusu sagatavojot pētniecības darbam.

- **Tekstu nedalāmība** – korpusā ir iekļauti pilni teksti. Daļā tekstu gan anonimizācijas gaitā ir veiktas izmaiņas vai izlaista daļa informācijas (par to

vairāk sk. 2.1.3. apakšnodaļā „Personas datu aizsardzība un autortiesības"), taču tas ir atsevišķi norādīts, un teksti nav dalīti.

- Pēc **valodu skaita** korpuss „Esam" ir divvalodīgs, jo tajā ir iekļauti teksti divās mērķvalodās: latviešu un lietuviešu valodā. Katrs no iepriekšminētajiem apakškorpusiem ir vienvalodīgs, t. i., tajā ir iekļauti teksti tikai vienā valodā.
- Korpusam nav noteiktas **specializācijas**, tas ir uzskatāms par vispārīga rakstura valodas apguvēju korpusu, un tā uzbūvē mēģināts ievērot līdzsvara un reprezentativitātes principus.
- **Dzimto valodu skaits** šajā korpusā ir tāds pats kā mērķvalodu skaits – tās ir divas: latviešu un lietuviešu (ar iespējamām papildu dzimtajām valodām). Katrā no iepriekšminētajiem apakškorpusiem ir pārstāvēta viena dzimtā valoda.
- **Valodas prasmes līmenis** korpusā „Esam" ir viens – tas atbilst A līmenim saskaņā ar Eiropas kopīgajām pamatnostādnēm (EKP 2006). Dalījums sīkāk A1 un A2 apakšlīmeņos nav veikts: lai arī ir zināms, kurā valodas apguves semestrī kurš no tekstiem ir tapis, autoru otrās baltu valodas prasmes ir attīstījušās dažādā ātrumā, un nevar apgalvot, ka visi pirmajā semestrī būtu bijuši A1 līmenī (daļa studentu ir sasnieguši A2) vai ka visi otrajā semestrī bijuši A2 līmenī (daļa studentu joprojām nav to sasnieguši).
- Pēc **kopējā apjoma** šis korpuss ir mazs – pagaidām tajā ir apmēram 52 000 vārdlietojumu, taču apjoms turpina palielināties. Korpusa potenciālo apjomu ierobežo salīdzinoši nelielais otrās baltu valodas apguvēju skaits.
- **Anotējums** – korpuss ir anotēts. Tajā ir anotētas pamatformas, vārdšķiras, sintaktiskie teikumu veidi un valodas lietojuma kļūdas.
- **Izmantotā programmatūra** – sākotnēji neanotēts izmēģinājuma korpuss bija piemērots darbam ar programmu *AntConc* (vairāk informācijas sk. Anthony 2014), un tas joprojām ir pieejams interesentiem. Darbam ar pilnu korpusu tika izvēlēta programma *TEITOK*, kas darbojas, balstoties uz CQP/CWB platformas.
- **Valodu pāru virzieni** šajā korpusā ir divi – tas ir divvirzienu korpuss, jo abas tajā pārstāvētās dzimtās valodas ir arī korpusa mērķvalodas.

4. Klasifikācija pēc parametriem, kas ir saistīti ar korpusa lietošanu, tās iespējām un līdzšinējo darbu ar attiecīgajiem korpusiem.

- Pēc **pedagoģiskā lietojuma** šis ir korpuss ar pastarpinātu pedagoģisko lietojumu. Tiešs tā pedagoģiskais lietojums nav paredzēts un nav arī iespējams, jo tekstu apkopošana korpusā notiek tad, kad otrās baltu valodas kurss ir jau beidzies.
- Pēc **korpusa izmantojuma jomas** šis ir akadēmisks korpuss – tas ir paredzēts pētījumu veikšanai un to ir aizliegts izmantot komerciālos nolūkos.
- Visbeidzot pēc **pieejamības** „Esam" ir publiski pieejams korpuss. Materiāls ir ievietots internetā brīvi pieejams ikvienam interesentam.

Tā kā korpuss primāri paredzēts valodniecības, nevis pedagoģiskiem mērķiem, metadati galvenokārt pievienoti tādi, kas varētu palīdzēt raksturot korpusā esošos tekstus no valodnieciskā viedokļa (piem., kāda starpniekvaloda ir izmantota mācību procesā), nevis valodas apguves metodes u. tml. Arī vairāki citi faktori, kas var ietekmēt valodas apguvi, piem., dzīvesvieta (iespējama dialekta ietekme un/vai dzīve netālu no Latvijas un Lietuvas robežas), personas datu aizsardzības nolūkā nav iekļauti korpusa papildinformācijā (par personas datu aizsardzību sk. promocijas darba 2.1.3.1. punktu).

2.1.3. Personas datu aizsardzība un autortiesības

Veidojot valodas apguvēju korpusu, jāņem vērā arī tajā iekļaujamo tekstu autoru tiesības. Padarot tekstus pieejamus pētniecībai, jāpievērš uzmanība diviem tiesību aspektiem: personas datu aizsardzībai un autortiesībām. Tā kā materiāli ir iegūti gan Latvijā, gan Lietuvā, turpmāk īsumā raksturots šo jautājumu regulējums Latvijas Republikas un Lietuvas Republikas tiesību aktos, ciktāl tas skar otrās baltu valodas apguvēju korpusa izveidi.

2.1.3.1. Personas datu aizsardzība

Kā nosaka Latvijas Republikas Fizisko personu datu aizsardzības likuma (FPDAL) 2. pants, personas dati ir „jebkāda informācija, kas attiecas uz identificētu vai identificējamu fizisko personu". Lietuvas Republikas Personas datu tiesiskās aizsardzības likuma (*Asmens duomenų teisinės apsaugos įstatymas*, ADTAI) 2. pantā tie definēti mazliet izvērstāk: „Jebkāda informācija, kas ir saistīta ar fizisku personu – datu subjektu, kura identitāte ir zināma vai var tikt tieši vai netieši noteikta". Tātad tajos tekstos, kuros studenti atklāj kādu informāciju par

sevi un/vai citām fiziskām personām, tie ir personas dati šī likuma izpratnē, ja vien attiecīgās personas ir iespējams identificēt (t. sk. arī tad, ja identifikācijai nepieciešama citā ceļā iegūstama papildinformācija: kādas augstskolas studentu saraksti, kāda uzņēmuma darbinieku saraksti utt.).

FPDAL 7. pants un ADTAI 5. pants norāda, ka fiziskās personas datu apstrāde (t. i., „jebkuras ar personas datiem veiktas darbības" – FPDAL 2. pants) ir atļauta tiem mērķiem, kuriem attiecīgā persona ir piekritusi, vai arī gadījumos, kurus tieši paredz tiesību akti. Uzreiz jāpiebilst, ka tiesību aktos nav paredzēts, ka augstskolas studiju darbos norādīto personas datu apstrāde būtu atļauta. Rakstīdami studiju darbus, studenti nodod šos datus augstskolai ar mērķi nokārtot attiecīgo kursu prasības un iegūt nepieciešamos kredītpunktus, taču par tālāku šo darbu izmantošanu ir jāvienojas atsevišķi. Tas nozīmē – lai šādus datus iekļautu korpusā (proti – apstrādātu), ir nepieciešams slēgt līgumu ar katru no datu subjektiem.

Konkrētajā gadījumā datu subjektu skaits ir ļoti liels, jo daudzi teksti ir par tādām tēmām kā „mana ģimene", „mani draugi" u. c. Šādā gadījumā datu subjekts ir ne tikai students, kas uzrakstījis tekstu, bet arī visas personas, par kurām viņš tajā stāsta (draugi, radinieki, ģimenes locekļi, kolēģi utt.). Līdz ar to personas datu apstrādes līguma slēgšana ar katru no datu subjektiem ir neadekvāti darbietilpīga attiecībā pret iegūto materiālu. Tādēļ ir nolemts šajā korpusā iekļaujamo tekstu autorus marķēt ar unikāliem kodiem[28], bet tekstos reālos personas datus aizstāt ar izdomātiem, pēc iespējas cenšoties saglabāt katra teksta īpatnības. Īpaši gadījumos, kuros marķējamās novirzes parādās aizstājamā teksta daļā, ir mēģināts saglabāt īpatnības, kas būtu varējušas ietekmēt šo noviržu rašanos. Ir saglabāta visu personvārdu un vietvārdu deklinācija un dzimte, kā arī vienskaitļa vai daudzskaitļa attiecīgā locījuma forma (piem., ja ir kļūdaini veidots vienskaitļa ģenitīvs no 5. deklinācijas sieviešu dzimtes personvārda *Ilze*, tad šis personvārds aizstāts tāpat ar 5. deklinācijas sieviešu dzimtes personvārdu *Dace* atbilstoši kļūdainā vienskaitļa ģenitīva formā[29]). Lai samazinātu identificēšanas iespēju, vienam un tam pašam personvārdam vai vietvārdam dažādos tekstos var būt dažādi substitūti. Nav aizstāta informācija par teksta autoram personīgi nepazīstamām personām (piem., students raksta par kādu dziedātāju vai komponistu, kuru apbrīno).

Korpusā iekļautajos tekstos aizstāti šādi sensitīvi dati:

- teksta autora un teksta autoram pazīstamu personu vārdi un uzvārdi;
- pilsētu un apdzīvotu vietu nosaukumi, mājvārdi, adreses;

[28] Katram viena un tā paša autora tekstam piešķirts autora kods, kas ļauj atlasīt viena autora vairākus darbus, taču nesniedz nekādu informāciju par autoru.
[29] Šis un nākamie piemēri ir izdomāti, un tajos minētie dati ir izvēlēti pēc nejaušības principa.

- profesijas;
- mācību iestāžu un/vai to struktūrvienību nosaukumi.

Ikviena no šīm informācijas vienībām tiek individuāli izvērtēta. Lai nezaudētu pārlieku daudz no sākotnējā teksta, katrā atsevišķā gadījumā tiek lemts, vai attiecīgā informācijas vienība var būt pietiekama, lai ar pārējās tekstā esošās informācijas palīdzību identificētu aprakstīto personu. Šim nolūkam nav konkrētu vadlīniju, jo nav iespējams precīzi paredzēt papildinformāciju, kas varētu būt lasītāja rīcībā un veicināt atpazīšanu. Līdz ar to anonimizēšana ir samērā subjektīva darbība.

Ir vēl kāds aspekts, kas jāņem vērā – vienai un tai pašai informācijas vienībai var būt dažāds personas identitātes atklāšanas potenciāls atkarībā no tā, kāda vēl informācija tekstā ir atrodama. Tā ir, piem., ar specifiskām profesijām, mazām mācību iestādēm un dzīvesvietām. Līdz ar to viena un tā paša veida informācija pat vienā un tajā pašā tekstā var parādīties gan aizstāta, gan neaizstāta. Piem., ja teksta autors raksta, ka viņam ir divas tantes, no kurām viena strādā ļoti specifiskā profesijā vai amatā, piem., par ministri, savukārt otra – ļoti bieži sastopamā, piem., par pārdevēju, tad pirmajā gadījumā profesijas nosaukums tiek aizstāts, bet otrajā – ne.

Arī vietvārdi tiek aizstāti tādā gadījumā, ja tie būtiski sašaurina iespējamo autoru loku. Piem., ja autors raksta, ka dzīvo Līgatnē, tad šis vietvārds tiktu aizstāts, taču, ja ir rakstīts, piem., „es studēju Rīgā", tad, ņemot vērā, ka lielākā daļa tekstu ir tapusi Latvijas Universitātē, kura atrodas Rīgā, šī informācija netiek uzskatīta par aizstājamu.

Ja tekstā ir vēl cita informācija, kas var ļaut atpazīt konkrētu personu, tās aizstāšanai vēlamais risinājums piemeklēts individuāli.

Tomēr šāda anonimizēšana ir iejaukšanās tekstā, kas var ietekmēt tajā balstītu pētījumu rezultātus, ja noviržu cēloņi saskatīti nepilnīgi vai neatbilstoši. Līdz ar to visas vietas, kurās veikti labojumi, ir īpaši marķētas, lai būtu iespējams tās atlasīt gadījumos, ja rodas šaubas par šīs korpusa datu daļas atbilstību pētījuma mērķiem. Ja konkrēti dati kādā no tekstiem nav aizstājami vai ir grūti aizstājami, tie ir izlaisti, to attiecīgi norādot teksta marķējumā (par marķējuma formātu vairāk sk. 2.3.1. apakšnodaļā „Korpusa izveidē izmantotā programmatūra").

Jāņem vērā arī korpusa tehniskā izveide un datu uzglabāšanas principi. Datorprogrammas mēdz saglabāt papildinformāciju par datnēm jeb metadatus, piem., informāciju par datoru, uz kura noteiktā datne tikusi izveidota, kas var padarīt autorus identificējamus pret viņu gribu. Šajā gadījumā nav izveidota tik droša sistēma, lai būtu

iespējams ar pilnīgu pārliecību garantēt, ka nepiederošām personām nav nekādas iespējas no korpusa vietnes iegūt visus tā datus, ieskaitot datnes ar visiem metadatiem. Tā kā korpusa veiksmīgai lietošanai datņu formātu jebkurā gadījumā nepieciešams vienādot, visi teksti tika pārkopēti jaunās vienāda formāta datnēs.

Šāds risinājums izvēlēts, lai aizsargātu tekstu autorus un citus datu subjektus no viņu datu neatļautas apstrādes un izvairītos no nepieciešamības slēgt līgumu ar ikvienu tekstā pieminētu personu. Tomēr uz tekstu autoriem attiecas vēl viens tiesību veids – autortiesības, kuru tiesiskais regulējums un pieņemtais risinājums skaidrots tālāk.

2.1.3.2. Autortiesības

Latvijas Republikas Autortiesību likums (AL) aizsargā radošu darbu autoru tiesības, un tām, kā norādīts AL 2. panta 5. punktā, „ir personisku un mantisku tiesību saturs". Par autortiesību aizsargātu darbu ir uzskatāms jebkāds radošs darbs „literatūras, zinātnes vai mākslas jomā neatkarīgi no tā izpausmes veida, formas un vērtības" (AL 1. panta 2. punkts). Analoģiski tas definēts arī Lietuvas Republikas Autortiesību un blakustiesību likuma (*Autorių teisių ir gretutinių teisių įstatymas*, ATGTI) 2. panta 19. punktā. Studentu rakstītie teksti tātad ietilpst šajā grupā kā zinātnes jomā tapuši radoši darbi. Tātad to autortiesības ir aizsargātas Latvijas Republikas AL un Lietuvas Republikas ATGTI noteiktajā kārtībā.

Neviena no augstskolām, no kurām līdz šim ir saņemti teksti iekļaušanai korpusā, augstskolas līgumā ar studentu nav noteikusi, ka studiju procesā tapušo darbu autortiesības pieder augstskolai. Līdz ar to šajā gadījumā katra atsevišķa teksta autortiesību subjekts ir students, kas tekstu ir rakstījis.

Studentu darbu izmantošana publiski pieejamā korpusā bez attiecīgo studentu ziņas un atļaujas nav pieļaujama, jo, kā norādīts, AL 14. panta 2. punktā un ATGTI 15. panta 2. punktā, tikai darba autors izlemj, vai darbs tiks izziņots[30] un kad tas notiks, ja autors lēmis par labu izziņošanai. Tā kā studentu teksti ir tikuši iesniegti tikai attiecīgā kursa pasniedzējam, tie nav izziņoti. Līdz ar to šo darbu ievietošana korpusā ir arī to izziņošana, jo šādi tie pirmoreiz kļūst pieejami sabiedrībai.

AL 20. panta 1. punktā gan minēts, ka zinātniskos un pētniecības nolūkos drīkst darbus izmantot bez autora atļaujas, taču tas attiecas uz iepriekš publiskotiem un publicētiem darbiem. Līdzīgas normas ir spēkā Lietuvas Republikā (ATGTI 21. un 22. pants); šeit gan ir

[30] Kā skaidrots AL 1. panta 12. punktā, izziņošana ir „darbība, ar kuras palīdzību darbs pirmo reizi kļūst pieejams sabiedrībai neatkarīgi no šīs darbības veida".

atļauts izglītības un pedagogu kvalifikācijas paaugstināšanas nolūkos izmantot darbus, kurus audzēkņi ir radījuši mācību sasniegumu vērtēšanai (ATGTI 22. panta 2. punkts). Tomēr šādā gadījumā jānorāda katra teksta autora vārds un uzvārds. Ņemot vērā, ka korpusa pētniecībā liela vērība tiks pievērsta valodas apguvēju pieļautajām kļūdām, tas var kaitēt autoru cieņai un reputācijai, tāpēc nebūtu vēlams. Turklāt, padarot korpusu publiski pieejamu, nav iespējams garantēt, ka tā auditorija būs tikai izglītības un pedagogu kvalifikācijas paaugstināšanas jautājumos tieši ieinteresētās personas. Līdz ar to ATGTI 22. pantu šajā gadījumā nevajadzētu piemērot.

Jāpievērš uzmanība arī kādam aspektam, kas šajā gadījumā savā ziņā pretnostata personas datu aizsardzības un autortiesību principus. Pat ja tekstu autori piekristu, ka viņu vārds tiek norādīts pie viņu tekstiem, tādā veidā paši uzņemoties atbildību par iespējamo kaitējumu savai reputācijai, daļā tekstu ir norādītas konkrētas personas (piem., ģimenes locekļi, kolēģi utt.), kuras, zinot teksta autoru, būtu precīzi identificējamas. Tas būtu personas datu aizsardzības principu pārkāpums (sīkāk sk. 2.1.2.1. punktā „Personas datu aizsardzība"), lai arī no autortiesību viedokļa tā būtu normāla un pat vēlama prakse. Līdz ar to tika nolemts, ka autoriem, kas vēlas, lai viņu vārds būtu minēts, ir iespēja vārdu iekļaut visa korpusa kopējā autoru sarakstā. Savukārt tie, kas to nevēlas, var saglabāt anonimitāti, jo autoru sarakstā tiek norādīts tikai anonīmo autoru kopskaits bez sīkākas informācijas par šiem autoriem.

Ņemot vērā minētos normatīvo aktu ierobežojumus, ar visu korpusā iekļaujamo tekstu autoriem nepieciešams vienoties par atļauju darbus izmantot valodas apguvēju korpusā, kā arī par veidu, kādā tiek vai netiek norādīts katra atsevišķa teksta autora vārds un uzvārds. Atļaujas autoriem ir tikušas lūgtas arī citu valodas apguvēju korpusu izveides procesā – šis jautājums ir aktuāls dažādās valstīs (Wible u. c. 2001, O'Sullivan, Chambers 2006), ir izveidoti arī atļauju paraugi (CLARIN 2010, META-NET), kas gan šajā gadījumā netika izmantoti to sarežģītības dēļ. Tā kā šim gadījumam būtiskas atšķirības normatīvo aktu prasībās Lietuvas Republikā un Latvijas Republikā netika atrastas, tika nolemts, ka gan Lietuvas, gan Latvijas studentu atļaujas saturs var būt viens un tas pats. Konsultējoties ar juristiem autortiesību speciālistiem, tika sastādīts atļaujas standarta teksts, kurā ir iekļauts:

- atļaujas devēja un atļaujas saņēmēja vārds, uzvārds un personas kods;
- darbu, kuru izmantošanai tiek dota atļauja, nosaukumi un gads, kurā šie darbi tika iesniegti kā mājasdarbi otrās baltu valodas kursā (tajos gadījumos, kad gads nav precīzi zināms, tas atļaujas tekstā nav iekļauts);
- datums, atļaujas devēja paraksts;

- lauks, kurā autors atzīmē, vai vēlas tikt iekļauts atsevišķā autoru sarakstā. Tajā piedāvātas šādas izvēles iespējas:
    - Piekrītu, ka mans vārds un uzvārds tiek iekļauts autoru sarakstā.
    - Vēlos palikt anonīms.
- Darbu izmantošanas nosacījumi:
    - Darbi tiek iekļauti otrās baltu valodas apguvēju korpusā un kā šī korpusa daļa kļūst publiski pieejami dažādās formās, pilnībā vai daļēji.
    - Korpuss ir pieejams bez maksas un ir paredzēts lietošanai mācību un pētniecības nolūkos. Tā lietošana komerciālos nolūkos ir aizliegta.[31] Autori par tekstu iekļaušanu korpusā materiālu atlīdzību nesaņem.
    - Lai aizsargātu tekstos pieminēto cilvēku personas datus, tekstos sniegtā informācija var tikt mainīta, norādot vietu, kurā veiktas izmaiņas, bet nenorādot sākotnējo informāciju. Šī iemesla dēļ arī autora vārds pie katra konkrēta teksta netiek norādīts. Katram autoram tiek piešķirts anonīms kods, ar kura palīdzību ir iespējams atpazīt viena un tā paša autora vairākus darbus, bet nav iespējams noteikt autora identitāti.
    - Korpusā iekļautie dati dažādās formās var tikt citēti mācību līdzekļos un pētnieciskos darbos.
    - Korpuss un visi tajā iekļautie materiāli var būt publiski pieejami neierobežotu laiku un pētīti neierobežotu skaitu reižu.
    - Visiem korpusā iekļautajiem tekstiem var tikt pievienota lingvistiska informācija (piem., kļūdu labojumi, vārdšķiru anotējums u. c.).

Atļaujas teksts sākotnēji tika sagatavots latviešu valodā un tad iztulkots lietuviešu valodā. Atļaujas paraugs latviešu un lietuviešu valodā ir pievienots arī promocijas darbam kā 1. un 2. pielikums.

Korpusa izveides nolūkā katram autoram tika elektroniski nosūtīta vai klātienē mutvārdos sniegta informācija par iespēju piekrist vai nepiekrist darbu izmantošanai un par to, kā tie tiks izmantoti, ieskaitot personas datu aizsardzības nolūkos veiktās darbības. Pēc tam autori tika aicināti izlemt, vai piekrīt savu darbu izmantošanai attiecīgajā veidā ar minētajiem nosacījumiem un, ja tika pausta piekrišana, parakstīt atļaujas veidlapu. Saņemtās atļaujas tiek glabātas korpusa sastādītājas privātajā arhīvā.

[31] Atverot korpusa vietni, lietotāji tiek informēti, ka, lietojot korpusu, viņi apliecina, ka šim nosacījumam piekrīt.

Apstrādāti un korpusā iekļauti tika tikai tie teksti, par kuriem ir saņemta to autoru rakstveida piekrišana tipveida atļaujas formā. Teksti, kuru autori piekrišanu ir pauduši, taču kāda iemesla dēļ nav parakstījuši tipveida atļauju, korpusā nav iekļauti.

Dažkārt pētnieki, vācot materiālus valodas apguvēju korpusiem, izvēlas nevis tikai iegūt atļauju teksta publicēšanai korpusā, bet gan visu attiecīgā teksta autortiesību pārņemšanu no studenta. Šādi rīkojušies, piem., angļu valodas apguvēju korpusa veidotāji Japānā (Nagata u. c. 2011, 1212). Šādi var rīkoties, ja ir vēlme nodrošināties gadījumam, kurā tie paši pētnieki tos pašus tekstus nākotnē varētu vēlēties izmantot vēl citādi. Šajā gadījumā tas nešķiet nepieciešams, tāpēc atļaujas teksts veidots tā, lai autortiesības saglabātu tekstu īstie autori.

Jāpiebilst, ka, tekstus anonimizējot, rodas atvasināts darbs, taču šis nav iemesls jebkādu papildu atļauju saņemšanai: sākotnējo tekstu autori atļaujās ir piekrituši anonimizēšanai un anonimizēto tekstu publiskai pieejamībai, savukārt atvasināto darbu autore vienlaikus ir arī korpusa sastādītāja un anonimizāciju veic tieši tādā nolūkā, būdama informēta par atvasināto darbu plānoto izmantošanu un piekrizdama tai.

2.1.3.3. Korpusa lietošanas nosacījumi

Lietotājs, atverot korpusa vietni, sākumlapā tiek informēts, ka, lietojot korpusu, viņš apliecina savu piekrišanu un apņemas ievērot lietošanas noteikumus, kas ir izstrādāti, lai aizsargātu iepriekš norādītās tekstu autoru tiesības. Tie ir šādi:

- korpusu ir atļauts lietot izglītības un pētniecības nolūkos;
- korpusā iekļautos tekstus drīkst citēt izglītības un/vai pētījumu publikācijās;
- korpusa izmantošana komerciāliem mērķiem ir aizliegta;
- korpusu var pētīt ar jebkādu programmatūru pēc pētnieka izvēles;
- korpusa veidotāja nav atbildīga par korpusa tekstos paustajiem viedokļiem;
- korpusa datnes ir pieejamas tikai pēc saites, kas ir norādīta šajā vietnē, un tās nedrīkst izplatīt citādi.

2.2. Tekstu apstrāde, marķēšana un anotēšana

Viena no korpusa lingvistikas metožu galvenajām priekšrocībām ir datora spēja liela apjoma datu kopā ātri atrast noteiktus elementus un veikt ar tiem saistītos aprēķinus. Taču, tā kā dators jebkādus datus uztver kā simbolu virkni, ierīce nespēj tos interpretēt tā, kā to spēj cilvēks. Tāpēc, pirmkārt, tekstiem jābūt mašīnlasāmā formātā, un, otrkārt, iespējas apstrādāt

tekstu korpusu ir ierobežotas, ja vien netiek pievienota papildu informācija, kas datoram ļauj atpazīt noteiktas parādības arī tad, ja tās ir grūtāk nosakāmas vai formāli neatšķiras.

Līdz ar to viens no svarīgākajiem lēmumiem, kas jāpieņem, lai valodas paraugus pētītu ar korpusa lingvistikas metodēm, ir – vai un kā šos datus nepieciešams marķēt un/vai anotēt. To var darīt pēc dažādiem principiem – piem., korpusā var marķēt tekstu autoru dzimto valodu vai anotēt tekstos vārdšķiras, noteiktas sintaktiskas konstrukcijas, semantiskas pazīmes utt. Korpusa anotēšanas kategoriju sistēma ļoti lielā mērā ietekmē analīzes kvalitāti. To pētījumā ar vācu valodas apguvēju korpusu *Falko* parāda arī A. Līdelinga (Lüdeling 2007, 28–29).

Šajā nodaļā aprakstīta korpusa „Esam" tekstu digitalizēšana, marķēšana un anotēšana.

2.2.1. Digitalizēšana

Visi korpusā iekļaujamie teksti, kurus nepieciešams digitalizēt, ir pārrakstīti ar visām kļūdām. Dažos gadījumos ir redzams, ka tekstā ir veikti labojumi, piem., atšķiras pildspalvas tonis atsevišķiem vārdiem. Ja autors ir veicis kādus labojumus, tad pārrakstot tiek saglabāts pēdējais variants. Lai arī dažādās formas, no kurām autors beigās izvēlējies vienu, var sniegt vēl papildu ieskatu valodas apguvēja domāšanā, šādu parādību marķēšana būtu sarežģīta un tās uzticamība būtu samērā neliela. Autoru pašu labojumu daudzums ir ļoti atšķirīgs dažādos tekstos un nesniedz viendabīgu ainu, jo daļa no studentiem acīmredzot ir rakstījuši atsevišķus melnrakstus, tāpēc iesniegtajā tekstā nekādu labojumu nav, kamēr citi, šķiet, iesnieguši to pašu papīra lapu, uz kuras parādās viss teksta tapšanas process. Pat tad, ja ir redzamas iepriekšējās versijas, tās ne vienmēr ir salasāmas. Turklāt jāņem vērā, ka korpusā „Esam" rokrakstā un datorrakstā iesniegtie teksti tiek pielīdzināti cits citam, un datorrakstā autora paša labojumi nav redzami. Tāpēc korpusā iekļauta teksta gala versija, kādu autors ir uzskatījis par pareizu vai vistuvāko pareizajai. Rokraksti gan tiek saglabāti korpusa dokumentu arhīvā – ja kādreiz rastos interese tieši par rokraksta īpatnību izpēti, to būs iespējams darīt. Tomēr tos nevar iekļaut korpusa publiski pieejamajā daļā, jo tie nav anonimizēti, t. i., satur personas datus (vairāk par personas datu aizsardzību sk. 2.1.3. apakšnodaļā „Personas datu aizsardzība un autortiesības").

Ja sākotnējais variants ir salasāms un ja labojums veikts pēc tam, kad pasniedzējs ir to atzīmējis kā kļūdu, tad transkripcijā tiek rakstīts sākotnējais, t. i., kļūdainais variants. Ja sākotnējais variants nav salasāms (piem., bijis rakstīts ar parasto zīmuli un labojot izdzēsts) vai ja labojums acīmredzot veikts pirms nodošanas pasniedzējam (pasniedzējs attiecīgo kļūdu nav atzīmējis), tad transkripcijā tiek rakstīts labotais teksts.

Nedaudzos gadījumos teksta autors ir dalījis vārdu pārnešanai jaunā rindā. Lielākoties tomēr vārds, kas rindā neietilpst, rakstīts jaunā rindā vai arī pāri malai (ja papīra lapai ir platākas malas). Turklāt arī šī parādība nav sastopama datorrakstā iesniegtajos tekstos, tāpēc pārnesums jaunā rindā nav saglabāts, t. i., teksts ir pārrakstīts bez defises, turpinot tajā pašā rindā.

2.2.2. Marķēšana

Kā jau iepriekš skaidrots, ar korpusa marķēšanu šajā darbā tiek saprasta ekstralingvistiskas papildinformācijas, kura nav atkarīga no lingvistiskas analīzes, jeb metadatu iekļaušana teksta datnēs.

Korpusā „Esam" ir divu veidu marķējums:

- strukturālais marķējums:
  - teksta sākums un beigas marķētas, izmantojot *TEI* standartu (TEI 2015);
  - rindkopas sākums un beigas marķētas teksta struktūras saglabāšanai;
  - izteikuma sākums un beigas marķētas teksta struktūras saglabāšanai un sintaktiskās anotēšanas nodrošināšanai;
  - marķētas anonimizētās vietas tekstā (par to vairāk sk. 2.1.3. nodaļā „Personas datu aizsardzība un autortiesības").
- metadatu marķējums:
  - marķēts teksta autors – norādīts autora unikālais kods (par to vairāk sk. 2.1.3. nodaļā „Personas datu aizsardzība un autortiesības");
  - marķēta augstskola, kurā teksta autors apguvis otro baltu valodu: Liepājas Universitāte, Latvijas Universitāte, Vītauta Dižā universitāte vai Viļņas Universitāte;
  - marķēts otrās baltu valodas apguves semestris: pirmais vai otrais;
  - marķēta katra teksta mērķvaloda: latviešu vai lietuviešu;
  - marķēts teksta nosaukums.

Marķējuma formāts ir aprakstīts 2.3.1. apakšnodaļā „Korpusa izveidē izmantotā programmatūra".

2.2.3. Anotēšanas veidu izvēle

Pasaulē strauji pieaug anotētu valodas korpusu skaits, jo tie palīdz atklāt daudz plašāku valodas parādību klāstu nekā neanotētie. Tomēr anotēšana prasa laiku un ir darbietilpīga, tāpēc neanotēti korpusi dažkārt var izrādīties pat piemērotāki noteikta pētījuma

mērķa sasniegšanai. Valodas apguvēju korpusi tādi ir samērā reti, jo materiāla vākšana prasa gana lielas pūles, un tāpēc no tā parasti cenšas iegūt pēc iespējas izsmeļošu informāciju. Taču, tā kā mūsdienās valodas apguvēji nereti tekstus rada un pasniedzējiem iesniedz elektroniski, neanotētu valodas apguvēju korpusu izveidot ir kļuvis ievērojami vieglāk. Arī otrās baltu valodas apguvēju korpusu var izvēlēties lietot neanotētu. Piemēri, kā strādāt ar šādu neanotētu korpusu, ir doti gan šī darba 2.4. nodaļā „Pētījumu iespējas otrās baltu valodas apguvēju korpusā", gan rakstā „Pētniecības iespējas neanotētā baltu valodu apguvēju korpusā" (Znotiņa 2015).

Izvēlē par labu valodas apguvēju korpusa anotēšanai vai – gluži otrādi – neanotēšanai galvenā opozīcija ir: no anotēšanas iegūstamais labums pret anotēšanai nepieciešamajiem resursiem.

Anotēšanā ieguldāmo resursu apjomu ietekmē dažādi faktori. Lūk, daļa no tiem:

- zināšanas, prasmes, pieejamie tehnoloģiskie risinājumi. Ja anotētāja prasmes darbā ar datoru vai zināšanas par anotēšanas kategorijām ir zemas, arī anotēšanas efektivitāte ir zema. Turklāt ir svarīgi, lai anotēšanai izvēlētā programma būtu piemērota anotējamajiem datiem;
- anotējuma variāciju (interpretācijas iespēju) daudzums, it īpaši, anotējot kļūdas valodas apguvēju korpusā. Anotētājs veic lingvistisku analīzi un bieži vien pieņem lēmumus par labu vienai interpretācijai no vairākām iespējamām. Ir nepieciešams laiks, lai pienācīgi izanalizētu attiecīgo piemēru un pieņemtu lēmumu par piešķiramo pazīmi;
- anotējuma sistēmas pieejamība vai nepieciešamība tādu radīt. Anotēšanas pamatā ir kāda klasifikācijas sistēma. Vēl ir nepieciešams datorlasāmu zīmju komplekts katras klasifikācijas grupas atšķiršanai, piem., lietvārdus apzīmējot ar <LTV> vai tamlīdzīgi. Dažkārt šādas sistēmas ir jārada pašam anotētājam, taču mēdz būt arī universālas sistēmas (Granger 2003a), un, ja tādas ir pieejamas, to izmantošana vai pielāgošana nereti prasa mazāk resursu nekā jaunas sistēmas izveide;
- korpusā iekļaujamo tekstu apjoms. Jo vairāk ir tekstu, ko anotēt, jo vairāk cilvēkstundu šajā darbā jāpatērē. Tomēr, tā kā apjoms ietekmē arī iegūstamo labumu (par to sk. tālāk), tā uzskatīšana par pamatkritēriju varētu tikt apstrīdēta.

Līdzīgi arī anotēšanas rezultātā iegūstamo vērtību ietekmē vairāki faktori, piem.:

- korpusā iekļaujamo tekstu apjoms. Tas ietekmē arī nepieciešamo resursu apjomu, taču, kā to atzinuši daudzi pētnieki, korpusa lielums tieši ietekmē iegūstamo rezultātu ticamību, sevišķi, ja runa ir par statistiskiem aprēķiniem. Lai gan valodas apguvēju korpusos arī salīdzinoši neliels apjoms nereti tiek uzskatīts par pietiekamu (piem., nedaudzi simti tūkstošu vārdu kā korpusā *ICLE* – sk. Granger 2003b; vairāk par valodas apguvēju korpusu apjomu sk. šī promocijas darba 1.1.2. apakšnodaļā „Valodas apguvēju korpusa raksturīgie parametri"), orientācija uz tā palielināšanu saglabājas arī šādos korpusos. Līdz ar to apjoms pats par sevi nav noteicošais izvēlē par vai pret anotēšanu, jo tas palielina gan nepieciešamo resursu apjomu, gan gūstamo labumu, kas turklāt ir atkarīgs no daudziem citiem blakus faktoriem;
- pētījumu skaits, apmērs un nozīmīgums. Attiecīgās pētījuma problēmas un to iecerētais risinājums jāizvērtē – kāds ir ieguvums, cik būtiskas atziņas tas sniegs? Ieguvumam būtu jāatsver materiālie un cita rakstura zaudējumi tekstu anotēšanā;
- iespēja anotējumu izmantot dažādos pētījumos. Jo plašāks ir noteikta anotēšanas veida lietojums dažādos pētniecības virzienos, jo lielāks ir ieguvums no tā paša darba daudzuma. Piem., ja korpusā tiek anotēti tikai transitīvie un intransitīvie verbi, tad šāda anotējuma veida lietojums ir šaurāks nekā, piem., pamatformu anotējumam.

Lemjot par anotēšanu, ir arī daudz citu vērā ņemamo faktoru, piemēram, pētnieka aizspriedumi vai personīgā motivācija; pieejamie palīgi – darbaspēks ar salīdzinoši zemām izmaksām; īpašas prasmes vai to trūkums darbā ar noteiktu programmatūru u. c. Līdz ar to lietderība nav precīzi izmērāma, bet gan daļēji subjektīva, un kritēriji ir izvērsti jāskaidro katrā atsevišķā gadījumā.

Galvenokārt jāņem vērā, kādus pētījumus ar attiecīgo korpusu ir paredzēts veikt. Tā kā otrās baltu valodas apguvēju korpuss nav paredzēts viena vai dažu konkrētu pētījumu veikšanai, bet gan pēc iespējas daudzpusīgai savākto valodas paraugu analīzei, tika nolemts par vienu no svarīgākajiem kritērijiem anotējuma veidu izvēlē uzskatīt to iederību esošajā pētniecības kontekstā. Tāpēc ir jānoskaidro valodas apguvēju korpusu izveides un pētniecības tendences Latvijā, Lietuvā un citur.

2.2.3.1. Valodas apguvēju korpusa anotējuma tendences Latvijā un Lietuvā

Valodas apguvēju korpusu pētniecība strauji attīstās, visā pasaulē tiek veidoti dažādi valodas apguvēju korpusi, un ir daudz arī anotējuma veidu variāciju. Dažkārt, veidojot jaunus valodu apguvēju korpusus, tiek izmantoti jau zināmi anotēšanas principi un pazīmju kopas, lai esošo un jaunveidoto korpusu datus varētu salīdzināt. Nereti gan korpusu veidotāji izveido arī savu anotējumu, taču salīdzināmība tomēr nereti tiek uzskatīta par priekšrocību, jo tā ļauj izvērtēt līdzšinējos priekšstatus par apguvēju valodu.

Ne vienmēr ir skaidrs, kuriem no līdz šim tapušajiem korpusiem būtu jācenšas pielīdzināt jaunveidojamos. Gadījumos, kad nav viena dominējoša modeļa, pētnieki dažkārt izvēlas apvienot piemērotākās īpatnības no dažādiem pastāvošiem korpusiem, nevis vienu no tiem pilnībā atdarināt, it īpaši tad, ja jaunajā korpusā ievietojamie dati nav piemēroti tiešam salīdzinājumam ar jau esošajiem korpusiem. Šāds ir korpusa „Esam" gadījums. Lai arī teksti, kas ir savākti otrās baltu valodas korpusam, nav sevišķi līdzīgi citu, jau eksistējošu korpusu materiālam, to var anotēt arī pēc citos korpusos izmantotām pazīmju kopām.

Otrās baltu valodas apguvēju korpusa primārā mērķauditorija ir Latvijas un Lietuvas pētnieki, tāpēc jāzina, kā ar valodas apguvēju korpusiem tiek strādāts šajās valstīs, lai iesākto darbu varētu turpināt ar jauno materiālu. Tomēr baltu valodas nav izolētas, un to apguvē var būt ar citu valodu apguvi salīdzināmi jautājumi, tāpēc jāņem vērā arī citviet aktuālie pētījumi. Līdz ar to šeit ne vien īsi raksturots katrs anotēšanas veids, bet arī aplūkots pētniecības konteksts galvenokārt (bet ne tikai) Latvijā un Lietuvā.

Liela daļa Latvijā un Lietuvā veidoto valodas apguvēju korpusu nav anotēti (Bikelienė 2009b; Burneikaitė 2009a; Juknevičienė 2009) un šeit netiek raksturoti. Savukārt šai darbā sīkāk aplūkotie anotētie korpusi ir veidoti dažādām valodām, tāpēc piemēri būtu pārāk dažādi, lai tos salīdzinātu. Tika nolemts, ka anotēšanas veidu ilustrēšanai vispiemērotākais būtu viens un tas pats teksts, shematiski anotēts pēc dažādiem klasifikācijas veidiem, tāpēc izmantots īss teikums latviešu un lietuviešu valodā, kā arī grūti atšifrējamu tagu vietā izmantoti apraksti, kas sastāv no pilniem vārdiem.

2.2.3.1.1. Valodas līmeņos balstīta anotēšana

Valoda tiek aplūkota dažādos līmeņos: fonētiskajā, morfoloģiskajā u. c. (Crystal 1993, 15). Tajā balstoties, valodnieki attiecīgi veic fonētisku analīzi, morfoloģisku analīzi utt., no kā izriet attiecīgi korpusa anotēšanas veidi, no kuriem populārākā šķiet morfoloģiskā un sintaktiskā anotēšana. Šādi ir anotēti dažādu veidu korpusi gan Latvijā, gan Lietuvā (Levāne-Petrova 2011; Rimkutė u. c. 2009).

Morfoloģiskā anotēšana

Korpusu morfoloģiskā anotēšana Latvijā ir pietiekami izplatīta, taču mazāk – valodas apguvēju korpusos. Morfoloģiski anotējot datus, vārdam tiek pievienota tam atbilstošā morfoloģiskā informācija (vairāk sk. Fitschen, Gupta 2008). Tiek izmantotas dažāda apjoma pazīmju kopas, piem., tādas, kas iekļauj informāciju par vārdformām (Levāne 2001; Paikens 2016; Paikens u. c. 2013), vai tādas, kurās iekļauta tikai viena pazīme – vārdšķira (līdz šim – valodas apguvēju korpusos). Anotējot pēc vārdšķirām, katram vārdam tiek piešķirts attiecīgās vārdšķiras kods, piem.:

<pos="p">*Viņa*</pos> <pos="v">*ir*</pos> <pos="a">*skaista*</pos>.

<pos="p">*Ji*</pos> <pos="v">*yra*</pos> <pos="a">*graži*</pos>.[32]

Anotējot vārdšķiras, nav iespējamas lielas variācijas, jo katrā valodā dalījums vārdšķirās ir diezgan stabils. Ir vārdi, par kuriem var raisīties diskusijas (sk. 2.2.5. apakšnodaļā „Morfoloģiskā anotēšana otrās baltu valodas apguvēju korpusā"), taču vairums ir skaidri identificējami kā lietvārdi, darbības vārdi, vietniekvārdi utt.

Pēc šāda principa ir anotēts LVASA pētījuma latviešu valodas apguvēju korpuss (Lokmane u. c. 2009, 91–92). Šeit ir izmantota ļoti vienkārša klasifikācija: ir identificētas vārdšķiras, kādas tiek šķirtas latviešu valodā, un katrai no tām ir piešķirts tags, kurā ir viens mazais burts:

- lietvārdi <n>;
- darbības vārdi <v>;
- īpašības vārdi <a>;
- apstākļa vārdi <r>;
- prievārdi <s>;
- saikļi <c>;
- skaitļa vārdi <m>;
- izsauksmes vārdi <i>;
- partikulas <q> (Lokmane u. c. 2009, 91–92).

Šis iedalījums ir jau tradicionāli ierasts, gandrīz tādu pašu min gan Dzintra Paegle (Paegle 2003, 25), gan Daina Nītiņa (LVG 2013, 320) u. c. Abu minēto autoru darbos vēl tiek šķirti arī vietniekvārdi, kas nav norādīti LVASA pētījuma korpusa anotēšanas sistēmas

[32] Tagi atbilst Latvijas Universitātes Matemātikas un informātikas institūtā izstrādātajai latviešu valodas morfoloģisko pazīmju kopai: *http://www.semti-kamols.lv/doc_upl/TagSet.html*

informācijā (Lokmane u. c. 2009, 91–92), taču uzreiz pēc tās sniegtajā anotētā teksta paraugā arī šī vārdšķira ir pārstāvēta un anotēta ar atsevišķu tagu <p>, tāpēc tā uzskatāma tikai par maznozīmīgu kļūdu aprakstošajā materiālā, bet ne korpusa izveidē.

LVASA pētījuma korpusā vārdšķiras nav sīkāk sadalītas apakšgrupās. Pētījuma autores gan nav norādījušas, kā tieši (automātiski vai manuāli) anotēšana tikusi veikta, taču, tā kā darbā nekur nav norādīts anotēšanas rīks, ticams, ka ir anotēts manuāli. līmeņa apguvēju teksti (tie, kas ieguvuši vērtējumu F) nav morfoloģiski anotēti, jo šajā līmenī „ir pārāk daudz nenosakāmu vārdu formu" (Kalnbērziņa u. c. 2011, 18). Augstāku līmeņu tekstos neskaidrajos gadījumos par primāro kritēriju izvēlēta vārda fukcija teikumā, nevis forma.

Latvijā ir tapis vēl viens valodas apguvēju korpuss, kurā ir anotētas vārdšķiras – tas ir Z. Vinčelas pētījumos izmantotais angļu valodas apguvēju korpuss (Vinčela 2011c, 2). Šis korpuss ir anotēts, izmantojot automātisko anotēšanas rīku angļu valodai *CLAWS* (Vinčela 2014). Šis rīks piedāvā dažādus anotējuma formātus, kā arī anotēšanas klasifikācijas veidus (Leech u. c. 1994), un vairums no tiem ir daudz sarežģītāki par to, kāds ir lietots iepriekšminētajā latviešu valodas apguvēju korpusā. Tā kā *CLAWS* darbojas tikai ar angļu valodas tekstiem, tas nav piemērots baltu valodu anotēšanai, un tajā izmantotā klasifikācijas sistēma ir pārāk sarežģīta, lai to izmantotu, manuāli anotējot korpusu „Esam", tāpēc tā šim korpusam nav izvēlēta.

Par valodas apguvēju korpusu morfoloģisku anotēšanu Lietuvā publikācijas nav atrastas.

Morfosintaktiskā anotēšana

Vispārīgi morfosintaktiskā anotēšana attiecas gan uz morfoloģiskām, gan sintaktiskām parādībām un dažkārt pat tiek saprasta kā vienkārši dziļāka morfoloģiskā anotēšana, iekļaujot informāciju par locījumu, dzimti utt. (Trushkina, Hinrichs 2004). Tomēr biežāk ar to saprot vārdu anotēšanu pēc to funkcijām teikumā (EAGLES 1996; Rögnvaldsson 2006 u. c.), teikuma locekļu anotējumu (teikuma priekšmets, izteicējs, papildinātājs, apstāklis u. c.). Līdz ar to šī anotēšanas veida piemērs shematiski varētu izskatīties šādi:

<pp, loma="teikuma priekšmets">*Viņa*</pp> <v, loma="saitiņa">*ir*</v> <a, loma="izteicējs">*skaista*.

<pp, loma="teikuma priekšmets">*Ji*</pp> <v, loma="saitiņa">*yra* <a, loma="izteicējs">*graži*.

No vienas puses, morfosintaktiskajā anotēšanā, tāpat kā morfoloģiskajā anotēšanā, nav jābūt sevišķi plašam tagu klāstam. No otras puses, tā ir daudz sarežģītāka par morfoloģisko anotēšanu, it sevišķi valodas apguvēju korpusos, kuros teikumi var būt neveiksmīgi veidoti vai neskaidri. Morfosintaktiski anotēt vieglāk ir tādu apguvēju valodas paraugus, kuru prasmju līmenis ir augsts, nevis iesācēju producētos tekstus.

Latvijā un Lietuvā valodas apguvēju korpuss ir ticis morfosintaktiski anotēts, šķiet, tikai vienā pētījumā. Vitalija Kazlauskiene (*Vitalija Kazlauskienė*) no Viļņas Universitātes šādi anotējusi franču valodas eksāmenu darbu korpusu (Kazlauskienė 2015). Sīkākas informācijas par anotējumu gan pagaidām nav.

Sintaktiskā anotēšana

Sintaktiskā anotēšana var tikt veikta dažādi (vairāk sk. Palmer, Xue 2010), taču viens no populārākajiem tās veidiem balstās teikumu veidu sadalījumā:

<s type="vienkāršs teikums">*Viņa ir skaista*.</s>

<s type="vienkāršs teikums">*Ji yra graži*.</s>

Tāpat kā citi anotēšanas veidi, kas balstās valodas līmeņos, katras valodas sistēma padara būtisku variāciju iespēju diezgan mazticamu. Jo sarežģītāka ir izmantojamā klasifikācija, jo lielāka ir iespēja, ka anotēt nepieciešams manuāli, vismaz valodas apguvēju korpusos, kuros anotēšanas programmatūras veikto automātisko analīzi var traucēt apguvēju pieļautās kļūdas.

Sintaktisko anotēšanu izmanto divas latviešu valodnieces: V. Rūtenberga un V. Kalnbērziņa. V. Rūtenberga to izmanto disertācijā (Rūtenberga 2014) un ar to saistītajās publikācijās. Te ir izmantota ļoti vienkārša klasifikācija: ir trīs tagi, kas atbilstoši tiek lietoti vienkāršu teikumu, saliktu sakārtotu teikumu un saliktu pakārtotu teikumu anotēšanai. Pētniece pati šo uzskata par problēmorientētu anotējumu, jo tas ir cieši saistīts ar viņas pētījuma jautājumu un korpusā nav izmantoti nekādi citi anotēšanas veidi (Rūtenberga 2014, 108). Korpuss anotēts manuāli, jo anotējamajos tekstos ir kļūdas, kuras apgrūtina automātisku anotēšanu.

V. Rūtenbergas un V. Kalnbērziņas pētījumā ir salīdzināti angļu un franču valodas dati. Arī klasifikācija, pēc kuras teksti ir anotēti, šeit ir sarežģītāka. Tā joprojām balstās tajos pašos trijos teikumu veidos, taču kategorijas ir sadalītas sīkāk, lai analizētu teikuma daļu veidus un to savstarpējo novietojumu (Kalnbērziņa, Rūtenberga 2012; Rūtenberga, Kalnbērziņa 2013; Kalnbērziņa 2015).

Vēl viens sintaktiski anotēts valodas apguvēju korpuss Latvijā ir iepriekšminētais LVASA latviešu valodas apguvēju korpuss (Kalnbērziņa u. c. 2011). Izmantotajā klasifikācijā ir nodalīti seši teikumu veidi, apzīmēti ar lielajiem burtiem:

- vienkāršs teikums <VT>;
- salikts sakārtots teikums <SST>;
- salikts pakārtots teikums <SPT>[33];
- jaukts salikts teikums <JST>;
- reducēts teikums <RT>;
- neskaidrs teikums ar nenosakāmu struktūru <?> (Lokmane u. c. 2009, 92).

Iepriekšminētajiem trim teikumu veidiem pievienojušies vēl divi, kā arī tags neskaidro gadījumu anotēšanai. Sīkāka kategorizācija šajā pētījumā nav veikta. LVASA veidotajā latviešu valodas apguvēju korpusā zemākā līmeņa ieguvēju teksti (ieguvuši vērtējumu F) nav sintaktiski anotēti, jo šajā līmenī „teikumu veidi daudzos gadījumos nebija nosakāmi" (Kalnbērziņa u. c. 2011, 17).

Par valodas apguvēju korpusu sintaktisku anotēšanu Lietuvā publikācijas nav atrastas.

Korpuss „Esam" ir anotēts gan sintaktiski, gan morfoloģiski. Kā vispiemērotākais morfoloģiskās anotēšanas veids izvēlēta vārdšķiru anotēšana, jo šis anotēšanas veids ir samērā vienkāršs. Tomēr bieži ir sastopamas neviennozīmīgas formas, it īpaši, ja apguvēja valodas prasmju līmenis ir zemāks. Tā kā gan latviešu valoda, gan lietuviešu valoda ir dažādām formām bagātas, otrās baltu valodas apguvēju korpusā šis ir ievērojams izaicinājums. Tekstos ir sastopamas gan kļūdas, gan neskaidras gramatiskās konstrukcijas, tāpēc tie ir anotēti manuāli. Gan latviešu, gan lietuviešu valodai ir radīti automātiski anotēšanas rīki, taču tie vislabāk darbojas tekstos, kuros morfoloģiskā daudznozīmība nav plaši sastopama (vairāk sk.: Rimkutė, Daudaravičius 2007; Paikens 2007). Valodas apguvēju tekstos daudznozīmība ir bieži konstatējama un nereti – negaidītās izpausmēs (konstruēti neeksistējoši vārdi, piem., la. *vazināties* 'vizināties, braukāties', lie. *rankdarbininkė* 'rokdarbniece'; vārdi, kas mērķvalodā pārstāv vienu vārdšķiru, lietoti kā citas vārdšķiras vārdi, piem., la. *izskats* (lietvārds; lietots kā darbības vārda *izskatīties* forma), lie. *meilužis* ('mīļākais', lietvārds; lietots kā īpašības vārda *mīļš* vispārākās pakāpes formas *mīļākais* lietuviskā atbilsme); utt.). Tāpēc, pat ja analīze tiktu veikta automātiski, tās rezultātu pārskatīšana vienalga būtu darbietilpīga. Iesācēju rakstīto

[33] Pētījumā sniegtajā pazīmju kopā gan saliktam pakārtotam teikumam norādīts tags <SPK>, taču anotētā teksta paraugā ir redzams tags <SPT>, un pēc analoģijas ar pārējiem tagiem ticams, ka tas ir pareizais taga formāts.

tekstu īpatnību dēļ pazīmju kopa ir arī piemērota neskaidru piemēru anotēšanai – lai arī parasti, dažādus korpusus anotējot, tiek mēģināts klasifikācijā skaidri nodalīt dažādas klasifikācijas grupas (zināmā mērā to mēģināts darīt, arī anotējot korpusu „Esam"), viena no pētniekiem interesantākajām parādībām valodas apguvēju valodā varētu būt tieši neskaidrība, kas ne vien sniedz ieskatu valodas apguvē sastopamajās grūtībās, bet arī var palīdzēt rekonstruēt valodas apguvēja domu gaitu. Un, tā kā pat vārdšķiru anotēšanā rodas šķēršļi, dziļāka morfoloģiskā anotēšana (piem., morfēmu anotēšana) būtu vēl lielāks izaicinājums, tāpēc šādi anotēšanas veidi korpusam „Esam" nav izvēlēti. Tas pats attiecas arī uz morfosintaktisko anotēšanu – tā balstās vārdšķiru anotēšanā, taču klasifikācija noteiktos gadījumos ir vēl sarežģītāka par vārdšķiru anotēšanu, līdz ar to arī ir diezgan grūts uzdevums iesācēju valodas apguvēju korpusā.

Otrās baltu valodas apguvēju iesācēju teksti ir arī sintaktiski anotēti. Latvijā valodas apguvēju korpusu anotēšanai lietotas dažādu sarežģītības pakāpju klasifikācijas. Korpusā „Esam" anotēts tikai teikuma veids (vienkāršs, salikts utt.), tomēr klasifikācija nav pārņemta no V. Rūtenbergas darbiem, bet gan (ar pielāgojumiem) no latviešu valodas apguvēju korpusa anotējuma. Pārlūkojot korpusā „Esam" iekļautos tekstus, ir atrodami dažādu veidu teikumi, un trīs grupu klasifikācija to pilnībā neatspoguļo, kamēr sešu grupu klasifikācija iekļauj arī retāk, tomēr sastopamos jauktos saliktos teikumus, kā arī reducētos teikumus. Tāpat kā morfoloģiskas un morfosintaktiskas anotēšanas gadījumā, iesācēju valodas apguvēju teksti anotēti manuāli.

2.2.3.1.2. Kļūdu anotēšana

Anotēt kļūdas korpusā nozīmē pievienot tekstam vajadzīgajās vietās informāciju, kā attiecīgā vieta būtu labojama, t. i., mērķa hipotēzi (par to vairāk sk. 2.2.7.1. „Kļūdas definīcija"), tad, salīdzinot to ar oriģināla formu, noteikt kļūdas tipu un arī to pievienot anotējumam. Vienkāršas kļūdas gadījumā, kad nav saskaņota vietniekvārda un īpašības vārda dzimte, tas varētu izskatīties tā:

Viņa ir <tok cform="skaista" error="MD">*skaists*</tok>.

Ji yra <tok cform="graži" error="MD">*gražus*</tok>.

Kamēr dažādi korpusi gan Baltijas valstīs, gan citur bieži tiek anotēti atbilstoši valodas līmeņiem, kļūdu anotēšana nereti tiek asociēta galvenokārt ar valodas apguvēju korpusiem. Tas gan nenozīmē, ka šis ir vienīgais korpusa veids, kurā var anotēt kļūdas – piem., šāds anotējuma veids izmantots arī latviešu valodas kā dzimtās valodas korpusā, kas tiek izmantots gramatikas pārbaudes rīka izstrādē (Deksne, Skadiņa 2014). Tomēr, runājot par valodas apguvēju korpusiem, Latvijas un Lietuvas publikācijās, šķiet, nav atrodama informācija par kļūdu

anotēšanu. Kļūdas tiek pieminētas, dažkārt arī nedaudz analizētas – tā rīkojas, piem., Z. Vinčela (Vinčela 2014). Viņa piemin ļoti vispārīgu klasifikācijas veidu, runājot par „vārdu" kļūdām (izvēlēts nepareizais vārds) un „ne-vārdu" (vārds izvēlēts pareizi, bet lietots nepareizi) kļūdām (angļu val. *word errors* un *non-word errors*), kur pēdējās var iedalīt sīkāk pareizrakstības kļūdās un morfoloģiskajās kļūdās (angļu val. *spelling errors* un *morphological errors*) (Vinčela 2014, 125). Pētniece gan nav sīkāk strādājusi ar visaptverošu kļūdu klasifikāciju.

Tas, kā kļūdas tiek anotētas, dažādos darbos var ļoti atšķirties, jo tas ir atkarīgs no veida, kā kļūdas tiek klasificētas. Klasifikācijas izvēli savukārt ietekmē vairāki faktori, t. sk. korpusa veids un plānotais klasifikācijas izmantojums. Iepriekšminētajā Daigas Deksnes un Ingunas Skadiņas publikācijā ir nosaukti 22 kļūdu tipi, kas ir iedalīti piecās grupās: formatējuma kļūdas, pareizrakstības kļūdas, morfoloģijas un sintakses kļūdas, interpunkcijas kļūdas un stila kļūdas (Deksne, Skadiņa 2014, 164). Lai arī nav iemesla apstrīdēt šīs klasifikācijas noderīgumu latviešu valodas kā dzimtās valodas tekstu anotēšanā, iesācēju valodas apguvēju tekstiem tā nav piemērota. Pastāv kļūdas, kas ir ļoti raksturīgas iesācēju tekstiem, taču ir reti atrodamas dzimtās valodas runātāju tekstos, piem., nepareizs vārds vai vārds, kura nozīme ne vien nav tāda, kā tekstā iecerēts, bet pat nav vēlamajai līdzīga. Turklāt minētās klasifikācijas autores ir nošķīrušas dažus diezgan specifiskus kļūdu veidus, piem., nepareizs lietvārda locījums, ja darbības vārds ir lietots vajadzības izteiksmē. Lai arī šāda kļūda droši vien nav retums latviešu valodas kā dzimtās valodas tekstos, tomēr diezin vai nesen mācīties sākušu valodas apguvēju tekstos tā būtu tik nozīmīga, lai tiktu nodalīta atsevišķā grupā.

Līdz ar to otrās baltu valodas apguvēju korpusā nav tieši pārņemta ne latviešu valodas kā dzimtās valodas korpusā izmantotā, ne citām valodām izveidotās kļūdu klasifikācijas, jo tajās var būt kategorijas, kuru baltu valodās nav, piem., artikuli (sk., piem., klasifikāciju, kuru piedāvā Granger 2003a). Tā vietā izveidota klasifikācija, kas atbilst baltu valodām un spēj raksturot iesācēju tekstiem raksturīgo plašo kļūdu daudzveidību (sk. 2.2.7. apakšnodaļā „Kļūdu anotēšana otrās baltu valodas apguvēju korpusā"). Tā kā, šķiet, nevienā latviešu un/vai lietuviešu valodas apguvēju korpusā līdz šim vēl nav anotētas kļūdas, klasifikācija, kas atbilst abām valodām, var palīdzēt ne vien kļūdu analīzē atsevišķi, bet arī salīdzināmos abu valodu apguvēju kļūdu pētījumos.

2.2.3.1.3. Problēmorientēta anotēšana

Dažkārt, it īpaši plašāka apmēra pētījumos, pētnieks sastāda individuālu kritēriju sistēmu, pēc kuras anotēt iegūtos datus. Tā ir problēmorientēta (*problem oriented*) anotēšana,

un šādā gadījumā kritēriju izvēle ir cieši saistīta ar pētījuma mērķi. Problēmorientēta anotēšana nav homogēna valodas apguvēju korpusu anotēšanas veidu kategorija. Tas drīzāk būtu uzskatāms var veidu, kā kopā aplūkot tos anotēšanas veidus, kuru uzdevums nav padarīt korpusu pēc iespējas noderīgāku vairumam mērķu. Tā vietā problēmorientētā anotēšana pievēršas šauru, specifisku pētījumu problēmu risināšanai.

Problēmorientētai korpusa anotēšanai nevar sniegt vispārīgu piemēru, jo variācijas ir gandrīz neierobežotas. Jebkuru lingvistisku elementu, ko ir iespējams identificēt, ir arī iespējams anotēt korpusā un pētīt ar korpusa lingvistikas metodēm. Piem., ja kādu pētnieku konkrēti interesētu konstrukcijas, kas sastāv no darbības vārda *būt* un īpašības vārda, anotētais teksts varētu izskatīties apmēram tā:

Viņa <ir+īpašības vārds> *ir skaista* </ir+īpašības vārds>.

Ji <ir+īpašības vārds> *yra graži* </ir+īpašības vārds>.

Noteikumus un ierobežojumus šeit nosaka pats pētnieks, tāpēc problēmorientēta anotēšana ir sevišķi elastīga. Fakts, ka šāda anotēšana bieži vien ir jāveic manuāli un ka tās izmantojums ir ierobežots, liek īpaši apsvērt, vai ieguvums no šāda anotējuma atsver anotēšanā patērēto resursu apjomu. Ir gandrīz neiespējami zināt, kāda veida problēmorientēts anotējums varētu būt nepieciešams pētniekiem nākotnē, un šo anotējuma veidu parasti izvēlas, tikai kad ir identificēta konkrēta pētījuma problēma. Šī iemesla dēļ arī otrās baltu valodas apguvēju korpusā problēmorientēts anotējums nav paredzēts.

Piemēram, lietuviešu valodnieces J. Grigaļūniene un R. Juknevičiene analizē divdabja *-ing* lietojumu angļu valodas apguvēju korpusā (Grigaliūnienė, Juknevičienė 2012). Līdz ar to šādi divdabji ir vienīgie valodas elementi, kurus viņas izmantotajā korpusā anotē.

Ir vairāki valodnieki, kas veikuši pētījumus, kuros, iespējams, ir izmantojuši problēmorientētu anotējumu, taču attiecīgajās publikācijās trūkst informācijas, lai droši to apgalvotu. Z. Vinčela vienā no darbiem apraksta lingvistiskā variatīvuma pētījumu tekstos, kuros ir ne tikai anotētas vārdšķiras (sk. iepriekš), bet ir identificētas arī citas parādības, kas īpaši izvēlētas šai analīzei (Vinčela 2011c, 2–3). Pētniece nepaskaidro, vai attiecīgās parādības tekstos anotētas vai meklētas bez anotējuma. Ja tās ir anotētas, tad šis ir tipisks problēmorientētas anotēšanas piemērs.

ICLE korpusa Lietuvas daļā ir pētīti diskursa marķieri; publikācijā gan nav īsti skaidri norādīts, vai tie aplūkoti, meklējot pēc konkrētām vārdformām, vai arī korpusa materiāls iepriekš anotēts (Šimčikaitė 2012).

Irina Surkova promocijas darbā „Kreativitāte mērķvalodas lietošanā" (Surkova 2008) un viņas vadībā – arī Svetlana Živjuka diplomdarbā „Kreativitāte rakstīšanas procesā"

(Živjuka 2008) arī, iespējams, anotē noteiktas parādības. I. Surkova promocijas darbā raksturo, kādi valodas elementi liecina par kreativitāti, un tos aplūko savāktajos valodas apguvēju tekstos (Surkova 2008, 144), un pēc šī parauga savā pētījumā vadījusies arī S. Živjuka (Živjuka 2008, 22). Arī šeit nav skaidri norādīts, ka teksti ir tikuši anotēti, tāpēc iespējams arī, ka materiāls pētīts bez anotēšanas.

Savukārt Inesa Šeškauskiene pētījumā par piesardzīgu formulējumu un piesardzīga formulējuma līdzekļiem (angļu val. *hedging, hedges*) studentu akadēmiskajos darbos tieši nenorāda, vai korpusu ir anotējusi ar tagu palīdzību, taču min, ka ir „atzīmējusi" attiecīgos valodas elementus, lai tos pēc tam varētu saskaitīt (Šeškauskienė 2008, 73). Var pieņemt, ka arī šeit ir runa par korpusa anotēšanu pēc individuāli noteiktas pazīmes konkrētam pētījuma mērķim.

2.2.3.1.4. Pamatformu anotēšana

Pamatformu anotēšana jeb lemmatizēšana (lietuviešu val. *lemavimas,* angļu val. *lemmatization*) ir morfoloģiskā anotējuma sastāvdaļa. Šeit tā gan tiek izdalīta atsevišķi, jo korpusā tehniski funkcionē kā atsevišķs anotējuma veids. Līdz šim, cik zināms, tā nav izmantota valodas apguvēju korpusos Lietuvā un Latvijā, taču var būt noderīga apguvēju valodas izpētē un ir sastopama citās valstīs tapušos valodu apguvēju korpusos (piem., De Mönnink 1999; De Haan 1998).

Pamatforma ir „sintaktiski neatkarīga forma, kuras celms parasti ir formveidošanas celms citām paradigmas formām, piem., lietvārda pamatforma ir nominatīvs, darbības vārda pamatforma – nenoteiksme" (VPSV 2007, 276). Kā jau redzams pēc nosaukuma, pamatformu anotēšanas mērķis ir katram vārdlietojumam piešķirt atbilstošu pamatformu jeb *vārdnīcas formu* (McEnery, Hardie 2012, 245). Šeit izmantotajā piemērā tas shematiski varētu izskatīties tā:

Viņa<viņa> *ir*<būt> *skaista*<skaists>.

Ji<ji> *yra*<būti> *graži*<gražus>.

Šāda anotējuma galvenais mērķis ir ļaut pētniekam atrast visus noteikta vārda dažādu formu lietojumus. Tās var atrast arī citādi – izmantojot aizstājējzīmes (angļu val. *wildcards*), piem., ja aizstājējzīme * aizstāj jebkādu skaitu zīmju, tad formas *gražus, graži, gražaus* utt. var atrast, meklējot pēc *graž**. Taču ar šādu vaicājumu var tikt atrastas arī vārdformas, kas nepieder tam pašam vārdam – šajā gadījumā, piem., lie. *gražuolis*. Turklāt, šādi meklējot, var neatrast kļūdainas formas, kādas ir īpaši bieži sastopamas iesācēju valodas apguvēju tekstos. Pretstatā literārajām formām, kļūdainās formas ir grūti paredzamas un tāpēc apgrūtina meklēšanu ar

aizstājējzīmēm. Tāpēc īpaši iesācēju valodas apguvēju korpusiem, tādi kā „Esam", pamatformu anotēšana ir nepieciešama. Turklāt, anotējot vārdšķiras, pamatforma tiek noteikta jebkurā gadījumā, tātad vienīgā papildu darbība ir noteiktās pamatformas pierakstīšana.

Otrās baltu valodas apguvēju korpuss nav veidots vienam konkrētam pētījumam, bet gan pēc iespējas pilnīgākai baltu starpvalodas izpētei, tāpēc ir skaidrs, ka šajā gadījumā nebūtu veicama problēmorientēta anotēšana, bet gan pēc iespējas daudzveidīga vispārīga anotēšana. Lai veicinātu sadarbību starp jau esošajiem valodas apguvēju korpusu pētniekiem Lietuvā un Latvijā, veidots anotējums, kas ir līdzīgs un, cik iespējams, salīdzināms ar līdzšinējiem pētījumiem citu valodu apguvēju korpusos. Vadoties pēc šāda principa, nolemts otrās baltu valodas apguvēju korpusu anotēt:

- morfoloģiski – pēc vārdšķirām;
- sintaktiski – pēc teikumu veidiem.

Savukārt morfosintaktiska anotēšana būtu sarežģīti veicama, it īpaši iesācēju tekstos, no kādiem galvenokārt sastāv šis korpuss. Tā kā daļā teikumu trūkst skaidri saprotamas struktūras, teikuma locekļu noteikšana daudzos gadījumos būtu atsevišķi veicams pētījums. Līdz ar to šis anotējuma veids, lai arī varētu sniegt vērtīgu ieskatu, nav izvēlēts otrās baltu valodas apguvēju korpusam.

Runājot par kļūdu anotēšanu, pēc šī principa līdz šim nav anotēti baltu valodu apguvēju korpusi – strādāts ir ar citām valodām –, līdz ar to lietotā kļūdu klasifikācija nav īsti piemērota tiem tekstiem, no kuriem sastāv otrās baltu valodas apguvēju korpuss. Tomēr kļūdu anotēšana kopumā ir ļoti nozīmīga korpusam „Esam" kaut vai tādēļ, ka tajā ir iesācēju producētā valoda, kurā pirmās baltu valodas ietekme ir izteikti jūtama un līdz ar to arī kļūdu apjoms un dažādība ir liela. Lai šajā korpusā anotētu kļūdas, ir sastādīta atbilstoša kļūdu klasifikācija (vairāk par to sk. 2.2.7. apakšnodaļā „Kļūdu anotēšana otrās baltu valodas apguvēju korpusā").

Tas, ka noteikts anotējuma veids Latvijā un Lietuvā vēl nav ieguvis popularitāti, gan nenozīmē, ka šo anotējuma veidu nedrīkstētu vai nevajadzētu izmantot otrās baltu valodas apguvēju korpusā. Gluži otrādi – ja ir kāds mazāk lietots anotējuma veids, kas var ļaut iegūt nozīmīgu informāciju, nav svarīgi, vai tas šeit ir ticis izmantots iepriekš. Šajā gadījumā tāds ir pamatformu anotējums. Pamatformas arī līdz šim valodas apguvēju korpusos ir anotētas – piem., Kembridžas valodas apguvēju korpusā (*Cambridge learner corpus,* Hawkins, Buttery 2010, 8), vācu valodas apguvēju korpusā *Falko* (Reznicek u. c. 2013, 109) u. c. Šis

anotējuma veids kopumā nav jaunums arī Latvijas un Lietuvas korpusa lingvistu vidū – pamatformas ir anotētas vairākos morfoloģiski anotētos latviešu (Levāne 2001) un lietuviešu valodas (Rimkutė 2006) korpusos. Taču abās valstīs tapušajos valodas apguvēju korpusos tas nav novērots. Tas tomēr ir īpaši noderīgs tieši valodas apguvēju iesācēju producētās valodas pētīšanā, jo iesācēju valodā ir daudz ne tikai konvencionālo vārdformu, bet arī nekonvencionālo, resp. kļūdaino formu, kuras pētniekam ne vienmēr ir iespējams paredzēt un atrast.

Anotējot korpusu „Esam", viena no svarīgākajām problēmām ir morfoloģiskā daudznozīmība. Ierasts, ka tā parādās, piem., pamatformu anotēšanas vai morfoloģiskās anotēšanas gadījumā, taču parasti nerada tik daudz grūtību cilvēkam anotētājam kā automātiskajiem anotēšanas rīkiem, kuru spējai atpazīt un pareizi anotēt daudznozīmīgus vārdus un konstrukcijas veltīti dažādi pētījumi (Bārzdiņš, Grūzītis u. c. 2007; Rimkutė 2002; 2006 u. c.). Tā kā korpusā „Esam" ir iekļauti iesācēju līmeņa teksti, tajos ir daudz noviržu no valodas normas, un tas ievērojami paplašina iespējamo variāciju klāstu. Tam ir divējādas sekas:

1. Korpusu nevar anotēt tikai ar automātiskiem anotēšanas rīkiem – pārāk daudz laika un resursu būtu jāpatērē, pārbaudot iznākumu un labojot automātiskā rīka pieļautās kļūdas. Atsevišķos fragmentos automātiskie rīki var noderēt procesa paātrināšanai, tāpēc, ja daļa teksta šķiet tam piemērota, automātiskie rīki ir izmantoti pusautomātiski – t. i., rezultātus pārbaudot un vajadzības gadījumā koriģējot. Kad tas kļūst pārāk laikietilpīgi, anotēšana tiek veikta manuāli.
2. Arī manuāli anotējot, sevišķi tekstos ar lielāku kļūdu skaitu, ne vienmēr ir iespējams noteikt, ko autors ir vēlējies pateikt – nereti ir vairāki ticami varianti. Piem., lie. *Kadangi aš gyvenu Kandavoje*[34] *netrukus, tik dveji metai,..* var interpretēt divējādi: (a) autors noteiktajā pilsētā drīz būs nodzīvojis divus gadus; (b) autors noteiktajā pilsētā nav dzīvojis ilgi, tikai divus gadus. Arī la. *Man arī ir ļoti daudz somu, vācu, angļu mūzikas grupju* var interpretēt dažādi: (a) autors pats ir daudzu grupu dalībnieks (maz ticams); (b) autoram patīk daudzas grupas; (c) autors stāsta, ka pastāv daudzas grupas; (d) autoram ir daudzu grupu ieraksti. Tas nozīmē, ka šādos gadījumos ir jāizlemj, kā attiecīgo teksta daļu anotēt un ko norādīt kā vēlamo labojumu. Lai sniegtu pēc iespējas korektu informāciju par korpusā iekļautajiem tekstiem, var anotēšanas sistēmu veidot tā, lai vienam un tam pašam elementam varētu paralēli pievienot vairāku viena veida

[34] Anonimizējot aizstāts pilsētas nosaukums.

kategoriju apzīmējumus, tomēr tas ir tehniski sarežģītāk un palielina anotēšanā patērējamo laika ilgumu. Korpusā „Esam" katrai kategorijai ir atzīmēts viens, pēc anotētāja domām, ticamākais variants (minētajos piemēros – lie. piemērā izvēlēts (b) variants; la. piemērā – (d) variants). Ja tālāka korpusa datu analīze liecinātu, ka šī izvēle nav bijusi pareiza, to var mainīt.

2.2.4. Pamatformu anotēšana otrās baltu valodas apguvēju korpusā

Kā jau iepriekš norādīts, pamatformu anotēšana valodas apguvēju korpusos ir īpaši noderīga, jo ļauj korpusa izmantotājam atrast dažādas nestandarta formas vienam un tam pašam vārdam, pat ja tās ir grūti paredzamas un uzminamas, piem.:

- la. *skaņoja* (acīmredzot domāts – *skanēja*);
- la. *galvapilsētā* (*galvaspilsētā*);
- la. *skatījam* (*skatāmies*);
- lie. *kikvieną* (*kiekvieną*) 'ikvienu, katru';
- lie. *kurious* (*kuriuos*) 'kurus';
- lie. *dalikų* (*dalykų*) 'lietu' u. c.

Pamatformu anotēšana sākotnēji nešķiet sevišķi sarežģīta, taču, kā atzīst Dž. Sinklērs, „tā nav vienkārša darbība; patiesībā tā ir procedūra, kas datoram sagādā lielas grūtības" (Sinclair 2004, 17), it īpaši gadījumos, kad vienai formai ir iespējamas vairākas pamatformas. Latviešu un lietuviešu valodā šādas daudznozīmīgas formas ir plaši sastopamas (Rimkutė u. c. 2009, 63). Turklāt tieši valodas apguvēju korpusos ir izaicinājumi, kuru iemesls ir bieži sastopamās kļūdas, tāpēc, lai arī gan latviešu valodai, gan lietuviešu valodai ir izveidoti automātiski rīki (Grūzītis 2012; Zinkevičius 2000), šajā gadījumā tajos var balstīties tikai daļēji un anotēšana jāveic pusautomātiski vai manuāli, jo, tāpat kā citos anotēšanas veidos, automātiskie rīki netiek galā ar plašu nekonvencionālu formu klāstu.

Redzams, ka arī šķietami vienkāršās pamatformu anotēšanas gadījumā nepieciešamas vadlīnijas, saskaņā ar kurām šo anotēšanu veikt. Tās izstrādājot, ņemti vērā divi aspekti, kas pamatformu anotēšanu šajā darbā padara atšķirīgu no pamatformu anotēšanas citos korpusos:

- vadlīnijām jābūt piemērotām baltu valodās rakstītu tekstu anotēšanai;
- vadlīnijām jābūt piemērotām iesācēju apguvēju rakstītu tekstu anotēšanai.

Korpusā „Esam" vienlaikus ir anotētas pamatformas un vārdšķiras (sk. 2.2.5. apakšnodaļā „Morfoloģiskā anotēšana otrās baltu valodas apguvēju korpusā"), jo, nosakot vienu, ir jānosaka arī otra, dažkārt pat skatot sīkāk, nekā nepieciešams anotēšanai.

Piem., latviešu valodas formai *mani* var tikt piešķirta viena no divām pamatformām (*es*, *mans*) atkarībā no konteksta:

- Tas var būt personas vietniekvārds akuzatīvā: *Vai neviens <u>mani</u> nemeklēja*?
- Tas var būt piederības vietniekvārds daudzskaitļa nominatīvā: *Šie ir <u>mani</u> studenti*.

Šajā korpusā netiek anotētas vietniekvārdu grupas, taču, lai noteiktu, kura no pamatformām ir atbilstoša konkrētajā gadījumā, pamatformu anotētājam šis šķīrums ir jāievēro.

Ņemot vērā visu iepriekš minēto, nolemts, anotējot pamatformas, vadīties pēc līdzšinējiem darbiem attiecīgās valodas pamatformu noteikšanā, par pamatu ņemot abu valodu morfoloģiskās anotēšanas rīku (SemTi 2009; KLC$_{\text{-e}}$) sniegtos variantus un nepieciešamības gadījumā (piem., ja tekstā ir kļūdas, kas traucē analīzei) tos labojot ar pārbaudi vārdnīcās (it īpaši Tēzaurs$_{\text{-e}}$; DLKŽ 2011). Tomēr, anotēšanas procesā sastopoties ar sarežģītākiem piemēriem, nepieciešams radīt tiem noteikumus, kas turklāt būtu saskanīgi abās mērķvalodās, lai atvieglotu korpusa lietošanu pat tad, ja līdzšinējā prakse abās valstīs atšķiras. Daži no tādā veidā līdz šim tapušajiem noteikumiem:

- vietniekvārdiem, ieskaitot personu vietniekvārdus un norādāmos vietniekvārdus, par pamatformu tiek uzskatīta vīriešu dzimtes vienskaitļa forma (t. i., par pamatformām tiek uzskatīti vietniekvārdi *viņš, tas,* bet ne *viņi, viņas, viņa, tie, tās, tā; jis, tas,* bet ne *ji, jie, jos, ta, tie, tos*);
- piederības vietniekvārdi la. *viņa, viņas, mūsu, jūsu, viņu,* lie. *jo, jos, mūsų, jūsų, jų* tiek uzskatīti par personas vietniekvārdu ģenitīva formām;
- dažādām divdabju formām ir viena pamatforma – attiecīgā darbības vārda nenoteiksme. Robežgadījumos (piem., *protams*; *mylimasis* ‘mīļotais’) pamatforma tiek noteikta, vadoties pēc vārdnīcās norādītā. Ja konkrētais vārds vārdnīcās tiek uzskatīts par citai vārdšķirai piederīgu – lietvārdu vai īpašības vārdu, tam arī korpusā tiek piešķirta atsevišķa pamatforma;
- deminutīvi tiek uzskatīti par atsevišķiem vārdiem, nevis tā paša vārda atsevišķām formām (tātad pamatforma ir deminutīvs);
- īpašības vārdu un apstākļa vārdu salīdzināmo pakāpju gadījumā par pamatformu tiek uzskatīta pamata pakāpes forma. Robežgadījumos (piem., *agrāk*) pamatforma tiek noteikta, vadoties pēc vārdnīcās norādītā;

- ōpašības vārdiem, vietniekvārdiem (izņemot personu un norādāmos vietniekvārdus) un skaitļa vārdiem pamatforma ir vīriešu dzimtes forma (piem., *mans,* nevis *mans, mana*).

Īpaši valodas apguvēju tekstu īpatnību dēļ tapuši šādi noteikumi:

- gadījumā, ja vārds ir citā, nevis mērķvalodā (piem., *zimols* tekstā, kura mērķvaloda ir lietuviešu), pamatformas netiek anotētas, taču kā pamatforma tiek norādīts kods CTV (abreviatūra veidota no izceltajiem burtiem vārdu savienojumā *cita valoda*);
- pamatformas netiek anotētas bezmorfoloģijas elementos (sk. tālāk par morfoloģiskās anotēšanas kategorijām);
- pieturzīmes tiek attēlotas kā atsevišķas teksta vienības, taču, tā kā tās ir rakstzīmes, tām pamatformas nav;
- pamatformas netiek anotētas arī anonimizētajās (t. i., mainītajās vai izlaistajās) teksta daļās;
- personvārdiem pamatforma netiek norādīta, jo bieži nav zināms, kāda ir bijusi oriģinālforma, no kuras varētu secināt, kā personvārds būtu jāatveido mērķvalodā. Personvārdiem kā pamatforma tiek norādīts kods PSV (abreviatūra veidota no izceltajiem burtiem vārdu savienojumā personvārds);
- gadījumā, ja, balstoties citas valodas un/vai mērķvalodas leksikā, ir izveidots vārds, kāda mērķvalodā nav, pēc tā gramatiskās uzbūves un lietojuma tekstā tiek piemeklēta pēc iespējas atbilstošāka pamatforma (piem., *rankdarbininkė* ‘rokdarbniece’ lietuviešu valodas tekstā, atvasināts no *rankdarbiai* ‘rokdarbi’ pēc latviešu valodas vārda *rokdarbnieks* parauga, pamatforma – *rankdarbininkė*; *nepapasakįs* < *ne* + *papasakoti* ‘pastāstīt’, *pasakyti* ‘pateikt’, pamatforma – *nepapasakyti*; *vazinātes* < *vizināties* + *važinėti* ‘braukāties’, pamatforma – *vazināties*);
- gadījumā, ja tekstā lietotā vārdforma būtiski atšķiras no tā, kas, spriežot pēc konteksta, bijis domāts, pamatforma ir tas vārds, kas, spriežot pēc konteksta un izvērtējot kļūdu iespējamību, bijis iecerēts (piem., *ura* < *yra* ‘ir’, pamatforma – *būti* ‘būt’; *paslepintu* < *palepinu* ‘palutinu’, pamatforma – *palepinti* ‘palutināt’; *mirtis* ‘nāve’ < *mintis* ‘doma’, pamatforma – *mintis* ‘doma’). Ja tekstā lietotā vārdforma līdzinās vairāku vārdu formām un nav nosakāms, kuru no tām autors sākotnēji ir iecerējis, par primāro kritēriju pamatformas noteikšanā tiek uzskatīta

forma – kā pamatforma tiek norādīts tas vārds, kuram vairāk līdzinās lietotā forma. Ja tekstā nav pareizi saskaņota dzimte un mērķvalodā pastāv arī otras dzimtes vārds, tad, anotējot pamatformas, dzimte netiek labota – piem., izteikumā *viņas ir uzņēmējs* pirmā vārda pamatforma ir *viņa*, lai arī saskaņā ar latviešu valodas normām šeit būtu lietojams vīriešu dzimtes vietniekvārds.

Šie noteikumi palīdz orientēties anotēšanas procesā, kā arī pēc tam, meklējot korpusā vēlamā vārda dažādas vārdformas. Tomēr nepieciešams anotēt piesardzīgi, paturot prātā variāciju iespējamību. Īpaši sarežģīti ir gadījumi ar līdzīgi rakstāmiem saikļiem, apstākļa vārdiem un prievārdiem. Piem., šādā teikumā:

lie. *Kad esu kartu su drauges,..* 'Kad esmu kopā ar draudzenēm,..'

Lietuviski būtu lietojams *kai* vai vismaz *kada*, taču lietuviešu valodā ir arī vārds *kad* 'ka'. No vienas puses, lietots nepareizais vārds, no otras puses – dažkārt iespējama arī vienkārša pārrakstīšanās – atšķirība starp lietoto un vēlamo ir tikai vienā burtā. Iespējams arī, ka autors ir domājis pareizo vārdu, taču nav zinājis, kā attiecīgo vārdu raksta. Variāciju dažādības dēļ nolemts neskaidros piemēros vadīties pēc oriģināla formas, nevis pēc mērķa hipotēzes (par to vairāk sk. 2.2.7.1. „Kļūdas definīcija"). Līdz ar to minētajā piemērā pirmajam vārdam tiek piešķirta pamatforma *kad* 'ka'.

Līdzīgi problemātisks gadījums sastopams latviešu valodā rakstītā tekstā:

Es esmu līdzīga tēvam, taču nē gara, nē resna.

Šeit ir noprotams, ka autors ir iecerējis lietot saikli *ne*, taču kļūdas dēļ tā vietā ir forma, kas līdzinās partikulai *nē*. Iespējams, kļūda ir radusies tikai aiz nezināšanas, ka saiklī uz patskaņa nav garumzīmes, vismaz lietojums vedina par to domāt. Tā kā konteksts un teikuma struktūra ļauj diezgan droši spriest par kļūdas veidu, šajā gadījumā kā pamatforma norādīts saiklis *ne*.

2.2.5. Morfoloģiskā anotēšana otrās baltu valodas apguvēju korpusā

Kā iepriekš jau norādīts, no morfoloģiskajām pazīmēm korpusā „Esam" anotētas tiek tikai vārdšķiras, proti, katrai vārdformai tiek norādīta atbilstošā vārdšķira, taču ne cita morfoloģiska rakstura informācija. Vārdšķiru anotēšanai otrās baltu valodas apguvēju korpusā tiek izmantota pazīmju kopa, kas balstās projekta *Semti-Kamols* latviešu valodas morfoloģisko pazīmju kopā (LVMPK 2009). Šī pazīmju kopa ir redzama 3. tabulā (sk. 104. lpp.).

3. tabula. Vārdšķiras.

Vārdšķira	**Piezīmes**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Lietvārds		n	mājas	stebuklą ‘brīnumu’
Darbības vārds	Ieskaitot divdabjus	v	gribu	supratau ‘sapratu’
Īpašības vārds		a	brūnas	lėtais ‘lēniem’
Vietniekvārds		p	man	aš ‘es’
Apstākļa vārds		r	ļoti	labai ‘ļoti’
Prievārds		s	pie	į ‘uz’
Saiklis	Ieskaitot saliktos saikļus	c	un	jei ‘ja’
Skaitļa vārds	Ieskaitot daļskaitļus un vairākvārdu skaitļa vārdus; vairākvārdu skaitļa vārda gadījumā katram no vārdiem vārdšķira ir norādīta atsevišķi	m	viens	dvi ‘divas’
Izsauksmes vārds		i	labdien	laba diena ‘labdien’
Partikula		q	nē	ne ‘nē’
Bezmorfoloģijas elements	Simbols vai simbolu virkne, kam nav mērķvalodas morfoloģiskās struktūras: cipari, abreviatūras, saīsinājumi, vārdi citās valodās, formulas u. tml.[35]	x	Varniukų	07:07

Bez jau nosauktajām kategorijām vārdšķiru anotējumā ir sastopams vēl viens kods – z. Tas tiek norādīts īpaši pieturzīmēm, jo pieturzīmes programma nodala kā atsevišķas teksta vienības, un, korpusu lietojot, var būt nepieciešamība atlasīt tieši pieturzīmes.

Minētās desmit vārdšķiras ir nošķiramas gan latviešu, gan lietuviešu valodā. Tā kā lietuviešu un latviešu valodniecībā var būt atšķirīga izpratne par to, vai konkrēts vārds ir piederīgs vienai vai citai vārdšķirai (it sevišķi, runājot par partikulām), anotējot mēģināts vadīties pēc katras valodas šķīruma tradīcijām, nevis tās apvienot.

Tomēr arī vienas valodas tekstos ir diezgan daudz problemātisku gadījumu. Pat nerunājot par korpusu anotēšanu, uz grūtībām noteikt dažādu vārdu vārdšķiras saskaņā ar pašreizējo klasifikāciju norāda dažādi valodnieki (Paegle 2003, 27; LVG 2013, 321–323 u. c.).

[35] Formulējums pielāgots no LVMPK 2009.

Var pieņemt, ka vismaz ar kādu no robežgadījumiem saskaras ikviens korpusa anotētājs, taču, lai būtiski nesamazinātu korpusa anotēšanas efektivitāti, nevar gaidīt, ka anotētājs katru atsevišķu gadījumu, ja to ir daudz, sīki analizēs pēc dažādām valodniecības teorijām, lai pieņemtu vispareizāko lēmumu. Ja korpusa aprakstā ir norādīts, kāda veida gadījumos kāds variants izvēlēts, korpusa lietotājs tos vēlāk atrast var arī tad, ja izvēlei nepiekrīt.

LVASA latviešu valodas apguvēju korpusa veidotājas norāda, ka vārdšķiru anotējumā problēmas var rasties gan valodas sistēmas robežgadījumu dēļ, gan arī apguvēju valodas kļūdu ietekmē. Kā valodas sistēmas izraisītas grūtības tiek minētas šādas:

1) neskaidras apstākļa vārdu un partikulu robežas;
2) neskaidras apstākļa vārdu un lietvārdu locījuma formu robežas;
3) neskaidras divdabju un adjektivējušos divdabju robežas (Kalnbērziņa u. c. 2011, 19).

Šādos gadījumos ir izvēlēts viens variants katram no gadījumiem un tas ievērots visā korpusā.

Savukārt kā piemērs apguvēju valodas kļūdas ietekmētām grūtībām LVASA korpusa anotēšanā tiek minēts lietvārda lietojums teikumā darbības vārda vietā vai otrādi. Šādos gadījumos par primāro ir uzskatīta vārda funkcija teikumā, un līdz ar to katrs šāds piemērs anotēts saskaņā ar tā funkciju, nevis formu (Kalnbērziņa u. c. 2011, 19).

Korpusa „Esam" morfoloģiskais anotējums veidots pusautomātiski, izmantojot divus rīkus: Latvijas Universitātes Matemātikas un informātikas institūtā tapušo tekstu korpusu marķēšanas rīku (SemTi 2009) un Vītauta Dižā universitātes Datorlingvistikas centrā izveidoto morfoloģisko anotētāju (KLC_{-e}). Tā kā šie rīki nav paredzēti valodas apguvēju korpusu anotēšanai, tie pieļauj ievērojamu daudzumu kļūdu nekonvencionālu formu vai konstrukciju gadījumos, taču vienlaikus ievērojami atvieglo un paātrina darbu ar tām formām, kuras atbilst mērķvalodas likumībām. Tas ļauj arī samazināt grūtības, ar kurām jāsaskaras mērķvalodas sistēmas robežgadījumu dēļ – tā kā rīku sastādītāji ar tām ir jau saskārušies un rēķinājušies, korpusā „Esam" tiek izvēlēts risinājums, ko piedāvā rīks, un netiek veidotas jaunas papildu noteikumu sistēmas. Gadījumos, ja rodas šaubas, tiek izmantotas dažādas vārdnīcas un citi uzziņas avoti, lai izvēlētos piemērotāko variantu.

Tomēr, kā jau minēts, valodas apguvēju korpusus anotējot, jārēķinās arī ar novirzēm no mērķvalodas normas, tāpēc rīku sniegtais rezultāts tiek pārbaudīts un vajadzības gadījumā koriģēts. Lai noteiktu vārdšķiru, tiek izmantota pamatforma, kas tiek izvēlēta iepriekš aprakstītajā kārtībā. Kā norādīts 2.2.4. apakšnodaļā „Pamatformu anotēšana otrās baltu valodas apguvēju korpusā", pamatformu anotēšana un morfoloģiskā anotēšana tiek veikta vienlaicīgi,

tādēļ visi iepriekš nosauktie noteikumi, kas attiecas uz pamatformu piešķiršanu noteiktiem vārdiem un šķīrumu starp līdzīgiem dažādu vārdšķiru vārdiem, attiecas arī uz vārdšķiru anotēšanu.

2.2.6. Sintaktiskā anotēšana otrās baltu valodas apguvēju korpusā

Korpusu anotējot sintaktiski, tajā iekļautie teksti ir sadalīti izteikumos un izsacījumos, un izteikumi ir iedalīti vairākās pamatgrupās pēc tā, kādam teikuma modelim tie atbilst. Izteikumos teikuma modelis var būt pilnībā vai daļēji realizēts – ir iespējama redukcija vai parcelācija. Ja izteikumā ir novērojama kādas gramatiskā centra daļas vai cita teikuma locekļa redukcija, tad šis izteikums tik un tā tiek klasificēts atbilstoši teikuma modelim. Parcelācijas gadījumā viena teikuma struktūra tiek dalīta vairākos izteikumos (VPSV 2007, 282), līdz ar to acīmredzamas parcelācijas gadījumā parcelētie izteikumi tiek anotēti kā viens teikums, nenodalot šos izteikumus atsevišķi. Gadījumos, kad rodas šaubas, vai izteikums ir parcelāts vai tajā ir daļēji realizēts atsevišķa teikuma modelis, tas tiek uzskatīts par atsevišķa teikuma modeļa daļēju realizāciju un anotēts atbilstoši šai izpratnei.

Pazīmju kopa šim anotējumam ir redzama 4. tabulā (sk. 107. lpp.), un tālāk ir aprakstīti dalījuma principi.

Sintaktiskajā anotējumā izmantota šāda klasifikācija:

VNT – vienkāršs nepaplašināts teikums. Šajā grupā ir izteikumi, kas atbilst vienkārša teikuma minimālajam teikuma modelim (par to sk. LVG 2013, 710) – t. sk. arī tad, ja daļa no gramatiskā centra ir reducēta.

VPT – vienkāršs paplašināts teikums. Šajā grupā ir izteikumi, kas atbilst vienkārša teikuma paplašinātajam teikuma modelim (par to sk. LVG 2013, 711) – t. sk. arī tad, ja daļa no gramatiskā centra vai viss gramatiskais centrs ir reducēts.

SPT – salikts pakārtots teikums. Šajā grupā ietilpst izteikumi, kuros ir vairākas daļas, starp kurām ir pakārtojuma attiecības – t. sk. arī tad, ja vienā vai vairākās daļās gramatiskais centrs vai daļa no tā ir reducēta.

SST – salikts sakārtots teikums. Šajā grupā ietilpst izteikumi, kuros ir vairākas daļas, starp kurām ir sakārtojuma attiecības – t. sk. arī tad, ja vienā vai vairākās daļās gramatiskais centrs vai daļa no tā ir reducēta.

SJT – jaukts salikts teikums. Šajā grupā ir tādi izteikumi, kuros ir vairākas daļas, starp kurām ir gan pakārtojuma, gan sakārtojuma attiecības – t. sk. arī tad, ja vienā vai vairākās daļās gramatiskais centrs vai daļa no tā ir reducēta.

IS – izsacījums. Gadījumā, ja teksta vienībā nav saskatāmas teikuma modeļa pazīmes, tā tiek ieskaitīta šajā grupā.

NT – neskaidri veidots izteikums. Šajā grupā tiek ieskaitītas tās teksta vienības, kuras autors acīmredzami ir iecerējis veidot kā izteikumus, taču izteikta neatbilstība mērķvalodas normām neļauj noteikt teikuma veidu.

4. tabula. Teikuma veidi.

Teikuma veids		**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Vienkāršs teikums	Vienkāršs nepaplašināts teikums	VNT	*Es esmu studente.*	*Niekas neįvyko.* ‘Nekas nenotika.’
	Vienkāršs paplašināts teikums	VPT	*Es gribu pastāstīt par mana ģimeni un mājas.*	*Man labai patinka saldainiai.* ‘Man ļoti patīk konfektes.’
Salikts teikums	Salikts sakārtots teikums	SST	*Viņas ir uzņemējs un viņam ir uzņēmums.*	*Pusiaukelėje man pasidariau sunkiai ir aš jaučiau labai pavargęs.* ‘Pusceļā man kļuva grūti, un es jutos ļoti noguris.’
	Salikts pakārtots teikums	SPT	*Draugi saka, ka es esu smieklīga, draudzīga.*	*Kartais dėviu pedkelnes po kelnėmis, kada yra labai šalta.* ‘Reizēm valkāju zeķbikses zem biksēm, kad ir ļoti auksti.’
	Jaukts salikts teikums	SJT	*Fotoaparātā bija manas skaistas fotogrāfijas, man ļoti patik fotografēt, īpaši kad esmu jūrmalā.*	*Man nepatinka kepurė, bet kada yra labai šalta aš nešioju kepures, irgi pirštines.* ‘Man nepatīk cepure, bet, kad ir ļoti auksti, es nēsāju cepures, arī cimdus.’
Izsacījums		IS	*Labdien!*	*Ačiū!* ‘Paldies!’
Neskaidri veidots teikums		NT	Latviešu valodā rakstītajos tekstos nav konstatēts.	*Po naktinis mainos džiaugiuosi kad pirkti naujus bandelu pusryčiams.* ‘Pēc nakts maiņas priecājos, ka (?) pirkt jaunas smalkmaizītes brokastīm’

Par problēmgadījumiem, sintaktiski anotējot latviešu valodas apguvēju korpusu, runā jau LVASA pētījuma (Kalnbērziņa u. c. 2011) autores. Norādītas šādas grūtības teikumu veidu šķirumā:

1) grūtības noteikt teikuma veidu, ja ir bezsaikļa saistījums starp teikuma daļām. Autores nolēmušas šaubu gadījumā šādus teikumus uzskatīt par saliktiem sakārtotiem teikumiem: *Arī Viktora izteiktajā domā ir daļa patiesības – no skolotāja ir daudz kas atkarīgs*[36];
2) grūtības noteikt, vai teikumā ir vienlīdzīgi virsteikumi vai vienlīdzīgi palīgteikumi. Šādā gadījumā jāņem vērā teikuma semantika. *Pilnīgi piekrītu arī Dmitrija teiktajam, ka vētras „epicentram" ir vajadzīga vide, kurā varētu sākt darboties, ja tādas vides nav, tad arī nav kur uzsākt vētru*. Šis piemērs anotēts kā salikts pakārtots, nevis jaukts salikts teikums, jo viss, par ko tiek runāts, ir Dmitrija teiktā atstāsts;
3) gadījumi, kuros viena vai cita vietniekvārda lietojums mainītu teikuma struktūru. Pētnieces sniedz šādu piemēru: *Laikam, katrā klasē ir cilvēki, kuri sava rakstura dēļ strīdās ar skolotājiem un viņiem tā ir īkdieņišķa attiecību forma*. Teikums ir anotēts kā jaukts salikts, taču ar norādi, ka, ja viņiem vietā būtu kuriem, tas būtu uzskatāms par saliktu pakārtotu teikumu;
4) teikumi, kuru jēga nav skaidra. Pētnieces izvēlējušās anotēt tos teikumus, kuru struktūra tomēr ir saprotama, bet gadījumiem, kuros nav saprotama ne teikumu jēga, ne struktūra, izmantot īpašu tagu <?> (Kalnbērziņa u. c. 2011, 18).

Attiecīgajos problēmgadījumos tāda pati pieeja izvēlēta arī korpusā „Esam"; 3. punktā minētajos gadījumos gan nekādas papildu norādes par hipotētisku vietniekvārdu maiņu nav – šādos gadījumos tiek anotēta teikuma struktūra, kāda tā ir saskatāma, pat ja šķiet, ka labāk iederētos cits vietniekvārds, kurš attiecīgi mainītu teikuma struktūru. Savukārt 4. punktā norādītajos gadījumos, kā jau iepriekš norādīts, netiek izmantots tags <?>, bet gan šādiem potenciāliem izteikumiem ir paredzēta sava grupa klasifikācijā – NT.

Vēl viens problemātisks aspekts, nosakot teikumu veidus, ir tiešā runa. Kā norāda Maigone Beitiņa, tiešā runa „īsteni ir teksta sintakses daļa, ne teikuma sintakses daļa, jo vēro vairāku teikumu saistījumus" (Beitiņa 2009, 209). Tomēr, analizējot korpusā iekļautos tekstus teikuma līmenī, ir jābūt skaidrībai, kā rīkoties gadījumā, ja tekstā parādās tiešā runa.

[36] Šis un divi turpmākie piemēri sniegti LVASA pētījumā, nevis iegūti no korpusa „Esam".

M. Beitiņa, zīmējot teikuma shēmu, tiešo runu apzīmē kā neatkarīgu teikuma daļu, uzsverot, ka „jāparāda gan piebildes, gan tiešās runas struktūra" (Beitiņa 2009, 205). Šāda pieeja var radīt grūtības, ja tiešā runa ir ļoti apjomīga, taču citādi šķiet atbilstoša.

Tā kā otrās baltu valodas apguvēju korpusā nav novēroti piemēri, kuros tiešā runa būtu garāka par 2 teikumiem, tādos gadījumos, kad tekstā ir tiešā runa, tā tiek pēc M. Beitiņas sniegtā parauga uzskatīta par tā paša izteikuma neatkarīgu(-ām) daļu(-ām).

Korpusā iekļauto tekstu sintaktiska anotēšana veikta manuāli.

2.2.7. Kļūdu anotēšana otrās baltu valodas apguvēju korpusā

Runājot par valodas apguvēju producēto valodu, bieži tiek pieminētas kļūdas, un to anotēšana ir īpašs izaicinājums valodas apguvēju korpusu izveidē ne vien tāpēc, ka kļūdas mēdz būt dažādi interpretējamas, bet arī tāpēc, ka pētnieku vidū nav vienprātības par jēdziena *kļūda* saturu un piemērotākajiem kļūdu klasifikācijas veidiem.

2.2.7.1. Kļūdas definīcija

Lai arī *kļūda* ir nozīmīgs jēdziens valodu apguvē, tā bieži vien nav definēta (VPSV 2007; LTSV 2011; Glück 2005, 189, 358) vai ir definēts tikai cits, radniecīgs termins (Crystal 1992, 125; Matthews 1997, 117; Bussmann 1996, 153; Ахманова 1969, 305; Ulrich 2002, 87). Kļūda parasti tiek izprasta kā „novirzīšanās no normas, pildot kādu darbību, uzdevumu u. tml." (PTSV 2000, 81; sal. arī Freimane 1993, 180; LLVV 1980, 276; LLV 2006, 513). *Lietuviešu valodas lingvodidaktikas terminu vārdnīcā* termins *klaida* 'kļūda' ir definēts kā „tāds valodas vienību (vārdu, gramatisko formu, runas aktu utt.) lietojums, kuru dzimtās valodas runātājs vai runātājs ar labām valodas prasmēm uzskata par kļūdainu; valodas un komunikācijas normu pārkāpums" (LTŽ 2012, 92). Angļu valodā ir vairāki vārdi, ko varētu uzskatīt par latviešu valodas vārda *kļūda* atbilsmēm: *slips, mistakes, errors, solecisms.* To šķīrumam pamatā ir kļūdas rašanās iemesli. Valodnieki mēdz īpaši uzsvērt atšķirību starp terminiem *error* un *mistake,* taču ne vienmēr ir vienisprātis par to, kādi principi ir šī šķīruma pamatā. Viens no uzskatiem ir, ka *error* ir pastāvīgas kļūdas jeb *persistent mistakes*, kas atklāj tām pamatā esošos modeļus (Field 2011, 66) – tātad, saskaņā ar šo uzskatu, *mistake* ir plašāks jēdziens. Citur lingvistiskajā literatūrā abi termini minēti kā viena līmeņa dažādas kļūdu grupas: *mistake* ir snieguma kļūdas, savukārt *error* – kļūdas, kuru cēlonis meklējams apguvēja zināšanās par mērķvalodu (Cherrington 2004, 198).

Vairākās citās valodās savukārt konsekventi tiek lietots virstermins, neiesaistot uzreiz sīkāku dalījumu: ne latviešu valodas terminam *kļūda*, ne lietuviešu valodas terminam *klaida,*

ne vācu valodas terminam *Fehler*, ne krievu valodas terminam *ошибка* nav noteikti jāsakrīt ar angļu valodas sazaroto terminu sistēmu. Angļu val. *error*, vācu val. *Fehler*, krievu val. *ошибка* definēti līdzīgi: kā „novirze[37] spontānā runā vai rakstos" (Crystal 2008, 173), „novirze no mērķvalodas normas" (Bußmann 2002, 214), „atkāpe no pareiza valodas vienību un formu lietojuma" (Щукин 2007, 198).

Jāpiebilst gan, ka pētnieki brīdina nejaukt termina *kļūda* izpratni pedagoģijā un valodniecībā (Crystal 2008, 173), atgādinot, ka pedagoģijā kļūda var būt ne vien neatbilstība valodas normai, bet arī uzdevuma prasībām.

Balstoties iepriekšminētajās definīcijās, korpusa „Esam" anotēšanas nolūkā kļūda tiek saprasta kā neatbilstība priekšstatam par to, kā attiecīgajai valodas struktūrai pareizi jābūt veidotai. Šo priekšstatu vācu valodnieki piedāvā saukt par mērķa hipotēzi (vācu val. *Zielhypothese,* angļu val. *target hypothesis*), definējot to kā, R. Elisa vārdiem runājot, „valodas apguvēju izteikumu rekonstrukciju mērķvalodā" (Ellis 1994, 54; sk. arī Lüdeling u. c. 2005; Siemen u. c. 2006). Tātad šajā darbā kļūda ir neatbilstība mērķa hipotēzei, kādu izvirza teksta labotājs.

2.2.7.2. Kļūdu pazīmju kopas izstrādes process

Lai korpusā anotētu kļūdas, nepieciešams izstrādāt klasifikāciju, kurā balstoties, kļūdas tiks anotētas. S. Greindžere nosauc četras īpašības, kurām būtu jāpiemīt kļūdu anotēšanas sistēmai, lai tā būtu efektīva:

1) detalizēta. Anotēšanas sistēmai jāatrodas līdzsvarā starp detalizāciju un izmantojumu: lai no anotēšanas gūtu maksimāli daudz labuma, kategorijām vēlams būt pēc iespējas detalizētākām, ļaujot atklāt īpatnības dažādos kļūdu tipos un apakštipos. No otras puses, pārmērīga detalizācija ievērojami apgrūtina anotēšanas procesu, jo sarežģītas sistēmas ieviešana prasa daudz vairāk uzmanības no anotētāja;
2) atkārtoti izmantojama. Optimāla anotēšanas sistēma būtu lietojama ne tikai tās valodas datiem, kurai tā ir radīta, bet arī (iespējams, ar pielāgojumiem) citu valodu apguvēju korpusos;
3) elastīga. Anotēšanas laikā anotētājam būtu jābūt iespējai pievienot vai dzēst kādas kategorijas, ja tas izrādās vajadzīgs;

[37] Šeit – *mistake*.

4) konsekventa. Lai nerastos situācija, kurā līdzīgas kļūdas ir anotētas atšķirīgi, anotēšanas sistēmai jābūt izstrādātām skaidrām vadlīnijām, pēc kurām vadās visi attiecīgā korpusa anotētāji (Granger 2003a, 467).

Jāņem vērā, ka kļūdu anotēšana vienmēr ir noteiktā mērā subjektīva. Bieži vien pastāv vairākas iespējamas mērķa hipotēzes, un katrs labojums jāuzskata nevis par vienīgo pareizo, bet gan par iespējamu pareizās formas variantu (Dagneaux u. c. 1998, 166). Pastāv pat uzskats, ka nereti anotētāji *pārcenšas*, kā kļūdu vai dzimtās valodas lietojumam neatbilstošu tendenci atzīmējot kādu objektīvi kultūrspecifisku faktoru dēļ radušos atšķirīgu lietojuma tendenci (Tan 2005, 127), tāpēc anotēšana ir jāplāno tā, lai šādu gadījumu skaits būtu pēc iespējas mazāks. Luvēnas Katoļu universitātē ir izstrādāta kļūdu anotēšanas metodika, kas sastāv no trim soļiem:

1) korpusā iekļauto tekstu manuāla labošana, pierakstot, kā attiecīgajam izteikumam būtu jābūt veidotam pareizi;
2) pazīmju sistēmas izstrāde attiecīgajai valodai;
3) kļūdām atbilstošo tagu un labojumu ievietošana teksta datnēs (Dagneaux u. c. 1998, 167; Granger 2003a, 466–467).

Pētnieki norāda, ka ieteicams darbu veikt pārī, cieši sadarbojoties: labošanu vēlams uzticēt cilvēkam, kuram teksta mērķvaloda ir dzimtā valoda, savukārt kļūdu tipus noteikt un anotēt ieteicams cilvēkam, kuram teksta mērķvaloda nav dzimtā valoda, taču ir ļoti labas šīs valodas prasmes, ieskaitot gramatikas zināšanas, un ideālā gadījumā dzimtā valoda sakrīt ar teksta autora dzimto valodu (Dagneaux u. c. 1998, 167). Šāds darbu sadalījums palīdz sasniegt ticamāku rezultātu: tā kā kļūdas tiek noteiktas attiecībā pret mērķvalodas sistēmu, dzimtās valodas runātājs visdrīzāk izveidos tādas mērķa hipotēzes, kas mērķvalodā skan dabiski. Savukārt anotētājam, kuram mērķvaloda nav dzimtā valoda un kurš pats mērķvalodu ir mācījies, ir labāk pazīstamas grūtības, ar kādām apguvējs saskaras, un vienlaikus ir vieglāk izprast kļūdu rašanās mehānismu, līdz ar to – arī katras kļūdas veidu atbilstoši izvēlētajai klasifikācijai.

Vadoties pēc šī principa, tika anotēts arī otrās baltu valodas apguvēju korpuss „Esam". Promocijas darba autores dzimtā valoda ir latviešu, un lietuviešu valodas prasmes ir nepieciešamajā līmenī, savukārt darbam pārī piekrita D. Puškorjute-Riduliene, kuras dzimtā valoda ir lietuviešu valoda un kura ir latviešu valodas kursa docētāja Vītauta Dižā universitātē Kauņā, Lietuvā. Tā kā šim nolūkam piemēroti tekstu analīzes rīki baltu valodām vēl nav izveidoti, teksti tika anotēti manuāli. Darba gaita aprakstīta 5. tabulā „Kļūdu anotēšana korpusā „Esam"" (sk. 112. lpp.).

5. tabula. Kļūdu anotēšana korpusā „Esam”

Darba uzdevumi	Uzdevumu sadalījums	
	Inga Znotiņa	**Daiva Puškorjute-Riduliene**
tekstu labošana	teksti latviešu valodā	teksti lietuviešu valodā
kļūdu klasifikācijas izveide	klasifikācijas izveide visam korpusam	klasifikācijas pārskatīšana un ieteikumi
laboto tekstu anotēšana	teksti lietuviešu valodā	teksti latviešu valodā

Kā redzams 5. tabulā, anotējot korpusu „Esam”, tāpat kā Luvēnas pētnieku darba gaitā, nav atsevišķas stadijas, kurā anotētais materiāls tiktu atsevišķi pārbaudīts. Šāds lēmums pieņemts tāpēc, ka, abām anotētāju pārī iesaistītajām personām cieši sadarbojoties, visi gadījumi, kuros rodas neskaidrības, tiek savstarpēji pārrunāti un papildu pārbaude prasītu daudz laika, dodot relatīvi mazu pienesumu. Korpusa turpmākās pilnveides gaitā varētu tikt piesaistīti vairāki savstarpēji neatkarīgi anotētāju pāri, lai atklātu iespējamās nesakritības anotējumā un gūtu iespēju to pilnveidot. Tas palīdzētu arī samazināt subjektivitātes ietekmi: piesaistot vairākus anotētājus un nosakot, cik lielā mērā sakrīt viņu viedokļi (angļu val. *inter-annotator agreement*; sk., piem., Bermingham, Smeaton 2009), tiktu palielināta kļūdu anotējuma ticamība.

2.2.7.3. Kļūdu tipi pazīmju kopā

Nākamais solis darba gaitā ir pazīmju kopas izstrāde. Kļūdas, kas rodas, apgūstot valodu, mēdz klasificēt dažādi. Lai kļūdas anotētu sistemātiski, ir nepieciešams sākumā nolemt, pēc kāda kļūdu klasifikācijas principa attiecīgo klasifikāciju veidot. Piem., lietuviešu valodnieki nošķir trīs kļūdu grupas: kompetences kļūdas (lietuviešu val. *kompetencijos klaida*), snieguma kļūdas (lietuviešu val. *atlikties klaida*) un nejaušības kļūdas (lietuviešu val. *riktas*) (LTŽ 2012, 92). Kompetences kļūdas rodas valodas lietotāja kompetences trūkuma dēļ (LTŽ 2012, 102), snieguma kļūdas ir tās kļūdas, kuras „rodas, valodas lietotājam pienācīgi neizmantojot savu esošo kompetenci” (LTŽ 2012, 26), savukārt nejaušības kļūdas vispār nav saistītas ar valodas kompetenci, bet gan rodas fizisku un psiholoģisku faktoru, piem., noguruma, stresa vai neuzmanības dēļ (LTŽ 2012, 174). Jāatzīst gan, ka dalījums nav lietderīgs korpusa anotēšanai, jo ir nepieciešamas paplašinātas zināšanas par kļūdas cēloņiem un/vai dziļa katras kļūdas analīze, lai noteiktu, pie kāda kļūdu tipa attiecīgā kļūda iederētos. Tas drīzāk būtu

noderīgs tālākajā darbā, kad kļūdu anotējums korpusā tiek izmantots pētījumos, piem., izmantojot kļūdu analīzes metodi.

Līdz šim tapušajos valodas apguvēju korpusos konstatējama:

- kļūdu klasifikācija pēc lingvistiskās kategorijas, kādai kļūda ir piederīga;
- kļūdu klasifikācija pēc izmaiņām attiecībā pret mērķvalodas sistēmu vai, precīzāk, mērķa hipotēzi;
- kļūdu klasifikācija, kurā apvienoti abi šie principi.

Iedalījumā pēc lingvistiskās kategorijas, kādai kļūda ir piederīga, kļūdu klasifikācijas detalizācijas pakāpe var būt dažāda. Dažkārt pētnieki tās iedala tikai pēc valodas līmeņa, kuram noteiktā novirze piemīt, piem., fonētikas, morfoloģijas, sintakses u. c. (Rascon 2013). Pēc vajadzības kategorijas var arī sašaurināt, runājot, piem., par kļūdu nomenu dzimtes lietojumā (Laizāne 2014b).

Kļūdas iedalot pēc izmaiņām attiecībā pret mērķa hipotēzi, viens no klasifikācijas veidiem nosaka trīs kļūdu grupas: izlaidums (angļu val. *omission*), lieks elements (angļu val. *addition*) un nepareiza forma (angļu val. *misformation*); kā nepareiza forma tiek interpretēta arī, piem., nepareizās leksēmas izvēle (šādu klasifikāciju izmanto, piem., Tono, Satake, Miura 2014, 152). Citi pētnieki šīm kļūdu grupām pievieno vēl arī nepareizu secību (angļu val. *misordering*), kā arī hibridizācijas kļūdas (angļu val. *blend*; *hybridization error*), kuras rodas, ja valodas apguvējs, svārstīdamies starp divām vai vairākām pareizām mērķa hipotēzēm, izveido no tām kļūdainu hibrīdformu.

Daļa pētnieku iesaka abas šīs klasifikācijas nevis izmantot atsevišķi, bet gan apvienot (James 1998, 114). To ir mēģinājusi darīt, piem., S. Greindžere (2003a). Šķiet, ka klasifikācija ir izdevusies atbilstoša valodas apguvēju korpusa vajadzībām, ļaujot nodalīt izteikti bieži sastopamus kļūdu veidus valodas apguvē, tātad šajā gadījumā par veiksmīgu ir uzskatāms lēmums nevis abus klasifikācijas principus ievērot visās grupās vienādā apmērā, bet gan piemērot tos atkarībā no lietderības kļūdu analīzē.

Viens no galvenajiem faktoriem, kas jāņem vērā, izstrādājot kļūdu anotēšanas sistēmu, ir tas, ka kļūdu klasifikācija ir valodspecifiska. Tā kā katras valodas sistēma noteiktā mērā atšķiras, arī attiecīgās valodas apguvēju tipiskākās kļūdas var būt atšķirīgas. Piem., latviešu literārajā valodā debitīvu lieto ar papildinātāju nominatīvā (*jāsaka vārds*), nevis, kā mēdz gadīties sarunvalodā, akuzatīvā (*jāsaka vārdu*) (LVG 2013, 487). Līdz ar to šo noteiktā situācijā varētu uzskatīt par atsevišķu kļūdu kategoriju latviešu valodā (šāda pieeja ir izvēlēta, piem., Deksne, Skadiņa 2014, 164). Taču, salīdzinājumam, tā kā angļu valodā nav tāda

debitīva, kāds ir latviešu valodā, angļu valodas apguvēju darbos piemēri kļūdas kategorijai „nepareizs locījums pēc debitīva" nav atrodami. Dažādām valodām pastāv arī rīki automātiskai kļūdu atpazīšanai tekstos (Fujishima, Ishizaki 2011, Deksne, Skadiņa 2014 u. c.).

Uzreiz gan jāpiebilst, ka tieši angļu valodai kļūdu anotēšanas sistēmas ir krietni plašāk izstrādātas nekā latviešu un lietuviešu valodai. Kļūdu analīze rakstu darbos Latvijā un Lietuvā nereti tiek veikta, neiesaistot korpusa lingvistikas metodes, un ne vien automātiski vai pusautomātiski rīki, bet pat visaptveroša kļūdu klasifikācija līdz šim nav izveidota.

Tā kā korpusa anotēšana nav pašmērķis, bet gan centieni padarīt korpusā iekļautos datus vieglāk pētāmus, izstrādājot klasifikāciju, ņemts vērā līdzšinējais pētniecības konteksts. Ne Latvijā, ne Lietuvā kļūdas nav primārais valodas apguves pētnieku intereses objekts, tomēr daži valodnieki ir pievērsušies arī šim valodas apguves aspektam. I. Laizāne vairākos rakstos ir raksturojusi dažas brīvi izvēlētas latviešu valodas apguvēju kļūdu kategorijas tekstos, kuru autori ir Latvijā studējoši ārvalstu studenti. Teksti analizēti, neizmantojot korpusa lingvistikas metodes, un publikācijās raksturoti dažādi kļūdu tipi: nomenu dzimtes (Laizāne 2014b) un locījuma (Laizāne 2014h) kategorijas kļūdas, kā arī kļūdas atsevišķi lokatīva (Laizāne 2014d, 2014c, 2014f), akuzatīva (Laizāne 2012), ģenitīva (Laizāne 2013), datīva un instrumentāļa (Laizāne 2014e) locījuma formu lietojumā. Runāts arī par grūtībām valodas apguvē vispār (Laizāne 2014g). Bieži sastopamajām kļūdām latviešu valodas apguvē, lai arī ne tik izvērsti, tomēr pievērsusies arī Veneta Žīgure (Žīgure 1999).

Arī Lietuvā vairāki pētnieki pievērsušies kļūdu analīzei dažādu valodu apguvēju tekstos, piem., spāņu valodas (Rascon 2013), lietuviešu valodas (Dabašinskienė, Čubajevaitė 2009) un latviešu valodas (Zujevaitė, Žilinskaitė 2012) apguvēju rakstu darbos; latviešu valodas studentu sacerējumos atrodamās kļūdas analizētas arī bakalaura darbā (Kazlauskaitė 2015).

D. Deksne un I. Skadiņa (Deksne, Skadiņa 2014) latviešu valodas kļūdu klasifikācijā izdala sīkus kļūdu tipus, tādējādi padarot šo savu klasifikāciju piemērotu automātiskas kļūdu pārbaudes sistēmas izveidei, kas ir paredzēta tipiskāko kļūdu noteikšanai augsta valodas prasmju līmeņa tekstos. Tomēr valodas apguvēji, it sevišķi iesācēji, pieļauj daudz plašāku klāstu dažādu kļūdu, tāpēc tādā korpusā, kurā ir iekļauti iesācēju apguvēju valodas paraugi, D. Deksnes un I. Skadiņas piedāvātā un izmantotā tagu sistēma neatbilst S. Greindžeres izvirzītajiem kritērijiem (Granger 2003a), jo ļoti sīkā sadalījuma dēļ ir grūti pārskatāma, būtiski papildināma un līdz ar to samērā neērti lietojama anotētājam. Tāpēc šajā korpusā, ņemot vērā arī D. Deksnes un I. Skadiņas veiktos novērojumus, izmantota S. Greindžeres (Granger 2003a) izstrādātā klasifikācija, adaptējot to baltu valodu sistēmai. Par pamatu adaptācijai izmantota

latviešu valodas gramatika (LVG 2013), lietuviešu valodas gramatika (DLKG 1994), kā arī dažādas publikācijas lingvodidaktikā.

Adaptētajā kļūdu klasifikācijā ir pieci kļūdu tipi ar apakštipiem. 6. tabulā (sk. 115.–118. lpp.) redzama kļūdu klasifikācija un katram apakštipam sniegts piemērs no korpusa materiāla. Tālāk raksturots katrs no tipiem un skaidrots klasifikācijas izmantojums kļūdu anotēšanā.

6. tabula. Kļūdu klasifikācija

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Forma	F	Kopā vai šķirti rakstāmi vārdi	FK		*širdyje kaž kas suvirpėja* (suvirpa)[38] ‘sirdī kaut kas ietrīsas’
		Lielie/mazie burti	FL	*...un Viņai patīk...*	*Olimpinėje* (Olympinėse) *Žaidynėse* ‘Olimpiskajās spēlēs’
		Diakritiskās zīmes	FD	*Viņas (Viņš) ir uzņemējs*	*dažnai nera pakankamai laiko* ‘bieži nav pietiekami daudz laika’
		Citas pareizrakstības kļūdas (ieskaitot pārrakstīšanos)	FP	*man patīk tiktis (tikties) ar draugiem*	*kikvieną dieną* ‘katru dienu’
Morfoloģija un vārddarināšana	M	Atvasināšana	MA	*patīk futbols, basketbols, vazinātes*[39] *ar ritēni (riteni)*	*todėl užmiegojome anksti* ‘tāpēc aizmigām agri’
		Salikteņdarināšana	MS		*aerouostas (oro uostas)* ‘lidosta’
		Locījums	ML	*Es gribu pastāstīt par mana ģimeni*	*didelė dalis drabužiai yra tokios spalvos* ‘liela daļa apģērbu ir tādā krāsā’

[38] Piemēros, kur nepieciešams, iekavās sniegts labojums; pasvītrota tā kļūda, kas atbilst attiecīgajam kļūdu apakštipam.
[39] Šķiet, šis vārds darināts no diviem vārdiem: lie. *važinėti* ‘braukāt’ un la. *vizināties.*

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Morfoloģija un vārddarināšana	M	Dzimte	MD	*Mans acis ir brūnas.*	*Jos visi yra šalia* ‘viņi visi ir līdzās’
		Skaitlis	MN	*es biju ļoti skumīga šoreiz pār (par) atvaļinājumiem*	*įvariuose gyvenimo valandą* ‘dažādās dzīves stundās’
		(Ne)noteiktā galotne	MG	*Fotoaparātā bija manas skaistas fotogrāfijas*	*žmonių kamšatis ir ilgoji (ilgos) valandos viešajame transporte* ‘cilvēku saspiestība un ilgās stundas sabiedriskajā transportā’
		Salīdzināmās pakāpes	MQ		*Aš esu jaunesnioji (jaunausia).* ‘es esmu visjaunākā’
		Persona	MP	*Tēvs interesējies par automobiliem (automobiļiem)*	*aš nebuvo name* ‘es nebiju mājās’
		Laiks	MT	*viņai patīk ceļot(,) un māte apmeklēja (ir apmeklējusi) Krieviju, Franciju...*	*aš pasibundu (pasibudau), nes buvau labai alkana* ‘es pamodos, jo biju ļoti izsalkusi’
		Izteiksme	MI		*Aš esu dėkinguma (dėkinga), ka (kad) sutikčiau (sutikau) jai (ją)* ‘es esmu pateicīga, ka satiku viņu’
		Kārta	MK		*Į ją galėtų įeiti ir iš lauko* ‘tajā varētu ieiet arī no āra’
		Refleksivitāte	MR		*netrukdėme ir neriejomės* ‘netraucējām un nebārāmies’

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Morfoloģija un vārddarināšana	M	Divdabis	MV		*vairuotojas, matytint, kad bėgtu (bėgu), (..) pristabdau (pristabdo)* ‘vadītājs, redzot, ka skrienu, piebremzē’
		Pabeigtība	MB		*Kada ji ėjo (atėjo) iš darbo...* ‘kad viņa atnāca no darba’
		Iterativitāte	MX		*mama man (mane) išmokydavo nekada nepasiduoti* ‘mamma man iemācīja nekad nepadoties’
Sintakse	S	Vārdu secība	SV	*..radoša tik (*tikai*) dēļ naudas (*naudas dēļ*)*	*Vieta, kur visada aš galiu grįžti yra...* ‘vieta, kur es vienmēr varu atgriezties’
		Izlaists vārds	SI	*tā vārds (*vārds ir*) Džekis*	*Biologijos fakultete yra labai daug (daug* ko?[40]*)* ‘Bioloģijas fakultātē ir ļoti daudz (kā?)’
		Lieks vārds	SL	*..ceļiauju (ceļoju) uz Klaipēdu būt brivdienās (brīvdienās)*	*ji pasiūlė man kartu su ja reikėjo ruošti pjesę* ‘viņa piedāvāja man kopā ar viņu vajadzēja gatavot lugu’
		Saistījums	SS	*Mans mātes vārds ir...*	*aš nešioju kepures, irgi pirštines* ‘es nēsāju cepures, arī cimdus’
Leksika	L	Nozīme	LN	*pelēks, biezs (resns) un ļoti labs kaķis Benas*	*Ne tik katris* (kiekvienas) *latvis* ‘ne tikai katrs latvietis’
		Saderība	LV	*zils paklājs, kurš der (piestāv) pie sienu (sienām)*	*nes esame tiek įvairios* ‘jo esam tik dažādas’

[40] Šajā piemērā tekstā nav kontekstuāla saistījuma.

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Leksika	L	Stabili vārdu savienojumi	LS	*..braukšu uz ciemus (ciemos)*	*Aš tą* (tai) *labai įvertinu* ‘es to ļoti novērtēju’
Interpunkcija	I	Nepiemērota pieturzīme	IN		*dar kartą užmigau..* ‘vēlreiz aizmigu’
		Lieka pieturzīme	IL	*Tāpēc, es biju ļoti skumīga*	*Trečią valandą naktį* (nakties), *aš...* ‘trijos naktī es...’
		Pieturzīmes trūkums	IT	*Viņai patīk ceļot(,) un māte apmeklēja...*	*Viskas būtu(,) kaip aš norėčiau.* ‘viss būtu, kā es gribētu’

Adaptācijas process veikts divās daļās: vispirms pārskatīti kļūdu tipi un izvērtēta to atbilstība baltu valodu sistēmai, kā arī korpusā iekļauto tekstu specifikai, vajadzības gadījumā noteiktus tipus mainot, dzēšot vai apvienojot. Pēc tam, atbilstoši S. Greindžeres noteiktajam elastības principam, veikta anotēšana, izvērtējot katra apakštipa iederību un tos vajadzības gadījumā pievienojot vai mainot. Apzīmējumi nav saglabāti tādi paši kā S. Greindžeres sākotnējā klasifikācijā; lai vienkāršotu korpusa izmantošanu un veicinātu darbu ar aizstājējzīmēm, tie veidoti pēc šāda principa:

1) pirmais burts apzīmē kļūdu tipu;
2) otrais burts apzīmē kļūdu apakštipu;
3) ja vēlāk korpusa pilnveides gaitā būs nepieciešams veidot vēl sīkākas apakškategorijas, to apzīmējumus veidot varēs, pievienojot vēl vienu burtu.

Pirmais kļūdu tips ir „Forma”, un tajā ietilpst tās kļūdas, kuras var attiecināt uz vārdformu veidošanu un rakstību. Šajā tipā ir četri apakštipi:

- apakštipā „Kopā vai šķirti rakstāmi vārdi” ietilpst piemēri, kuros vārds, kas saskaņā ar mērķvalodas normām būtu rakstāms kopā, ir rakstīts šķirti vai pretēji – vārdi, kas saskaņā ar mērķvalodas normām būtu rakstāmi šķirti, ir rakstīti kopā;
- apakštipā „Lielie/mazie burti” tiek iekļautas kļūdas lielo un mazo burtu lietojumā. Tie var būt gan īpašvārdi – personvārdi, toponīmi, iestāžu un organizāciju nosaukumi u. c. –, gan sugas vārdi, piem., ja lielais burts nav lietots

teikuma sākumā vai ja teikuma vidū kāds vārds nepamatoti tiek rakstīts ar lielo burtu;

- apakštipā „Diakritiskās zīmes" ir kļūdas, kurās kā neatbilstība mērķvalodas normai ir konstatēts diakritisko zīmju neatbilstošs lietojums, trūkums vai pārdaudzums;
- apakštipā „Citas pareizrakstības kļūdas" ietilpst dažādas pārrakstīšanās kļūdas, kļūdaini veidotas formas, kā arī kļūdainas personvārdu atveides gadījumi. Liela daļa personvārdu tekstos gan ir mainīti anonimizācijas gaitā, taču to formālās īpatnības ir mēģināts saglabāt, un kā mērķa hipotēze tiek izmantota tā vārda atveide, kurš ir izmantots kā aizstājējs. Ja atveide aizstājēja formā neatbilst mērķa hipotēzei, tā līdz ar to tiek norādīta kā šā tipa kļūda.

Otrais kļūdu tips ir morfoloģija un vārddarināšana, un tajā ietilpst kļūdas, kas ir attiecināmas uz tādām gramatiskajām kategorijām kā dzimte un skaitlis, kā arī uz jaunu vārdu darināšanas īpatnībām. Tā kā liela daļa kļūdu, kas apgrūtina komunikāciju, uzsākot mācīties svešvalodu, ietilpst šajā tipā, tas ir ievērojami vairāk sazarots nekā pārējie, galvenokārt balstoties baltu valodām raksturīgajās gramatiskajās kategorijās. Šajā tipā ir piecpadsmit apakštipu:

- apakštipā „Atvasināšana" ietilpst piemēri, kuros atvasinot ir radīts jauns vārds, kāds netiek lietots mērķvalodā;
- apakštipā „Saliktenņdarināšana" ir līdzīgi piemēri, taču tie ir tādi, kuros jauns, mērķvalodā nepastāvošs vārds tiek lietots kā saliktenis;
- apakštipā „Locījums" tiek iekļautas tās kļūdas, kurās ir lietota neiederīga locījuma forma. Ja autors ir acīmredzami mēģinājis veidot to locījuma formu, kura ir nepieciešama, taču kļūda ir pašā formā (piem., *man ir tumši brūni garši mati*), tad šī kļūda ietilpst tipā „Forma", apakštipā „Citas pareizrakstības kļūdas"[41];
- apakštipā „Dzimte" ir iekļaujamas kļūdas, kuras raksturo neiederīgas dzimtes formas lietojums;
- apakštipā „Skaitlis" ietilpst piemēri, kuros lietota vārdforma neatbilstošā skaitlī;

[41] Tas pats attiecas arī uz apakštipiem „Dzimte", „Skaitlis" utt.

- apakštipā „(Ne)noteiktā galotne" ir tādi gadījumi, kuros ir lietota noteiktā galotne nenoteiktās galotnes vietā vai otrādi – lietota nenoteiktā galotne noteiktās galotnes vietā;
- apakštipā „Salīdzināmās pakāpes" iekļauj piemērus, kuros īpašības vārds vai apstākļa vārds lietots neiederīgā salīdzināmās pakāpes formā;
- apakštipā „Persona" ir kļūdas ar neiederīgas darbības vārda personas formas lietojumu;
- apakštipā „Laiks" ietilpst piemēri, kuros ir lietota nepiemērota darbības vārda laika forma. Tas attiecas gan uz tādām kļūdām, kurās ir sajaukta nākotne, tagadne un pagātne, gan uz tādām, kurās vienkāršā laika vietā ir lietots saliktais laiks vai otrādi;
- apakštipā „Izteiksme" iekļaujamajos piemēros ir lietota neiederīga darbības vārda izteiksme, piem., vēlējuma izteiksme īstenības izteiksmes vietā vai tml.;
- apakštipā „Kārta" ir kļūdas ciešamās un darāmās kārtas lietojumā;
- apakštipā „Refleksivitāte" ietilpst gadījumi, kuros ir neiederīgs atgriezenisku vārdformu lietojums vai tā trūkums, vai nepareizi veidotas atgriezeniskās vārdformas;
- apakštipā „Divdabis" ietilpst neiederīgi lietotas un/vai nepareizi veidotas divdabju formas;
- apakštipā „Pabeigtība" iekļaujami tie piemēri, kuros neatbilstoši mērķvalodas normām ir lietots priedēkļverbs, kas norāda uz darbības pabeigtību, vai arī tas nav lietots, lai arī būtu nepieciešams;
- apakštipā „Iterativitāre" ir tādi gadījumi, kuros ir lietota vārdforma, kas izsaka vienkārtēju darbību, lai arī bijusi domāta daudzkārtēja darbība, vai otrādi – ir lietota vārdforma, kas izsaka daudzkārtēju darbību, kaut gan ir domāta vienkārtēja.

Trešais kļūdu tips ir „Sintakse", un tajā ietilpst tādas neatbilstības mērķvalodas normai, kuras ir attiecināmas uz vārdu savstarpējo saistījumu, secību, pārdaudzumu vai trūkumu. Šajā tipā A līmenī ir būtiski labojot ievērot mērenību. Pamatlīmenim ir ļoti raksturīgas neveiklas teikumu konstrukcijas (Šalme, Auziņa 2013, 12), un tās pārlieku sīkumaini labojot, var nākties pārrakstīt par jaunu visu tekstu, zaudējot autentiskā valodas lietojuma pazīmes. Šajā kļūdu tipā ir četri apakštipi:

- apakštipā „Vārdu secība" ir piemēri, kuros vārdi lietoti nepiemērotā secībā;

- apakštipā „Izlaists vārds” iekļauj gadījumus, kuros ir nepieciešams lietot vēl kādu vārdu, kuru autors ir izlaidis;
- apakštipā „Lieks vārds” ir tādas kļūdas, kuras raksturo valodas līdzekļu pārdaudzums – lietota kāda vārdforma, kas attiecīgajā teikumā uzskatāma par lieku;
- apakštipā „Saistījums” ietilpst piemēri, kuros neatbilstība mērķvalodas normām vērojama vārdu savstarpējā saistījumā, ja tā neatbilst kļūdu tipa „Morfoloģija un vārddarināšana” nošķīruma principiem (piem., nepiemērota locījuma forma) un ja šo neatbilstību nosaka gramatiski faktori. Ja neatbilstības cēlonis ir vārda leksiskā nozīme, tad šāds piemērs uzskatāms par piederīgu kļūdu tipa „Leksika” apakštipam „Saderība”.

Ceturtais kļūdu tips ir „Leksika”, kurā ietilpst vārdu leksiskās nozīmes neatbilstības. Tajā ir trīs apakštipi:

- apakštipā „Nozīme” iekļaujami tie piemēri, kuros lietoti vārdi ar neatbilstošu leksisko nozīmi (pilnībā vai niansēs), ja izvēli nosaka teikuma semantika, bet ne konkrēti lietoti vārdi – piem., kontekstā neiederīgs sinonīms;
- apakštipā „Saderība” piemērs ietilpst tad, ja tajā ir vārdi, kuru atsevišķās leksiskās nozīmes atbilst teikuma kopējai iecerētajai nozīmei, taču šie vārdi nesader savā starpā un to nosaka negramatiski faktori;
- apakštipā „Stabili vārdu savienojumi” ir tādi gadījumi, kuros neveiksmīgi mēģināts veidot konstrukciju, kas mērķvalodā pastāv kā stabils vārdu savienojums. Šādā gadījumā, pat ja starp autora lietotajiem vārdiem šķietami nav nozīmīgas neatbilstības, teksta labotājs izvēlas mērķa hipotēzē lietot citu vārdu.

Piektais kļūdu tips ir „Interpunkcija”. Pie tā piederas neatbilstības mērķvalodas pieturzīmju lietojuma normām. Arī šis kļūdu tips jāmarķē ar apdomu, jo iesācēju līmenī apguvēji interpunkcijas principus gandrīz vēl nezina. Šajā tipā ir trīs apakštipi:

- apakštipā „Nepiemērota pieturzīme” ietilpst tie piemēri, kuros ir lietota pieturzīme vietā, kurā pieturzīme ir nepieciešama, taču saskaņā ar mērķvalodas normām tai būtu jābūt citai pieturzīmei;
- apakštipā „Lieka pieturzīme” iekļaujami gadījumi, kuros teksta autors ir lietojis kādu pieturzīmi, taču attiecīgajā vietā nebūtu lietojama nekāda pieturzīme;

- apakštipā „Pieturzīmes trūkums” ir tādas kļūdas, kurās pieturzīmes nav, lai gan to būtu nepieciešams lietot.

Katrai kļūdai ir izvēlēta tikai viena, pēc labotāja domām, iederīgākā mērķa hipotēze un viens, pēc anotētāja domām, iederīgākais kļūdas tips. Tā rīkoties nolemts tādēļ, lai pārlieku nepaildzinātu korpusa izveidi un nesarežģītu tā lietošanu, kā tas notiktu, ja tiktu aplūkotas vairākas mērķa hipotēzes katram gadījumam. Reizēm ir gadījumi, kuros kļūdu var uzskatīt par piederīgu dažādām kategorijām – piemērā *mans acis ir brūnas* var saskatīt dzimtes, skaitļa vai pārrakstīšanās kļūdu. Tomēr arī šādos gadījumos anotētājs izvēlas to kļūdas tipu, kurš viņam šķiet iederīgāks. Tā kā katra gadījuma padziļināta izpēte uz nenoteiktu laiku paildzinātu korpusa anotēšanu, šādu piemēru anotēšanā tika nolemts vadīties pēc anotētāja pieredzes un intuīcijas. Šajā gadījumā par primāro tiek uzskatīta dzimtes lietojuma kļūda, jo, pēc anotētāja domām, sākotnēji tiek noteikts, kādā dzimtē piederības vietniekvārds ir lietojams, un tikai pēc tam, ja nepieciešams, tiek izveidota daudzskaitļa forma. Par šādiem jautājumiem gan ir iespējamas plašākas diskusijas un, ja tiktu nonākts pie secinājuma, ka šī kļūda drīzāk pieder pie cita tipa, anotējumu ir iespējams labot.

Ir iespējams, ka, izlabojot vienu kļūdu, rodas jaunas, piem., šajā gadījumā: *Viņas uzbūve ir smalka, spēcīga [..]*. Teksta autors ir pareizi saskaņojis vairākus īpašības vārdus ar lietvārdu sieviešu dzimtē, taču teksta labotājs lietvārdu ir nomainījis pret citu lietvārdu vīriešu dzimtē – *augums*. Šādā gadījumā jāmaina arī īpašības vārdu dzimte. Labojumā tas arī tiek darīts, un tiek anotēts atbilstošais kļūdas veids. Tomēr šādu piemēru iespaidā, nekritiski vērtējot kļūdu veidu skaitu, var nonākt pie nepareiziem secinājumiem par tekstu autoru prasmi (vai prasmes trūkumu) saskaņot lietvārdus ar īpašības vārdiem. Tāpēc nepieciešams uzsvērt, ka kļūdas klātbūtne ne vienmēr liecina par trūkumiem teksta autora prasmēs.

2.3. Programmatūra

Tā kā korpuss ir elektronisks, tā veiktspēja ir atkarīga no izmantotās programmatūras. Šajā nodaļā aprakstīts, kādas programmas ir izmantotas korpusa izveidē un kādas ir paredzēts izmantot, lietojot korpusu.

2.3.1. Korpusa izveidē izmantotā programmatūra

Daļa tekstu gan latviešu, gan lietuviešu valodā sākotnēji ir rakstīti rokrakstā, tāpēc, lai ievietotu korpusā, nepieciešams tos **digitalizēt**, t. i., pārrakstīt datorrakstā. Tas darīts tā, lai pēc iespējas saglabātu rakstības īpatnības. Šim nolūkam izmantoti populāri teksta redaktori

Windows un *Ubuntu* vidē: *Microsoft Word* un *LibreOffice Writer*. Teksti tālāk apstrādāti ar programmām *Notepad++* un *Gedit* atkarībā no tā, kāda operētājsistēma darbojas datorā, ar kuru apstrādāts konkrētais teksts. Tekstu apstrāde abos gadījumos veikta pēc identiskiem principiem.

Datņu nosaukumi ir unikālas sešu ciparu virknes, kas ir ģenerētas pēc nejaušības principa ar nejaušo skaitļu virkņu ģeneratoru *Random.org*[42]. Publiski nepieejamā korpusa arhīvā **.xlsx* formāta datnē (paraugu sk. 5. pielikumā) glabājas visa par tekstiem un to autoriem ievāktā informācija, ieskaitot katra teksta datnes nosaukumā lietoto ciparu virkni, lai būtu iespējams, piem., nepieciešamības gadījumā izdzēst kādu konkrētu tekstu. Šāda rīcība gan tiek pieļauta tikai īpašos gadījumos – piem., ja izrādītos, ka kāds no tekstiem ir plaģiāts vai tml. (par plaģiāta problēmu sk. arī 2.1.1. apakšnodaļā „Tekstu ieguve un atlases kritēriji"). Gadījumos, kuros nav radušās pamatotas aizdomas par plaģiātu, tiek uzskatīts, ka plaģiāta nav. Par tekstu saturu ir atbildīgi to autori, tāpēc gadījumā, ja atklātos plaģiāts vai citi pārkāpumi, korpusa veidotāja un publiskotāja atbildību neuzņemas.

Katra autora individuālais kods, kas ļauj noteikt viena un tā paša autora vairākus darbus, sastāv no 4 cipariem un, tāpat kā datņu nosaukumi, ir ģenerēts pēc nejaušības principa ar ģeneratoru *Random.org*. Arī šie individuālie kodi un to atbilstība konkrētām personām tiek glabāti iepriekšminētajā nepubliskotajā datnē.

Izmēģinājuma korpusā, kas ir paredzēts apstrādei ar programmu *AntConc* vai līdzīgu, nav anotējuma, un vienīgais mērķtiecīgi pievienotais marķējuma veids ir anonimizācija. Šie teksti ir saglabāti **.txt* formāta datnēs. Anonimizācijas iezīmes ir *<anon>* un *</anon>*, bet izlaiduma iezīme – *<izlaid>*. Piem., izteikumā *mani sauc Inga* personvārdu aizstājot ar *Aija*, tekstā tas izskatās šādi: *mani sauc <anon>Aija</anon>*. Savukārt informācijas izlaiduma gadījumā tekstā ir tikai viena iezīme: *mana bakalaura darba temats ir <izlaid>*. Šis marķējums tekstos ir ievietots manuāli: tekstus apstrādājot ar kādu no iepriekšminētajām tekstu rediģēšanas programmām, ir dzēsta sākotnējā informācija, atbilstošie tagi ir iekopēti attiecīgajās vietās un vajadzības gadījumā ir pievienots aizstājošs teksts.

Cits marķējums tekstiem nav pievienots, taču līdz ar izmēģinājuma korpusa datnēm lejupielādei ir pieejama **.xlsx* formāta datne, kas satur šādus datus:

- teksta kods konkrētā teksta atpazīšanai;
- autora kods vairāku viena autora tekstu atrašanai;
- tēma – teksta nosaukums;

[42] Pieejams tiešsaistē: https://www.random.org/

- vārdu skaits tekstā (vārdu skaitīšana veikta pirms anonimizācijas)[43];
- teksta mērķvaloda (izmēģinājuma korpusā visiem tekstiem tā ir lietuviešu valoda);
- valodas apguves semestris;
- mācību valoda (izmēģinājuma korpusā visiem tekstiem tā ir latviešu valoda)[44].

Izmantojot šo tabulu, izmēģinājuma korpusa lietotājs var manuāli atlasīt tekstus pēc noteiktiem parametriem, izmantojot datņu nosaukumos ietvertos sešzīmju kodus.

Korpusa pilnajā versijā risinājums ir citāds – tās izveidei, marķēšanai un anotēšanai izmantota programma *TEITOK* (vairāk informācijas sk. Janssen 2016). Šī programma ir paredzēta samērā nelielu, specifisku korpusu izveidei un pētīšanai, balstoties CWB/CQP sistēmā (vairāk informācijas sk. CQPweb 2016). Tā līdz šim ir izmantota dažādu korpusu izveidē (TEITOK Projects$_{-e}$); tai skaitā arī portugāļu valodas apguvēju korpusā (COPLE2$_{-e}$). Viena no galvenajām *TEITOK* priekšrocībām ir tā, ka šī programma atvieglo gan korpusa izveidi, gan tā izmantošanu, jo piedāvā dažādas funkcijas ar samērā vienkāršas tiešsaistes saskarnes starpniecību.

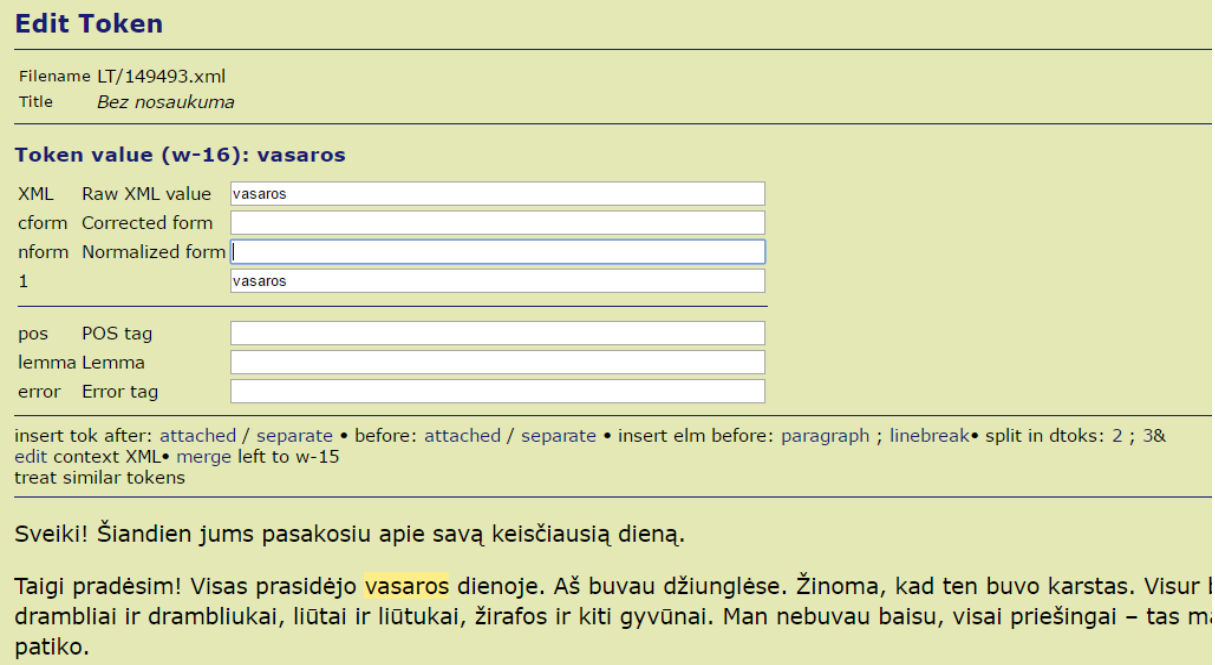
Pamatformu un vārdšķiru anotēšanā kā palīgrīki izmantoti arī divi automātiski baltu valodu analīzes rīki: *SemTi-Kamols marķētājs* (SemTi 2009) un *KLC morfoloģiskais anotētājs* (KLC $_{-e}$). Tā kā *TEITOK* izmanto formātu, kas atšķiras no *SemTi-Kamols marķētāja* un *KLC morfoloģiskā anotētāja* veidotajiem formātiem, tieša tekstu anotēšana ar šiem rīkiem nav iespējama. Arī konvertēt abu rīku rezultātus vēlamajā formātā nav lietderīgi, jo tekstu apjoms nav tik liels, lai darbs, kas ieguldīts konvertācijā, atmaksātos.

Korpusa datnes ir **.xml* formātā. Nav atsevišķas metadatu datnes, pēc kuras datu atlase būtu javeic manuāli – visa informācija ir iekļauta vai nu pašā teksta datnē, vai atsevišķā datnē, kuru programma aptaujā automātiski, kad korpusa lietotājs tai uzdod attiecīgu vaicājumu, izmantojot programmas saskarni. Arī anotēšana lielākoties notiek, izmantojot saskarni. Daļai anotējuma un marķējuma veidu ir izveidoti īpaši lauki (sk. 6. attēlā 125. lpp.), savukārt citi tiek anotēti tieši **.xml* datnē, izmantojot *TEITOK* iebūvēto datņu apskates un rediģēšanas lauku (sk. 7. attēlā 125. lpp.).

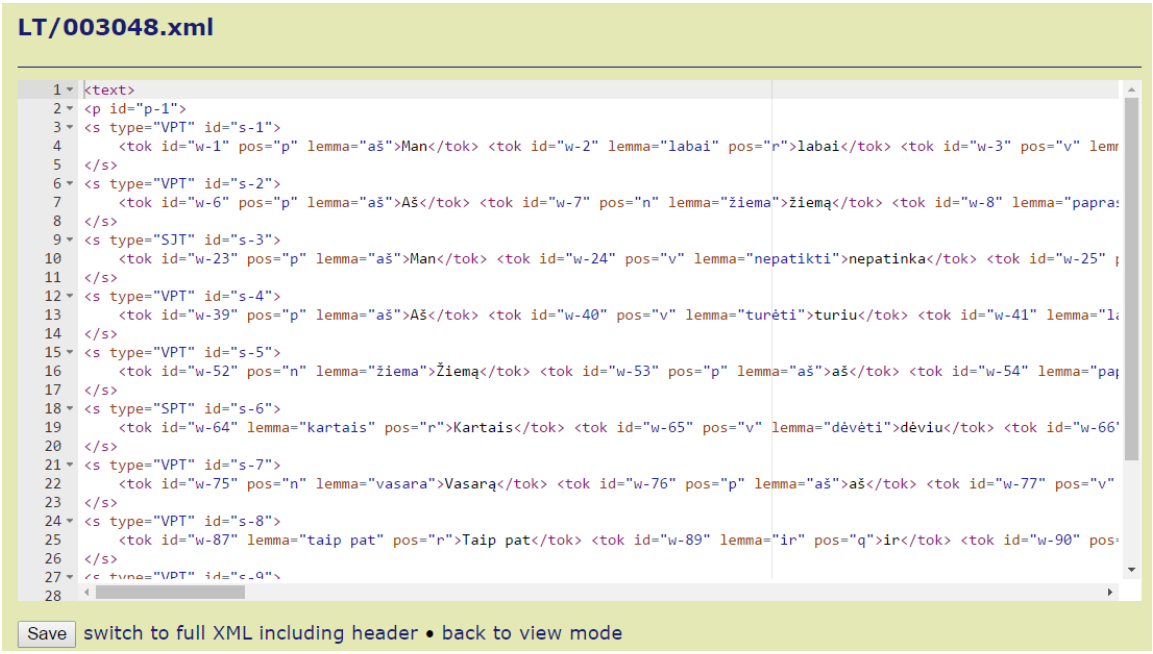
[43] Izmantojams tikai lietojamā korpusa aptuvena apjoma noteikšanai gadījumā, ja tiek atlasīta tikai daļa no izmēģinājuma korpusā iekļautajām datnēm.

[44] Šī informācija korpusa pilnajā versijā vismaz pagaidām nav marķēta, jo visu līdz šim iegūto tekstu gadījumā tā ir bijusi latviešu valoda lietuviešu valodas apguvējiem un lietuviešu valoda latviešu valodas apguvējiem. Ja nākotnē korpusā tiks iekļauti arī teksti, kuru autori mērķvalodu mācījušies ar citas valodas starpniecību, šādu marķējumu ir iespējams ieviest.

6. attēls. Anotēšana, izmantojot TEITOK iebūvētos laukus.

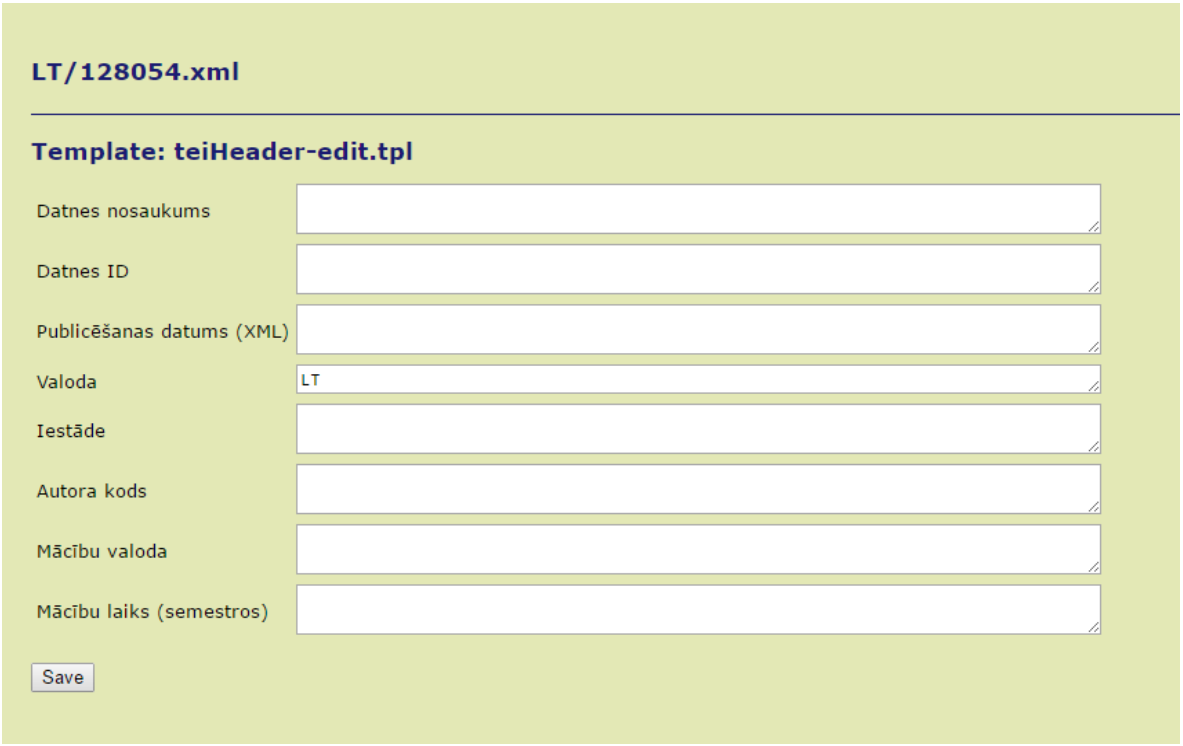


7. attēls. TEITOK iebūvētais datņu apskates un rediģēšanas lauks.



Marķēšanas atvieglošanai izveidota *TEI* formātam atbilstošas galvenes izveides veidlapa, kurā ir norādīts, kādu informāciju ir nepieciešams sniegt par korpusā iekļaujamajiem tekstiem (sk. 8. attēlā 126. lpp.).

8. attēls. TEI galvenes izveides veidlapa.



Tālāk raksturots, kā ir veikta dažādu parādību marķēšana un anotēšana un kā tas ir atveidots **.xml* datnēs.

- Marķēšana.
  - Teksta sākums un beigas tiek marķēts automātiski, izveidojot jaunu **.xml* datni. Šī marķējuma iezīmes ir *<text>* un *</text>*.
  - Rindkopas sākums un beigas marķētas manuāli tieši **.xml* datnē. Šim nolūkam tiek izmantotas iezīmes * * un * *.
  - Izteikuma sākums un beigas marķētas manuāli tieši **.xml* datnē. Šim nolūkam tiek izmantotas iezīmes *<s>* un *</s>*.
  - Teksta mērķvaloda tiek ievadīta tai paredzētā laukā iepriekš izveidotajā *TEI* galvenes izveides veidlapā. *.xml datnē šī informācija ir atainota šādās iezīmēs: *<language><lang n="LT"/></language>*.
  - Teksta nosaukums tiek ievadīts tam paredzētā laukā minētajā *TEI* galvenes izveides veidlapā. *.xml datnē šī informācija ir atainota šādās iezīmēs: *<title></title>*.
  - Teksta autora kods tiek ievadīts tam paredzētā laukā *TEI* galvenes izveides veidlapā. *.xml datnē šī informācija ir atainota pēc šāda parauga: *<persName><name id="2409" n="2409"/></persName>*.

- Augstskola (augstākās izglītības iestāde), kurā teksta autors ir mācījies otro baltu valodu, kad ir tapis attiecīgais teksts, tiek norādīta kā abreviatūra, kas tiek ievadīta tai paredzētā laukā *TEI* galvenes izveides veidlapā. *.xml datnē šī informācija ir redzama šādi: *<institution><institAbbrev n="VDU"/></institution>*.
- Informācija par to, kurš otrās baltu valodas apguves semestris šis ir teksta autoram, tiek ievadīta tai paredzētā laukā *TEI* galvenes izveides veidlapā. *.xml datnē šī informācija ir atainota šādi: *<time dur="semesters" n="1">1</time>*.
- Anonimizētās vietas marķētas, izmantojot *TEITOK* iebūvēto iespēju teksta elementiem norādīt vairākas formas: šeit norādītas divas – transkripcija un normalizēta forma. Transkripcijā visu anonimizēto elementu vietā ir simbolu virkne *xxx*. Savukārt kā normalizētā forma vieglākai teksta uztveršanai un analīzei ir norādīts aizstājošais teksta elements (gadījumā, ja anonimizācija veikta, aizstājot teksta daļu) vai arī iezīme *[izlaid]* (gadījumā, ja anonimizācijas nolūkos teksta daļa ir izlaista). *.xml datnē aizstāts teksta fragments izskatās šādi: *<tok id="w-149" nform="Lina" pos="n">xxx</tok>*. Izlaiduma gadījumā *.xml datnē attiecīgā teksta daļa ir noformēta pēc šāda parauga: *<tok id="w-54" nform="[izlaid]">xxx</tok>*.

- Anotēšana.
  - Sintaktiskā anotēšana veikta, balstoties jau esošajā izteikumu marķējumā. Katram atsevišķi marķētajam izteikumam tiek anotēts teikuma tips, iezīmei *<s>* pievienojot atribūtu *type* un tā vērtību, kas atbilst 2.2.5. apakšnodaļā „Sintaktiskā anotēšana" norādītajiem teikumu veidu kodiem. Tātad *.xml datnē tas izskatās, piem., šādi: *<s type="VNT">Es rakstu.</s>*
  - Pamatformu anotēšanai un vārdšķiru anotēšanai *TEITOK* ir izveidota īpaša saskarne: katram vārdlietojumam (*token*) var norādīt vairāku veidu papildinformāciju, un starp definētajiem laukiem ir arī pamatforma un vārdšķira. Pamatformu un vārdšķiru anotējums *.xml datnē izskatās šādi: *<tok id="w-40"* ***pos="v" lemma="turėti"****>turiu</tok>* (izcēlums mans – I. Z.).

- Morfoloģiskās anotēšanas vienīgais veids korpusā „Esam” ir vārdšķiru anotēšana, kas jau ir aprakstīta iepriekšējā punktā kopā ar pamatformu anotēšanu.
- Kļūdu anotēšana *TEITOK* ir iespējama divos veidos. Viens no tiem ir analogs vārdšķiru un pamatformu anotēšanai – katram vārdlietojumam var pievienot labotu formu un kļūdas tipu. Tas ir iekļautā anotējuma (angl. *inline annotation*) veids. Otrs variants paredz, ka anotētājs vispirms norāda kļūdas robežu, atzīmējot, uz kurām teksta vienībām anotējums attiecas, un tikai pēc tam ieraksta labotu formu un kļūdas tipu. Šādā gadījumā tas ir nevis iekļautais, bet gan savrupais anotējums (angl. *stand-off annotation*). Gan anotēšanā, gan lietošanā ērtāks ir pirmais variants, jo tas ļauj kļūdu anotējumu ērti pārlūkot ne tikai tad, ja tiek meklēts konkrētais kļūdas tips, bet arī tad, ja apskatei tiek atvērts konkrēts teksts. Tas sniedz arī tādu priekšrocību, ka ērti ir aplūkojams teksts nelabotā un labotā formā. Savukārt savrupās anotēšanas priekšrocība ir iespēja kā kļūdu anotēt dažāda garuma teksta daļas. Līdz šim korpusā iekļautajos tekstos nav atrastas kļūdas, kurām tas būtu tik nozīmīgi, lai tā dēļ būtu vērts atteikties no iekļautā anotējuma sniegtajām priekšrocībām, tāpēc „Esam” tekstos kļūdas ir anotētas pēc iekļautā anotējuma principa. Katram vārdlietojumam ir norādīts kļūdas tips (ja kļūda ietver vairākus vārdlietojumus, tad katram no šiem vārdlietojumiem ir norādīts viens un tas pats kļūdas tips) un labotā forma. *.xml datnē tas tiek parādīts šādi: *<tok id="w-5" pos="p" lemma="mans" **cform="manu" error="ML"**>mana</tok>* (izcēlums mans – I. Z.). Tehniski ir saglabāta arī iespēja veidot savrupo anotējumu, ja nākotnē tiktu nolemts, ka tas ir vajadzīgs.

Marķēta un anotēta teksta attēlojums *.xml datnē ir redzams šī teksta 3. pielikumā (teksts latviešu valodā) un 4. pielikumā (teksts lietuviešu valodā). Korpusā iekļauto tekstu, marķējuma un anotējuma aplūkošana un izmantošana ir aprakstīta 2.3.2. apakšnodaļā „Korpusa lietošanai nepieciešamā programmatūra”.

Korpusa pilnās versijas datnes atbilst *TEI* standartam (TEI 2015), un vajadzības gadījumā darbā ar tām var izmantot citas *TEI* standartam pielāgotas programmas (Janssen 2016, 4037).

Lietotājiem korpuss ir pieejams divos veidos: izmēģinājuma korpuss ir pieejams lejupielādei kā datņu kopums. Pieejamības nodrošināšanai datnes ir augšupielādētas datņu uzglabāšanas vietnē *failiem.lv*, un saites uz lejupielādi ir norādītas vietnē, kas izveidota, izmantojot *Wordpress* interneta mājaslapu izveides servisu. Tā pati vietne ir izmantota arī, lai sniegtu pamatinformāciju par korpusu un tā lietošanu.

Pilna korpusa pieejamība ir nodrošināta, izveidojot atsevišķu mājaslapu. Tās uzturēšanai izveidots privāts serveris, kurš darbojas uz operētājsistēmas *Linux Ubuntu* bāzes. Serveris atrodas Rīgā un ir fiziski pieejams tikai korpusa veidotājai.

2.3.2. Korpusa lietošanai nepieciešamā programmatūra

Pirmā publicētā šī korpusa daļa – **izmēģinājuma korpuss** aptuveni 15 000 tekstvienību apmērā – tika publicēta 2015. gada 14. jūnijā vietnē *https://esamkorpuss.wordpress.com/*[45] kā 68 **.txt* formāta datnes, kuras ir pieejamas lejupielādei un apstrādājamas ar jebkādu programmu pēc pētnieka izvēles. Katru no tām var arī atvērt kā pilnu tekstu, izmantojot jebkuru teksta redaktoru (ir pārbaudīta datņu saderība ar *Notepad* jeb *Piezīmjbloku*[46] un *Notepad++*[47]). Vietnē ir sniegts arī ekrānšāviņš, lai parādītu, kā datnes lejupielādēt.

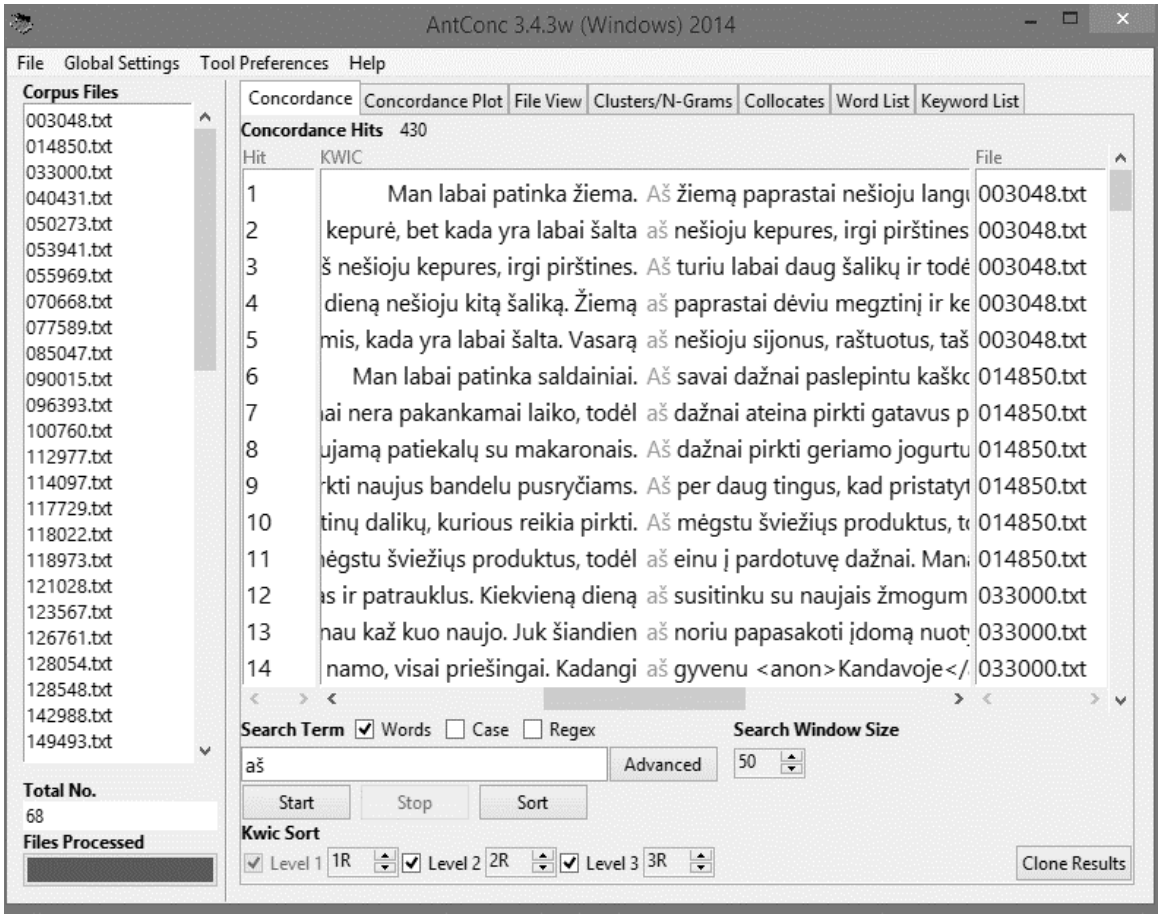
Ieteiktā korpusa izpētes programma ir *AntConc*[48], un visu datņu saderība ar šo programmu ir pārbaudīta[49]. Izmēģinājuma korpusā iekļautie teksti nav anotēti, taču ir anonimizēti – lai pasargātu tekstu autoru un citu tekstos minēto personu datus, tie tika izlaisti vai mainīti, izmaiņu vietu norādot ar atbilstošu tagu (vairāk par anonimizēšanu sk. 2.1.3. nodaļā „Personas datu aizsardzība un autortiesības"). Šis izmēģinājuma korpuss joprojām ir pieejams minētajā formātā.

Lai korpusu pētītu ar programmu *AntConc*, gan tekstus, gan programmu nepieciešams lejupielādēt. Programma nav jāinstalē, to var aktivizēt no **.exe* datnes. Programma ir viegli apgūstama, tā piedāvā samērā plašas iespējas (konkordanču rindu atlasi, kolokāciju meklēšanu, atslēgvārdu noteikšanu u. c.), un tās lietošana ir uzskatāmi aprakstīta dažādos palīglīdzekļos (Anthony 2014 un Tang 2011). Par īpaši noderīgu funkciju var uzskatīt konkordanču rindu

[45] Alternatīvās adreses: *https://esamtekstynas.wordpress.com/* (vietne lietuviešu valodā) un *https://esamcorpus.wordpress.com/* (vietne angļu valodā).
[46] Pieejama visās Windows versijās kopš Windows 1.0 kā iebūvēta programma (vairāk sk. Windows$_{-e}$ un Piezīmjbloks$_{-e}$).
[47] Pieejama bez maksas tiešsaistē: *https://notepad-plus-plus.org/*
[48] Pieejama bez maksas tiešsaistē: *http://www.laurenceanthony.net/software/antconc/*
[49] Lai varētu strādāt ar *AntConc*, jāņem vērā, ka datnēm jāatrodas mapē, kuras nosaukumā nav burtu ar garumzīmēm, mīkstinājuma zīmēm u.tml., citādi programma var nespēt datus nolasīt.

atlasi. 9. attēlā redzams, kādas izskatās vārda *aš* konkordanču rindas „Esam" izmēģinājuma korpusā.

9. attēls. Vārda aš *konkordanču rindas „Esam" izmēģinājuma korpusā.*



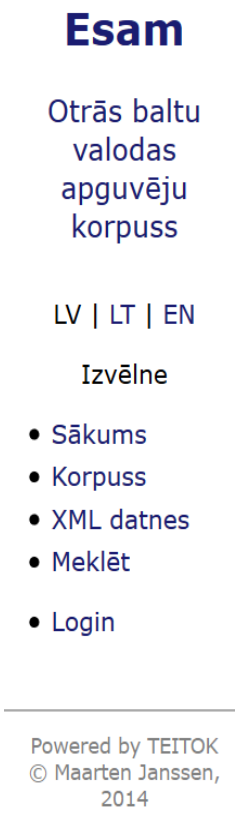
Kā redzams attēlā, kopumā šis vārds izmēģinājuma korpusā atrasts 430 reižu, un, izmantojot sakārtošanas funkciju (*Kwic Sort*), ir iespējams rezultātus aplūkot citā secībā pēc pētnieka izvēles. Izvēloties citu cilni, var aplūkot datus pēc citiem parametriem. Tomēr pilna korpusa apstrādei ir izvēlēta cita programma, tāpēc darbs ar korpusu tiks aprakstīts, balstoties tās, nevis *AntConc* funkcionalitātē, jo sevišķi ņemot vērā to, ka par *AntConc* jau ir pieejama plaša dokumentācija, ieskaitot ekrānšāviņus, un darbs ar „Esam" izmēģinājuma korpusu tajā būtiski neatšķiras no darba ar citiem korpusiem.

AntConc ir nevis galvenais, bet gan tikai viens no rīkiem, kas tika izvēlēti darbam ar korpusu „Esam", vairāku iemeslu dēļ:

- programmu nepieciešams lejuplādēt katrā datorā, kurā iecerēts strādāt ar korpusu. Atkarībā no tā, kādas tiesības ir piešķirtas konkrētā datora attiecīgajam lietotājam, **.exe* formāta datņu atvēršana var nebūt iespējama;

- lai arī programma nav jāinstalē, atkarībā no drošības uzstādījumiem datorā to var neizdoties palaist[50]. Šādā gadījumā var būt nepieciešamas samērā augstas datorprasmes, lai problēmu atrisinātu.

10. attēls. Galvenā izvēlne



Tāpēc pilnajai anotētā korpusa versijai ir izvēlēts citāds risinājums: korpuss ir pieejams tiešsaistē, un to nav nepieciešams lejupielādēt. Lietotāja saskarne ir redzama interneta pārlūkā, vietnē *www.esamkorpuss.lv*. Saskarne ir Mārtena Jansena (*Maarten Janssen*) izstrādātā *TEITOK* (Janssen 2016), un īss tās apraksts attiecībā uz korpusu „Esam" sniegts tālāk.

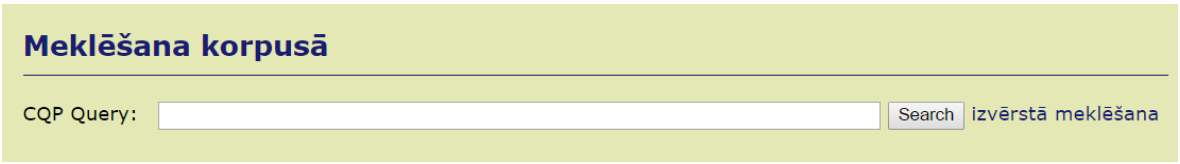
Atverot korpusa vietnes sākumlapu, parādās korpusa lietošanas nosacījumi un saites uz īsu korpusa aprakstu, papildinformāciju (piem., autoru sarakstu, sastādītājas kontaktinformāciju) un izmēģinājuma korpusa vietni. Ir arī saite „Uz korpusu", un, to atverot, kreisajā pusē ir redzama izvēlne (sk. 10. attēlā), kas ļauj:

- izvēlēties vietnes valodu (lielākā daļa vietnē redzamo tekstu ir latviešu, lietuviešu un angļu valodā; retāk redzamie paziņojumi var būt tikai angļu valodā. Ja lietotājs programmā sastopas ar informāciju valodā, kas neatbilst izvēlētajai vietnes valodai, viņam ir iespēja par to paziņot korpusa sastādītājai, izmantojot kontaktinformācijas sadaļā norādīto e-pastu. Iespēju robežās tulkojumi tiek pastāvīgi pilnveidoti);
- no jebkuras vietnes sadaļas atgriezties sākumlapā;
- no jebkuras vietnes sadaļas doties uz korpusu;
- aplūkot korpusā iekļautos *.xml failus;
- atvērt meklēšanas sadaļu;
- autorizēties, lai piekļūtu korpusa administratora funkcijām (autorizācijas dati tiek piešķirti tikai korpusa izveidē tieši iesaistītām personām uz nepieciešamo darbību veikšanas laiku; citiem korpusa lietotājiem autorizācija nav nepieciešama).

[50] Sk., piem., tiešsaistes diskusiju par *AntConc* palaišanu operētājsistēmā *Maverick*: *https://groups.google.com/forum/#!topic/antconc/46Gzyd0omBU*

Korpusa pētniecība galvenokārt notiek, izmantojot meklēšanas iespējas. Lai tām piekļūtu, izvēlnē ir jāizvēlas sadaļa „Meklēt", kas atver vienkāršās meklēšanas lauku (sk. 11. attēlā). Meklēšana korpusā darbojas, izmantojot CQP (*Corpus Query Processor*) vaicājumu sintaksi. Zem meklēšanas lauka ir sniegts īss skaidrojums par meklēšanas galvenajiem principiem un bieži lietotajiem parametriem: *word* (vārda rakstītā forma); *cform* (kļūdas labojums); *nform* (anonimizēto fragmentu aizstājējelementi); *lemma* (vārda pamatforma); *pos* (vārdšķira) un *error* (kļūdas tips). Lietotājs var veidot arī dažādus kompleksus vaicājumus, izmantojot CQP sniegtās iespējas (aizstājējzīmes, loģiskos operatorus u. c.).

11. attēls. Vienkāršās meklēšanas lauks



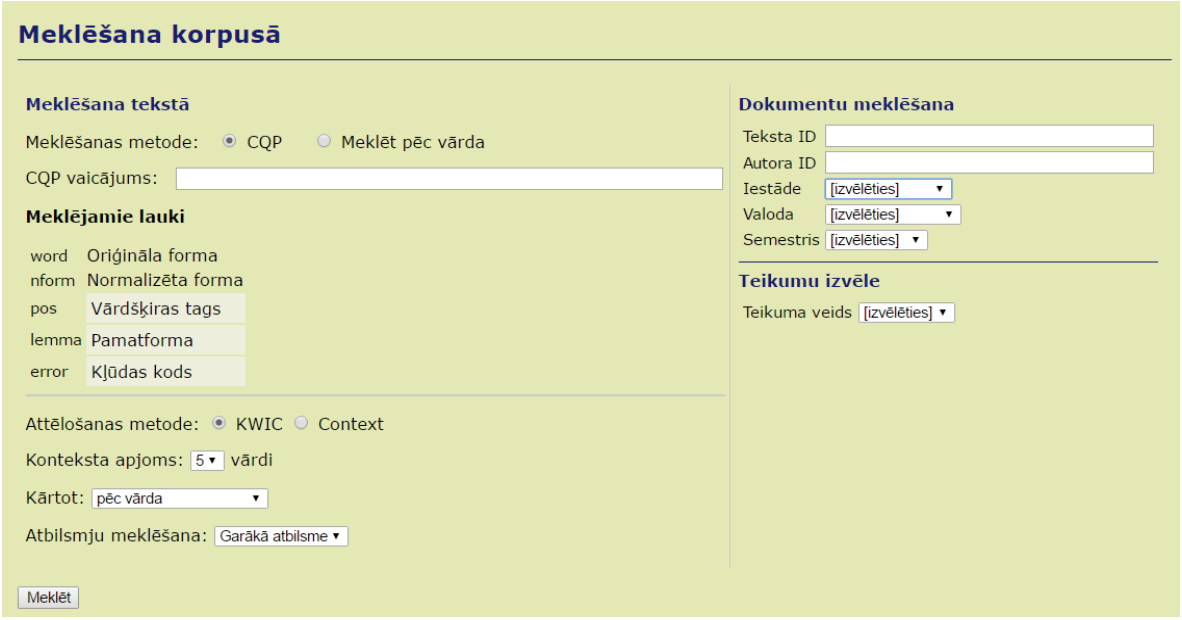
Līdzās meklēšanas laukam atrodas saite uz izvērsto meklēšanu. To nospiežot, lietotājs nonāk izvērstās meklēšanas sadaļā (sk. 12. attēlā 133. lpp.), kurā iespējams ne tikai ievadīt CQP sintaksei atbilstošu vaicājumu, bet arī izvēlēties vairākus papildu parametrus (teikuma veids, kļūdu anotējuma kods), kā arī mainīt to, kā tiek attēloti meklēšanas rezultāti.

Lietotājiem, kam trūkst zināšanu vai pieredzes darbā ar CQP vaicājumiem, ir pieejama arī cita veida saskarne: izvērstajā meklēšanā izvēloties meklēšanas metodi „Meklēt pēc vārda" (sk. 13. attēlā 133. lpp.), vaicājumu var veidot, izvēloties atbilstošo parametru un norādot, vai meklējamo vārdu attiecīgā parametra vērtībai ir pilnībā vai daļēji jāsakrīt ar attiecīgajā laukā ierakstīto simbolu virkni. Lietotājam izvēloties sakritības līmeni un ierakstot meklējamo simbolu virkni, programma pati automātiski izveido vaicājumu. Pārējo anotējuma un marķējuma veidu izvēle šajā skatā saglabājas nemainīga.

Jāpiebilst gan, ka CQP vaicājumu sniegtās iespējas ir plašākas, nekā pieejams sadaļā „Meklēšana pēc vārda" – piem., pareizi konstruējot vaicājumu, var ne tikai atrast visus fragmentus, kas pilnībā vai daļēji atbilst kādai simbolu virknei, bet arī visus tos, kas kādai simbolu virknei neatbilst; meklēšana pēc vārda šādu iespēju nepiedāvā. Tāpēc atkarībā no korpusa izmantošanas mērķa un vēlamā rezultāta tomēr var būt nepieciešams arī apgūt aizstājējzīmju un loģisko operatoru lietošanu.

Kad lietotājs ir ierakstījis meklējamos parametrus un izvēlējies iestatījumus, jāspiež poga „Meklēt” lapas apakšā.

12. attēls. Izvērstā meklēšana



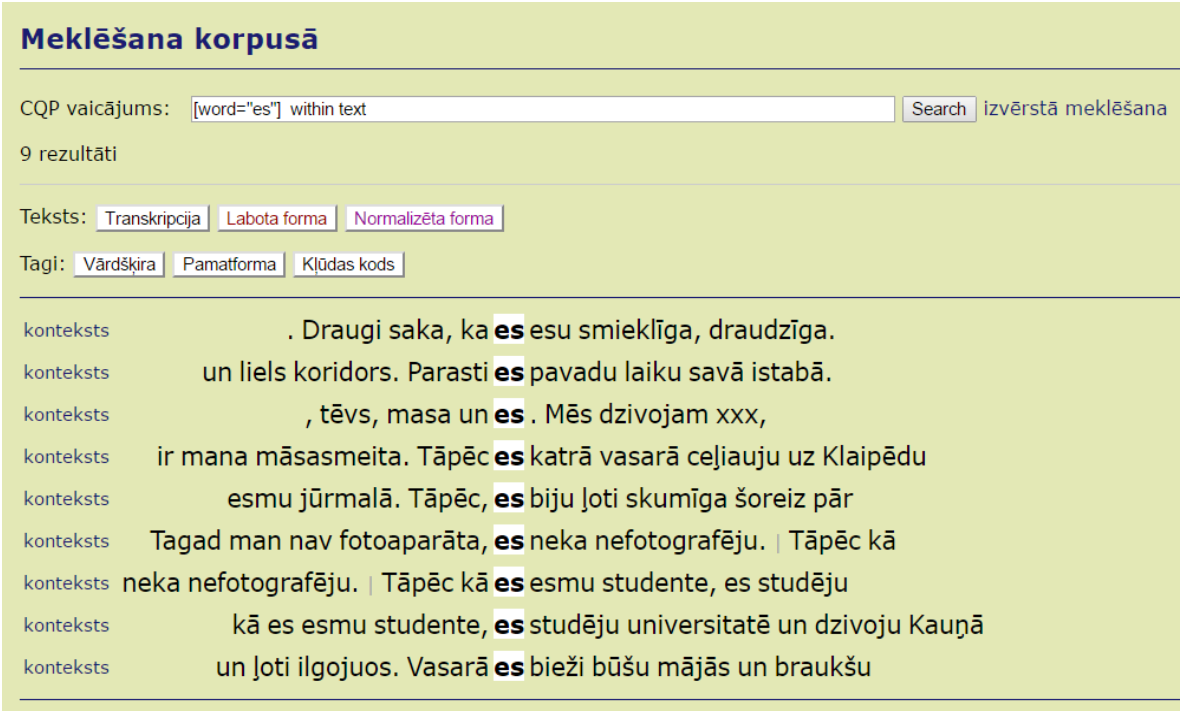
13. attēls. Meklēšana pēc vārda

Parādot meklēšanas rezultātus (sk. 14. attēlā[51] 134. lpp.), lapas augšdaļā tiek parādīts atrasto rezultātu skaits, iespēja izvēlēties transkribēto tekstu („Transkripcija”), laboto tekstu („Labota forma”) vai tekstu, kurā anonimizētās daļas ir aizstātas („Normalizēta forma”). Vēl

[51] Ekrānšāviņi uzņemti dažādos korpusa tapšanas posmos, tāpēc iegūstamie rezultāti var atšķirties no ekrānšāviņos redzamajiem rezultātiem.

šeit ir pieejama iespēja parādīt vārdšķiru, pamatformu un/vai kļūdu anotējumu katram vārdam. Lapas apakšpusē zem rezultātiem ir iespēja lejupielādēt rezultātus **.txt* formātā, kā arī saskaitīt vārdšķiru un/vai pamatformu biežumu rezultātos. Ir iespējams arī izveidot savu papildu CQP vaicājumu noteiktu īpatnību biežuma aprēķināšanai.

14. attēls. Meklēšanas rezultāti



Neatkarīgi no tā, kā lietotājs ir izvēlējies attēlot rezultātus, katru no rezultātiem ir iespējams aplūkot arī pilnā tekstā. Lai to izdarītu, ir jānospiež saite *konteksts* rezultāta kreisajā pusē. Šī saite atver pilnu tekstu, kurā ir attēlots konkrētais rezultāts, un attiecīgais teksta elements ir iekrāsots gaiši dzeltenā krāsā, lai būtu vieglāk pamanāms. Turpat iespējams arī norādīt, lai tiktu parādīts kāds anotējums (vārdšķiras, pamatformas). Zem teksta ir iespēja tekstu lejupielādēt pašreizējā izskatā **.txt* formātā, lejupielādēt pilnu **.xml* datni vai skatīt tekstu sadalītu pa izteikumiem.

TEITOK nav īpaši paredzēts kvantitatīvo aprēķinu veikšanai, jo programma ir izveidota neliela apjoma korpusiem, kuros kvantitatīvi pētījumi varētu būt ar relatīvi zemu ticamību. Tomēr meklēšanas funkcija uzrāda atrasto rezultātu skaitu, un, pareizi formulējot vaicājumu, ar to var iegūt dažādus kvantitatīvos datus. Kā jau norādīts, pēc meklējuma rezultātu iegūšanas ir iespējams iegūt to skaitu, grupējot pēc dažādiem kritērijiem, piem., pamatformas vai vārdšķiras.

Vaicājumu izveide ar regulāro izteiksmju palīdzību ir aprakstīta dažādos avotos (piem., Evert 2009), tomēr, lai atvieglotu darbu tiem pētniekiem, kuriem nav daudz pieredzes ar regulārajām izteiksmēm, 7. tabulā ir sniegti vaicājumu paraugi dažu vispārīgu kvalitatīvo un kvantitatīvo datu iegūšanai. Daži operatori, piemēram, VNT vai ML ir specifiski tieši šī korpusa anotācijai.

7. tabula. Vaicājumu paraugi

Vaicājums	Iegūstamā informācija
[] within text	Visas vienības tekstā (gan vārdi, gan pieturzīmes u.tml.).
[lemma=".+"] within text	Visi vārdi, kuriem ir norādīta pamatforma viena vai vairāku simbolu garumā.
[lemma!=".+"] within text	Visas vienības, kurām nav norādīta pamatforma (pieturzīmes, anonimizētās teksta vienības utt.).
<s_type = "VNT"> [] within text	Katra izteikuma, kas ir anotēts kā atbilstošs vienkārša nepaplašināta teikuma modelim, pirmais vārds. Dod iespēju saskaitīt viena veida izteikumus.
<s_type = ".*"> [] within text	Katra anotētā izteikuma pirmais vārds. Ļauj saskaitīt izteikumus korpusā.
[pos="v"] [error="ML"] within text	Visi gadījumi, kuros darbības vārdam seko kļūdaina locījuma vārdforma. Līdzīgi modelē vaicājumu dažādu citu vairākvārdu konstrukciju atrašanai.

Sīkāks programmas lietošanas apraksts ar ekrānšāviņiem ir sniegts korpusa lietošanas instrukcijā (6. pielikums; pagaidām pieejama tikai latviešu valodā).

Saite uz korpusa vietni ir ievietota programmas *TEITOK* mājaslapā un *TEITOK* projektu sarakstā (TEITOK Projects[-e]), kā arī Valodas apguvēju korpusu asociācijas mājaslapā publicētajā valodas apguvēju korpusu sarakstā (Dumont, Granger 2017).

2.4. Pētījumu iespējas otrās baltu valodas apguvēju korpusā

Korpusa noderīgumam ir vairāki aspekti. Pirmkārt, tas sniedz plašam pētnieku lokam iespēju izmantot materiālus, kas citādi ir pieejami tikai konkrētajam pasniedzējam, turklāt, izmantojot šo korpusu, nav nepieciešamas papildu rūpes par personas datu aizsardzību un autortiesības regulējošo normatīvo aktu ievērošanu. Otrkārt, korpuss sniedz jaunu veidu, kā meklēt un aplūkot noteiktas valodiskās parādības (kļūdas, vārdšķiras u.c.) kvalitatīvas analīzes nolūkos. Vēl korpuss ļauj veikt arī kvantitatīvu analīzi. Tālāk šajā nodaļā raksturots darbs ar korpusu „Esam".

Kā atzīst T. Makenerijs un Endrū Vilsons (*Andrew Wilson*), „korpusa lingvistika ir metodoloģija, kuru var izmantot gandrīz jebkurā valodniecības apakšnozarē" (McEnery, Wilson 2001, 2). Arī valodas apguvēju korpusos var pētīt ļoti dažādām valodniecības apakšnozarēm piederīgas parādības visos valodas līmeņos – no fonētikas līdz semantikas un, iespējams, pat teksta līmenim.

Jāpiebilst, ka korpuss nevar pierādīt kaut kā neesamību, tikai kaut kā esamību, tomēr, piem., kvantitatīvi skaitot, var rēķināt attiecības starp vārdiem, kuros ir noteikta tipa kļūdas un tāda paša veida vārdiem, kuros šo kļūdu nav. Jāatceras gan – ja tas tiek darīts, tad pētniekam ir jāpamato sava izvēle, vai aprēķinos iekļaut aizstātos personas datus vai nē.

Šajā nodaļā īsumā parādītas dažas otrās baltu valodas apguvēju korpusa lietošanas iespējas pētniecībā. Jāpiebilst, ka šeit nav aprakstīts darbs ar visām iespējamajām metodēm, jo to klāsts nav ierobežots. Tāpat arī metodes, kuras šeit ir skaidrotas, atkarībā no pētījuma mērķa var izmantot citādi, nekā šeit norādīts.

Tā kā korpuss „Esam" tiek pastāvīgi papildināts un promocijas darba izstrādes laikā vēl ir ļoti ierobežotā apjomā, šeit minētie rezultāti nebūtu uzskatāmi par pētījumam drošticamiem – drīzāk gan par ilustrāciju, kā pētnieks var izmantot korpusu, kad ir gūta pārliecība, ka konkrētā jautājuma aplūkošanai korpusā tekstu pietiek un tie ir atbilstoši.

2.4.1. Kontrastīvā starpvalodas analīze otrās baltu valodas apguvēju korpusā

Kā jau norādīts šī darba 1.3.2. apakšnodaļā „Valodas apguvēju korpusu izpētes metodes", kontrastīvajai starpvalodu analīzei ir divi veidi: valodas apguvēju un dzimtās valodas runātāju producēto valodas paraugu salīdzinājums un divu dažādu starpvalodas veidu salīdzinājums. Korpusā „Esam" iekļautie teksti pārstāv divu starpvalodas veidu paraugus, kas gan nav vienā valodā, tomēr ir salīdzināmi.

Piem., var aplūkot darbības vārda la. *būt*, lie. *būti* formu daudzveidību. Izvērstajā meklēšanā norādot, ka jāmeklē visi vārdi, kuru pamatforma ir *būt*, un atzīmējot, ka meklēšana jāveic tikai tekstos latviešu valodā[52], atrastajos rezultātos ir redzamas šādas formas (sk. 8. tabulā):

8. tabula. Darbības vārda būt *formas*

		Īstenības izteiksme			
		Tagadne	Pagātne	Nākotne	
Vienskaitlis	1. pers.	esmu – 2,7 (3,38 %), esu – 0,3 (0,42 %)	biju – 2 (2,53 %)	būšu – 0,3 (0,42 %)	
	2. pers.	–	–	–	
Vienskaitlis / daudzskaitlis	3. pers.	ir – 55 (69,62 %)	bija – 5,7 (7,17 %)	būs – 9,3 (11,81 %), bus – 0,3 (0,42 %)	
Daudzskaitlis	1. pers.	esam – 0,3 (0,42 %)	bijām – 0,3 (0,42 %), bijam – 0,7 (0,84 %)	–	
	2. pers.	–	–	–	
Vēlējuma izteiksme	–				būtu – 1 (1,27 %)
Nenoteiksme	–				būt – 1 (1,27 %)
KOPĀ					79 (100%)

Atbilstoši meklējot lietuviešu valodas paraugos darbības vārda *būti* formas, iegūstamie rezultāti ir šādi (sk. 9. tabulā 138. lpp.):

[52] Šajā gadījumā tas nav noteikti jādara, jo meklējamā pamatforma katrā valodā ir citāda, taču, ja tā sakrīt (piem., la. *diena*, lie. *diena*), valodu norādīt ir nepieciešams.

9. tabula. Darbības vārda būti *formas*

		Īstenības izteiksme			
		Tagadne	Pagātne	Nākotne	
Vienskaitlis	1. pers.	esu – 2,5 (9,05 %), esme (0,35 %), esmu (0,35 %)	buvau (6,26 %)	–	
	2. pers.	esi (0,17 %)	buvai (0,17 %)	–	
Vienskaitlis / daudzskaitlis	3. pers.	yra – 8,4 (30,09 %), ure (0,17 %), ir (0,17 %)	buvo – 6,6 (23,83 %)	bus (4,17 %), būs (0,87 %)	
Daudzskaitlis	1. pers.	esame (4,70 %)	buvome (1,39 %), buvame (0,17 %)	būsime (0,52 %)	
	2. pers.	esate (0,17 %)	–	–	
		Vēlējuma izteiksme			
Vienskaitlis	1. pers.	–			
	2. pers.	–			
Vienskaitlis / daudzskaitlis	3. pers.	būtų – 3,1 (11,30 %), būtu (0,17 %)			
Daudzskaitlis	1. pers.	–			
	2. pers.	–			
Divdabji		–			esą (0,35 %)
Nenoteiksme		–			būti (4,87 %), buti (0,35 %)
KOPĀ					27,6 (100 %)

Redzams, ka lietuviešu valodas datos ir vērojama lielāka formu dažādība, taču to lielā mērā ietekmē fakts, ka apskates brīdī ir anotēta daudz lielāka daļa lietuviešu valodas tekstu

nekā latviešu valodas tekstu, un arī atrasto rezultātu skaits ir ievērojami atšķirīgs (575 rezultāti lietuviešu valodā, 237 rezultāti latviešu valodā). Taču salīdzināt var lietojumu uz tūkstoš vārdiem (tabulās norādīts formas lietojuma biežums uz 1000 vārdiem korpusā, aprēķināts manuāli, dalot absolūto skaitu ar korpusā iekļauto atbilstošās valodas tekstu vārdu kopskaitu), kā arī noteiktas formas lietojuma attiecību pret citām lietotajām tā paša vārda formām (to parāda programma, kad rezultāti tiek grupēti pēc formas; tabulās norādīta procentuāli). Ja tiek nolemts veikt sīkākus statistiskus aprēķinus (piem., Hī-kvadrāta testu), tas jādara atsevišķi, jo, kā jau minēts, programma, nebūdama paredzēta galvenokārt kvantitatīviem pētījumiem, šādu iespēju nepiedāvā.

Šie dati var ļaut izvirzīt dažādas hipotēzes. Piem., daudz biežākais vārda *būt* lietojums latviešu valodā (79 uz 1000 vārdiem) pretstatā lietuviešu valodas tekstiem (27,6 uz 1000 vārdiem) varētu norādīt uz plašu piederības konstrukciju lietojumu (lietuviešu valodā piederību izsaka ar citu vārdu – *turėti*), salikto laiku formu lietojumu (piem., *ir rakstījis*), salīdzinoši biežāku nomināla izteicēja lietojumu (piem., *aš esu Inga*) vai citām parādībām. Tos pašus datus aplūkojot detalizētāk, var arī gūt atbildi, kura(-as) no šīm hipotēzēm ir atbilstošāka(-as). Tālāk var mēģināt spriest par to, kādi varētu būt atbilstošās parādības cēloņi (piem., kāpēc nomināls izteicējs attiecībā pret verbālu izteicēju ir sastopams biežāk), vai to, cik lielā mērā atbilstošā parādība otrās baltu valodas mācību procesā ir vērtējama pozitīvi vai negatīvi (piem., piederības izteikšanas gadījumā).

Ja pētniekam ir vēlme salīdzināt starpvalodas un dzimtās valodas parādības, tad atsevišķi ir jāizveido atbilstošs dzimtās valodas tekstu korpuss.

2.4.2. Datorizēta kļūdu analīze otrās baltu valodas apguvēju korpusā

Kā jau iepriekš norādīts, kļūdu analīze, lai arī nereti tiek kritizēta par neviendabīgu datu lietošanu un nekonsekventu kļūdu kategorizāciju, var būt noderīga, ja arī ne kā pilnīga vispusīga apguvēju valodas pētniecības metode, tad noteiktu parādību konstatēšanai gan. Tā kā korpusā ir labotas un anotētas kļūdas, šī anotējuma izmantošana paver dažādas iespējas.

Piem., var aplūkot, kādos gadījumos lietuviešiem, mācoties latviešu valodu, rodas grūtības ar pareiza locījuma lietošanu, meklējot pēc kļūdas koda ML un norādot latviešu valodu kā meklējamo tekstu valodu:

[error="ML"] :: match.text_lang = "LV" within text

Piemērs, kā izskatās iegūstamie rezultāti, ir redzams 15. attēlā (sk. 140. lpp.), savukārt izvēloties aplūkot teksta laboto formu, var pārliecināties par anotēšanas gaitā izvēlēto mērķa

hipotēzi katram no piemēriem, lai gadījumā, ja korpusa lietotājs savā pētījumā vēlas oponēt anotētāja viedoklim, varētu to izskaidrot un pamatot.

15. attēls. Tekstu fragmenti, kas ir anotēti kā piederīgi locījuma kļūdu apakštipam

konteksts	noteikts, taču gudrs.	**Mana**	mates vārds ir xxx.
konteksts	, ir brūnas. | Mana	**brāli**	vārds ir xxx. Viņam
konteksts	ir skolnieks, tagad mācās	**divpadsmitais**	klasē. Tam patīk futbols
konteksts	laiku savā istabā. Mana	**istabā**	ir maza. Viņš ir
konteksts	mana māsasmeita. Tāpēc es	**katrā**	vasarā ceļiauju uz Klaipēdu būt
konteksts	maza. Viņš ir oranža	**krāsas**	. Uz grīdas ir zils
konteksts	Es gribu pastāstīt par	**mana**	ģimeni un mājas. Es
konteksts	staigāju jūrmalā, zaglis nozag	**mana**	fotoaparātu un tālruņu. Fotoaparātā
konteksts	Viņa acis, kā	**manu**	, ir brūnas. | Mana
konteksts	pastāstīt par mana ģimeni un	**mājas**	. Es esmu xxx.
konteksts	un brāli. Man nav	**māsa**	taču ir daudz māsīcu
konteksts	acis ir zilas, kā	**mātes**	, viņam ir īsi māti
konteksts	lielā mūrā mājā. Pie	**mūsū**	arī dzivo veca vecmāmiņa,
konteksts	patīk adīt, klausīties	**mūzikas**	un dejat. Mātes mati
konteksts	man nav fotoaparāta, es	**neka**	nefotografēju. | Tāpēc kā es
konteksts	ir maza. Viņš ir	**oranža**	krāsas. Uz grīdas ir
konteksts	paklājs, kurš der pie	**sienu**	. Pie loga ir sarkanas

Aplūkojot iegūtos rezultātus, var izvirzīt dažādas hipotēzes par kļūdu cēloņiem, iespējām mācību procesā iemācīties izvairīties no tām. Piem., 15. attēlā redzams, ka samērā liela daļa locījuma kļūdu rodas, lietojot piederības vietniekvārdu *mans*. Tāpēc var izvirzīt hipotēzi, ka tam iemesls ir fakts, ka lietuviešu valodā atbilstošais piederības vietniekvārds *mano* ir nelokāms. Attiecīgi viens nākamā soļa piemērs varētu būt – aplūkot visus vietniekvārda *mans* lietojuma piemērus tekstos latviešu valodā un mēģināt atrast iemeslu, kāpēc tieši konkrētajos gadījumos radušās kļūdas.

2.4.3. Baltu starpvaloda

Šķiet, ka pirms „Esam" nav bijis valodas apguvēju korpusu, kurus varētu uzskatīt par divvirzienu korpusiem. Korpusā „Esam" tekstus ir iespējams atlasīt pēc to valodas, bet, pētniekam izvēloties tekstus **ne**atlasīt, bet skatīt visus kopā, var atklāt parādības, kas ir raksturīgas baltu valodu pratējiem, mācoties otro baltu valodu, taču nenodalot tos pēc konkrētas valodas. Tātad šajā gadījumā runa ir par baltu starpvalodu.

Termins *baltu starpvaloda* ir jauns. Līdz šim parasti tiek runāts vai nu par *mērķvalodas* starpvalodu (piem., angļu starpvaloda – starpvaloda, kas veidojas, mācoties angļu valodu; igauņu starpvaloda – starpvaloda, kas veidojas, mācoties igauņu valodu), vai par *dzimtās valodas – mērķvalodas* starpvalodu, ja tiek īpaši uzsvērta dzimtās valodas ietekme (piem., ķīniešu – angļu starpvaloda, angļu – franču starpvaloda).

Retāk tiek runāts par to, cik lielā mērā starpvaloda, kas rodas, apguvējam ar dzimto *valodu A* apgūstot *valodu B*, ir salīdzināma ar *valodas B – valodas A* starpvalodu. Šāds skatpunkts varētu būt īpaši noderīgs, pētot apguvēja dzimtajai valodai tuvi radniecīgu valodu apguves īpatnības.

Viena no iespējām, kā aplūkot korpusa „Esam" datus šādā griezumā, ir – neatlasīt datus pēc valodas, bet gan skatīt visa korpusa materiālu kopā kā starpvalodas paraugus, kuru autori ir baltu valodas runātāji, kas apgūst otro baltu valodu. Tas var būt noderīgi divos aspektos:

- pētot, kā valodas apguvējs iesācējs raksta tekstus valodā, par kuru ir zināms, ka tā ir līdzīga apguvēja jau pārvaldītai valodai;
- atklājot un aprakstot līdzības un atšķirības starp abām baltu valodām kopumā.

Līdzīgi varētu būt lietderīgi runāt ne tikai par baltu starpvalodu, bet arī par, piem., ģermāņu starpvalodu, slāvu starpvalodu utt., lai runātu par starpvalodu, kas veidojas, kad apguvējs mācās jaunu valodu, kas ir tuvi radniecīga viņa dzimtajai valodai vai, iespējams, valodai, kuras prasmes konkrētajam apguvējam ir samērā augstā līmenī un kura tiek izmantota kā starpniekvaloda mācību procesā.

Piem., baltu valodu aplūkojot kopumā, var atlasīt visus izteikumus, kas ir anotēti kā atbilstoši vienkārša nepaplašināta teikuma modelim. Lai to izdarītu, meklēšanas ailē jāievada šāds vaicājums:

<s_type = "VNT"> [] within text

Pēc šāda vaicājuma atrastie rezultāti ir katra atbilstoši anotētā izteikuma pirmais vārds, un līdz ar to var aplūkot arī katru atbilstošo izteikumu (rezultātu skatījuma paraugu sk. 16. attēlā 142. lpp.).

Kā redzams attēlā redzamajos rezultātos, lielākoties vienkārša nepaplašināta teikuma modelī izmantots nomināls izteicējs. Līdz ar to, balstoties šādos datos, var izvirzīt hipotēzi, ka baltu starpvalodai nepaplašinātos teikumos nomināls izteicējs ir vairāk raksturīgs nekā verbāls vai adverbiāls izteicējs. Salīdzinot datus ar citur iegūstamiem latviešu un lietuviešu valodas kā dzimtās valodas datiem, var spriest, vai tā ir baltu valodām kopumā raksturīga īpatnība vai arī

drīzāk piemīt tieši starpvalodai. Šo informāciju var salīdzināt arī ar citu valodu apguvēju korpusos atrodamajiem datiem, lai kontrastīvās starpvalodas analīzes ceļā noskaidrotu, vai tā ir universāla parādība vai arī raksturīga tieši baltu valodām.

16. attēls. Vienkārša nepaplašināta teikuma modelim atbilstoši izteikumi

konteksts	mates vārds ir xxx.	**Māmiņa**	ir maza, smalka un
konteksts	Viņas raksturs ir labs.	**Viņa**	ir godīga, bikla,
konteksts	per mūsu dzīvī Lietuvā.	**Virtuve**	nav liela, bet
konteksts	un padzirdējās durvju zvans.	**Draudzenes**	! Ķērās pie manis,
konteksts		**Es**	esmu xxx. Es esmu
konteksts	Mans tēvs ir xxx.	**Viņš**	ir inženieris. Viņš strādā
konteksts	Mana māte ir xxx.	**Viņa**	ir xxx. Viņa strādā
konteksts	Mans brālis ir xxx.	**Viņš**	ir skolnieks.
konteksts	braucām atpakaļ uz Lietuvu.	**Bijām**	noguri, bet apmierināti.
konteksts	. Mani sauc xxx.	**Es**	esmu studente. Es studēju
konteksts	ir motina gyvena xxx.	**Jiē**	dirba. Tėtė vardù xxx
konteksts	yra labai gerai senelė.	**Ji**	yra didinga. Senelė mezga
konteksts	pietavome parke prie upės.	**Diena**	buvo šilta ir saulėta.

Savukārt lai atklātu atšķirības starp abām baltu valodām, īpaši lietderīgs ir kļūdu anotējums. Piem., meklējot pēc kļūdas tipa *LV* (leksikas tips, saderības apakštips), var atklāt, kādās konstrukcijās valodas apguvēji neiederīgi lietojuši kādu konkrētu vārdu, piem.:

Mana māsa dzīvo ne [nevis] *Kupišķos, bet Klaipēdā* [..]

Šāds piemērs atklāj atšķirību vārda *ne* lietojumā latviešu un lietuviešu valodā. Lietuviešu valodā tas tiek lietots, lai izteiktu pretstatu saikļos *ne... bet* un *ne... o* (sk. arī Bielinskienė 2003, 99), savukārt latviešu valodā šādā konstrukcijā tiešā pārcēluma *ne... bet* vietā ierastāk lietot *nevis... bet*, līdz ar to pamatu arī atšķirībām starp starpvalodu un mērķvalodu.

Kā redzams piemēros, baltu starpvalodas pētīšana nav atsevišķa metode, drīzāk gan aspekts, kādā ir iespējams aplūkot korpusa datus, izmantojot citas metodes, ieskaitot jau nosauktās.

Nobeigums

Promocijas darba mērķis – izveidot otrās baltu valodas apguvēju korpusu un aprakstīt tā izveidi – ir sasniegts, veicot izvirzītos uzdevumus.

Promocijas darba pirmajā daļā ir raksturots valodas apguvēju korpusa jēdziens, šādu korpusu veidi un akadēmiskajā literatūrā sniegtā informācija par valodas apguvēju korpusa izveidi un lietošanu. Lai arī valodas apguvēju korpusi ir tikai viens valodas korpusu paveids, tas pēdējos gados strauji gūst popularitāti. Tomēr šādu korpusu veidošana un pētniecība joprojām ir fragmentēta. Tas redzams arī Latvijā un Lietuvā – abās valstīs darbs ar šādiem korpusiem galvenokārt notiek individuāli, nesadarbojoties ar kaimiņvalstu pētniekiem.

Darbā ar valodas apguvēju korpusiem izmantojamas dažādas korpuslingvistikā pazīstamas procedūras, tās izvēloties atbilstoši pētījuma metodei. Īpaši valodas apguvēju korpusu izpētē pazīstamas divas metodes: kontrastīva starpvalodas analīze un datorizēta kļūdu analīze. Otrās baltu valodas apguvēju korpusā ir iespējams izmantot tās abas.

Veidojot otrās baltu valodas apguvēju korpusu, ar divu Latvijas universitāšu un divu Lietuvas universitāšu pasniedzēju palīdzību ir savākti, raksturoti, marķēti un ievietoti korpusā otrās baltu valodas 1. un 2. semestra studentu rakstīti teksti mērķvalodā. Ir nodrošināts tehniskais risinājums, kas sniedz pieeju korpusam tiešsaistē ikvienam, kas piekrīt tā lietošanas nosacījumiem. Darba gaita ir aprakstīta promocijas darba 2. daļas 1. un 3. nodaļā.

Izveidotais korpuss ir neliels (nedaudz vairāk par 50 000 vārdlietojumiem), taču uzskatāms par reprezentatīvu, ņemot vērā, ka arī otrās baltu valodas apguvēju skaits kopumā nav liels. Lai nākotnē paplašinātu korpusa apjomu, būtu vēlams iesaistīt pasniedzējus un studentus arī no pārējām izglītības iestādēm, kurās tiek apgūta otrā baltu valoda.

Korpusa apjomu ietekmē arī autortiesību un personas datu aizsardzības jautājumi. Tie visvairāk apgrūtina korpusa izveidi tad, ja teksti korpusā tiek iekļauti, kad kopš teksta tapšanas brīža ir pagājis ilgāks laiks: nepieciešams atrast veidu, kā sazināties ar autoru un iegūt atļauju. Nereti autora kontaktinformācija ir grūti atrodama vai arī autors neuzticas svešam cilvēkam, kas viņu uzrunājis sociālajā tīklā. Ja atļauju lūdz pasniedzējs uzreiz pēc teksta tapšanas, process ir abām pusēm vienkāršāks un ērtāks.Vēl iespējams, jau uzdodot uzdevumu rakstīt tekstu, norādīt, ka tas tiks iekļauts korpusā, un informēt autorus par nosacījumiem. Šādā gadījumā gan var raisīties diskusijas par to, vai teksta tapšanu (vārdu, konstrukciju izvēli u. tml.) varētu būt ietekmējusi apziņa, ka teksts būs publiski pieejams.

Korpuss ir anotēts, daļēji izmantojot vai pielāgojot jau esošas pazīmju kopas, kas ir tikušas izmantotas arī citos baltu valodu tekstu korpusos. Morfoloģiskajai anotēšanai izmantota

jau esoša pazīmju kopa, taču sintaktiskajai anotēšanai tā pielāgota atbilstoši mūsdienu izpratnei par teikumu un tā veidiem, lai atbilstoši anotētu arī izsacījumus, parcelātus, tiešo runu. Kļūdu anotēšanai izveidotā klasifikācija veidota no jauna, balstoties esošos citu valodu kļūdu klasifikācijas paraugos. Darba gaita ir aprakstīta promocijas darba 2. daļas 2. nodaļā.

Kļūdas jēdziens nav viennozīmīgs. Korpusu anotējot, tekstu labotājs katrā atsevišķā gadījumā izvirza mērķa hipotēzi, un par kļūdu korpusā uzskatāma neatbilstība mērķa hipotēzei. Korpusa anotēšanas nolūkos izveidotā kļūdu anotēšanas klasifikācija veidota pēc iespējas efektīva un visaptveroša, lai būtu izmantojama ne tikai šajā korpusā, bet dažādos baltu valodu apguvēju korpusos.

Anotējot iesācēju rakstītos tekstus, kļūdu dēļ ir diezgan daudz gadījumu, kuros rodas grūtības saprast, ko autors ir vēlējies teikt. Tas arī apgrūtina automātisku anotēšanas rīku izmantošanu – šeit tādi rīki izmantoti tikai kā palīglīdzekļi atsevišķos gadījumos. Lai nesarežģītu anotēšanas procesu, katrā gadījumā izvēlēts šķietami ticamākais variants, taču par to var turpināties diskusijas un, ja tajās tiktu secināts, ka sākotnējā izvēle nav bijusi atbilstoša, to var labot.

Korpusa veiktspējas nodrošināšanai izvēlēta programma *Teitok*, kas ir īpaši izstrādāta nelieliem anotētiem korpusiem. Programmas izvēlē un saskarnes izveidē viens no galvenajiem faktoriem ir lietošanas ērtums, lai korpusa lietošanā nebūtu nepieciešamas augsta līmeņa datorprasmes vai specifiskas zināšanas.

Promocijas darba 2. daļas 4. nodaļā ir īsumā aprakstītas būtiskākās korpusa sniegtās iespējas un lietošanas veidi. Tas gan nav uzskatāms par pilnīgu un ierobežojošu aprakstu, jo korpusu akadēmiskos nolūkos drīkst izmantot ar jebkādām metodēm. Ierobežotais apjoms gan nozīmē, ka ne vienmēr iespējams iegūt pietiekami daudz datu par interesējošo jautājumu, lai izdarītu ticamus secinājumus. Korpusa atbilstība katram atsevišķam pētījuma jautājumam jāatstāj attiecīgā pētnieka ziņā.

Runājot par korpusa lietojumu, jāņem vērā, ka tas primāri ir pētījumu materiāls, nevis pedagoģiskas ievirzes rīks, un tā tiešais pedagoģiskais izmantojums ne visiem otrās baltu valodas pasniedzējiem var šķist iederīgs vai nepieciešams. Tomēr šī korpusa materiāla izpētes rezultāti var sniegt papildu zināšanas, kas palīdzētu veidot atbilstošākus mācību materiālus un ietekmēt mācību procesu pastarpināti. Savukārt pasniedzēji šajā procesā var sniegt būtisku ieguldījumu, turpinot vākt tekstus ievietošanai korpusā. Tātad ļoti liela nozīme ir pētnieku un pasniedzēju savstarpējai sadarbībai – ieguvējas būtu abas puses.

Korpusa izveidē ņemtas vērā citu valodas apguvēju korpusus veidojušu pētnieku atziņas, taču ir ieviesti arī jaunājumi. Promocijas darbā tiek runāts par diviem jauniem

jēdzieniem, kurus būtu vērts īpaši uzsvērt. Viens no tiem ir *divvirzienu valodas apguvēju korpuss*. Virziens kā korpusa raksturojuma parametrs jau līdz šim ir bijis pazīstams darbā ar paralēlajiem, proti, tulkojumu korpusiem, taču, ņemot vērā, ka arī valodas apguvē par ļoti svarīgu faktoru ir uzskatāma ne vien apgūstamā valoda, bet arī autora dzimtā valoda, šajā darbā dzimtās valodas un mērķvalodas opozīcija tiek uzskatīta par virzienu. Atbilstoši par divvirzienu valodas apguvēju korpusu tiek uzskatīts tāds korpuss, kurā iekļauto tekstu autoru dzimtās valodas ir pārstāvētas arī kā tekstu mērķvalodas, un otrādi – tekstu mērķvalodas ir pārstāvētas arī kā tekstu autoru dzimtās valodas.

No tā izriet arī otrs piedāvātais jaunais jēdziens – *baltu starpvaloda*. Strādājot ar divvirzienu korpusu, rodas likumsakarīgs jautājums, kā būtu saucams tā materiāls. Līdz ar to tiek piedāvāts starpvalodu, kas rodas, vienas baltu valodas runātājam apgūstot otru baltu valodu (latvietim – lietuviešu valodu, lietuvietim – latviešu valodu), saukt par baltu starpvalodu. Tā kā abas valodas ir tuvi radniecīgas, ir ticams, ka šai starpvalodai varētu būt citādas īpatnības nekā starpvalodai, kas rodas, kādu no baltu valodām apgūstot cilvēkam, kura dzimtā valoda nav baltu valodām piederīga.

Pētījuma turpinājumā kā primāra noteikti būtu jāuzsver otrās baltu valodas apguvēju mērķvalodas producēšanas īpatnību pētniecība. Tieši ar šādu nolūku korpuss ir ticis veidots. Atklājot parādības, kas valodas apguvējiem sagādā grūtības vai – gluži otrādi – nemēdz tās izraisīt, atbilstoši var izstrādāt vai uzlabot esošos mācību materiālus, vārdnīcas, mācību metodes u. c. Vērtīgi būtu arī pētījumi par korpusa datu tiešu izmantošanu valodu mācīšanas un mācīšanās procesā.

Visbeidzot jāuzsver, ka korpuss ir noderīgs tikai tikmēr, kamēr akadēmiskajā vidē ir interese par otro baltu valodu un tās apguvi. Pēdējā laikā tendences otrās baltu valodas izvēlē augstākās izglītības iestādēs nav īpaši iepriecinošas – vairākās mācību iestādēs tā tiek pasniegta arvien retāk. Cerams, ka otrās baltu valodas apguvēju korpusa izveide varētu veicināt otrās baltu valodas popularitātes atgriešanos gan studentu, gan pētnieku lokā.

Promocijas darba aizstāvēšanai izvirzītās tēzes

Promocijas darbā izpētītais un secinātais ļauj aizstāvēšanai izvirzīt šādas tēzes:

1. Valodas apguvēju korpusi gūst popularitāti daudzviet pasaulē, taču Latvijā un Lietuvā to lietojums šobrīd nav plaši izplatīts un ir samērā fragmentārs. Šeit izveidotie korpusi lielākoties ir pieejami tikai to veidotājiem, vai arī tiek izmantoti citviet tapušie valodas apguvēju korpusi. Būtu ieteicama plašāka sadarbība, veidojot un pētot dažādu valodu apguvēju korpusus ar latviešu un lietuviešu valodu kā dzimto valodu, it īpaši savstarpēji salīdzināmus valodas apguvēju korpusus. Ņemot vērā baltu valodu savstarpējo radniecību un līdzību, līdzīgais vai – gluži pretēji – atšķirīgais starpvalodu datos var palīdzēt izprast katras baltu valodas kā dzimtās valodas ietekmi uz citas valodas apguves procesu.
2. Promocijas darba gaitā ir izveidots tiešsaistē publiski pieejams otrās baltu valodas apguvēju korpuss un īsi raksturota tā lietošana pētījumos. Tas ir pirmais publiski pieejamais valodas apguvēju korpuss Latvijā un Lietuvā, pirmais publiski pieejamais baltu valodu apguvēju korpuss un pirmais divvirzienu valodas apguvēju korpuss. Divvirzienu korpusa jēdziens šajā darbā tiek piedāvāts pirmoreiz, ar to saprotot tādu korpusu, kurā iekļauti teksti divos valodas apguvēju pāros un katra no valodām vienā pārī ir dzimtā valoda, bet otrā – mērķvaloda.
3. Korpuss ir anotēts četros līmeņos: sintaktiskā, morfoloģiskā, leksiskā, kā arī tajā ir anotētas kļūdas. Korpusa sintaktiskai, morfoloģiskai un leksiskai anotēšanai ir izmantotas jau iepriekš pētnieku izstrādātas klasifikācijas. Kļūdu anotēšanai ir izstrādāta jauna klasifikācija, balstoties S. Greindžeres izveidotajā klasifikācijā franču valodas apguvēju tekstu anotēšanai un pielāgojot to baltu valodām un konkrētā korpusa vajadzībām. Par kļūdu šādos tekstos ir uzskatāma atšķirība no mērķa hipotēzes, ko izvirza labotājs ar mērķvalodu kā dzimto valodu. Kļūdas anotējamas arī tādos aspektos, kādus attiecīgajā līmenī vēl neapgūst vai apgūst daļēji (piem., A līmenī – interpunkcija, darbības vārda saliktie laiki u. c.).
4. Runājot par starpvalodu, kas rodas, vienas baltu valodas runātājam apgūstot otru baltu valodu, tiek piedāvāts lietot terminu *baltu starpvaloda*; šo starpvalodu pētīt kā kopumu ļauj divvirzienu korpusa uzbūve. Ja turpmākos pētījumos tiktu atklātas īpatnības, kas ievērojami atšķir baltu starpvalodu no starpvalodas, kas rodas, citu valodu runātājiem apgūstot baltu valodas, tas varētu palīdzēt raksturot baltu valodas sastatījumā ar citām valodām.

5. Turpinot pētījumu, vērtīga būtu šajā korpusā pieejamā materiāla papildināšana ar jauniem tekstiem un/vai vēl citus valodas prasmes līmeņus pārstāvošiem tekstiem, lai veicinātu valodas prasmes attīstības izpēti. Noderīgi būtu arī izveidot salīdzināmu korpusu, kurā būtu atbilstoša līmeņa teksti, kuru autoriem nav nevienas baltu valodas priekšzināšanu. Turpmākie pētījumi šajā jomā būtu saistāmi ar diviem virzieniem:
    a. pētījumi otrās baltu valodas apguvēju korpusā – izmantojot šo korpusu, būtu atklājamas baltu starpvalodas īpatnības; tālāk šādu pētījumu rezultāti būtu izmantojami otrās baltu valodas mācību līdzekļu, baltu valodu vārdnīcu izstrādē un pilnveidē u. c.;
    b. papildu datu ievietošana korpusā (piem., augstāku valodas prasmes līmeni atspoguļojoši teksti vai mutvārdu teksti) vai jaunu salīdzināmu korpusu veidošana (piem., baltu valodu apguvēju korpuss, kurā informanti būtu dažādu citu, ne baltu, valodu runātāji).

Pateicība

Promocijas darba autore sirsnīgi pateicas visiem, kas ir palīdzējuši un atbalstījuši darba tapšanas gaitā.

Pirmkārt, liels paldies darba vadītājai Ilzei Auziņai, kura ļāva nepazīstamam *fruktam* sevi pierunāt uz avantūru vairāku gadu garumā. Paldies arī visiem doktora studiju programmas pārstāvjiem par pacietību un atbalstu.

Paldies otrās baltu valodas pasniedzējiem un studentiem par sadarbību un entuziasmu – bez jums šis darbs nebūtu varējis tapt!

Paldies kampaņas „Misija: Lenkastera 2015" vadītājai Dacei Znotiņai un atbalstītājiem, it īpaši Robertam Brežģim, Diānai Kazinai, Artūram Klaivam, Ojāram Krūmiņam, Ievai Ozolai, Zitai Priedītei, Anitai Rašmanei, Laimai Sadovičai, Maijai Strautmanei, Mārītei Znotiņai un Rutai Znotiņai.

Paldies doktorantūras cīņubiedriem, it īpaši Ingai Laizānei, par nerimstošu atbalstu un draudzīgu sacensību.

Paldies Harijam Kaijam, manai ģimenei un visiem pārējiem palīgiem un atbalstītājiem. Negaidīju, ka jūsu būs tik daudz!

Saīsinājumi

ADTAI – Lietuvas Republikas Personas datu tiesiskās aizsardzības likums

AL – Latvijas Republikas Autortiesību likums

ATGTI – Lietuvas Republikas Autortiesību un blakustiesību likums

FPDAL – Latvijas Republikas Fizisko personu datu aizsardzības likums

la. – latviešu valoda

LCA – Valodas apguvēju korpusu asociācija

lie. – lietuviešu valoda

LVASA – Latvijas Valodu skolotāju asociācija

Tabulas un attēli

Tabulas

Attēli

Literatūra

1. **Abel u. c. 2014a – Abel, Andrea, Wisniewski, Katrin, Nicolas, Lionel, Boyd, Adriane, Hana, Jirka, Meurers, Detmar.** A Trilingual Learner Corpus Illustrating European Reference Levels. *Ricognizioni.* Vol. 1, No. 2, 2014, pp. 11–126.
2. **Abel u. c. 2014b – Abel, Andrea, Glaznieks, Aivars, Nicolas, Lionel, Stemle, Egon.** KoKo: An L1 Learner Corpus for German. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).* Calzolari, Nicoletta et al. (eds). Reykjavik : European Language Resource Association, 2014, pp. 2414–2421.
3. **ACTFL 2015** – *ACTFL Proficiency Guidelines 2012.* Alexandria : ACTFL, 2012.
4. **ADTAI –** *Asmens duomenų teisinės apsaugos įstatymas* [skatīts 2014. gada 5. oktobrī]. Pieejams: *http://www3.lrs.lt/pls/inter3/dokpaieska.showdoc_l?p_id=400103*
5. **Aijmer 2011 – Aijmer, Karin**. Valodas apguvēju korpuss – tā veidošana un izmantošana valodu apguvē, mācību materiālu izveidē. *Konferences „Valoda. Izglītība. Tehnoloģijas" materiāli* [tiešsaiste]. Rīga : LVA, 2012 [skatīts 2014. gada 2. aprīlī]. Pieejams: *http://www.bilingvals.lv/uploads_docs/Karin_Aijmer_PPT_ 1318592822.ppt*
6. **Akadterm-e –** Akadēmiskā terminu datubāze AkadTerm [skatīts 2016. gada 19. jūnijā]. Pieejams: *http://termini.lza.lv/term.php*
7. **AL –** *Autortiesību likums* [skatīts 2015. gada 16. martā]. Pieejams: *http://likumi.lv/doc.php? id=5138*
8. **Altay, Tilfarlioğlu 2012 – Altay, Mehmet, Tilfarlioğlu, Filiz Yalçin.** Building up a Learner Corpus Through Creative Nonfiction Prose: An Experimental Research. *Electronic Journal of Social Sciences,* Vol. 11, No. 39, Winter 2012.
9. **Andronova 2009 – Andronova, Everita.** Mūsdienu latviešu valodas korpuss un tā izmantošana. CLARIN praktiskā semināra „Mūsdienu latviešu valodas korpuss un tā izmantošana" (4.-5.02.2009.) materiāli [tiešsaiste]. Rīga : b. i., 2009 [skatīts 2014. gada 2. aprīlī]. Pieejams: *http://www.clarin.lv/materiali/Latviesu_valodas_ korpuss050209.ppt*
10. **Andronova, Andronovs 2011 – Andronova, Everita, Andronovs, Aleksejs.** Latviešu valodas korpuss un tā izmantošana. *Valodas prakse: vērojumi un ieteikumi.* Populārzinātnisku rakstu krājums. O. Bušs (red.). Nr. 6. Rīga : Latviešu valodas aģentūra, 2011, 41.–57. lpp.
11. **Anthony 2014 – Anthony, Lawrence.** *AntConc (Windows, Macintosh OS X, and Linux).* Tokyo : Waseda University, 2014 [skatīts 2015. gada 8. novembrī]. Pieejams: *http://www.laurenceanthony.net/software/antconc/releases/AntConc343/help.pdf*
12. **Arhire 2011 – Arhire, Mona.** Verb plus verbal noun collocations in a translational learner corpus. *Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies,* Vol. 6, Issue 55, No.1, 2013, pp. 65–70.
13. **Aston 2011 –** Applied Corpus Linguistics and the learning experience : An interview with Guy Aston. Viana, Vander, Zyngier, Sonia, Barnbrook, Geoff (eds.). *Perspectives on corpus linguistics*. Amsterdam, Philadelphia : John Benjamins Publishing Company, 2011, pp. 1–16.
14. **ATGTI –** *Autorių teisių ir gretutinių teisių įstatymas* [skatīts 2015. gada 17. martā]. Pieejams: *http://www3.lrs.lt/pls/inter2/dokpaieska.showdoc_l?p_id=471807*
15. **Atkins 1992 – Atkins, Sue, Clear, Jeremy, Ostler, Nicholas.** Corpus Design Criteria. *Literary and Linguistic Computing* Vol. 7/1, 1992, pp. 1–16.

16. **Auziņa u. c. 2015 – Auziņa, Ilze, Darģis, Roberts, Rābante-Buša, Guna.** Fonētiski marķēts latviešu valodas runas korpuss. *XII Starptautiskais baltistu kongress Viļņas Universitātē 2015. gada 28.–31. oktobrī : Referātu tēzes.* Viļņa : Viļņas Universitāte, 2015, 151. lpp. [skatīts 13.12.2015.]. Pieejams: *http://www.baltistukongresas.flf.vu.lt/failai/XII_Tarptautinio_baltistu_kongreso_tezes.pdf*
17. **Baker 2006 – Baker, Paul.** *Using Corpora in Discourse Analysis*. London, New York : Continuum, 2006.
18. **Baker u. c. 2006 – Baker, Paul, Hardie, Andrew, McEnery, Tony.** A Glossary of Corpus Linguistics. Edinburgh : Edinburgh University Press, 2006.
19. **Bankava, Vinčela 1999 – Bankava, Rota, Vinčela, Zigrīda.** Integration of New Technologies into Modern Language Studies. *Latvijas Universitātes 80 gadu jubilejai veltītās 57. konferences materiāli*. Latvijas Universitātes zinātniskie raksti, 622. sēj. Rīga : Latvijas Universitāte, 1999, 38.–44. lpp.
20. **Barlow 2005 – Barlow, Michael.** Computer-based analyses of learner language. *Analysing Learner Language*. Rod Ellis, Gary Barkhuizen. Oxford : Oxford University Press, 2005.
21. **Barnbrook 2008 – Barnbrook, Geoff.** Uncovering the Secret Life of Language. *The Third Baltic Conference on Human Language Technologies, October 4–5, 2007 : Proceedings*. Vilnius : Vytautas Magnus University, Institute of the Lithuanian Language, 2008, pp. 23–32.
22. **Bārzdiņš u. c. 2007 – Bārzdiņš, Guntis, Grūzītis, Normunds, Nešpore, Gunta, Saulīte, Baiba.** Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu : Association for Computational Linguistics, 2007, pp. 13–20.
23. **Benz, Newman 2008 – Benz, Carolyn R., Newman, Isadore.** *Mixed Methods Research: Exploring the Interactive Continuum*. Carbondale : Southern Illinois University Press, 2008.
24. **Bermingham, Smeaton 2009 – Bermingham, Adam, Smeaton, Alan F.** A Study of Inter-Annotator Agreement for Opinion Retrieval. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Boston : ACM, 2009.
25. **Bernardini u. c. 2003 – Bernardini, Silvia, Stewart, Dominic, Zanettin, Federico.** An Introduction. *Corpora in Translator Education*. Federico Zanettin, Silvia Bernardini, Dominic Stewart (eds.). Abingdon, New York : Routledge, 2003, pp. 1–13.
26. **Bērziņa 2013 – Bērziņa, Agnese.** *Teikumu konstrukciju variācijas angļu valodas apguvēju rakstu valodā.* Bakalaura darbs. Rīga : Latvijas Universitāte, 2013.
27. **Biber u. c. 2006** – **Biber, Douglas, Conrad, Susan, Reppen, Randi.** *Corpus Linguistics.* Investigating Language Structure and Use. Cambridge : Cambridge University Press, 2006.
28. **Bielinskienė 2003 – Bielinskienė, Agnė.** Iš jungtukų *bet* ir *o* vartosenos. *Kalbos kultūra 76,* 98–102 psl.
29. **Bikelienė 2008a – Bikelienė, Lina.** Connector usage in advanced Lithuanian Learners' English Writing. *Corpus Linguistics, Computer Tools, and Applications – State of the Art.* B. Lewandowska-Tomaszczyk (ed.), Frankfurt :Peter Lang, 2008, pp. 741–755.

30. **Bikelienė 2008b – Bikelienė, Lina.** Resultive Connectors in Advanced Lithuanian Learners' English Writing. *Linguistics: Germanic and Romance Studies* (Kalbotyra: Germanų ir romanų studijos), Vol. 59, No. 3), 2008, pp. 30–37.
31. **Bikelienė 2009a – Bikelienė, Lina.** Insights from the Lithuanian Learner Corpus of English: Pilot Study on the Use of Resultive Connectors. *Language Forum: An International Journal of Language and Linguistics*, Vol. 35, No.2, 2009, pp. 113–126.
32. **Bikelienė 2009b – Bikelienė, Lina.** Priešpriešos konektorių vartojimas besimokančių anglų kalbos ir anglakalbių studentų rašto darbuose. *Kalbotyra*, Vol. 61, No. 3, 2009, 21.–35. psl.
33. **Bikelienė 2010 – Bikelienė, Lina.** Summative Connectors in Lithuanian Learners' and Native Speakers' English Essays. *Proceedings of the International Scientific Conference „Man in the Space of Language", Kaunas, Lithuania, 14-15 May 2010*, Vol. 6, pp. 546–552.
34. **Bikelienė 2012 – Bikelienė, Lina.** *Connector usage in native and non-native learner's English writing. Contrastive analysis.* Summary of doctoral dissertation. Vilnius : Vilnius University, 2012.
35. **Bikelienė 2013 – Bikelienė, Lina.** Sentence Initial Additive Linking Words in Lithuanian Learners' Language and British English. *Anglistics in Lithuania: Cross-Linguistic and Cross-Cultural Aspects of Study*. Newcastle upon Tyne : Cambridge Scholars Publishing, 2013, pp. 198–208.
36. **Bikelienė 2015a – Bikelienė, Lina.** Lithuanian Learners' English: British or American? *Verbum*, No. 16, 2015, pp. 29–40.
37. **Bikelienė 2015b – Bikelienė, Lina.** Person markers in non-native students' writing: A study of Lithuanian learner English. *Third international learner corpus research conference, 11–13 September 2015*. Book of abstracts. Nijmegen : Radboud University, 2015, pp. 28–30.
38. **Bikelienė 2016 – Bikelienė, Lina.** Evaluative Adjectives in Lithuanian and Native Students' English Writing. *Kalba ir kontekstai*, t. VII (1), 1 dalis, 2016, 197–206 psl.
39. **Blažek 2007 – Blažek, Václav.** From August Schleicher to Sergei Starostin: On the development of the tree-diagram models of the Indo-European languages. *Journal of Indo-European Studies*, No. 35 (1-2), 2007, pp. 82–109.
40. **Bohát u. c. 2015 – Bohát, Róbert, Horáková, Nina, Rödlingová, Beata.** Building COHAT: Corpus of High-School Academic Texts. *Corpus linguistics 2015.* Abstract Book. Federica Formato, Andrew Hardie (eds). Lancaster : UCREL, 2015, pp. 378–379.
41. **Boulton 2011 – Boulton, Alex.** What data for data-driven learning? *Abstracts for conference of European Association for Computer-Assisted Language Learning (EUROCALL).* Nottingham : University of Nottingham, 2011, pp. 23–27.
42. **Bowker, Bennison 2003 – Bowker, Lynne, Bennison, Peter.** Student Translation Archive. Design, Development and Application. *Corpora in Translator Education*. Federico Zanettin, Silvia Bernardini, Dominic Stewart (eds). Abingdon, New York : Routledge, 2003, pp. 103–118.
43. **Breckle 2015 – Breckle, Margit.** *Learner corpus research in Lithuania* [tiešsaiste]. Ziņojuma saņēmējs: **Inga Znotiņa**. 2015. g. 15. decembris [skatīts 2015. gada 15. decembrī]. Nepubliskota sarakste.
44. **Breckle, Zinsmeister 2010 – Breckle, Margit, Zinsmeister, Heike.** Zur lernersprachlichen Generierung referierender Ausdrücke in argumentativen Texten. *Textmuster: schulisch-universitär-kulturkontrastiv*. 2010, s. 79–101.

45. **Breckle, Zinsmeister 2012 – Breckle, Margit, Zinsmeister, Heike.** A corpus-based contrastive analysis of local coherence in L1 and L2 German. Karabalić, V., Varga, M. und Pon, L.(eds.), *Discourse and Dialogue/Diskurs-und Dialog*, 2012, pp. 235–250.
46. **Brent 1991 – Brent, Michael R.** Automatic acquisition of subcategorization frames from untagged text. *Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics*, 1991. pp. 209–214.
47. **Burneikaitė 2006 – Burneikaitė, Nida.** Evidentiality in graduate student writing: a study of Lithuanian students' master's theses in English. *Kalba ir kontekstai,* t. 1, 2006, pp. 97–105.
48. **Burneikaitė 2007 – Burneikaitė, Nida.** Text-Organizing Metadiscourse in MA Theses in English L1 and L2. *Kalba ir kontekstai*, t. 2, 2007, pp. 154–164.
49. **Burneikaitė 2008 – Burneikaitė, Nida.** Metadiscourse in Linguistics Master's Theses in English L1 and L2. *Kalbotyra,* No. 59 (3), 2008, pp. 38–47.
50. **Burneikaitė 2009a – Burneikaitė, Nida.** Endophoric markers in Linguistics Master's theses in English L1 and L2. *Žmogus ir žodis*. *Svetimosios kalbos,* No. 11(3), 2009, pp. 11–16.
51. **Burneikaitė 2009b – Burneikaitė, Nida.** Evaluative Metadiscourse in Linguistics Master's Theses in English L1 & L2. *Kalba ir kontekstai*, t. 3(1), 2009, pp. 87–95.
52. **Burneikaitė 2009c – Burneikaitė, Nida.** Metadiscoursal connectors in linguistics MA theses in English L1 & L2. *Kalbotyra* 61 (3), 2009, pp. 36–50.
53. **Burneikaitė 2011a – Burneikaitė, Nida.** *Note the difference*: the Use of Imperatives in Native and Non-Native English MA Theses. *Tekstas: lingvistika ir poetika : 18 tarptautinės mokslinės konferencijos medžiaga*, 2011 m. lapkričio 11 d. Šiauliai : Šiaulių universiteto leidykla, 2011. 22 psl.
54. **Burneikaitė 2011b – Burneikaitė, Nida.** Questions in Linguistics Master's Theses. *Kalba ir kontekstai,* t. 4(1),. 2011, pp. 108**–**122.
55. **Burneikaitė 2012 – Burneikaitė, Nida.** *Let~* Patterns in MA Theses in Linguistics. *Kalba ir kontekstai,* t. 5(1), 2012, pp. 117**–**124.
56. **Burneikaitė 2013** – **Burneikaitė, Nida.** Writer Positioning in Linguistics MA Theses in English L1 and L2. *Anglistics in Lithuania: Cross-Linguistic and Cross-Cultural Aspects of Study*. Newcastle upon Tyne : Cambridge Scholars Publishing, 2013, pp. 154–177.
57. **Burneikaitė, Zabiliūtė 2003 – Burneikaitė, Nida, Zabiliūtė, Jurgita.** Information Structuring in Learner Texts: a Possible Relationship Between the Topical Structure and the Holistic Evaluation of Learner Essays. *Kalbų studijos*, No. 4, 2003, pp. 67**–**72.
58. **Bussmann 1996 – Bussmann, Hadumond.** *Routledge Dictionary of Language and Linguistics*. London, New York : Routledge, 1996.
59. **Bußmann 2002 – Bußmann, Hadumond.** *Lexicon der Sprachwissenschaft*. 3., aktualisierte und erweiterte Auflage. Stuttgart : Kröner, 2002.
60. **Butkus 2008 – Butkus, Alvydas.** *Baltiškos impresijos.* Kaunas : Aesti, 2008.
61. **CALD 2003 –** *Cambridge Advanced Learner's Dictionary.* Cambridge : Cambridge University Press, 2003.
62. **Callies, Paquot 2015 – Callies, Marcus, Paquot, Magali.** An interview with Yukio Tono. In memoriam Geoffrey Leech. *International Journal of Learner Corpus Research,* Vol. 1, No. 1, 2015, pp. 160–171.

63. **Castagnoli u. c. 2006 – Castagnoli, Sara, Ciobanu, Dragos, Kunz, Kerstin, Kübler, Natalie, Volanschi, Alexandra.** Designing a learner translator corpus for training purposes. *Proceedings of TALC2006*. Paris : France, 2006.
64. **Cherrington 2004 – Cherrington, Ruth.** Error analysis. *Routledge Encyclopedia of Language Teaching and Learning*. Michael Byram (ed.). London, New York : Routledge, 2004.
65. **Cigankova 2009 – Cigankova, Nataļja.** Linguistic Variation in English Computer Mediated Academic Discourse. Promocijas darbs. Rīga : Latvijas Universitāte, 2009.
66. **Cigankova, Vinčela 2012 – Cigankova, Nataļja, Vinčela, Zigrīda.** Specialized Corpora Structural and Functional Variability in Sociolinguistic Studies of Variation in English. *Valoda 2012*. Valoda dažādu kultūru kontekstā. XXII. Daugavpils : Saule, 2012, 246.–256. lpp.
67. **Cigankova, Vinčela 2013 – Cigankova, Nataļja, Vinčela, Zigrīda.** Controlling Sociolinguistic Variables in Quantitative Corpus-based Research of Variation in English. *Valoda 2013*. Valoda dažādu kultūru kontekstā. XXIII. Daugavpils : Saule, 2013, 304.–314. lpp.
68. **CLARIN 2010 –** *CLARIN Model Contracts for Tools and Resources - Initial versions of licensing templates*. Version 3.1 [skatīts 2016. gada 17. augustā]. Pieejams: *www-sk.let.uu.nl%2Fu%2FM7S-2.4.pdf&usg=AFQjCNFy_G0_3RJMe_W0dU_-VuNB7BJTLQ&sig2=E4xbLMsP7yEIZKZJ5BFYMA&bvm=bv.131286987,d.bGs*
69. **Collins-e –** *The Collins Corpus* [skatīts 2015. gada 14. augustā]. Pieejams: *http://www.collins.co.uk/page/The+Collins+Corpus*
70. **COPLE2-e –** *Learner Corpus of Portuguese L2 - COPLE2* [skatīts 2015. gada 2. augustā]. Pieejams: *http://alfclul.clul.ul.pt/teitok/learnercorpus/*
71. **CQPweb 2016 –** CQPweb and other Web GUIs for CWB [skatīts 2016. gada 18. augustā]. Pieejams: *http://cwb.sourceforge.net/cqpweb.php*
72. **Crystal 1992** – **Crystal, David.** *An Encyclopedic Dictionary of Language and Languages*. Oxford, Cambridge, Mass. : Blackwell, 1992.
73. **Crystal 1993 – Crystal, David.** The structure of language. Teaching literacy: balancing perspectives. R. Beard (ed.). London : Hodder and Stoughton, 1993, pp. 15-21 [skatīts 2016. gada 21. janvārī]. Pieejams: *http://www.davidcrystal.com/?id=4586*
74. **Crystal 2008** – **Crystal, David.** *A Dictionary of Linguistics and Phonetics*. 6th edition. Malden, MA, Oxford : Blackwell Pub., 2008.
75. **Dabašinskienė, Čubajevaitė 2009 – Dabašinskienė, Ineta, Čubajevaitė, Laura.** Acquisition of Case in Lithuanian as L2: Error Analysis. *Eesti Rakenduslingvistika Ühingu aastaraamat* 5. Tallinn : Eesti Keele Sihtasutus, 2009. pp. 47–66.
76. **Dagneaux u. c. 1998 – Dagneaux, Estelle, Denness, Sharon, Granger, Sylviane.** Computer-aided error analysis. *System*, Vol. 26, No. 2, 1998, pp. 163–174.
77. **De Cock, Granger 2005 –** De Cock, Sylvie, Granger, Sylviane. Computer Learner Corpora and Monolingual Learners Dictionaries: the Perfect Match. *Lexicographica*, No. 20, 2005, pp. 72**–**86.
78. **De Haan 1998 – De Haan, Pieter.** How 'native-like' are advanced learners of English? *Explorations in Corpus Linguistics*. A. Renouf (ed.). Amsterdam : Rodopi, 1998, pp. 55–66.
79. **De Mönnink 1999 – De Mönnink, Inge.** Parsing a learner corpus? *Corpus Linguistics and Linguistic Theory*. C. Mair and M. Hundt (eds.). Berlin : Walter de Gruyter, 1999, pp. 81–90.

80. **Deksne, Skadiņa 2014** – **Deksne, Daiga, Skadiņa, Inguna.** Error-Annotated Corpus of Latvian. Language Resources and Technology in Latvia (2010–2014). *Human Language Technologies – The Baltic Perspective.* Andrius Utka et al. (eds.). Amsterdam, Berlin, Tokyo, Washington : IOS Press, 2014, pp. 163–166.
81. **Delais-Roussarie, Yoo 2011 – Delais-Roussarie, Elisabeth, Yoo, Hi-Yon.** Learner Corpora and Prosody: From the *COREIL* Corpus to Principles on Data Collection and Corpus Design. *Poznań Studies in Contemporary Linguistics*, Vol. 47, No. 1, 2011, pp. 26–39.
82. **Díaz-Negrillo 2012 – Díaz-Negrillo, Ana.** Learner corpora: the case of the NOSE corpus. *Journal of Systemics, Cybernetics & Informatics*, Vol. 10, Issue 1, 2012, pp. 42-47 [skatīts 18.12.2015.]. Pieejams: *http://www.oalib.com/paper/2891896#.Vp0RP1JN-mJ*
83. **Díez-Bedmar, Papp 2008 – Díez-Bedmar, María Belén, Papp, Szilvia.** The use of English article system by Chinese and Spanish learners. *Linking up contrastive and learner corpus research.* Gaëtanelle Gilquin, Szilvia Papp, María Belén Díez-Bedmar (eds.). Amsterdam, New York, NY, 2008, XI, 282 pp.
84. **DLKG 1994 –** Dabartinės lietuvių kalbos gramatika. Vytautas Ambrazas (red.). Vilnius : Mokslo ir enciklopedijų leidykla, 1994.
85. **DLKT-e –** *Dabartinės lietuvių kalbos tekstynas*[skatīts 2015. gada 6. decembrī]. Pieejams: *http://tekstynas.vdu.lt/tekstynas/*
86. **DLKŽ 2011 –** *Dabartinės lietuvių kalbos žodynas*: šeštas (trečias elektroninis) leidimas. Kompaktinė plokštelė. Stasys Keinys (red.). Vilnius : Lietuvių kalbos institutas, 2006; internetinė versija 2011 [skatīts 2016. gada 20. janvārī]. Pieejams: *http://dz.lki.lt*
87. **Doorslaer 1995 – van Doorslaer, Luc.** Quantitative and Qualitative Aspects of Corpus Selection in Translation Studies. *Target*, Vol. 7-2, 1995, pp. 245–260.
88. **Dörnyei 2007 – Dörnyei, Zoltán.** *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford, New York : Oxford University Press, 2007.
89. **Dubovičienė, Gulbinskienė 2014 – Dubovičienė, Tatjana, Gulbinskienė, Dalia.** Learning/Teaching EFL to Adult Learners at Language Courses. *Žmogus ir žodis,* Vol. 16, No.3, 2014, pp. 138–149.
90. **Dumont, Granger 2017 – Dumont, Amandine, Granger, Sylviane.** Learner corpora around the world [skatīts 2017. gada 13. jūlijā]. Pieejams: *https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html*
91. **Durian 2002 – Durian, David.** Corpus-Based Text Analysis from a Qualitative Perspective: A Closer Look at NVivo. *Style*, Vol. 36-4, 2002,, pp. 738–742.
92. **Dzērve 2013 – Dzērve, Elīna.** *Personifikācija studentu noslēguma darbos*. Bakalaura darbs. Rīga : Latvijas Universitāte, 2013.
93. **EAGLES 1996 –***Recommendations for the morphosyntactic annotation of corpora*. EAGLES Document EAG-TCWG-MAC/R, Version of Mar, 1996. Retrieved from *http://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/ annotate.pdf*
94. **EIC-e –** *What is Estonian Interlanguage Corpus (EIC)?* [skatīts 2014. gada 2. septembrī]. Pieejams: *http://evkk.tlu.ee/wwwdata/what_is_evk?language=en*
95. **EKP 2006** – **Eiropas Padome. Valodas politikas nodaļa.** *Eiropas kopīgās pamatnostādnes valodu apguvei: mācīšanās, mācīšana, vērtēšana*. Rīga : Madonas poligrāfists, 2006.
96. **ELE 2008** – *Encyclopedia of Language and Education*. Nancy H. Hornberger (ed.). New York : Springer, 2008.

97. **ELFA 2014** – *The ELFA project* [skatīts 2016. gada 17. janvārī]. Pieejams: *http://www.helsinki.fi/englanti/elfa/elfacorpus.html*
98. **ELL 2005** – *Encyclopedia of Language and Linguistics*. [n. v.] : Elsevier, 2005.
99. **Ellis 1994** – **Ellis, Rod.** *The Study of Second Language Acquisition*. Oxford : Oxford University Press, 1994.
100. **Ellis, Barkhuizen 2005 – Ellis, Rod, Barkhuizen, Gary.** *Analysing Learner Language*. Oxford : Oxford University Press, 2005.
101. **EP 2006-6 –** *Encyclopedia of Philosophy*. 2nd ed. Vol. 6. Donald M. Borchert (ed.). Detroit : Thomson Gale/Macmillan Reference USA, 2006.
102. **EP 2006-7 –** *Encyclopedia of Philosophy*. 2nd ed. Vol. 7. Donald M. Borchert (ed.). Detroit : Thomson Gale/Macmillan Reference USA, 2006.
103. **Eslon 2014 – Eslon, Pille.** Eesti vahekeele korpus. *Language and Literature (Keel ja Kirjandus)*, No. 6, 2014, pp: 436–451.
104. **Evert 2009 – Evert, Stefan.** *The CQP Query Language Tutorial* [skatīts 25.09.2016.]. N. v. : n. i., 2009. Pieejams: *http://cwb.sourceforge.net/temp/ CQPTutorial.pdf*
105. **Field 2011 – Field, Fredric W.** *Key Concepts in Bilingualism*. New York : Palgrave Macmillan, 2011.
106. **Fitschen, Gupta 2008 – Fitschen, Arne, Gupta, Piklu.** Lemmatizing and morphological tagging. *Corpus linguistics : an international handbook*, Vol. 1. Ed. by A. Lüdeling, M. Kytö. Berlin, New York : Walter de Gruyter, 2008, pp. 552–564.
107. **Flowerdew 2004** – **Flowerdew, Lynne.** The argument for using English specialized corpora to understand academic and professional language. *Discourse in the Professions : Perspectives from Corpus Linguistics*. Ulla Connor, Thomas A. Upton (eds.). Amsterdam, Philadelphia : John Benjamins Publishing, 2004, pp. 11–36.
108. **FPDAL –** *Fizisko personu datu aizsardzības likums* [skatīts 2014. gada 5. oktobrī]. Pieejams: *http://likumi.lv/doc.php?id=4042*
109. **Freimane 1993 – Freimane, Inta.** *Valodas kultūra teorētiskā skatījumā*. Rīga : Zvaigzne, 1993.
110. **Fujishima, Ishizaki 2011 – Fujishima, Satoru, Ishizaki, Shun.** Automated Detection of Usage Errors in non-native English Writing using One-Class Support Vector Machines. The 13th Industrial Electronics Seminar 2011.
111. **FV 1974 –** *Filozofijas vārdnīca.* Marks Rozentāls (red.). Rīga : Liesma, 1974.
112. **Gass, Selinker 1983** – **Gass, Susan M., Selinker, Larry.** *Language Transfer in Language Learning. Issues in Second Language Research.* Susan M. Gass, Larry Selinker (eds.). Rowley : Newbury House Publishers, 1983.
113. **Gilquin 2007 – Gilquin, Gaëtanelle.** To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik*, Vol. 55, no. 3, 2007, pp. 273–291.
114. **Glück 2005 – Glück, Helmut.** *Metzler Lexicon Sprache*. Dritte, neubearbeitete Auflage. Stuttgart, Weimar : J. B. Metzler, 2005.
115. **Granger 1993 – Granger, Sylviane.** The International Corpus of Learner English. *The European English Messenger*, Vol. 2(1), 1993, p. 34.
116. **Granger 1994 – Granger, Sylviane.** The learner corpus: a revolution in applied linguistics. *English Today*, Vol. 39, No. 10/3, 1994, pp. 25–29.
117. **Granger 1997 – Granger, Sylviane.** The computer learner corpus: a testbed for electronic EFL tools. In: Nerbonne J., Linguistic Databases, CSLI Publications : Stanford 1997, p. 175-188.

118. **Granger 1998a – Granger, Sylviane.** *Learner English on Computer*. Sylviane Granger (ed.). London, New York : Longman, 1998.
119. **Granger 1998b – Granger, Sylviane.** Prefabricated patterns in advanced EFL writing: collocations and lexical phrases. *Phraseology: Theory, Analysis and Applications.* Anthony Paul Cowie (ed.). Oxford : Clarendon Press, 1998, pp. 145–160.
120. **Granger 2002 – Granger, Sylviane.** A bird's-eye view of learner corpus research. In: *Computer learner corpora, second language acquisition and foreign language teaching*. S. Granger, J. Hung, and S. Petch-Tyson (eds). Amsterdam: John Benjamins, 2002, pp. 3–33.
121. **Granger 2003a – Granger, Sylviane.** Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, Vol. 20, No. 3, 2003, pp. 465–480.
122. **Granger 2003b – Granger, Sylviane.** International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, Vol. 37, No. 3, 2003, pp. 538–546.
123. **Granger 2004 – Granger, Sylviane.** Computer learner corpus research: current status and future prospects. *Language and Computers,* Vol. 52, No. 1, pp. 123–145. Available at: *http://www.ingentaconnect.com/content/rodopi/lang/2004/00000052/00000001/art0008.*
124. **Granger 2007 – Granger, Sylviane.** A bird's eye view of learner corpus research. *Corpus linguistics*, vol. VI. W. Teubert, R. Krishnamurthy (eds.). London, New York : Routledge, 2007, pp. 44–72.
125. **Granger 2008a – Granger, Sylviane.** Learner corpora. *Corpus Linguistics* : *An International Handbook*. Anke Lüdeling, Merja Kytö (eds.). Berlin, New York : Walter de Gruyter, 2008, pp. 259–275.
126. **Granger 2008b – Granger, Sylviane.** Learner Corpora in Foreign Language Education. *Encyclopedia of Language and Education*, Vol. 4. Second and Foreign Language Education. Nelleke van Deusen-Scholl, Nancy H. Hornberger (eds). New York : Springer, 2008, pp. 337–352.
127. **Granger 2009 – Granger, Sylviane.** The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. Corpora and Language Teaching. Karin Aijmer (ed.). Amsterdam, Philadelphia : John Benjamins Publishing Company, 2009, pp. 13–32.
128. **Granger 2013 – Granger, Sylviane.** *Learner English on Computer*. Sylviane Granger (ed.). London, New York : Routledge, 2013.
129. **Granger 2015 – Granger, Sylviane.** *Learner corpus research: A fast-growing interdisciplinary field*. Plenary lecture in the Corpus Linguistics 2015 conference. July 22, 2015, Lancaster University.
130. **Granger, Tribble 1998 – Granger, Sylviane, Tribble, Chris.** Learner Corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In: Learner English on Computer. S. Granger (ed). London, New York: Addisson Wesley Longman.
131. **Granger, Tyson 1996 – Granger, Sylviane, Tyson, Stephanie.** Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, Vol. 15, No. 1, 1996, pp. 17–27.
132. **Grigaliūnienė 2013a – Grigaliūnienė, Jonė.** *Corpora in Language Studies*. Vilnius : Vilnius University, 2013.
133. **Grigaliūnienė 2013b – Grigaliūnienė, Jonė.** *Corpora in the Classroom*. Vilnius : Vilnius University, 2013.

134. **Grigaliūnienė 2013c** – **Grigaliūnienė, Jonė.** The Status and Use of the Word *RIGHT* in Native Speaker and Learner Speech: A Case of Lithuanian Learners of English. *Anglistics in Lithuania: Cross-Linguistic and Cross-Cultural Aspects of Study*. Newcastle upon Tyne : Cambridge Scholars Publishing, 2013, pp. 209–227.
135. **Grigaliūnienė, Juknevičienė 2011 – Grigaliūnienė, Jonė, Juknevičienė, Rita.** Formulaic language, learner speech and the spoken corpus of learner English LINDSEILITH. *Žmogus ir žodis,* Vol. 13, No. 3, pp. 12–18.
136. **Grigaliūnienė, Juknevičienė 2012 – Grigaliūnienė, Jone, Juknevičienė, Rita.** Corpus-based learner language research: contrasting speech and writing. *Darbai ir Dienos,* No. 58, 2012, pp. 137.–150.
137. **Grigaliūnienė, Juknevičienė 2013 – Grigaliūnienė, Jone, Juknevičienė, Rita.** Recurrent formulaic sequences in the speech and writing of the Lithuanian learners of English. In: Twenty years of learner corpus research: Looking back, Moving ahead ; Proceedings of the First Learner Corpus Research Conference (LCR 2011). S. Granger, G. Gilquin, & F. Meunier (eds.), pp. 211–222.
138. **Grigaliūnienė u. c. 2008 – Grigaliūnienė, Jonė, Bikelienė, Lina, Juknevičienė, Rita.** The Lithuanian Component of the International Corpus of Learner English (LICLE): a resource for English language learning, teaching and research at Lithuanian institutions of higher education. *Žmogus ir žodis,* Vol. 10, No. 3, pp. 62.–66.
139. **Grīnberga 2004a – Grīnberga, Iveta.** Pirmās valodas semantiskā interference latviešu kā otrajā valodā. *Kalbos teorija ir praktika* (straipsnių rinkinys, parengtas 2003 m. spalio 17 d. vykusios konferencijos pranešimų pagrindu). Kaunas : Technologija, 2004, 60.–67. lpp.
140. **Grīnberga 2004b – Grīnberga, Iveta.** Starpvalodas gramatiskās sistēmas iezīmes latviešu kā otrās valodas apguves procesā. *Valoda 2004*. Valoda dažādu kultūru kontekstā. XIV. Daugavpils : Saule, 2004, 23.–29. lpp.
141. **Grīnberga 2004c – Grīnberga, Iveta.** Starpvalodas gramatiskās sistēmas raksturojums: kopīgās un individuālās iezīmes. *Valoda kā identitāte*. Zinātnisko rakstu krājums. Rīga : Valodu mācību centrs, 2004, 67.–70. lpp.
142. **Grūzītis 2012 – Grūzītis, Normunds.** Datorlingvistikas pētījumi LU Matemātikas un informātikas institūtā. *Latviešu valoda digitālajā vidē: datorlingvistika.* Informatīvi izglītojoša semināru cikla materiāli [tiešsaiste]. Rakstu krājums. Rīga : LVA, 2012, 15.–36. lpp. [skatīts 2015. gada 13. janvārī]. Pieejams: *http://valoda.lv/downloadDoc_648/mid_622*
143. **Hana u. c. 2010 – Hana, Jirka, Rosen, Alexander, Škodová, Svatava, Štindlová, Barbora.** Error-tagged Learner Corpus of Czech. Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, Uppsala, Sweden, 15-16 July 2010, pp. 11–19.
144. **Hana u. c. 2012 – Hana, Jirka, Rosen, Alexandr, Štindlová, Barbora, Jäger, Petr.** Building a learner corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. N. v.: European Language Resources Association, 2012.
145. **Hardie 2012 – Hardie, Andrew.** CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17:3, 2012, pp. 380–409.
146. **Hardy, Römer 2011 – Hardy, Jack. A., Römer, Ute.** Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora,* 8 (2), 2011, pp. 183–207.

147. **Hawkins, Buttery 2010 – Hawkins, John A., Buttery, Paula.** Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 2010, 1 (1), e5 [skatīts 2015. gada 13. decembrī]. Pieejams: *http://journals.cambridge.org/download.php?file=%2F6645_AEBC5CABD3405501E4CC1C74A4056E70_journals__EPJ_EPJ1_01_S2041536210000103a.pdf&cover=Y&code=5aa245645f91bfc4954f7b01a59f9543*
148. **Helviga 2012 – Helviga, Anita.** Ieskats datorlingvistikas terminoloģijas iezīmēs un attīstības tendencēs. *Latviešu valoda digitālajā vidē : datorlingvistika.* Informatīvi izglītojoša semināru cikla materiāli [tiešsaiste]. Rakstu krājums. Rīga : LVA, 2012, 104.–120. lpp. [skatīts 2014. gada 2. aprīlī]. Pieejams: *http://valoda.lv/Petijumi/Elektroniskie_izdevumi/mid_622*
149. **Heringer 2001 – Heringer, Hans Jürgen.** *Fehlerlexikon : Deutsch als Fremdsprache*. Berlin : Cornelsen Verlag, 2001.
150. **Hunston 2008** – **Hunston, Susan.** Collection strategies and design decisions. *Corpus Linguistics* : *An International Handbook*. Anke Lüdeling, Merja Kytö (eds.). Berlin, New York : Walter de Gruyter, 2008, pp. 154–168.
151. **ICLE 2015** – *ICLE*. [skatīts 2015. gada 16. decembrī]. Pieejams: *http://www.uclouvain.be/en-cecl-icle.html*
152. **ILR-e** – Descriptions of proficiency levels. [skatīts 2016. gada 5. janvārī]. Pieejams: *http://www.govtilr.org/Skills/ILRscale1.htm*
153. **James 1998 – James, Carl.** *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York : Longman, 1998.
154. **James 2013 – James, Carl.** *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York : Routledge, 2013.
155. **Janssen 2016 – Janssen, Maarten.** TEITOK: Text-Faithful Annotated Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation.* Paris : ELRA, 2016. [skatīts 2016. gada 4. septembrī]. Pieejams: *http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf*
156. **Jantunen 2011 – Jantunen, Jarmo Harri.** Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivőrdlusi. Lähivertailuja*, No. 21, 2011, pp. 86–105.
157. **Janulienė, Dziedravičius 2015 – Janulienė, Aušra, Dziedravičius, Justinas.** On the use of conjunctive adverbs in learners' academic essays. *Verbum*, No. 6, 2015, pp. 69–83.
158. **Johansson 2007 – Johansson, Stig.** *Seeing through Multilingual Corpora*. On the use of corpora in contrastive studies. Amsterdam, Philadelphia : John Benjamins, 2007.
159. **Johansson 2011** – A multilingual outlook of corpora studies : An interview with Stig Johansson. *Perspectives on corpus linguistics*. Viana, Vander, Zyngier, Sonia, Barnbrook, Geoff (eds.). Amsterdam, Philadelphia : John Benjamins Publishing Company, 2011, pp. 115–130.
160. **Jordens 2003 – Jordens, Peter.** Constraints on the Shape of Second Language Learner Varieties. Psycholinguistik. Ein internationales Handbuch. Gert Rickheit, Theo Herrmann, Werner Deutsch (Hrsg.). Berlin, New York : Walter de Gruyter, 2003, s. 819–833.
161. **Joyce, Burns 1999 – Joyce, Helen, Burns, Anne.** Focus on Grammar. Sydney : National Centre for English Language Teaching and Research. Macquarie University, 1999.
162. **Juknevičienė 2007 – Juknevičienė, Rita.** Analyzing topic-specific vocabulary in EFL student writing. *Innovations in language teaching and learning in the*

multicultural context : International Nordic-Baltic conference of the World Federation of Language Teacher Associations, Riga, Latvia, 15-16 June, 2007 Riga, 2007. [skatīts 2015. gada 18. decembrī]. Pieejams: *http://web.vu.lt/flf/r.jukneviciene/files/2009/06/topic-specific-vocabulary-20071.pdf*

163. **Juknevičienė 2008 – Juknevičienė, Rita.** Collocations with High-Frequency Verbs in Learner. English: Lithuanian Learners versus Native Speakers. *Kalbotyra,* Vol. 59, No. 3, pp. 119.–127.
164. **Juknevičienė 2009 – Juknevičienė, Rita.** Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra,* Vol. 61, No. 3, pp. 61–72.
165. **Juknevičienė 2011 – Juknevičienė, Rita.** Leksinės samplaikos svetimkalbių ir gimtakalbių vartotojų rašytinėje anglų kalboje. Daktaro disertacija. Mokslinė vadovė Jonė Grigaliūnienė. Vilnius : Vilniaus universitetas, 2011.
166. **Juknevičienė 2013a – Juknevičienė, Rita.** Insights from a corpus of secondary school English examination essays in Lithuania. *ICAME 34*. English corpus linguistics on the move: Applications and implications. Book of abstracts. Santiago de Compostela : University of Santiago de Compostela, 2013, pp. 55–56.
167. **Juknevičienė 2013b – Juknevičienė, Rita.** Recurrent word sequences in written learner English. *Anglistics in Lithuania: Cross-Linguistic and Cross-Cultural Aspects of Study*. Newcastle upon Tyne: Cambridge Scholars Publishing, 2013, pp. 178–197.
168. **Juknevičienė 2014 – Juknevičienė, Rita.** *Tracing the development of phraseological competence*. Abstract. Formulaic Language Research Network 2014 Conference at Swansea University [skatīts 2015. gada 18. decembrī]. Pieejams: *http://flarn.viviennerogers.info/wp-uploads/2014/02/Jukneviciene.pdf*
169. **Juknevičienė, Šeškauskienė 2014a – Juknevičienė, Rita, Šeškauskienė, Inesa.** Kriteriniai anglų kalbos pasiekimo lygių požymiai brandos egzamine: kaip rašo abiturientai? *Kalba ir Kontekstai,* Vol. 6, No. 1, Part 1&2, 2014, pp. 198-207.
170. **Juknevičienė, Šeškauskienė 2014b – Juknevičienė, Rita, Šeškauskienė, Inesa.** The National Examination of English in Lithuania: Searching for Evidence of CEFR Criterial Achievement Levels. *Kalbų studijos,* Vol. 25, 2014, pp. 88–96.
171. **Kalnbērziņa 2015 – Kalnbērziņa, Vita.** *Sakārtojuma un pakārtojuma attiecības valodu apguves līmeņos.* Stenda referāts Latvijas Universitātes 73. konferences Latviešu un vispārīgās valodniecības sekcijas sēdē 06.02.2015. Rīga : Latvijas Universitāte, 2015.
172. **Kalnbērziņa u. c. 2011 – Kalnbērziņa, Vita, Lokmane, Ilze, Kunda, Tatjana, Vinčela, Zigrīda, Baiža, Kristīne.** Pētījums „Latviešu valodas apguves kvalitāte mazākumtautību skolās" [tiešsaiste]. Rīga : LVASA, 2011. [skatīts 2014. gada 2. aprīlī]. Pieejams: *http://www.lvasa.lv/files/file/Petijums_Latv_val_apguves_kvalitate_14_02_11(1).pdf*
173. **Kalnbērziņa, Rūtenberga 2012 – Kalnbērziņa, Vita, Rūtenberga, Vineta**. Subordinate clauses as critical features in English and French learner examination corpora. *Baltic Journal of English Language, Literature and Culture,* Vol. 2, 2012, pp. 54–62.
174. **Karapetjana 2007 – Karapetjana, Indra**. *Language in Bachelor Papers as a Result of the Development of Linguo-Functional Research Competence.* Promocijas darbs. Rīga : Latvijas Universitāte, 2007.
175. **Kazlauskaitė 2015 – Kazlauskaitė, Karolina.** *Gramatinė kompetencija lietuvių studentų rašiniuose latvių kalba.* Bakalauro darbas. Vilnius : Vilniaus universitetas, 2015.

176. **Kazlauskienė 2015 – Kazlauskienė, Vitalija.** Daiktavardinis junginys kaip gramatinės kompetencijos elementas prancūzų kalbos baigiamojo egzamino rašto darbuose. *Darnioji daugiakalbystė*, 6. Kaunas : Vytauto Didžiojo universitetas, 2015, 134–158 psl.
177. **Kļaviņa 1980 – Kļaviņa, Sarma.** *Statistika valodniecībā.* Rīga : LVU, 1980.
178. **KLC-e –** *Morfologinis anotatorius* [skatīts 2016. gada 25. septembrī]. Pieejams: *http://tekstynas.vdu.lt/page.xhtml?id=morphological-annotator*
179. **Koduhovs 1987 – Koduhovs, Vitalijs.** *Vispārīgā valodniecība.* Rīga : Zvaigzne, 1987.
180. **Koester 2010 – Koester, Almut.** Building small specialized corpora. The Routledge Handbook of Corpus Linguistics. Anne O'Keefe, Michael McCarthy (eds.). London, New York : Routledge, 2010, pp. 66–79.
181. **Koo 2006 – Koo, Kyosung.** Effects of Using Corpora and Online Reference Tools on Foreign Language Writing: A Study of Korean Learners of English as a Second Language. PhD thesis. University of Iowa 2006.
182. **Laiveniece 2010 – Laiveniece, Diāna.** Teksta definīciju varianti lingvistiskajā literatūrā un latviešu valodas mācību grāmatās. *Filologija*, Nr. 15, 2010, 83.–90. lpp.
183. **Laizāne 2012 – Laizāne, Inga.** Akuzatīvs latviešu valodas kā svešvalodas apguvē. Valoda – 2012. *Valoda dažādu kultūru kontekstā.* Zinātnisko rakstu krājums XXII. Daugavpils : Saule, 2012, 194.–203. lpp.
184. **Laizāne 2013 – Laizāne, Inga.** Ģenitīva locījums un tā nozīmju lietojums latviešu valodas kā svešvalodas apguvē. *Valodu apguve: problēmas un perspektīva* : zinātnisko rakstu krājums, IX. Liepāja : LiePA, 2013, 82.–93. lpp.
185. **Laizāne 2014a – Laizāne, Inga.** Jēdzienu *pirmā valoda, otrā valoda* un *svešvaloda* izpratne Latvijā. *Vārds un tā pētīšanas aspekti :* rakstu krājums, 18 (2). Red. kolēģijas vadītāja Benita Laumane. Krājuma atb. red. Linda Lauze. Liepāja : LiePA, 2014. 136.–147. lpp.
186. **Laizāne 2014b – Laizāne, Inga.** Latviešu valodas apguvēju tipiskākās kļūdas, mācoties nomenu dzimtes kategoriju. *Via Scientiarum* : starptautiskās jauno lingvistu konferences rakstu krājums. 2. laidiens. Sastādītājas S. Sviķe un Z. Veidenberga. Ventspils, Liepāja : Ventspils Augstskola, Liepājas Universitāte, 2014, 125.–137. lpp.
187. **Laizāne 2014c – Laizāne, Inga.** Difficulties in Acquisition of Latvian as a Foreign Language Learning the Locative. Abstracts of Presentations of the 6th International Scientific Conference *Linguistic, Didactic and Sociocultural Aspects of Language Functioning*. Vilnius, 2014, 24. pp.
188. **Laizāne 2014d – Laizāne, Inga**. Difficulties in Acquisition of Latvian as a Foreign Language Learning the Locative. *Language in Different Contexts,* VI (1). Vilnius : Lithuanian University of Educational Sciences, 2014, pp. 218–225.
189. **Laizāne 2014e – Laizāne, Inga.** Datīva un datīva/instrumentāļa nozīmju apguve latviešu valodā kā svešvalodā. *Valodu apguve: problēmas un perspektīva* : zinātnisko rakstu krājums, X. Liepāja : LiePA, 2014, 154.–169. lpp.
190. **Laizāne 2014f – Laizāne, Inga.** Grūtības latviešu valodas kā svešvalodas apguvē: lietvārda lokatīvs. 50. prof. Artura Ozola dienas starptautiskā zinātniskā konference *Vispārīgā valodniecība: valodas sistēma un lietojums* referātu tēzes, Rīga, 2014, 40.–41. lpp.
191. **Laizāne 2014g – Laizāne, Inga.** Grūtības, apgūstot latviešu valodu kā svešvalodu. Rīgas Stradiņa universitāte. *2014. gada zinātniskā konference:* Tēzes, Rīga, 2014, 470. lpp.

192. **Laizāne 2014h – Laizāne Inga.** Lietvārda locījumu nozīmes latviešu valodas kā svešvalodas apguvē. *Zinātniski metodisks izdevums „Tagad",* LVAVA, 2014, 25.–31. lpp.
193. **Laizāne 2015 – Laizāne, Inga.** Latviešu valoda trimdā – pirmā, otrā valoda vai svešvaloda. Zinātnisko rakstu krājums XXII. Daugavpils : Saule
194. **LCA-e –** *Learner Corpus Association.* N. v. : n. i. [skatīts 2015. gada 13. decembrī]. Pieejams: *www.learnercorpusassociation.org*
195. **LDLTAL 2010 – Richards, Jack C., Schmidt, Richard.** *Longman Dictionary of Language Teaching and Applied Linguistics.* 4th ed. Harlow : Longman, 2010.
196. **LDOCE 2003** – Longman Dictionary of Contemporary English. Harlow (Essex) : Longman, 2003.
197. **Leech 1998 – Leech, Geoffrey.** Preface. *Learner English on Computer*. London, New York : Longman, 1998, pp. xiv–xx.
198. **Leech 2014 – Leech, Geoffrey.** The state of the art in corpus linguistics.English Corpus Linguistics. Karin Aijmer, Bengt Altenberg (eds.). New York, London : Routledge, 2014, pp. 8–29**.**
199. **Leech u. c. 1994 – Leech, Geoffrey, Garside, Roger, Bryant, Michael.** CLAWS4: The tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, 1994, pp. 622–628 [skatīts 2015. gada 13. jūnijā]. Pieejams: *http://ucrel.lancs.ac.uk/papers/coling1994paper.pdf*
200. **Leech, Onwuegbuzie 2009 – Leech, Nancy L., Onwuegbuzie, Anthony J.** A typology of mixed methods research designs. *Quality & Quantity*. Vol. 43-2, 2009, pp. 265–275.
201. **Lessard 1999 – Lessard, Greg.** *Learner English on Computer, Sylviane Granger (editor)*. Review. *Computational Linguistics*, Vol. 25.2, 1999, pp. 302–303.
202. **Levāne 2001 – Levāne, Kristīne.** *Paula Bankovska romāna „Plāns ledus" pirmās nodaļas morfoloģiskā anotēšana un statistiskā analīze*. Bakalaura darbs. Rīga : Latvijas Universitāte, 2001.
203. **Levāne-Petrova 2011 – Levāne-Petrova, Kristīne.** Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. *Vārds un tā pētīšanas aspekti :* rakstu krājums, 15 (1). Liepāja : LiePa, 2011, 187.–193. lpp.
204. **Levāne-Petrova 2012a – Levāne-Petrova, Kristīne.** Latviešu-lietuviešu-latviešu paralēlo tekstu korpusa izveide. *Vārds un tā pētīšanas aspekti* : rakstu krājums, 16 (2). Liepāja : LiePA, 2012, 180.–189. lpp.
205. **Levāne-Petrova 2012b – Levāne-Petrova, Kristīne.** Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. *Baltistica :* rakstu krājums 8 (2012). Vilnius : Vilniaus Universiteto leidykla, 2012, 89.–98. lpp.
206. **LINDSEI-e –** Louvain International Database of Spoken English Interlanguage (LINDSEI) [skatīts 2016. gada 17. janvārī]. Pieejams: *http://www.uclouvain.be/en-cecl-lindsei.html*
207. **Līduma 2013 – Līduma, Laura.** *Angļu valodas apguvēju rakstisko korpusu leksiskā analīze.* Bakalaura darbs. Rīga : Latvijas Universitāte, 2013.
208. **LLA 1993 –** *Longman Language Activator : The World's First Production Dictionary*. Harlow (Essex) : Longman, 1993.
209. **LLA 1996** – *Longman Language Activator : The World's First Production Dictionary*. Harlow (Essex) : Longman, 1996.
210. **LLV 2006 –** *Latviešu valodas vārdnīca*. Rīga : Avots, 2006.
211. **LLVV 1980 –** *Latviešu literārās valodas vārdnīca*. 4. sēj. Rīga : Zinātne, 1980.

212. **Ločmele 2015** – **Ločmele, Līga.** *Devītās klases angļu valodas eksāmena korpusa sistēmiski funkcionālā analīze*. Maģistra darbs. Rīga : Latvijas Universitāte, 2015.
213. **Lokastova 2007 – Lokastova, Jeļena.** *Kontrastējoša retorika: akadēmiskā angļu rakstu valoda Latvijas kontekstā.* Bakalaura darbs. Rīga : Latvijas Universitāte, 2007.
214. **Lokmane u. c. 2009 – Lokmane, Ilze, Kunda, Tatjana, Vinčela, Zigrīda, Baiža, Kristīne.** *Pētījuma veikšana par valsts valodas apguves kvalitāti mazākumtautību izglītībā*. Nepublicēts dokuments. Rīga : Valsts aģentūra „Latviešu valodas aģentūra", 2009.
215. **López-Lago, Saiz de Lobado García 2011 – López-Lago, José María, Saiz de Lobado García, Ester.** Online chat and the language learning classroom: synchronous computer mediated communication (SCMC) at the UEM LAB. VIII Jornadas Internacionales de Innovación Universitaria. Villaviciosa de Odón : Universidad Europea de Madrid, 2011 [skatīts 2015. gada 12. augustā]. Pieejams: *http://abacus.universidadeuropea.es/bitstream/handle/11268/1790/191_ONLINE.pdf?sequence=2*
216. **Lüdeling 2006 – Lüdeling, Anke.** Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. *Institut für Deutsche Sprache. Jahrbuch 2006.* Berlin : Walter de Gruyter, 2006, s. 28–48.
217. **Lüdeling 2007 – Lüdeling, Anke.** Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. (Hrsg. von W. Kallmeyer, G. Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007, s. 28–48.
218. **Lüdeling u. c. 2005 – Lüdeling, Anke, Walter, Maik, Adolphs, Peter.** Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*, Birmingham : University of Birmingham, 2005 [skatīts 2015. gada 18. augustā]. Pieejams: *https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf*
219. **Lüdeling et al. 2008 – Lüdeling, Anke, Doolittle, Seanna, Hirschmann, Hagen, Schmidt, Karin, Walter, Maik.** Das Lernerkorpus Falko. *Deutsch Als Fremdsprache,* 45 (2), 2008, s. 67–73.
220. **LTSV 2011 – Skujiņa, Valentīna, Anspoka, Zenta, Kalnbērziņa, Vita, Šalme, Arvils.** *Lingvodidaktikas terminu skaidrojošā vārdnīca*. Rīga : Latviešu valodas aģentūra, Latviešu valodas institūts, 2011.
221. **LTŽ 2012 – Ramonienė, Meilutė, Brazauskienė, Jelena, Burneikaitė, Nida, Daugmaudytė, Jurga, Kontutytė, Eglė, Pribušauskaitė, Joana**. *Lingvodidaktikos terminų žodynas*. Vilnius : Vilniaus universiteto leidykla, 2012.
222. **LVG 2013 – Auziņa, Ilze, Breņķe, Ieva, Grigorjevs, Juris, Indričāne, Inese, Ivulāne, Baiba, Kalnača, Andra, Lauze, Linda, Lokmane, Ilze, Markus, Dace, Nītiņa, Daina, Smiltniece, Gunta, Valkovska, Baiba, Vulāne, Anna**. *Latviešu valodas gramatika*. Rīga : LU Latviešu valodas institūts, 2013.
223. **LVKK 2005 –** Latviešu valodas korpusa koncepcija [tiešsaiste]. Rīga : LU Matemātikas un informātikas institūts, 2005 [skatīts 2014. gada 1. aprīlī]. Pieejams: *http://www.korpuss.lv/ uzzinas/koncepcija.pdf*
224. **LVMPK 2009 –** Latviešu valodas morfoloģisko pazīmju kopa. Rīga : n.i., 08.12.2009 [skatīts 2015. gada 22. novembrī]. Pieejams: *http://www.semti-kamols.lv/doc_upl/TagSet.pdf*
225. **Mair 1991 – Mair, Christian.** Quantitative or Qualitative corpus analysis? Infinitival complement clauses in the Survey of English Usage corpus. *English*

Computer Corpora: Selected Papers and Research Guide. Berlin : Walter de Gruyter, 1991, pp. 67–80.

226. **Marcinkevičienė 2000 – Marcinkevičienė, Rūta.** Tekstynų lingvistika: teorija ir praktika. *Darbai ir dienos*, Nr.. 24, 2000, 7–64 psl.
227. **Marcinkevičienė 2010 – Marcinkevičienė, Rūta.** Lietuvių kalbos kolokacijos. Kaunas : Vytauto Didžiojo universitetas, 2010.
228. **Matthews 1997 – Matthews, Peter Hugoe.** *The Concise Oxford Dictionary of Linguistics*. Oxford, New York : Oxford University Press, 1997.
229. **McEnery, Hardie 2012 – McEnery, Tony, Hardie, Andrew.** *Corpus Linguistics: Method, Theory and Practice*. Cambridge, New York : Cambridge University Press, 2012.
230. **McEnery u. c. 2006 – McEnery, Tony, Xiao, Richard, Tono, Yukio.** *Corpus-based Language Studies :* An Advanced Resource Book. London, New York : Routledge, 2006.
231. **McEnery, Wilson 2001 – McEnery, Tony, Wilson Andrew.** Corpus Linguistics : An Introduction. Edinburgh: Edinburgh University Press, 2001.
232. **META-NET –** *Licenses* [skatīts 2016. gada 17. augustā]. Pieejams: *http://www.meta-net.eu/meta-share/licenses*
233. **Meunier 2000 – Meunier, Fanny.** A computer corpus linguistics approach to interlanguage grammar: noun phrase complexity in advanced learner writing. Université Catholique de Louvain : Unpublished PhD dissertation.
234. **Meunier 2006 – Meunier, Fanny.** Review: Rod Ellis and Gary Barkhuizen. *Analysing Learner Language.* Oxford: Oxford University Press. 2005. viii + 404 pages. ISBN 0-19-431634-3. £22. *Int J Lexicography, Vol. 19, No.1* 2006, pp. 110–111.
235. **Meunier 2007 – Meunier, Fanny.** The pedagogical value of native and learner corpora in EFL grammar teaching. *Corpus linguistics*, vol. VI. W. Teubert, R. Krishnamurthy (eds.). London, New York : Routledge, 2007, pp. 22–43.
236. **Meyer 2004 – Meyer, Charles F.** *English Corpus Linguistics: An Introduction.* Cambridge : Cambridge University Press, 2004.
237. **MGUKM 2013 –** *Miestų gyventojų užsienio kalbų mokėjimas*. Vilnius : Vilniaus universitetas, 2013 [skatīts 2015. gada 16. oktobrī]. Pieejams: *http://www.kalbuzemelapis.flf.vu.lt/lt/zemelapiai/miestu-gyventoju-anglu-vokieciu-ir-prancuzu-kalbu-mokejimas/miestu-gyventoju-uzsienio-kalbu-mokejimas/*
238. **MiCASE 2007 –** Michigan Corpus of Academic Spoken English [skatīts 2016. gada 17. janvārī]. Pieejams: *http://quod.lib.umich.edu/m/micase/*
239. **Mihailova, 2015 – Mihailova, Arīna.** *Ortogrāfisko kļūdu biežums izglītojamo lietišķajā angļu rakstu valodā*. Bakalaura darbs. Rīga : Latvijas Universitāte, 2015.
240. **Milton, Tsang 1993 – Milton, John, Tsang, Elza Shuk-Ching.** A corpus-based study of logical connectors in EFL students' writing., *Studies in lexis.* Proceedings of a seminar on lexis organized by the Language Centre of the HKUST, Hong Kong, 6–7 July 1992, Language Centre, HKUST, Hong Kong. Hong Kong: Hong Kong University of Science and Technology, 1993, pp. 215–246. [skatīts 2015. gada 13. decembrī]. Pieejams: *http://repository.ust.hk/dspace/bitstream/1783.1/1083/2/studies04.pdf*
241. **MK 281 –** Noteikumi par valsts vispārējās vidējās izglītības standartu, mācību priekšmetu standartiem un izglītības programmu paraugiem. Ministru kabineta noteikumi Nr. 281 [skatīts 2015. gada 13. augustā]. Pieejams: *http://likumi.lv/doc.php? id=257229*

242. **Myles 2005 – Myles, Florence.** Interlanguage corpora and SLA research. *Second Language Research*, Vol. 21, No. 4, pp. 373–391.
243. **Nagata u. c. 2011 – Nagata, Ryo, Whittaker, Edward, Sheinman, Vera.** Creating a manually error-tagged and shallow-parsed learner corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies.* Volume 1. Stroudsburg : Association for Computational Linguistics, 2011, pp. 1210-1219.
244. **Nesselhauf 2004 – Nesselhauf, Nadja.** Learner corpora and their potential for language teaching. *How to Use Corpora in Language Teaching.* J. M. Sinclair (ed.). Asterdam: John Benjamins, pp. 125–152.
245. **Orol González, Alonso Ramos 2013 – Orol González, Ana, Alonso Ramos, Margarita.** A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia – Social and Behavioral Sciences*, Vol. 95, 25 October 2013, pp. 563–570.
246. **O'Sullivan, Chambers 2006 – O'Sullivan, Íde, Chambers, Angela.** Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, vol. 15 (1), 2006, pp. 49-68.
247. **Paegle 2003 – Paegle, Dzintra.** *Latviešu literārās valodas morfoloģija*. 1. daļa. Rīga : Zinātne, 2003.
248. **Paikens 2007 – Paikens, Pēteris.** Lexicon-based morphological analysis of Latvian language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT).* Vilnius : Vytautas Magnus University, Institute of the Lithuanian Language, 2007, pp. 235–240.
249. **Paikens 2016 – Paikens, Pēteris.** Deep neural learning approaches for Latvian morphological tagging. *Human Language Technologies – The Baltic Perspective.* Amsterdam : IOS Press, 2016, pp. 160–166.
250. **Paikens u. c. 2013 – Paikens, Pēteris, Rituma, Laura, Pretkalniņa, Lauma.** Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).* Linköping : Linköping University Electronic Press, 2013, pp. 267–277.
251. **Palmer, Xue 2010 – Palmer, Martha, Xue, Nianwen**. Linguistic annotation. *The Handbook of Computational Linguistics and Natural Language Processing*. Ed. by A. Clark, C, Fox, S. Lappin.Malden, Oxford, West Sussex : Wiley-Blackwell, 2012, pp. 238–270.
252. **Petryla 2015 – Petryla, Tomas.** *Apie universitetą : Istorija*. Vilnius : Lietuvos edukologijos universitetas, 2015 [skatīts 2015. gada 30. decembrī]. Pieejams: *https://leu.lt/lt/apie_universiteta/istorija.html*
253. **Piezīmjbloks-e –** Piezīmjbloka atvēršana [skatīts 2015. gada 8. novembrī]. Pieejams: *http://windows.microsoft.com/lv-lv/windows/open-notepad#1TC=windows-7*
254. **Pravec 2002** – **Pravec, Norma A.** Survey of learner corpora. *ICAME Journal*, Vol. 26, pp. 81–114.
255. **PTSV 2000** – *Pedagoģijas terminu skaidrojošā vārdnīca*. Inārs Beļickis, Dainuvīte Blūma, Tatjana Koķe, Dace Markus, Valentīna Skujiņa, Arvils Šalme. Rīga : Zvaigzne ABC, 2000.
256. **Puškoriutė-Ridulienė 2011 – Puškoriutė-Ridulienė, Daiva.** Latviešu valoda kā izvēles svešvaloda Lietuvā. Viļņa : Vytauto Didžiojo Universitetas, 2011.
257. **Ragan 2001 – Ragan, Peter H.** Classroom Use of a Systemic Functional Small Learner Corpus. *Small Corpus Studies and ELT: Theory and practice*. Amsterdam, Philadelphia : John Benjamins Publishing Company, 2001, pp. 207–236.

258. **Rascón 2012 – Rascón Caballero, Alfonso.** Análisis de errores e interlengua. Aplicación a las frases con el verbo gustar en estudiantes lituanos de español, I p. *Verbum*, No. 3, 2012, pp. 91–100.
259. **Rascón 2013a – Rascón Caballero, Alfonso.** Avance de análisis de errores de la interlengua en estudiantes lituanos de español, II p. *Verbum*, No. 4, 2013, pp. 95–105.
260. **Rascón 2013b – Rascón Caballero, Alfonso.** Lietuvių, besimokančių ispanų kalbos, veiksmažodžio *gustar* vartojimas. Santrauka konferencijai *Darnioji daugiakalbystė: tyrimai, mokslas, kultūra* 2013 m. rugsėjo 27–28d. Kaunas : Vytauto Didžiojo universitetas [skatīts 2016. gada 17. janvārī]. Pieejams: *http://daugiakalbyste.vdu.lt/wp-content/uploads/docs/antra/abstracts/rascon.pdf*
261. **Rastelli 2009 – Rastelli, Stefano.** Learner Corpora without Error Tagging. *Linguistik online*, Vol. 38, 2/2009. [skatīts 2014. gada 5. janvārī]. Pieejams: *http://www.linguistik-online.de/38_09/rastelli.html*
262. **Rābante-Buša 2012 – Rābante-Buša, Guna.** Runas korpuss: izveide un izmantošana. *Latviešu valoda digitālajā vidē : datorlingvistika*. Informatīvi izglītojoša semināru cikla materiāli [tiešsaiste]. Rakstu krājums. Rīga : LVA, 2012, 125.–129. lpp. [skatīts 2015. gada 3. februārī]. Pieejams: *http://valoda.lv/Petijumi/Elektroniskie_izdevumi/mid_622*
263. **RELTL 2004** – *Routledge Encyclopedia of Language Teaching and Learning*. Michael Byram (ed.). London, New York : Routledge, 2004.
264. **Reppen 2010 – Reppen, Randi.** Building a corpus: what are the key considerations? The Routledge Handbook of Corpus Linguistics. Anne O'Keefe, Michael McCarthy (eds.). London, New York : Routledge, 2010, pp. 31–39.
265. **Reznicek u. c. 2013 – Reznicek, Marc, Lüdeling, Anke, Hirschmann, Hagen.** Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. *Automatic Treatment and Analysis of Learner Corpus Data*, Vol. 59, 2013, pp. 101–123.
266. **Rimkutė 2002 – Rimkutė, Erika.** Homoformos dabartinės lietuvių kalbos tekstyne. *Lituanistica*, No. 2 (50), 2002, 86–101 psl.
267. **Rimkutė 2006 – Rimkutė, Erika.** *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne.* Daktaro disertacija. Kaunas : Vytauto Didžiojo universitetas, 2006.
268. **Rimkutė u. c. 2009 – Rimkutė, Erika, Valskys, Vidas, Vaskelienė, Jolanta.** Lietuvių kalbos leksemų morfologinis anotavimas: ypatumai ir sunkumai. *Kalbų studijos*, 15, 2009, 63–70 psl.
269. **Rimkutė u. c. 2013 – Rimkutė, Erika, Utka, Andrius, Levāne-Petrova, Kristīne.** Lietuvių-latvių ir latvių-lietuvių kalbų lygiagretusis tekstynas LILA. *Kalbų studijos*, No. 23. 2003, 70–77 psl.
270. **Rögnvaldsson 2006 – Rögnvaldsson, Eiríkur.** The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation. *Copenhagen studies in language*, 32, 2006, pp. 133–145.
271. **Römer, O'Donnell 2011 – Römer, Ute, O'Donnell, Matthew B.** From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6 (2), 2011, pp. 159-177.
272. **Rosen 2014 – Rosen, Alexandr.** *CzeSL-CGT – a corpus of non-native speakers' Czech with automatic annotation.* N.v. : n. i., 27 July 2014, updated 7 October 2015. [skatīts 2015. gada 22. novembrī]. Pieejams: *http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-en.pdf*

273. **Rosen u. c. 2013 – Rosen, Alexandr, Hana, Jirka, Štindlová, Barbora, Feldman, Anna.** Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation.* Springer Science+Business Media Dordrecht 2013.
274. **Rutenberga 2012 – Rutenberga, Vineta.** Contrastive analysis of complex sentences in English and French language learner corpora. *Learner Language, Learner Corpora*. Abstracts. Sisko Brunni, Jarmo Jantunen, Antti Tolonen (eds.). Oulu : University of Oulu, 2012, pp. 67–68.
275. **Rūtenberga 2012 – Rūtenberga, Vineta.** Sintaktiskās struktūras svešvalodu apguves līmeņu salīdzināšanā. Referāta kopsavilkums. Latvijas Universitātes, Liepājas Universitātes un Ventspils Augstskolas valodniecības nozares doktorantu semināra materiāli [tiešsaiste]. Rīga : LU Humanitāro zinātņu fakultāte, 2012 [skatīts 2014. gada 2. aprīlī]. Pieejams: *http://www.hzf.lu.lv/fileadmin/user_upload/lu_portal/projekti/hzf/zinas/LU_ Vineta_Rutenberga.pdf*
276. **Rūtenberga 2014 – Rūtenberga, Vineta**. *Sintaktiskās kriteriālās pazīmes angļu un franču valodas rakstveida snieguma vērtēšanā* : promocijas darbs filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares lietišķās valodniecības apakšnozarē. Latvijas Universitāte. Rīga : Latvijas Universitāte, 2014.
277. **Rūtenberga, Kalnbērziņa 2013 – Rūtenberga, Vineta, Kalnbērziņa, Vita**. Syntactic indicators of language acquisition levels in English and French written language learner corpora. *Lublin Studies in Modern Languages and Literature,* issue: 37 / 2013, pp: 111–126
278. **Savenkova 2011 – Savenkova, Tatjana.** *Rakstu valodas sarežģītības attīstība vidusskolēnu tīmekļa žurnālos.* Maģistra darbs. Rīga : Latvijas Universitāte, 2011.
279. **Savickienė 2006 – Savickienė, Ineta.** Linksnio kategorijos įsisavinimas: lietuvių kalba kaip gimtoji ir svetimoji. *Kalbotyra*, Vol. 56, No. 3, 2006, 122–129 psl.
280. **SemTi 2009 –** *Latviešu valodas tekstu korpusu morfoloģiskās un sintaktiskās marķēšanas rīks* [lejupielādēts 2016. gada 19. janvārī]. Pieejams: *http://semti-kamols.lv/doc_upl/annotator-r885.zip*
281. **Siemen u. c. 2006 – Siemen, Peter; Lüdeling, Anke; Müller, Frank Henrik.** Falko – ein fehlerannotiertes Lernerkorpus des Deutschen. *Vortrag für Konvens 2006*. [skatīts 2015. gada 12. augustā]. Pieejams: *https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/SiemenLuedelingMueller-Konvens06.pdf.*
282. **Sigott, Dobrić 2014 – Sigott, Günther, Dobrić, Nikola.** *Learner Corpus Annotation Manual.* Klagenfurt : Alpen-Adria Universität, 2014.
283. **Sīlis 2009 – Sīlis, Jānis.** *Tulkojumzinātnes jautājumi. Teorija un prakse*. Ventspils : Ventspils Augstskola, 2009.
284. **Sinclair 2004 – Sinclair, John.** *Trust the text.* London, New York : Routledge, 2004.
285. **Skadiņa u. c. 2014 – Skadiņa, Inguna, Auziņa, Ilze, Bārzdiņš, Guntis, Skadiņš, Raivis, Vasiļjevs, Andrejs.** Language Resources and Technology in Latvia (2010–2014). *Human Language Technologies – The Baltic Perspective*. A. Utka et al. (Eds.) Amsterdam : IOS Press, 2014, pp. 227–235.
286. **Skadiņa, Vasiļjevs 2013** – **Skadiņa, Inguna, Vasiļjevs, Andrejs.** Valodas tehnoloģijas. *Latviešu valoda.* Andrejs Veisbergs (red.). Rīga : Latvijas Universitāte, 2014, 453.–475. lpp.
287. **Skujiņa-e – Skujiņa, Valentīna.** *Pārmaiņas izglītības sistēmā un pedagoģijas terminoloģija* [skatīts 2016. gada 11. decembrī]. Pieejams: *http://www.vvk.lv/ index.php?sadala=147&id=305&PHPSESSID=f2f2*

288. **Smolovskaya u. c. 2015 – Smolovskaya, Evgeniya, Mescheryakova, Evgeniy, Kisselev, Olesya, Rakhilina, Ekaterina.** Russian in the English mirror: (non)grammatical constructions in learner Russian. *Corpus linguistics 2015.* Abstract Book. Federica Formato, Andrew Hardie (eds.). Lancaster : UCREL, 2015, pp. 239–241.
289. **Sõrmus 2014 – Sõrmus, Kadri.** EMMA – a native Estonian learner text corpus. 21-oji tarptautinė Jono Jablonskio konferencija „Šiuolaikinės kalbos tyrimai ir problemos" 03.10.2014. Tezės. Vilnius : Vilniaus universitetas, Lietuvių kalbos institutas, 2014, 38 psl. [skatīts 2014. gada 29. septembrī]. Pieejams: *http://www.lki.lt/LKI_LT/images/Instituto_darbai/Konferencijos/20140921%20TEZES_Jablonskiui.pdf*
290. **Spektors 2000 – Spektors, Andrejs.** Datorlingvistika un tās resursi. Baltistika IX, 2000. Starptautiskais baltistu kongress „Baltu valodas laikmetu griežos" 03.10.2000.–06.10.2000. Referātu tēzes. Rīga : LU Latviešu valodas institūts, 2000, 296.–298. lpp.
291. **Surkova 2008 – Surkova, Irina.** Kreativitāte mērķvalodas lietošanā. Promocijas darbs. Rīga : Latvijas Universitāte, 2008.
292. **Šalme, Auziņa 2013 – Šalme, Arvils, Auziņa, Ilze.** Latviešu valodas prasmes līmeņi: Pamatlīmenis A1, A2. Rīga : Latviešu valodas aģentūra, 2013.
293. **Šeškauskienė 2008 – Šeškauskienė, Inesa.** Hedging in ESL: a Case Study of Lithuanian Learners. *Studies About Languages (Kalbų Studijos)*, issue: 13 / 2008, pp: 71–76
294. **Šeškauskienė, Juknevičienė 2015 – Šeškauskienė, Inesa, Juknevičienė, Rita.** From spatial to non-spatial prepositional meaning: IN and ON in the language of Lithuanian learners of English. *Third international learner corpus research conference, 11–13 September 2015*. Book of abstracts. Nijmegen : Radboud University, 2015, pp. 151–152.
295. **Šimčikaitė 2012 – Šimčikaitė, Alė.** Spoken Discourse Markers in Learner Academic Writing. *Kalbų studijos,* No. 20, 2012, pp. 27–34.
296. **Tan 2005 – Tan, Melinda.** Authentic language or language errors? Lessons form a learner corpus. *ELT Journal*, Vol. 59/2, April 2005.
297. **Tang 2011 – Tang, Warren.** *A Simple Guide to Using Antconc*. Hiroshima : Hiroshima University, 2011 [skatīts 2015. gada 8. novembrī]. Pieejams: *http://www.laurenceanthony.net/software/antconc/resources/help_AntConc321_english.pdf*
298. **TEI 2015 –** *TEI P5 : Guidelines for Electronic Text Encoding and Interchange.* The TEI Consortium, 2015 [skatīts 2015. gada 30. novembrī]. Pieejams: *http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf*
299. **TEITOK Projects-e –** *TEITOK Projects* [skatīts 2016. gada 6. maijā]. Pieejams: *http://alfclul.clul.ul.pt/teitok/site/index.php?action=projects*
300. **Tēzaurs-e – Spektors, Andrejs.** *Tēzaurs*. Rīga : LU MII Mākslīgā intelekta laboratorija [skatīts 2016. gada 17. janvārī]. Pieejams: *http://tezaurs.lv/*
301. **Tognini-Bonelli 2001 – Tognini-Bonelli, Elena.** *Corpus linguistics at work.* Amsterdam : John Benjamins, 2001.
302. **Tono 2002 – Tono, Yukio.** *The role of learner corpora in SLA research and foreign language teaching. The multiple comparison approach.* Dissertation. Lancaster : Lancaster University, 2002.
303. **Tono 2003 – Tono, Yukio.** Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom, 28–31 March 2003, pp. 800–809.

304. **Tono u. c. 2014 – Tono, Yukio, Satake, Yoshiho, Miura, Aika.** The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, Vol. 26, Special issue 02, May 2014, pp. 147–162.
305. **Trushkina, Erhard 2004 – Trushkina, Julia S., Hinrichs, Erhard W.** A Hybrid Model for Morpho-Syntactic Annotation of German with a Large Tagset. *Conference on Empirical Methods in Natural Language Processing*. 2004, pp. 238-245 [skatīts 2016. gada 22. janvārī]. Pieejams: *http://clair.eecs.umich.edu/aan/paper.php?paper_id=W04-3231#pdf*
306. **Ulrich 2002 – Ulrich, Winfried.** *Wörterbuch linguistische Grundbegriffe*. Berlin, Stuttgart : Gebrüder Borntraeger, 2002.
307. **Utka u. c. 2012 – Utka, Andrius, Levāne-Petrova, Kristīne, Bielinskienė, Agnė, Kovalevskaitė, Jolanta, Rimkutė, Erika, Vēvere, Daira.** Lithuanian-Latvian-Lithuanian parallel corpus. *Human language technologies – the Baltic perspective :* the fifth international conference Baltic HLT, Tartu, Estonia, October 4–5, 2012: proceedings. Amsterdam : IOS Press, 2012, pp. 260–264.
308. **Vanhaegendoren 2002 – Vanhaegendoren, Koen.** Fremdsprachendidaktik in Theorie und Praxis : Deutsch als Fremdsprache. Lage : Hans Jacobs, 2002.
309. **Vinčela 2010a – Vinčela, Zigrīda.** *Student-Composed Electronic Discourse as a Result of Applied Linguistic Research.* Promocijas darbs. Rīga : Latvijas Universitāte, 2010.
310. **Vinčela 2010b – Vinčela, Zigrīda.** Vietniekvārdu lietojums studentu elektroniskajos tekstos angļu valodā. *Vārds un tā pētīšanas aspekti* : rakstu krājums, 14(1). Liepāja : LiePA, 2010, 346.–351. lpp.
311. **Vinčela 2011a – Vinčela, Zigrīda.** Cross-cultural Online Communication: from Interactional to Transactional Writing. *Hermeneia* : Journal of Hermeneutics, Art Theory & Criticism, Issue 11, 2011, pp. 28–37.
312. **Vinčela 2011b – Vinčela, Zigrīda.** Leksikas un gramatisko formu lietojuma īpatnības studentu elektroniskajos tekstos angļu valodā. *Vārds un tā pētīšanas aspekti*. [Nr.]15, 2.[sēj.]: Valoda un vide. Terminoloģija. Tulkošanas jautājumi (2011), 352.–359. lpp. : diagr.
313. **Vinčela 2011c – Vinčela, Zigrīda.** Linguistic Variation in EFL Students-Composed Virtual Texts in Different Registers. *Corpus Linguistics Conference 2011*, Birmingham, 20-22 July, 2011 : Proceedings [elektronisks resurss], p. 9. Pieejams: *http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-223.pdf*
314. **Vinčela 2013a – Vinčela, Zigrīda.** Generic pronouns in Latvian student-composed essays in English: A comparison of the BNC (British National Corpus) and BCML (Balanced Corpus of Modern Latvian). *Corpus Linguistics, 2013* : [7th International Conference CL2013, 23-26 July, 2013, Lancaster University] : abstract book / ed. Andrew Hardie, Robbie Love. Lancaster : UCREL, 2013. Pp. 360–362. Pieejams: *http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf*
315. **Vinčela 2013b – Vinčela, Zigrīda.** Linking Adverbials in the Corpus of Student-Composed Texts. *Žmogus ir Žodis*, No. 1, 2013, pp. 215–222.
316. **Vinčela 2014 – Vinčela, Zigrīda.** Tagging Errors in Non-Native English Language Student-Composed Texts of Different Registers. *Baltic Journal of English Language, Literature and Culture,* Vol. 4, 2014, pp. 122–129.
317. **Vinčela 2016 – Vincela, Zigrida.** Complex Sentences and their Punctuation in English Texts Composed by Latvian Students. *Žmogus ir Žodis*, Vol. 18, No.1, 2016, pp. 96–105.

318. **VL 2012 –** *Valodas Latvijā : Pētījuma kopsavilkums*. Pauls Balodis, Māris Baltiņš, Vineta Ernstsone, Valts Ernštreits, Gunta Kļava, Dite Liepa, Kristīne Motivāne, Inese Muhka, Jānis Oga, Evija Papule, Jānis Valdmanis, Anna Vulāne. Rīga : Latviešu valodas aģentūra, 2012 [skatīts 2016. gada 2. janvārī]. Pieejams: *http://www.izm.gov.lv/images/statistika/petijumi/30.pdf*
319. **VPSV 2007** – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Red. V. Skujiņa. Rīga : LU Latviešu valodas institūts, 2007.
320. **VTV 1963 – Grabis, Rūdolfs, Barbare, Dzidra, Bergmane, Anna.** Valodniecības terminu vārdnīca. Rīga : LVI, 1963.
321. **VUPA 2011** – Vidurinio ugdymo programos aprašas. Lietuvos Respublikos Švietimo ir mokslo ministerija, 2011 [skatīts 2015. gada 13. augustā]. Pieejams: *http://www.smm.lt/uploads/documents/veikla/Veiklos_sritys/Svietimas/pradinis_ugdymas/Vidurinio%20ugdymo%20aprasas(tvirtinimui)0628.doc*
322. **Wible u. c.** – **Wible, David, Kuo, Chin-Hwa, Chien, Feng-yi, Liu, Anne, Tsao, Nai-Lung.** A Web-based EFL writing environment: integrating information for learners, teachers, and researchers. *Computers & Education,* vol. 37 (3-4), 2001, pp. 297-315.
323. **Windows-e** – *A history of Windows*. [skatīts 2015. gada 8. novembrī]. Pieejams: *http://windows.microsoft.com/en-US/windows/history#T1=era1*
324. **Zauberga 2001 – Zauberga, Ieva.** Discourse interference in translation. *Across Languages and Cultures* 2 (2), 2001, pp. 265–276.
325. **Zevakhina u. c. 2015 – Zevakhina, Natalia, Dzhakupova, Svetlana, Mustakimova, Elmira.** Corpus of Russian Student Texts: goals, annotation, and perspectives. *Corpus linguistics 2015.* Abstract Book. Federica Formato, Andrew Hardie (eds.). Lancaster : UCREL, 2015, pp. 444–445.
326. **Zinkevičius 2000 – Zinkevičius, Vytautas.** *Lemuoklis* – morfologinei analīzei. *Darbai ir dienos,* 24 (2000), 245–274 psl.
327. **Zinsmeister, Breckle 2010 – Zinsmeister, Heike, Breckle, Margit.** Starting a sentence in L2 German–Discourse annotation of a learner corpus. *Semantic approaches in natural language processing: Proceedings of the Conference on Natural Language Processing,* 2010, pp. 181–185.
328. **Zinsmeister, Breckle 2012 – Zinsmeister, Heike, Breckle, Margit.** The ALeSKo learner corpus: design–annotation–quantitative analyses. *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John Benjamins, 2012, pp. 71–96.
329. **Znotiņa 2012 – Znotiņa, Inga.** *Parodomieji įvardžiai lietuvių–latvių lygiagrečiajame tekstyne.* Magistro darbas. Kaunas : Vytauto Didžiojo universitetas, 2012.
330. **Znotiņa 2014 – Znotiņa, Inga.** Valodas apguvēju korpuss: lietuviešu un latviešu termins un definīcija. *Vārds un tā pētīšanas aspekti.* Liepāja : LiePa, 2014, 265.–273. lpp.
331. **Znotiņa 2015 – Znotiņa, Inga.** Pētniecības iespējas neanotētā baltu valodu apguvēju korpusā. *Vārds un tā pētīšanas aspekti.* Liepāja : LiePa, 2015.
332. **Zujevaitė, Žilinskaitė 2012 – Zujevaitė, Agnė, Žilinskaitė, Eglė.** Latvių kalbos kaip užsienio kalbos tekstynas. *Studentų moksliniai tyrimai.* Konferencijos pranešimų santraukos. Vilnius : Lietuvos mokslo taryba, 2012, 55.–57. psl.
333. **Živjuka 2008 – Živjuka, Svetlana.** *Kreativitāte rakstīšanas procesā.* Diplomdarbs. Rīga : Latvijas Universitāte, 2008.
334. **Žīgure 1999 – Žīgure, Veneta.** Biežāk sastopamās kļūdas, apgūstot latviešu valodas elementārkursu. *Sastatāmā un lietišķā valodniecība. Kontrastīvie pētījumi.*

Zinātniskie raksti, VIII / A. Veisberga redakcijā. Rīga : Latvijas Universitāte, 1999. 107.–113. lpp.

335. **Ахманова 1969 – Ахманова, Ольга.** *Словарь лингвистических терминов*. Изд. 2-ое, стереотип. Москва : Советская Энциклопедия, 1969.
336. **Багироков, Блягоз 2012 – Багироков, Хазрет, Блягоз, Зулькарин.** К вопросу о понятиях «родной язык» и «неродной язык». Вестник Адыгейского государственного университета. Серия 2: Филология и искусствоведение (2/2012) [skatots 2016. gada 10. janvbro]. Pieejams: *http://cyberleninka.ru/article/n/k-voprosu-o-ponyatiyah-rodnoy-yazyk-i-nerodnoy-yazyk*
337. **Камшилова 2009 – Камшилова Ольга.** Специальный корпус как составляющая лингвистического обеспечения языкового образования. *Иностранные языки в дис-танционном обучении: мат-лы. III Междунар. науч.-практ. конф.* Т.2 – Пермь : ПГТУ, 2009.
338. **Камшилова 2013 – Камшилова, Ольга.** Учебный корпус текстов: работа над ошибками. *Труды Международной конференции «Корпусная лингвистика – 2013», 25–27 июня 2013*. Санкт-Петербург : Санкт-Петербургский гос. университет, Филологический факультет, 2013, с. 301–308.
339. **Мальцева 2011 – Мальцева, Марине.** Учебный корпус (learner corpus) как база для лингвистического и лингводидактического анализа в рамках методики преподавания иностранных языков. *Социально-экономические явления и процессы* 9 (031), 2011, с. 209–212.
340. **Савчук, Сичинава 2009 – Савчук, Светлана, Сичинава, Дмитрий.** Обучающий корпус русского языка и его использование в преподавательской практике. *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы*. Санкт-Петербург : Нестор-История, 2009, с. 317–334.
341. **Стойкова 2013 – Стойкова, Татьяна.** Грамматическая интерференция в русской речи латышей: к проблеме освоения неродного языка. *Kontaktilingvistikas un slāvu valodniecības aktuālās problēmas*. Rīga : LU akadēmiskais apgāds, 2013, 73.–81. lpp.
342. **Щукин 2007 – Щукин, Анатолий.** *Лингводидактический энциклопедический словарь : более 2000 единиц*. Москва : Астрель, 2007.

Pielikumi

1. pielikums. Atļaujas teksta paraugs (latviešu valodā).
2. pielikums. Atļaujas teksta paraugs (lietuviešu valodā).
3. pielikums. Marķēts un anotēts teksts latviešu valodā.
4. pielikums. Marķēts un anotēts teksts lietuviešu valodā.
5. pielikums. Korpusa metadatu arhīva formāta paraugs.
6. pielikums. Korpusa lietošanas instrukcija ar pielikumiem.

1. pielikums. Atļaujas teksta paraugs (latviešu valodā).

ATĻAUJA

Es, _____ (vārds, uzvārds), personas kods _____-_____, piekrītu, ka mani 20___. gadā lietuviešu valodas kursā iesniegtie mājasdarbi

- „_____”,
- „_____”,
- „_____”,
- „_____”,
- „_____” un
- „_____”

tiek iekļauti otrās baltu valodas apguvēju korpusā un kā šī korpusa daļa kļūst publiski pieejami dažādās formās, pilnībā vai daļēji, ar šādiem nosacījumiem:

1. Korpusa veidotāja un publiskotāja ir Inga Znotiņa, p. k. _____-_____.
2. Korpuss ir pieejams bez maksas un ir paredzēts lietošanai mācību un pētniecības nolūkos. Tā lietošana komerciālos nolūkos ir aizliegta. Autori par tekstu iekļaušanu korpusā materiālu atlīdzību nesaņem.
3. Lai aizsargātu tekstos pieminēto cilvēku personas datus, tekstos sniegtā informācija var tikt mainīta, norādot vietu, kurā veiktas izmaiņas, bet nenorādot sākotnējo informāciju. Šī iemesla dēļ arī autora vārds pie katra konkrēta teksta netiek norādīts. Katram autoram tiek piešķirts anonīms kods, ar kura palīdzību ir iespējams atpazīt viena un tā paša autora vairākus darbus, bet nav iespējams noteikt autora identitāti.
4. Korpusā iekļautie dati dažādās formās var tikt citēti mācību līdzekļos un pētnieciskos darbos.
5. Korpuss un visi tajā iekļautie materiāli var būt publiski pieejami neierobežotu laiku un tikt aplūkoti un pētīti neierobežotu skaitu reižu.
6. Visiem korpusā iekļautajiem tekstiem var tikt pievienota lingvistiska informācija (piem., kļūdu labojumi, vārdšķiru anotējums u. c.).

Korpusā iekļauto tekstu autoru vārdi un uzvārdi tiek norādīti atsevišķā autoru sarakstā. Ja autors nevēlas tajā tikt iekļauts, viņš var izvēlēties palikt anonīms.

☐ Piekrītu, ka mans vārds un uzvārds tiek iekļauts autoru sarakstā.
☐ Vēlos palikt anonīms.

Datums:_____ Vārds, uzvārds:_____

Paraksts:_____

2. pielikums. Atļaujas teksta paraugs (lietuviešu valodā).

SUTIKIMAS

Aš, _____ (vardas, pavardė), asmens kodas _____, sutinku, kad mano _____ metais įteikti latvių kalbos kurso namų darbai

- „_____“,
- „_____“,
- „_____“,
- „_____“ ir
- „_____“

būtų įtraukti į antrosios baltų kalbos besimokančiųjų tekstyną ir, kad šio tekstyno dalis taptų viešai prieinama įvairiomis formomis, visiškai arba iš dalies, laikantis šių sąlygų:

1. Tekstyno sudarytoja ir skelbėja yra Inga Znuotinia (Inga Znotiņa), a. k. _____.
2. Tekstynas yra nemokamas ir yra skirtas naudoti mokymo ir tyrimų tikslais. Jį draudžiama naudoti komerciniais tikslais. Už tekstų įtraukimą į tekstyną autoriams neatlyginama.
3. Siekiant apsaugoti tekstuose minimų žmonių asmens duomenis, tekstuose pateikta informacija gali būti keičiama, nurodant vietą, kurioje atlikti pakeitimai, bet nenurodant pirminės informacijos. Dėl šios priežasties prie kiekvieno konkretaus teksto nenurodomas ir autoriaus vardas. Kiekvienam autoriui priskiriamas anoniminis kodas, kurio dėka galima atpažinti keletą to paties autoriaus darbų, bet neįmanoma nustatyti autoriaus tapatybės.
4. Į tekstyną įtraukti duomenys įvairiomis formomis gali būti cituojami mokymo priemonėse ir tiriamuosiuose darbuose.
5. Tekstynas ir visa jame esanti medžiaga gali būti viešai prieinama neribotą laiką ir peržiūrima bei tiriama neribotą skaičių kartų.
6. Prie visų į tekstyną įtrauktų tekstų gali būti pridėta lingvistinė informacija (pvz., klaidų pataisymai, kalbos dalių anotavimas ir kt.).

Į tekstyną įtrauktų tekstų autorių vardai ir pavardės nurodomi atskirame autorių sąraše. Jeigu autorius nenori būti į jį įtrauktas, jis gali likti anoniminis.

☐ Sutinku, kad mano vardas ir pavardė būtų įtraukti į autorių sąrašą.
☐ Noriu likti anoniminis.

Data:_____ Vardas, pavardė:_____

Parašas:_____

3. pielikums. Marķēts un anotēts teksts latviešu valodā.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
   <teiHeader>
   <profileDesc><langUsage><language><lang
n="LV"/></language></langUsage><institution><institAbbrev
n="VDU"/></institution><particDesc><listPerson><person><persName><name
n="2409"/></persName><langInstruction>LT</langInstruction><learning><semester
n="1"/></learning></person></listPerson></particDesc></profileDesc><fileDesc><titleStmt><titl
e>Mana ģimene un mājas</title><id
n="114290"/></titleStmt><publicationStmt><date>2016</date></publicationStmt></fileDesc></t
eiHeader>
   <text>
   <p id="p-1">
<s type="VPT" id="s-1">
<tok id="w-1" pos="p" lemma="es">Es</tok> <tok id="w-2" pos="v"
lemma="gribēt">gribu</tok> <tok id="w-3" pos="v" lemma="pastāstīt">pastāstīt</tok> <tok
id="w-4" lemma="par" pos="s">par</tok> <tok id="w-5" pos="p" lemma="mans" cform="manu"
error="ML">mana</tok> <tok id="w-6" pos="n" lemma="ģimene">ģimeni</tok> <tok id="w-7"
pos="c" lemma="un">un</tok> <tok id="w-8" cform="mājām" pos="n" lemma="mājas"
error="ML">mājas</tok><tok id="w-9" pos="z">.</tok>
</s>
<s type="VNT" id="s-2">
<tok id="w-10" pos="p" lemma="es">Es</tok> <tok id="w-11" pos="v"
lemma="būt">esmu</tok> <tok id="w-12" nform="Agne" pos="n">xxx</tok><tok id="w-13"
pos="z">.</tok>
</s>
<s type="VPT" id="s-3">
<tok id="w-14" pos="p" lemma="es">Man</tok> <tok id="w-15" pos="v" lemma="būt">ir</tok>
<tok id="w-16" pos="a" lemma="liels">liela</tok> <tok id="w-17" pos="n"
lemma="ģimene">ģimene</tok><tok id="w-18" pos="z">.</tok>
</s>
<s type="VPT" id="s-4">
<tok id="w-19" pos="p" lemma="es">Es</tok> <tok id="w-20" pos="v"
lemma="dzīvot">dzīvoju</tok> <tok id="w-21" lemma="ar" pos="s">ar</tok> <tok id="w-22"
pos="n" lemma="māte">māti</tok><tok id="w-23" pos="z">,</tok> <tok id="w-24" pos="n"
lemma="tēvs">tēvu</tok> <tok id="w-25" pos="c" lemma="un">un</tok> <tok id="w-26"
pos="n" lemma="brālis">brāli</tok><tok id="w-27" pos="z">.</tok>
</s>
```

<s type="SST" id="s-5">
<tok id="w-28" pos="p" lemma="es">Man</tok> <tok id="w-29" pos="v"
lemma="nebūt">nav</tok> <tok id="w-30" pos="n" lemma="māsa" cform="māsas"
error="ML">māsa</tok> <tok id="w-31" cform="," pos="z" error="IT">--</tok> <tok id="w-32"
pos="c" lemma="taču">taču</tok> <tok id="w-33" pos="v" lemma="būt">ir</tok> <tok id="w-
34" lemma="daudz" pos="r">daudz</tok> <tok id="w-35" pos="n"
lemma="māsīca">māsīcu</tok> <tok id="w-36" pos="c" lemma="un">un</tok> <tok id="w-37"
pos="n" lemma="brālēns">brālēnu</tok><tok id="w-38" pos="z">.</tok>
</s>
<s type="SST" id="s-6">
<tok id="w-39" pos="p" lemma="mēs">Mums</tok> <tok id="w-40" cform="ir" pos="v"
lemma="būt" error="SV">--</tok> <tok id="w-41" pos="c" lemma="arī">arī</tok> <tok id="w-42"
pos="v" lemma="būt" cform="--" error="SV">ir</tok> <tok id="w-43" lemma="divi"
pos="m">divi</tok> <tok id="w-44" pos="n" lemma="mājdzīvnieks">mājdzīvnieki</tok><tok
id="w-45" pos="z">:</tok> <tok id="w-46" pos="a" lemma="pelēks">pelēks</tok><tok id="w-47"
pos="z">,</tok> <tok id="w-48" pos="a" lemma="biezs" cform="resns" error="LN">biezs</tok>
<tok id="w-49" pos="c" lemma="un">un</tok> <tok id="w-50" lemma="ļoti" pos="r">ļoti</tok>
<tok id="w-51" pos="a" lemma="labs">labs</tok> <tok id="w-52" pos="n"
lemma="kaķis">kaķis</tok> <tok id="w-53" pos="n" cform="Bens" error="FP">Benas</tok> <tok
id="w-54" pos="c" lemma="un">un</tok> <tok id="w-55" pos="a" lemma="augsts" cform="liels"
error="LN">augsts</tok><tok id="w-56" pos="z">,</tok> <tok id="w-57" pos="a"
lemma="brūns">brūns</tok><tok id="w-58" pos="z">,</tok> <tok id="w-59" lemma="ļoti"
pos="r">ļoti</tok> <tok id="w-60" pos="a" lemma="gudrs">gudrs</tok> <tok id="w-61" pos="n"
lemma="suns">suns</tok><tok id="w-62" pos="z">,</tok> <tok id="w-63" pos="p"
lemma="tas">tā</tok> <tok id="w-64" pos="n" lemma="vārds">vārds</tok> <tok id="w-65"
cform="ir" error="SI">--</tok> <tok id="w-66" pos="n" cform="Džekijs"
error="FP">Džekis</tok><tok id="w-67" pos="z">.</tok>
</s>

<s type="SST" id="s-7">
<tok id="w-68" pos="p" lemma="es">Man</tok> <tok id="w-69" pos="v" lemma="būt">ir</tok>
<tok id="w-70" lemma="divdesmit" pos="m">divdesmit</tok> <tok id="w-71" pos="n"
lemma="gads">gadu</tok><tok id="w-72" pos="z">,</tok> <tok id="w-73" pos="c"
lemma="taču">taču</tok> <tok id="w-74" lemma="ātri" pos="r" cform="drīz"
error="LN">ātri</tok> <tok id="w-75" pos="v" lemma="būt">būs</tok> <tok id="w-76"

lemma="divdesmit" pos="m">divdesmit</tok> <tok id="w-77" lemma="viens"
pos="m">viens</tok><tok id="w-78" pos="z">.</tok>
</s>
<s type="VNT" id="s-8">
<tok id="w-79" pos="p" lemma="es">Es</tok> <tok id="w-80" pos="v"
lemma="būt">esmu</tok> <tok id="w-81" pos="n" lemma="studente">studente</tok><tok
id="w-82" pos="z">.</tok>
</s>
<s type="VPT" id="s-9">
<tok id="w-83" pos="p" lemma="es">Es</tok> <tok id="w-84" pos="v"
lemma="mācīties">mācos</tok> <tok id="w-85" pos="n">Vītauta</tok> <tok id="w-86"
pos="a">Dižā</tok> <tok id="w-87" pos="n" lemma="universitāte">universitātes</tok> <tok
id="w-88" pos="a" lemma="humanitārs">Humanitāro</tok> <tok id="w-89" pos="n"
lemma="zinātne">zinātņu</tok> <tok id="w-90" pos="n" lemma="fakultāte">fakultātē</tok><tok
id="w-91" pos="z">.</tok>
</s>
<s type="VPT" id="s-10">
<tok id="w-92" pos="p" lemma="es">Es</tok> <tok id="w-93" pos="v"
lemma="studēt">studēju</tok> <tok id="w-94" pos="n" lemma="lietuvietis">lietuviešu</tok> <tok
id="w-95" pos="n" lemma="filoloģija">filoloģiju</tok> <tok id="w-96" lemma="trešais"
pos="m">trešajā</tok> <tok id="w-97" pos="n" lemma="kurss">kursā</tok><tok id="w-98"
pos="z">.</tok>
</s>
<s type="VPT" id="s-11">
<tok id="w-99" pos="p" lemma="es">Man</tok> <tok id="w-100" pos="v"
lemma="patikt">patīk</tok> <tok id="w-101" pos="v" lemma="tikties" cform="tikties"
error="FP">tiktis</tok> <tok id="w-102" lemma="ar" pos="s">ar</tok> <tok id="w-103" pos="n"
lemma="draugs">draugiem</tok><tok id="w-104" pos="z">,</tok> <tok id="w-105" pos="v"
lemma="zīmēt">zīmēt</tok> <tok id="w-106" pos="c" lemma="un">un</tok> <tok id="w-107"
pos="v" lemma="jokot">jokot</tok><tok id="w-108" pos="z">,</tok> <tok id="w-109" pos="v"
lemma="ēst">ēst</tok> <tok id="w-110" pos="n" lemma="banāns">banānus</tok> <tok id="w-
111" pos="c" lemma="un">un</tok> <tok id="w-112" pos="n"
lemma="mandarīns">mandarīnus</tok><tok id="w-113" pos="z">,</tok> <tok id="w-114"
pos="v" lemma="svētīt">svētīt</tok> <tok id="w-115" pos="n"
lemma="Ziemassvētki">Ziemassvētkus</tok> <tok id="w-116" pos="c" lemma="un">un</tok>
<tok id="w-117" pos="v" lemma="sajust">sajust</tok> <tok id="w-118" pos="n"
lemma="siens">siena</tok> <tok id="w-119" pos="c" lemma="un">un</tok> <tok id="w-120"

```
pos="n" lemma="ieva">ievas</tok> <tok id="w-121" pos="n"
lemma="smarža">smaržas</tok><tok id="w-122" pos="z">.</tok>
</s>
<s type="SPT" id="s-12">
<tok id="w-123" pos="n" lemma="draugs">Draugi</tok> <tok id="w-124" pos="v"
lemma="sacīt">saka</tok><tok id="w-125" pos="z">,</tok> <tok id="w-126" lemma="ka"
pos="c">ka</tok> <tok id="w-127" pos="p" lemma="es">es</tok> <tok id="w-128" pos="v"
lemma="būt" cform="esmu" error="FP">esu</tok> <tok id="w-129" pos="a"
lemma="smieklīgs">smieklīga</tok><tok id="w-130" pos="z">,</tok> <tok id="w-131" pos="a"
lemma="draudzīgs">draudzīga</tok><tok id="w-132" pos="z">.</tok>
</s>
<s type="VPT" id="s-13">
<tok id="w-133" pos="p" lemma="mans" cform="Manas" error="MD">Mans</tok> <tok id="w-
134" pos="n" lemma="acs">acis</tok> <tok id="w-135" pos="v" lemma="būt">ir</tok> <tok
id="w-136" pos="a" lemma="brūns">brūnas</tok><tok id="w-137" pos="z">.</tok>
</s>
<s type="VPT" id="s-14">
<tok id="w-138" pos="p" lemma="es">Man</tok> <tok id="w-139" pos="v"
lemma="būt">ir</tok> <tok id="w-140" lemma="tumši" pos="r">tumši</tok> <tok id="w-141"
pos="a" lemma="brūns">brūni</tok> <tok id="w-142" pos="a" lemma="garš" cform="gari"
error="FP">garši</tok> <tok id="w-143" pos="n" lemma="mats">mati</tok><tok id="w-144"
pos="z">.</tok>
</s>
</p>

<p id="p-3">
<s type="VPT" id="s-15">
<tok id="w-145" pos="p" lemma="mans" cform="Manas" error="SS">Mans</tok> <tok id="w-146"
pos="n" lemma="māte">mātes</tok> <tok id="w-147" pos="n" lemma="vārds">vārds</tok> <tok
id="w-148" pos="v" lemma="būt">ir</tok> <tok id="w-149" nform="Lina"
pos="n">xxx</tok><tok id="w-150" pos="z">.</tok>
</s>
<s type="VPT" id="s-16">
<tok id="w-151" pos="p" lemma="viņa">Viņai</tok> <tok id="w-152" pos="v"
lemma="būt">ir</tok> <tok id="w-153" lemma="četrdesmit" pos="m">četrdesmit</tok> <tok
id="w-154" lemma="deviņi" pos="m">deviņi</tok><tok id="w-155" pos="z">.</tok>
</s>
<s type="VNT" id="s-17">
```

<tok id="w-156" pos="n" lemma="māte">Māte</tok> <tok id="w-157" pos="v" lemma="būt">ir</tok> <tok id="w-158" pos="n" lemma="pavāre">pavāre</tok><tok id="w-159" pos="z">.</tok>
</s>
<s type="SST" id="s-18">
<tok id="w-160" pos="p" lemma="viņa">Viņai</tok> <tok id="w-161" pos="v" lemma="patikt">patīk</tok> <tok id="w-162" pos="v" lemma="ceļot">ceļot</tok> <tok id="w-163" cform="," pos="z" error="IT">--</tok> <tok id="w-164" pos="c" lemma="un">un</tok> <tok id="w-165" pos="n" lemma="māte">māte</tok> <tok id="w-166" cform="ir" error="MT">--</tok> <tok id="w-167" pos="v" lemma="apmeklēt" cform="apmeklējusi" error="MT">apmeklēja</tok> <tok id="w-168" pos="n" lemma="Krievija">Krieviju</tok><tok id="w-169" pos="z">,</tok> <tok id="w-170" pos="n" lemma="Francija">Franciju</tok><tok id="w-171" pos="z">,</tok> <tok id="w-172" pos="n" lemma="Vācija">Vāciju</tok> <tok id="w-173" pos="c" lemma="un">un</tok> <tok id="w-174" lemma="daudz" pos="r">daudz</tok> <tok id="w-175" lemma="cits" pos="p">citu</tok> <tok id="w-176" pos="n" lemma="valsts">valstu</tok> <tok id="w-177" cform="," pos="z" error="IT">--</tok> <tok id="w-178" pos="c" lemma="un">un</tok> <tok id="w-179" pos="p" lemma="viņa" cform="viņai" error="FL">Viņai</tok> <tok id="w-180" pos="v" lemma="patikt">patīk</tok> <tok id="w-181" pos="v" lemma="ceļot">ceļot</tok> <tok id="w-182" lemma="par" pos="s" cform="pa" error="LN">par</tok> <tok id="w-183" pos="n" lemma="Lietuva">Lietuvu</tok><tok id="w-184" pos="z">.</tok>
</s>
<s type="VPT" id="s-19">
<tok id="w-185" lemma="daudz" pos="r" cform="Vēl" error="LN">Vairāk</tok> <tok id="w-186" cform="viņai" error="SI">--</tok> <tok id="w-187" pos="v" lemma="patikt">patīk</tok> <tok id="w-188" pos="v" lemma="adīt">adīt</tok><tok id="w-189" pos="z">,</tok> <tok id="w-190" pos="v" lemma="klausīties">klausīties</tok> <tok id="w-191" pos="n" lemma="mūzika" cform="mūziku" error="ML">mūzikas</tok> <tok id="w-192" pos="c" lemma="un">un</tok> <tok id="w-193" pos="v" lemma="dejot" cform="dejot" error="FP">dejat</tok><tok id="w-194" pos="z">.</tok>
</s>
<s type="SST" id="s-20">
<tok id="w-195" pos="n" lemma="māte">Mātes</tok> <tok id="w-196" pos="n" lemma="mats">mati</tok> <tok id="w-197" pos="v" lemma="būt">ir</tok> <tok id="w-198" pos="a" lemma="brūns">brūni</tok> <tok id="w-199" pos=" c" lemma="un">un</tok> <tok id="w-200" pos="a" lemma="īss">īsi</tok><tok id="w-201" pos="z">,</tok> <tok id="w-202" pos="n" lemma="acs">acis</tok> <tok id="w-203" cform="–" error="IT">--</tok> <tok id="w-204" pos="a" lemma="zils">zilas</tok><tok id="w-205" pos="z">.</tok>

```
</s>
</p>

<p id="p-4">
<s type="VPT" id="s-21">
<tok id="w-206" pos="p" lemma="mans">Mana</tok> <tok id="w-207" pos="n"
lemma="tēvs">tēva</tok> <tok id="w-208" pos="n" lemma="vārds">vārds</tok> <tok id="w-209"
pos="v" lemma="būt">ir</tok> <tok id="w-210" nform="Gintaras" pos="n" cform="Gintars"
error="FP">xxx</tok><tok id="w-211" pos="z">.</tok>
</s>
<s type="VPT" id="s-22">
<tok id="w-212" pos="p" lemma="viņš">Viņam</tok> <tok id="w-213" pos="c"
lemma="arī">arī</tok> <tok id="w-214" pos="v" lemma="būt">ir</tok> <tok id="w-215"
lemma="četrdesmit" pos="m">četrdesmit</tok> <tok id="w-216" lemma="deviņi"
pos="m">deviņi</tok><tok id="w-217" pos="z">.</tok>
</s>
<s type="SST" id="s-23">
<tok id="w-218" pos="p" lemma="viņa" cform="Viņš" error="MD">Viņas</tok> <tok id="w-219"
pos="v" lemma="būt">ir</tok> <tok id="w-220" pos="n" lemma="uzņēmējs" cform="uzņēmējs"
error="FD">uzņemējs</tok> <tok id="w-221" cform="," error="IT">--</tok> <tok id="w-222"
pos="c" lemma="un">un</tok> <tok id="w-223" pos="p" lemma="viņš">viņam</tok> <tok id="w-
224" pos="v" lemma="būt">ir</tok> <tok id="w-225" pos="n"
lemma="uzņēmums">uzņēmums</tok><tok id="w-226" pos="z">.</tok>
</s>
<s type="SST" id="s-24">
<tok id="w-227" pos="n" lemma="tēvs">Tēvs</tok> <tok id="w-228" pos="v"
lemma="interesēties" cform="interesējas" error="MP">interesējies</tok> <tok id="w-229"
lemma="par" pos="s">par</tok> <tok id="w-230" pos="n" lemma="automobilis"
cform="automobiļiem" error="FD">automobiliem</tok> <tok id="w-231" pos="c"
lemma="un">un</tok> <tok id="w-232" pos="n" lemma="tehnika" cform="tehniku"
error="FP">techniku</tok> <tok id="w-233" cform="," error="IT">--</tok> <tok id="w-234"
pos="c" lemma="un">un</tok> <tok id="w-235" pos="p" lemma="tas" cform="viņam"
error="LN">tam</tok> <tok id="w-236" pos="v" lemma="patikt">patīk</tok> <tok id="w-237"
pos="v" lemma="makšķerēt">makšķerēt</tok> <tok id="w-238" pos="c" lemma="un">un</tok>
<tok id="w-239" pos="v" lemma="ēst">ēst</tok><tok id="w-240" pos="z">.</tok>
</s>
<s type="VPT" id="s-25">
```

<tok id="w-241" pos="n" lemma="tēvs">Tēva</tok> <tok id="w-242" pos="n" lemma="mats">mati</tok> <tok id="w-243" pos="v" lemma="būt">ir</tok> <tok id="w-244" pos="a" lemma="tumšs">tumši</tok> <tok id="w-245" pos="c" lemma="un">un</tok> <tok id="w-246" pos="a" lemma="īss">īsi</tok><tok id="w-247" pos="z">.</tok>
</s>
<s type="VPT" id="s-26">
<tok id="w-248" pos="p" lemma="viņš">Viņa</tok> <tok id="w-249" pos="n" lemma="acs">acis</tok><tok id="w-250" pos="z">,</tok> <tok id="w-251" cform="tāpat" error="SI">--</tok> <tok id="w-252" lemma="kā" pos="r">kā</tok> <tok id="w-253" pos="p" lemma="mans" cform="manas" error="ML">manu</tok><tok id="w-254" pos="z">,</tok> <tok id="w-255" pos="v" lemma="būt">ir</tok> <tok id="w-256" pos="a" lemma="brūns">brūnas</tok><tok id="w-257" pos="z">.</tok>
</s>

<s type="VPT" id="s-27">
<tok id="w-258" pos="p" lemma="mans">Mana</tok> <tok id="w-259" pos="n" lemma="brālis" cform="brāļa" error="ML">brāli</tok> <tok id="w-260" pos="n" lemma="vārds">vārds</tok> <tok id="w-261" pos="v" lemma="būt">ir</tok> <tok id="w-262" nform="Vytautas" pos="n" cform="Vītauts" error="FP">xxx</tok><tok id="w-263" pos="z">.</tok>
</s>
<s type="VPT" id="s-28">
<tok id="w-264" pos="p" lemma="viņš">Viņam</tok> <tok id="w-265" pos="v" lemma="būt">ir</tok> <tok id="w-266" lemma="astoņpadsmit" pos="m">astoņpadsmit</tok><tok id="w-267" pos="z">.</tok>
</s>
<s type="VPT" id="s-29">
<tok id="w-268" pos="n" lemma="brālis">Brālis</tok> <tok id="w-269" pos="v" lemma="būt">ir</tok> <tok id="w-270" pos="n" lemma="skolnieks">skolnieks</tok><tok id="w-271" pos="z">,</tok> <tok id="w-272" lemma="tagad" pos="r">tagad</tok> <tok id="w-273" pos="v" lemma="mācīties">mācās</tok> <tok id="w-274" lemma="divpadsmitais" pos="m" cform="divpadsmitajā" error="ML">divpadsmitais</tok> <tok id="w-275" pos="n" lemma="klase">klasē</tok><tok id="w-276" pos="z">.</tok>
</s>
<s type="VPT" id="s-30">
<tok id="w-277" pos="p" lemma="tas" cform="Viņam" error="LN">Tam</tok> <tok id="w-278" pos="v" lemma="patikt">patīk</tok> <tok id="w-279" pos="n"

lemma="futbols">futbols</tok><tok id="w-280" pos="z">,</tok> <tok id="w-281" pos="n"
lemma="basketbols">basketbols</tok><tok id="w-282" pos="z">,</tok> <tok id="w-283"
lemma="vazināties" pos="v" cform="braukāties" error="MA">vazinātes</tok> <tok id="w-284"
lemma="ar" pos="s">ar</tok> <tok id="w-285" pos="n" lemma="ritenis" cform="riteni"
error="FD">ritēni</tok><tok id="w-286" pos="z">,</tok> <tok id="w-287" pos="v"
lemma="būt">būt</tok> <tok id="w-288" lemma="ar" pos="s">ar</tok> <tok id="w-289"
pos="n" lemma="draugs">draugiem</tok><tok id="w-290" pos="z">.</tok>
</s>
<s type="SST" id="s-31">
<tok id="w-291" pos="p" lemma="viņš">Viņa</tok> <tok id="w-292" pos="n"
lemma="acs">acis</tok> <tok id="w-293" pos="v" lemma="būt">ir</tok> <tok id="w-294"
pos="a" lemma="zils">zilas</tok><tok id="w-295" pos="z">,</tok> <tok id="w-296" lemma="kā"
pos="r">kā</tok> <tok id="w-297" pos="n" lemma="māte" cform="mātei"
error="ML">mātes</tok><tok id="w-298" pos="z">,</tok> <tok id="w-299" pos="p"
lemma="viņš">viņam</tok> <tok id="w-300" pos="v" lemma="būt">ir</tok> <tok id="w-301"
pos="a" lemma="īss">īsi</tok> <tok id="w-302" pos="n" lemma="mats" cform="mati"
error="FD">māti</tok> <tok id="w-303" cform="," error="IT">--</tok> <tok id="w-304" pos="c"
lemma="un">un</tok> <tok id="w-305" pos="n" lemma="brālis">brālis</tok> <tok id="w-306"
pos="v" lemma="būt">ir</tok> <tok id="w-307" pos="a" lemma="garš" cform="gara"
error="FP">gera</tok> <tok id="w-308" pos="n" lemma="augums">auguma</tok><tok id="w-
309" pos="z">.</tok>
</s>

<s type="VPT" id="s-32">
<tok id="w-310" pos="p" lemma="mans">Manās</tok> <tok id="w-311" pos="n"
lemma="mājas">mājās</tok> <tok id="w-312" pos="v" lemma="būt">ir</tok> <tok id="w-313"
lemma="astoņi" pos="m" cform="astoņas" error="MD">astoņi</tok> <tok id="w-314" pos="n"
lemma="istaba">istabas</tok><tok id="w-315" pos="z">:</tok> <tok id="w-316" lemma="viens"
pos="m">viena</tok> <tok id="w-317" pos="n" lemma="virtuve">virtuve</tok><tok id="w-318"
pos="z">,</tok> <tok id="w-319" lemma="viens" pos="m">viena</tok> <tok id="w-320" pos="n"
lemma="vannasistaba">vannasistaba</tok><tok id="w-321" pos="z">,</tok> <tok id="w-322"
lemma="divi" pos="m">divas</tok> <tok id="w-323" pos="n"
lemma="tualete">tualetes</tok><tok id="w-324" pos="z">,</tok> <tok id="w-325" lemma="trīs"
pos="m">trīs</tok> <tok id="w-326" pos="n" lemma="guļamistaba">guļamistabas</tok><tok
id="w-327" pos="z">,</tok> <tok id="w-328" lemma="viens" pos="m">viena</tok> <tok id="w-
329" pos="n" lemma="viesistaba">viesistaba</tok> <tok id="w-330" pos="c"

```
lemma="un">un</tok> <tok id="w-331" pos="a" lemma="liels">liels</tok> <tok id="w-332"
pos="n" lemma="koridors">koridors</tok><tok id="w-333" pos="z">.</tok>
</s>
<s type="VPT" id="s-33">
<tok id="w-334" lemma="parasti" pos="r">Parasti</tok> <tok id="w-335" pos="p"
lemma="es">es</tok> <tok id="w-336" pos="v" lemma="pavadīt">pavadu</tok> <tok id="w-337"
pos="n" lemma="laiks">laiku</tok> <tok id="w-338" pos="p" lemma="savs">savā</tok> <tok
id="w-339" pos="n" lemma="istaba">istabā</tok><tok id="w-340" pos="z">.</tok>
</s>
<s type="VPT" id="s-34">
<tok id="w-341" pos="p" lemma="mans">Mana</tok> <tok id="w-342" pos="n" lemma="istaba"
cform="istaba" error="ML">istabā</tok> <tok id="w-343" pos="v" lemma="būt">ir</tok> <tok
id="w-344" pos="a" lemma="mazs">maza</tok><tok id="w-345" pos="z">.</tok>
</s>
<s type="VPT" id="s-35">
<tok id="w-346" pos="p" lemma="viņš" cform="Tā" error="LN">Viņš</tok> <tok id="w-347"
pos="v" lemma="būt">ir</tok> <tok id="w-348" pos="a" lemma="oranžs" cform="oranžā"
error="ML">oranža</tok> <tok id="w-349" pos="n" lemma="krāsa" cform="krāsā"
error="ML">krāsas</tok><tok id="w-350" pos="z">.</tok>
</s>
<s type="SPT" id="s-36">
<tok id="w-351" lemma="uz" pos="s">Uz</tok> <tok id="w-352" pos="n"
lemma="grīda">grīdas</tok> <tok id="w-353" pos="v" lemma="būt">ir</tok> <tok id="w-354"
pos="a" lemma="zils">zils</tok> <tok id="w-355" pos="n" lemma="paklājs">paklājs</tok><tok
id="w-356" pos="z">,</tok> <tok id="w-357" pos="p" lemma="kurš">kurš</tok> <tok id="w-358"
pos="v" lemma="derēt" cform="piestāv" error="LV">der</tok> <tok id="w-359" lemma="pie"
pos="s">pie</tok> <tok id="w-360" pos="n" lemma="siena" cform="sienām"
error="ML">sienu</tok><tok id="w-361" pos="z">.</tok>
</s>
<s type="VPT" id="s-37">
<tok id="w-362" lemma="pie" pos="s">Pie</tok> <tok id="w-363" pos="n"
lemma="logs">loga</tok> <tok id="w-364" pos="v" lemma="būt">ir</tok>  <tok id="w-365"
pos="a" lemma="sarkans">sarkanas</tok> <tok id="w-366" pos="n"
lemma="žalūzija">žalūzijas</tok><tok id="w-367" pos="z">.</tok>
</s>
</p>
   </text>
</TEI>
```

4. pielikums. Marķēts un anotēts teksts lietuviešu valodā.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
   <teiHeader>
   <profileDesc><langUsage><language><lang
n="LT"/></language></langUsage><institution><institAbbrev
n="LU"/></institution><particDesc><listPerson><person><persName><name
n="1445"/></persName><langInstruction>LV</langInstruction><learning><semester
n="2"/></learning></person></listPerson></particDesc></profileDesc><fileDesc><titleStmt><titl
e>Drabužiai žiemą ir vasarą</title><id
n="003048"/></titleStmt><publicationStmt><date>2016</date></publicationStmt></fileDesc></t
eiHeader>
   <text>
<p id="p-1">
<s type="VPT" id="s-1">
   <tok id="w-1" pos="p" lemma="aš">Man</tok> <tok id="w-2" lemma="labai"
pos="r">labai</tok> <tok id="w-3" pos="v" lemma="patikti">patinka</tok> <tok id="w-4"
pos="n" lemma="žiema">žiema</tok><tok id="w-5" pos="z">.</tok>
</s>
<s type="VPT" id="s-2">
   <tok id="w-6" pos="p" lemma="aš">Aš</tok> <tok id="w-7" pos="n"
lemma="žiema">žiemą</tok> <tok id="w-8" lemma="paprastai" pos="r">paprastai</tok> <tok
id="w-9" pos="v" lemma="nešioti">nešioju</tok> <tok id="w-10" pos="a"
lemma="languotas">languotą</tok> <tok id="w-11" pos="a" lemma="juodas">juodą</tok> <tok
id="w-12" pos="n" lemma="paltas">paltą</tok> <tok id="w-13" lemma="iš" pos="s">iš</tok>
<tok id="w-14" pos="n" lemma="vilna">vilnos</tok><tok id="w-15" pos="z">,</tok> <tok id="w-
16" pos="n" lemma="džinsai">džinsus</tok> <tok id="w-17" lemma="arba" pos="c">arba</tok>
<tok id="w-18" pos="n" lemma="kelnės" cform="kelnes" error="ML">kelnės</tok> <tok id="w-19"
lemma="ir" pos="c">ir</tok> <tok id="w-20" lemma="šiltas" pos="a">šiltus</tok> <tok id="w-21"
pos="n" lemma="batas">batus</tok><tok id="w-22" pos="z">.</tok>
</s>
<s type="SJT" id="s-3">
   <tok id="w-23" pos="p" lemma="aš">Man</tok> <tok id="w-24" pos="v"
lemma="nepatikti">nepatinka</tok> <tok id="w-25" pos="n" lemma="kepurė">kepurė</tok><tok
id="w-26" pos="z">,</tok> <tok id="w-27" lemma="bet" pos="c">bet</tok> <tok id="w-28"
cform="," error="IT">--</tok> <tok id="w-29" lemma="kada" pos="r" cform="kai"
error="LV">kada</tok> <tok id="w-30" pos="v" lemma="būti">yra</tok> <tok id="w-31"
lemma="labai" pos="r">labai</tok> <tok id="w-32" lemma="šaltas" pos="a">šalta</tok> <tok
id="w-33" cform="," error="IT">--</tok> <tok id="w-34" pos="p" lemma="aš">aš</tok> <tok
```

id="w-35" pos="v" lemma="nešioti">nešioju</tok> <tok id="w-36" pos="n" lemma="kepurė">kepures</tok><tok id="w-37" pos="z" cform="--" error="IL">,</tok> <tok id="w-38" lemma="irgi" pos="q" cform="ir" error="SS">irgi</tok> <tok id="w-39" pos="n" lemma="pirštinė">pirštines</tok><tok id="w-40" pos="z">.</tok>
</s>
<s type="VPT" id="s-4">
<tok id="w-41" pos="p" lemma="aš">Aš</tok> <tok id="w-42" pos="v" lemma="turėti">turiu</tok> <tok id="w-43" lemma="labai" pos="r">labai</tok> <tok id="w-44" lemma="daug" pos="r">daug</tok> <tok id="w-45" pos="n" lemma="šalikas">šalikų</tok> <tok id="w-46" cform="," error="IT">--</tok> <tok id="w-47" lemma="ir" pos="c" cform="--" error="SL">ir</tok> <tok id="w-48" lemma="todėl" pos="r">todėl</tok> <tok id="w-49" pos="p" lemma="kiekvienas" cform="kiekvieną" error="FP">kikvieną</tok> <tok id="w-50" pos="n" lemma="diena">dieną</tok> <tok id="w-51" pos="v" lemma="nešioti">nešioju</tok> <tok id="w-52" pos="p" lemma="kitas">kitą</tok> <tok id="w-53" pos="n" lemma="šalikas">šaliką</tok><tok id="w-54" pos="z">.</tok>
</s>
<s type="VPT" id="s-5">
<tok id="w-55" pos="n" lemma="žiema">Žiemą</tok> <tok id="w-56" pos="p" lemma="aš">aš</tok> <tok id="w-57" lemma="paprastai" pos="r">paprastai</tok> <tok id="w-58" pos="v" lemma="dėvėti">dėviu</tok> <tok id="w-59" pos="n" lemma="megztinis">megztinį</tok> <tok id="w-60" lemma="ir" pos="c">ir</tok> <tok id="w-61" pos="n" lemma="kelnės" cform="kelnes" error="ML">kelnės</tok><tok id="w-62" pos="z">,</tok> <tok id="w-63" lemma="beveik" pos="r">beveik</tok> <tok id="w-64" lemma="niekada" pos="r">niekada</tok> <tok id="w-65" cform="nedėviu" error="SI">--</tok> <tok id="w-66" pos="n" lemma="suknelė" cform="suknelės" error="ML">suknėle</tok><tok id="w-67" pos="z">.</tok>
</s>
<s type="SPT" id="s-6">
<tok id="w-68" lemma="kartais" pos="r">Kartais</tok> <tok id="w-69" pos="v" lemma="dėvėti">dėviu</tok> <tok id="w-70" pos="n" lemma="pėdkelnės" cform="pėdkelnes" error="FD">pedkelnes</tok> <tok id="w-71" lemma="po" pos="s">po</tok> <tok id="w-72" pos="n" lemma="kelnės">kelnėmis</tok><tok id="w-73" pos="z">,</tok> <tok id="w-74" lemma="kada" pos="r" cform="kai" error="LV">kada</tok> <tok id="w-75" pos="v" lemma="būti">yra</tok> <tok id="w-76" lemma="labai" pos="r">labai</tok> <tok id="w-77" lemma="šaltas" pos="a">šalta</tok><tok id="w-78" pos="z">.</tok>
</s>
<s type="VPT" id="s-7">

<tok id="w-79" pos="n" lemma="vasara">Vasarą</tok> <tok id="w-80" pos="p" lemma="aš">aš</tok> <tok id="w-81" pos="v" lemma="nešioti">nešioju</tok> <tok id="w-82" pos="n" lemma="sijonas">sijonus</tok><tok id="w-83" pos="z" cform=":" error="IN">,</tok> <tok id="w-84" pos="a" lemma="raštuotas">raštuotus</tok><tok id="w-85" pos="z">,</tok> <tok id="w-86" pos="a" lemma="taškuotas">taškuotus</tok> <tok id="w-87" cform="," error="IT">--</tok> <tok id="w-88" pos="a" lemma="gėlėtas">gėlėtus</tok> <tok id="w-89" lemma="ir" pos="c">ir</tok> <tok id="w-90" pos="p" lemma="kitas" cform="kitokius" error="LN">kitus</tok><tok id="w-91" pos="z">.</tok>
</s>
<s type="VPT" id="s-8">
<tok id="w-92" lemma="taip pat" pos="r">Taip pat</tok> <tok id="w-93" lemma="ir" pos="q" cform="--" error="SL">ir</tok> <tok id="w-94" pos="n" lemma="suknelė" cform="sukneles" error="FD">suknėles</tok><tok id="w-95" pos="z">,</tok> <tok id="w-96" pos="n" lemma="marškinėliai">marškinėlius</tok> <tok id="w-97" lemma="ir" pos="c">ir</tok> <tok id="w-98" pos="n" lemma="basutė">basutes</tok><tok id="w-99" pos="z">.</tok>
</s>
<s type="VPT" id="s-9">
<tok id="w-100" pos="p" lemma="aš">Man</tok> <tok id="w-101" lemma="labai" pos="r">labai</tok> <tok id="w-102" pos="v" lemma="patikti">patinka</tok> <tok id="w-103" pos="v" lemma="dėvėti">dėvėti</tok> <tok id="w-104" pos="n" lemma="skrybėlė" cform="skybėlę" error="FD">skrybele</tok><tok id="w-105" pos="z">.</tok>
</s>
<s type="SST" id="s-10">
<tok id="w-106" pos="p" lemma="aš">Man</tok> <tok id="w-107" pos="v" lemma="patikti">patinka</tok> <tok id="w-108" pos="a" lemma="juodas">juoda</tok><tok id="w-109" pos="z">,</tok> <tok id="w-110" pos="a" lemma="pilkas">pilka</tok><tok id="w-111" pos="z">,</tok> <tok id="w-112" pos="a" lemma="baltas">balta</tok> <tok id="w-113" pos="n" lemma="spalva">spalva</tok> <tok id="w-114" cform="," error="IT">--</tok> <tok id="w-115" lemma="ir" pos="c" cform="--" error="SL">ir</tok> <tok id="w-116" lemma="todėl" pos="r">todėl</tok> <tok id="w-117" pos="a" lemma="didelis">didelė</tok> <tok id="w-118" pos="n" lemma="dalis">dalis</tok> <tok id="w-119" pos="n" lemma="drabužis" cform="drabužių" error="ML">drabužiai</tok> <tok id="w-120" pos="v" lemma="būti">yra</tok> <tok id="w-121" pos="p" lemma="toks">tokios</tok> <tok id="w-122" pos="n" lemma="spalva">spalvos</tok><tok id="w-123" pos="z">.</tok>
</s>
</p>
</text>
</TEI>

5. pielikums. Korpusa metadatu arhīva formāta paraugs.

Kods	Vārds	Uzvārds	Iepr. uzvārds	Augstskola	Pasniedzējs	Materiālu iedeva
0000	Pēteris	Pēternieks	-	SU	Miķelis Miķelnieks	Miķelis Miķelnieks

Gads	Digitalizēja	Teksta kods	Vārdu skaits	Tēma	Valoda	Iekļaušana sarakstā	Semestris
2009	-	000000	123	Mano šeima	LT	jā	1

Filologs	Mācību valoda	Mācību valsts
jā	latviešu	LV

6. pielikums. Korpusa lietošanas instrukcija ar pielikumiem

Inga Znotiņa
inga.s.znotina@gmail.com

Otrās baltu valodas apguvēju korpusa "Esam" lietošanas instrukcija

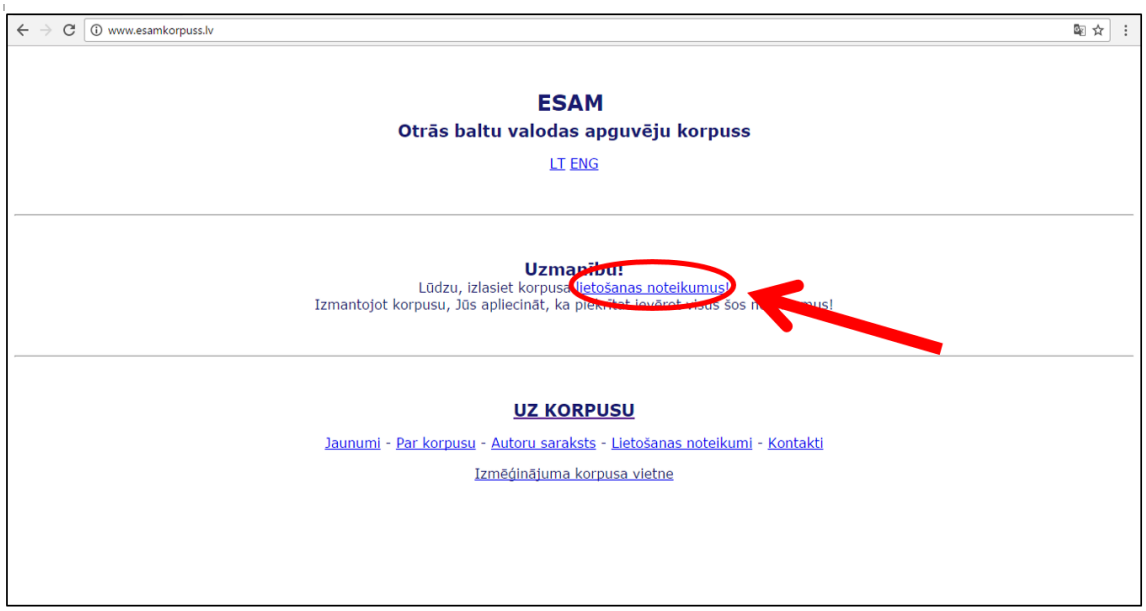
Versija 1.1.

Saturs

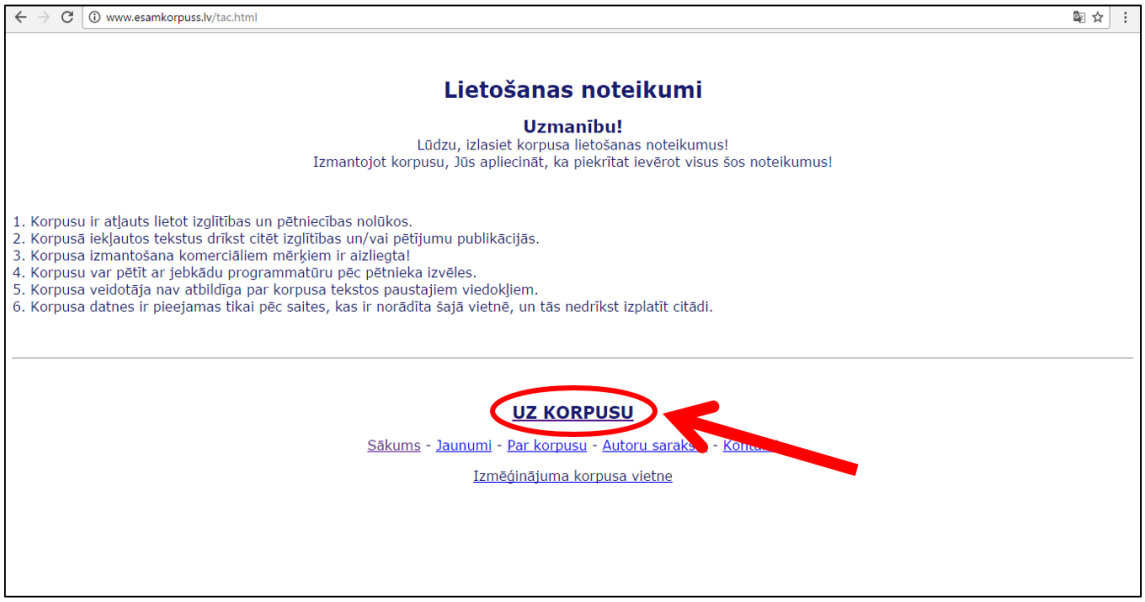
Ja darbā ar korpusu rodas kādas neskaidrības vai šaubas par darītā pareizību, lūdzu sazināties ar korpusa veidotāju pa e-pastu inga.s.znotina@gmail.com.

Piekļuve korpusam

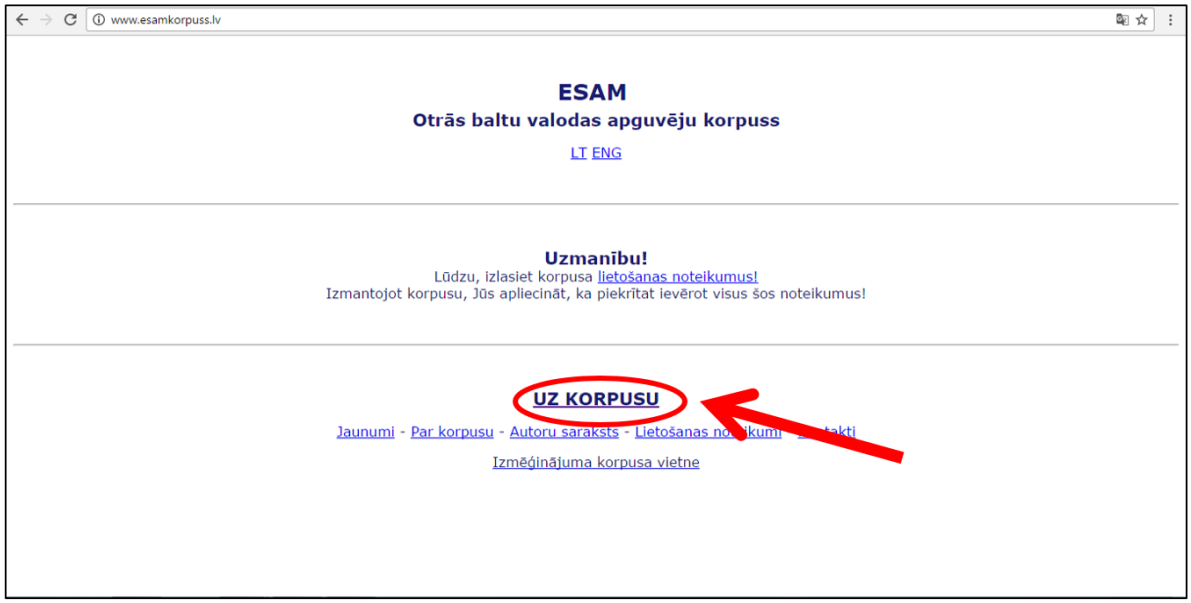
Lai piekļūtu korpusam, ir jāiet uz vietni *www.esamkorpuss.lv* un jāizlasa korpusa lietošanas noteikumi. Korpusu ir atļauts lietot tikai tad, ja piekrītat ievērot tā lietošanas noteikumus.



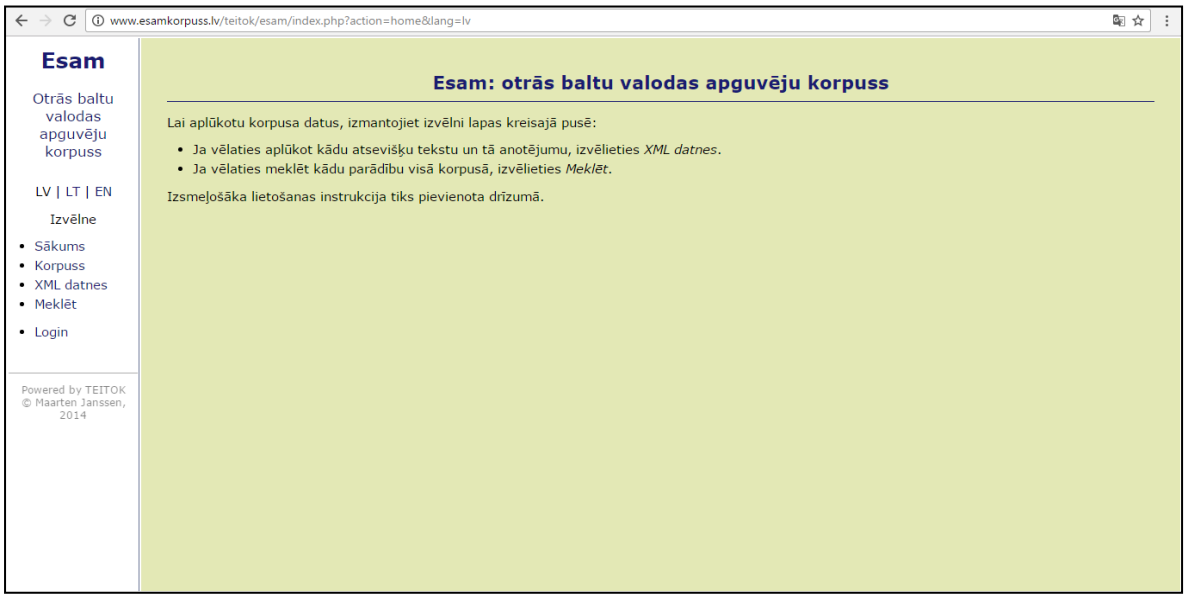
Kad lietotājs ir izlasījis lietošanas noteikumus, piekļūt korpusam var, atverot saiti "Uz korpusu". To atverot, lietotājs apņemas ievērot korpusa lietošanas noteikumus.



Ja lietotājs iepriekš jau ir izlasījis noteikumus un tiem piekrīt, tad tie nav jālasa atkārtoti. Šādā gadījumā var uzreiz atvērt saiti “Uz korpusu”.

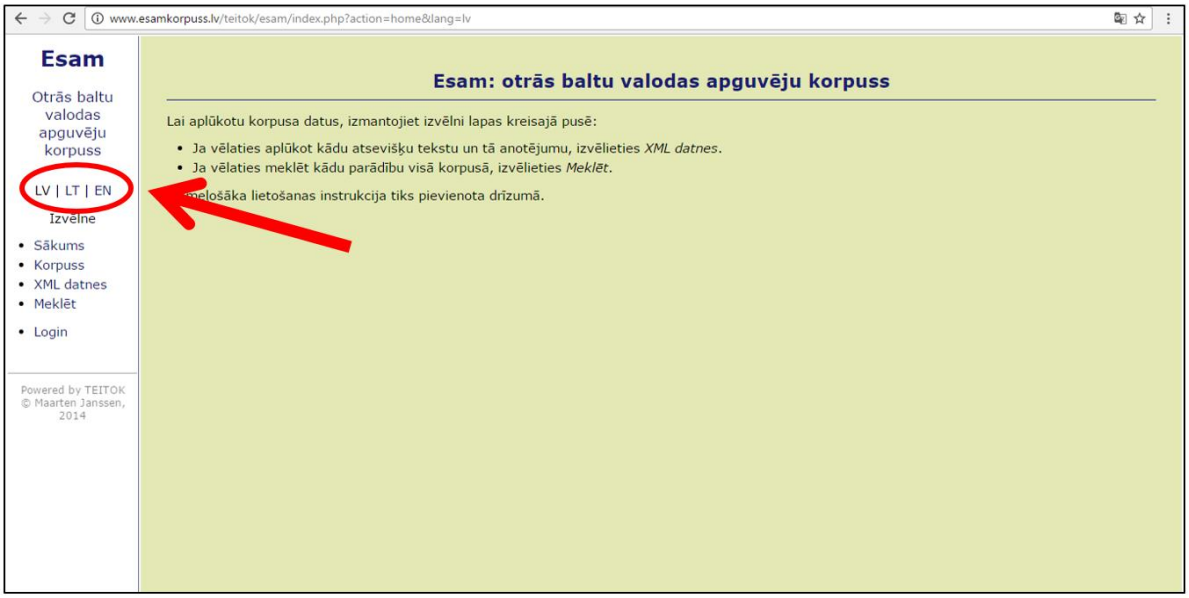


Nospiežot uz saites “Uz korpusu”, atveras korpusa sākumlapa.

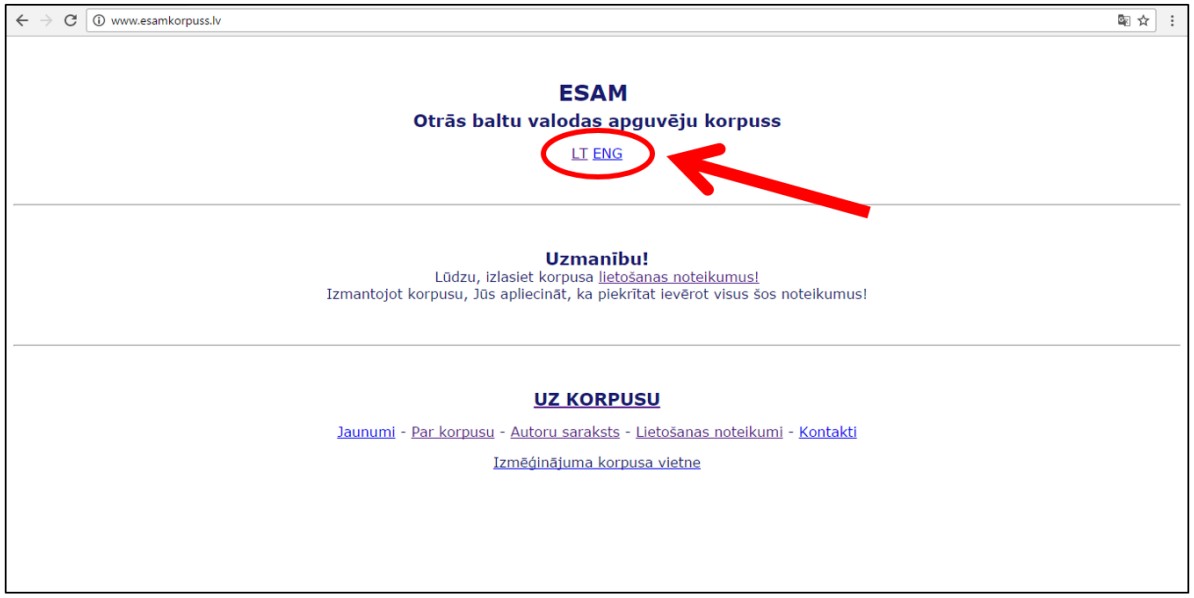


Valodas izvēle

Lai izvēlētos, kādā valodā aplūkot korpusa vietni, ir jāspiež uz saites ar atbilstošās valodas abreviatūru:
LV – lai vietni aplūkotu latviešu valodā;
LT – lai vietni aplūkotu lietuviešu valodā;
EN – lai vietni aplūkotu angļu valodā.
Pēc attiecīgās saites nospiešanas vietne tiks parādīta izvēlētajā valodā.



Vietnes valodu var mainīt arī sākumlapā, spiežot uz saites ar atbilstošās valodas abreviatūru.



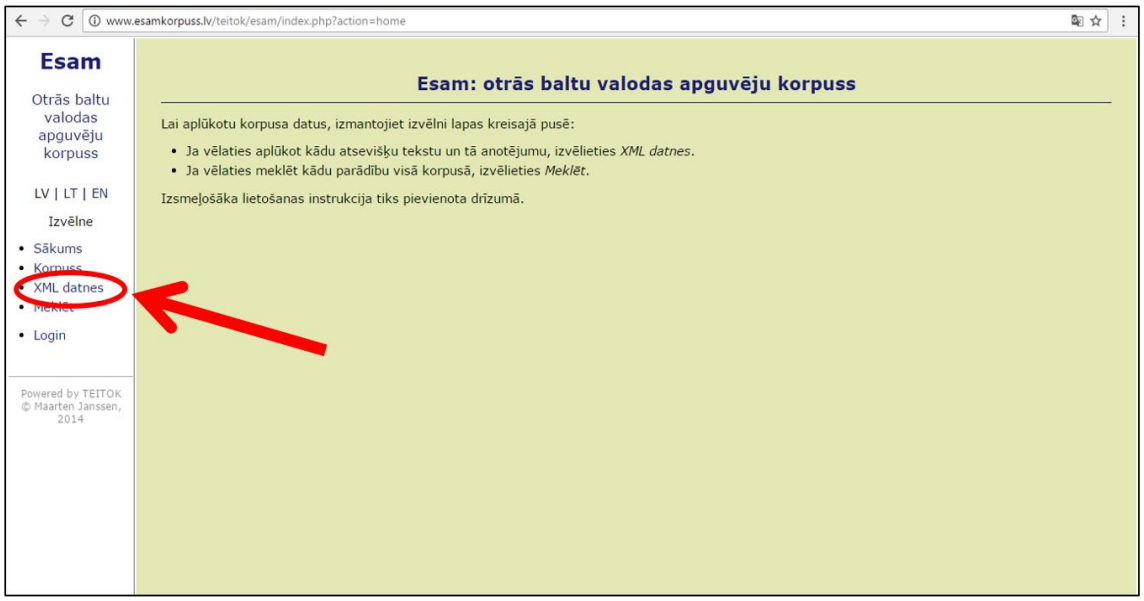
Korpusa lietošana

Korpusu var izmantot divējādi:

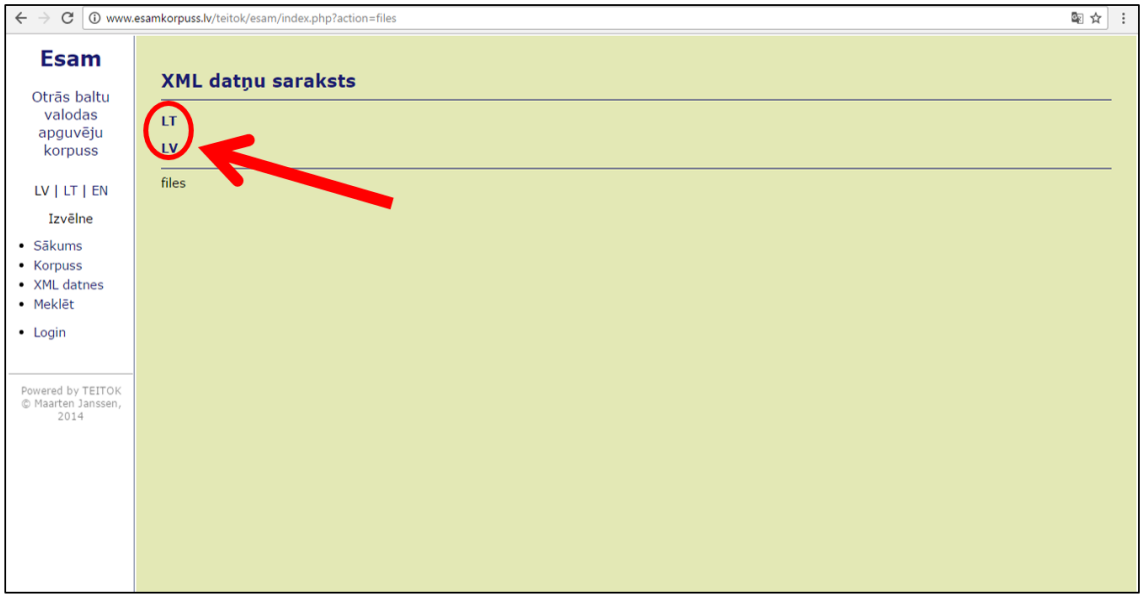
1. Aplūkojot atsevišķus tekstus ar anotējumu vai bez tā;
2. Meklējot noteiktas parādības visos korpusā iekļautajos tekstos vai daļā no tiem.

Atsevišķu tekstu aplūkošana

Lai aplūkotu atsevišķus korpusā iekļautos tekstus, izvēlnē pa kreisi ir jāizvēlas “XML datnes”.



Lapā, kas atveras pēc tam, ir jāizvēlas tās valodas abreviatūra, kurā rakstītus tekstus lietotājs vēlas aplūkot:
LT – lai aplūkotu tekstus lietuviešu valodā;
LV – lai aplūkotu tekstus latviešu valodā.



Kad ir izvēlēta valoda, tiek parādīts visu attiecīgās valodas datņu saraksts. Datņu nosaukumi ir skaitliski kodi. Lai atvērtu tekstu, jāspiež uz saites ar attiecīgās datnes nosaukumu.

Nospiežot uz saites ar datnes nosaukumu, tiek atvērts izvēlētās datnes teksts bez anotējuma. Tekstu var aplūkot trijos skatījuma veidos:

- transkripcija – sākotnējais teksts, anonimizētās vietas atzīmētas ar xxx;
- normalizēta forma – sākotnējais teksts, anonimizētās teksta daļas aizstātas ar līdzīgiem vārdiem atbilstošās formās vai ar iezīmi [izlaid];
- labota forma – tekstā izlabotas kļūdas.

Lai izvēlētos skatījuma veidu, jāspiež uz attiecīgās pogas sadaļā "Teksts".

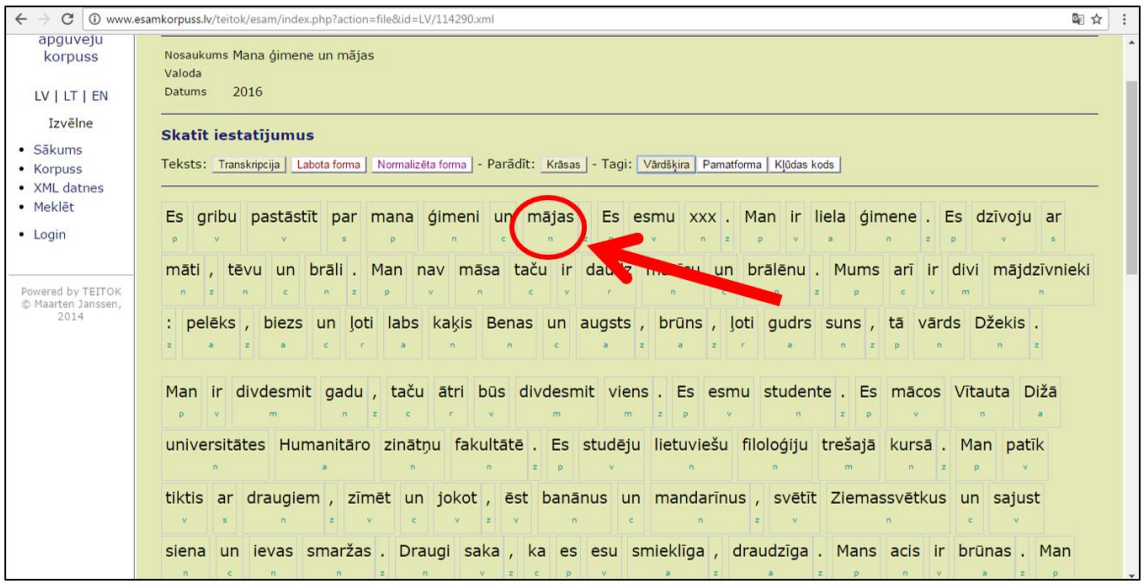
Izvēloties labotu vai normalizētu formu, ir iespējams norādīt, vai labotās (labotajā formā) vai anonimizētās (normalizētajā formā) teksta daļas parādīt citā krāsā. Lai šo funkciju ieslēgtu vai izslēgtu, jāspiež uz pogas “Krāsas” sadaļā “Parādīt”.

Izmantojot šo funkciju, labotās teksta daļas tiek parādītas tumši sarkanā krāsā, bet anonimizētās – violetā krāsā.

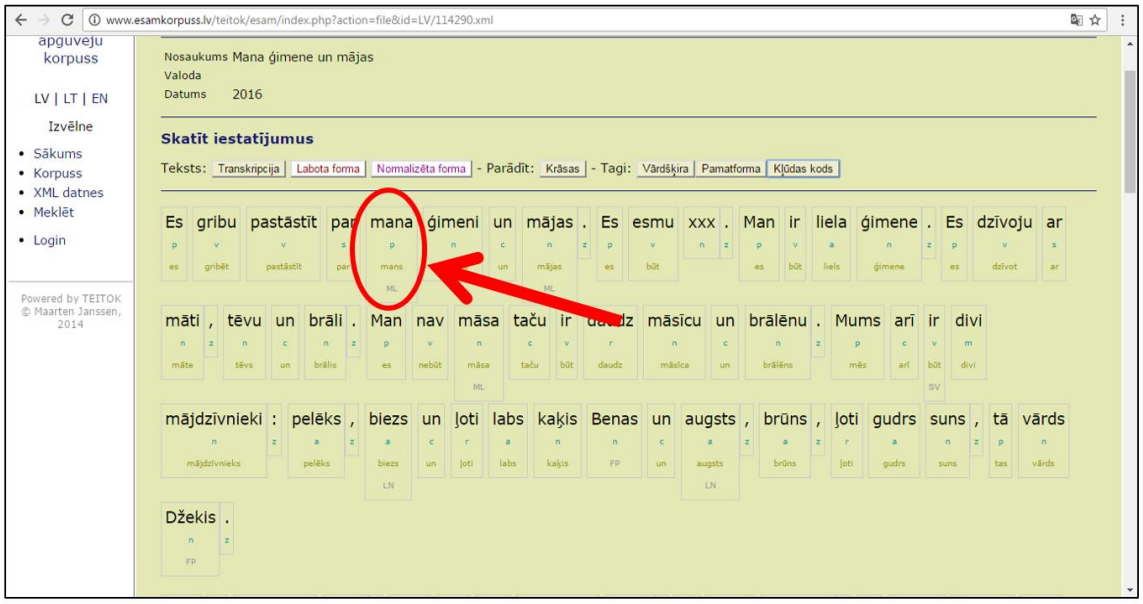
Tekstu var aplūkot arī ar anotējumu. Šādā nolūkā jāizvēlas attiecīgais anotējuma veids sadaļā “Tagi”:

- vārdšķira – vārdšķiru anotējums (vārdšķiru kodus skat. instrukcijas 1. pielikumā);
- pamatforma – pamatformu anotējums;
- kļūdas kods – kļūdu veidu anotējums (kļūdu kodus skat. instrukcijas 2. pielikumā).

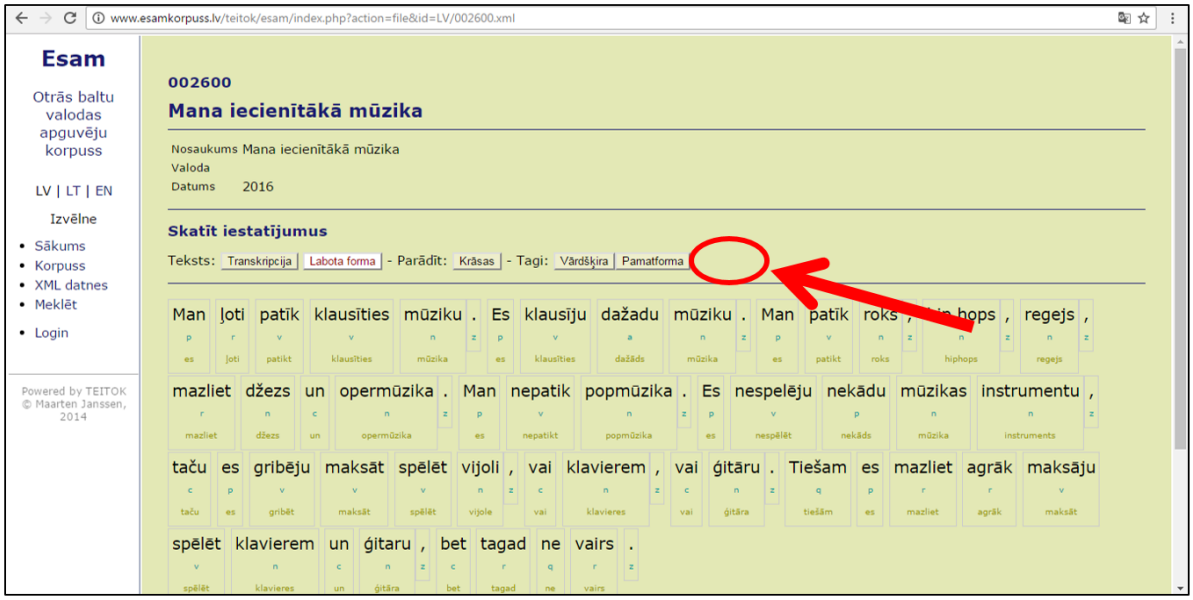
Izvēloties aplūkot kādu anotējuma veidu, attiecīgā informācija parādās zem katra vārda.



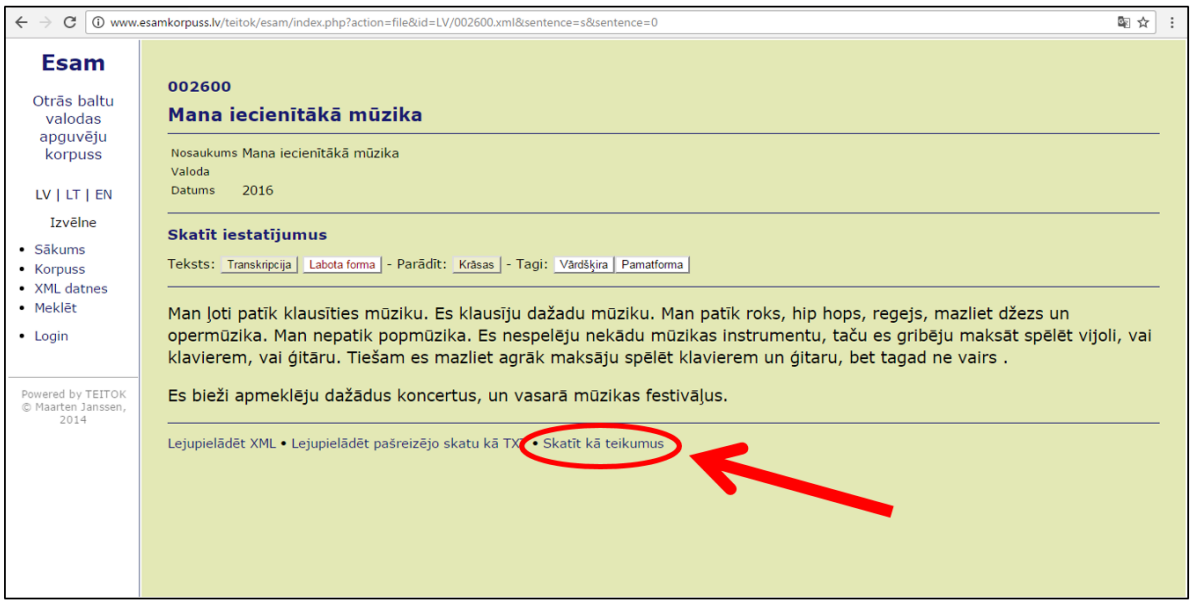
Vienlaikus var izvēlēties vienu vai vairākus anotējuma veidus.



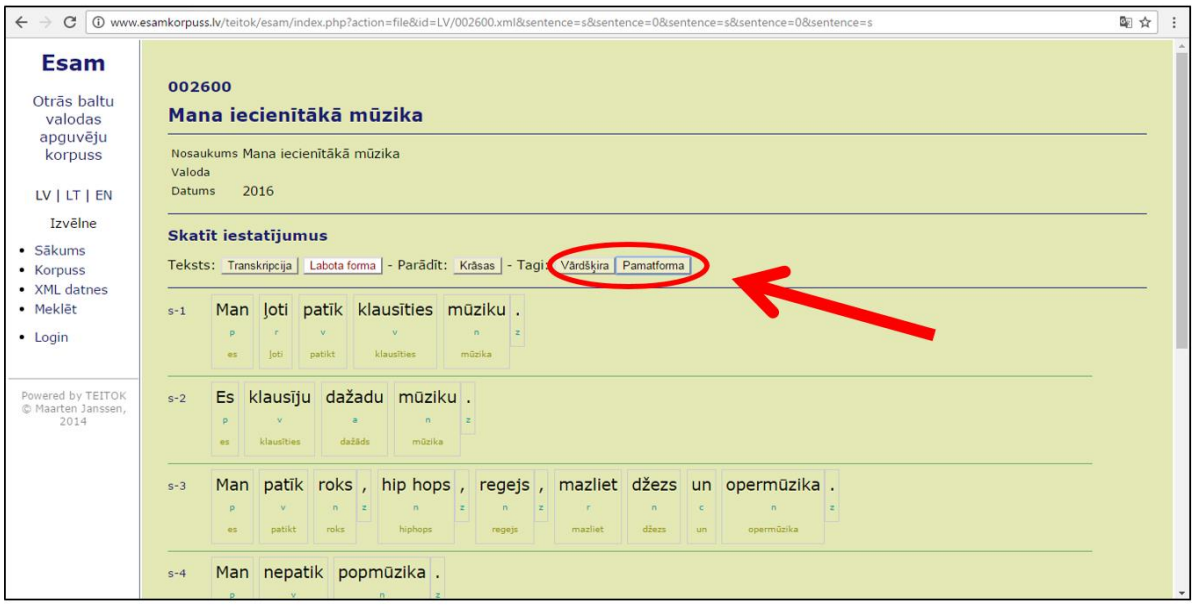
Ja konkrētais teksts pēc attiecīgās pazīmes nav anotēts, tad šīs pazīmes poga attiecīgajam tekstam netiek rādīta, piemēram, var nebūt pogas “Kļūdas kods”.



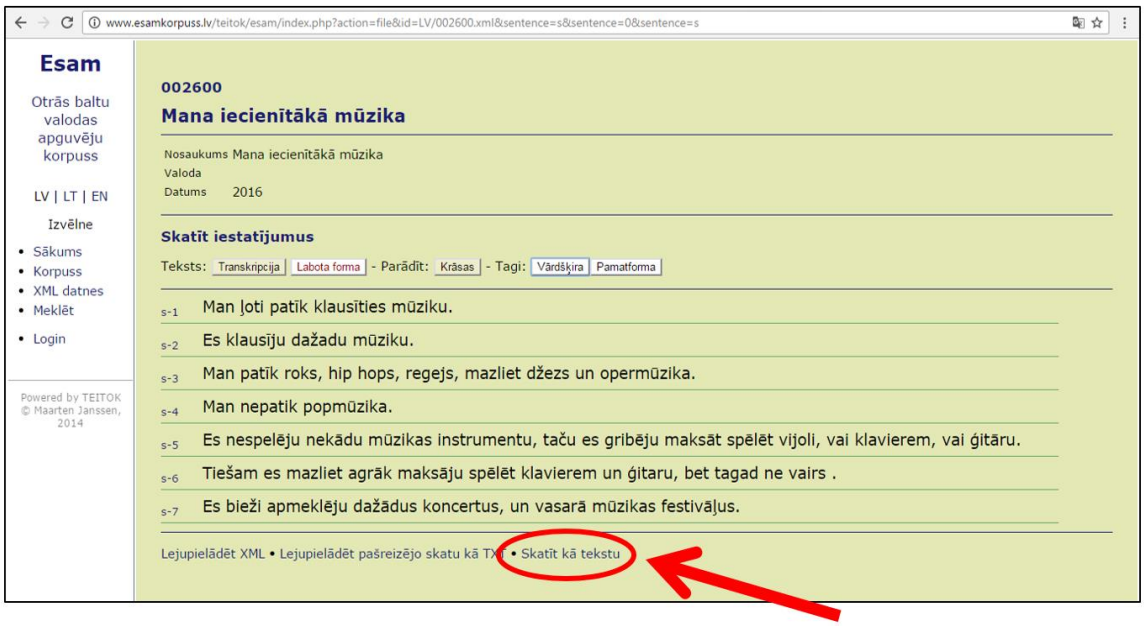
Tekstu var aplūkot arī sadalītu teikumos. Lai to izdarītu, lapas apakšā zem teksta jāspiež uz saites “Skatīt kā teikumus”.



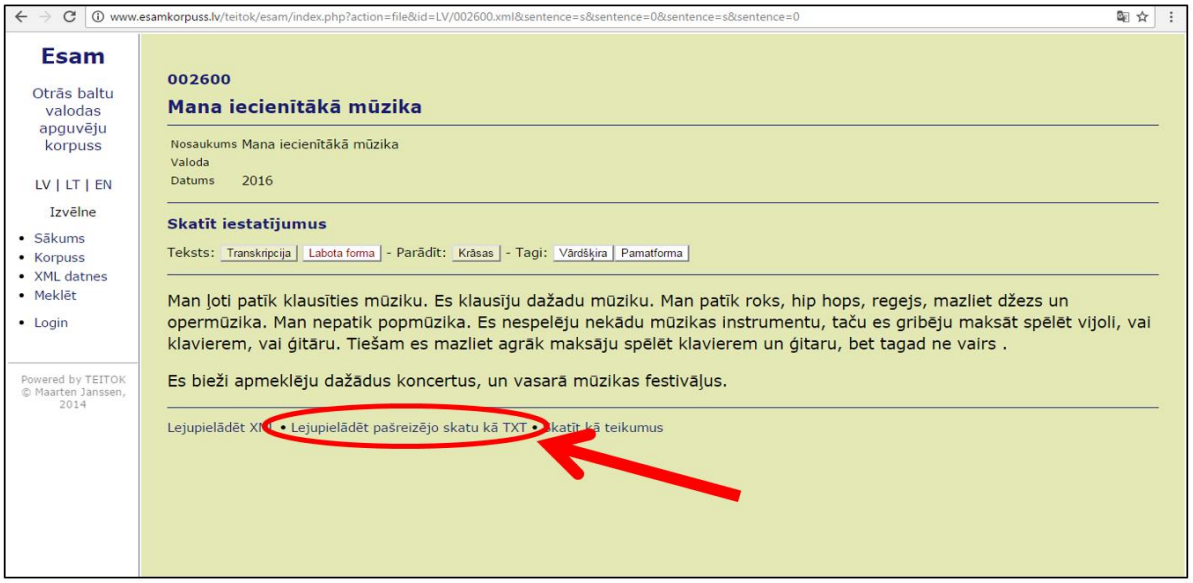
Arī teikumu skatījumā ir iespējams tekstu aplūkot gan anotētu, gan neanotētu. Anotējumu var ieslēgt un atslēgt ar atbilstošā anotējuma veida pogām sadaļā “Tagi” (sk. iepriekš).



Lai atkal mainītu skatījumu uz nesadalītu tekstu, lapas apakšā zem teksta jāspiež uz saites “Skatīt kā tekstu”.



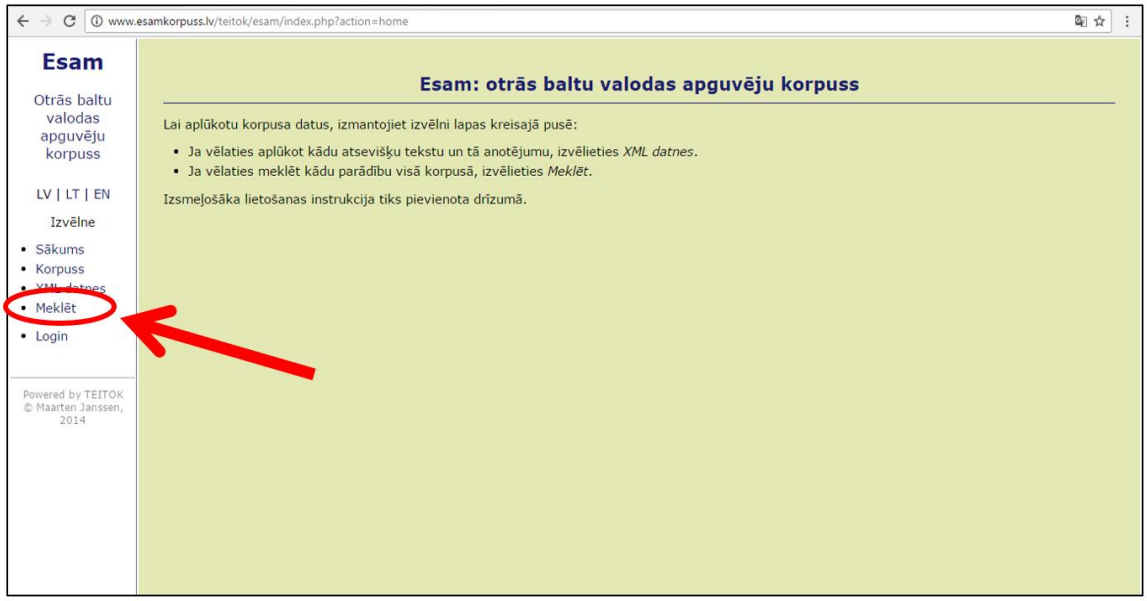
Ir iespējams arī lejupielādēt pašreizējo skatu kā TXT formāta datni. Lai to izdarītu, jāspiež uz saites “Lejupielādēt pašreizējo skatu kā TXT”.



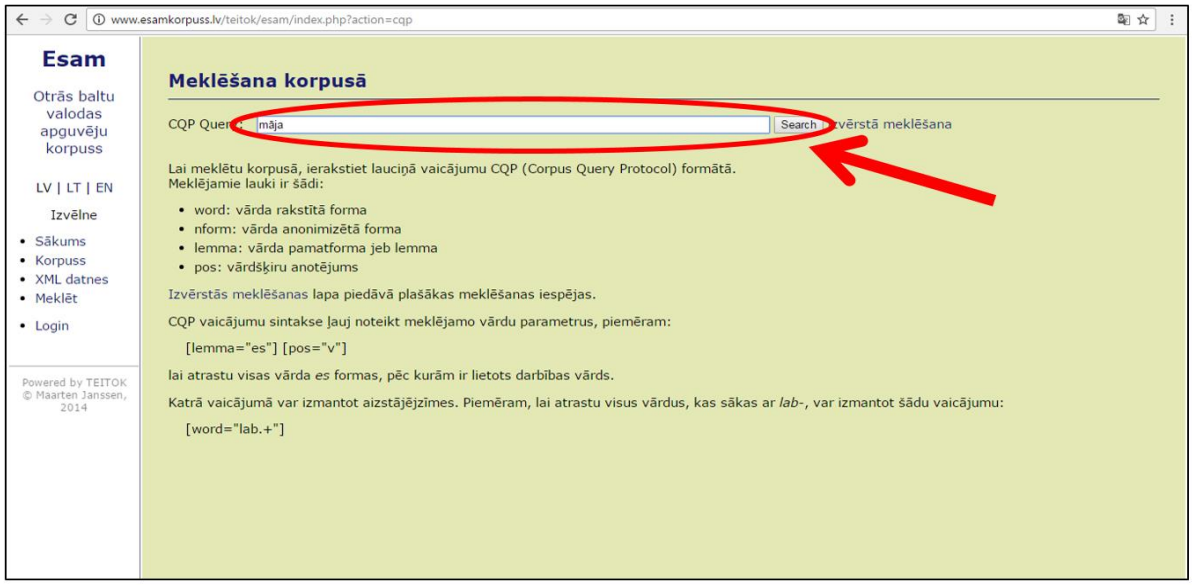
Ir iespējams arī lejupielādēt XML datni, kurā mašīnlasāmā veidā ir iekļauta visa pieejamā informācija par attiecīgo tekstu. Lai to izdarītu, jāspiež uz saites “Lejupielādēt XML”.

Meklēšana korpusā

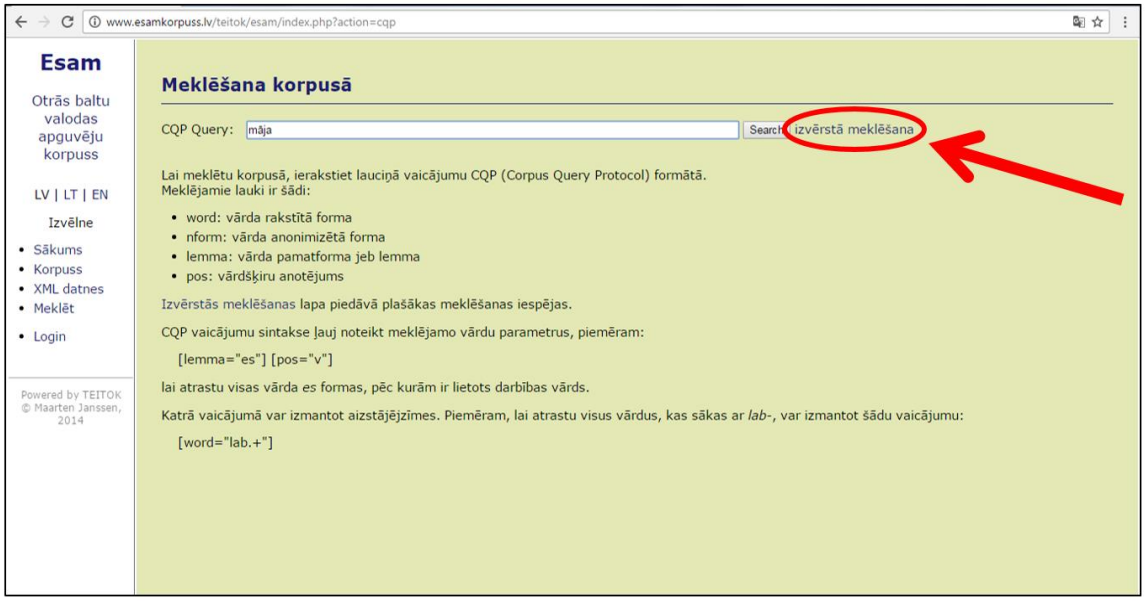
Lai meklētu korpusā, izvēlnē pa kreisi ir jāizvēlas “Meklēt”.



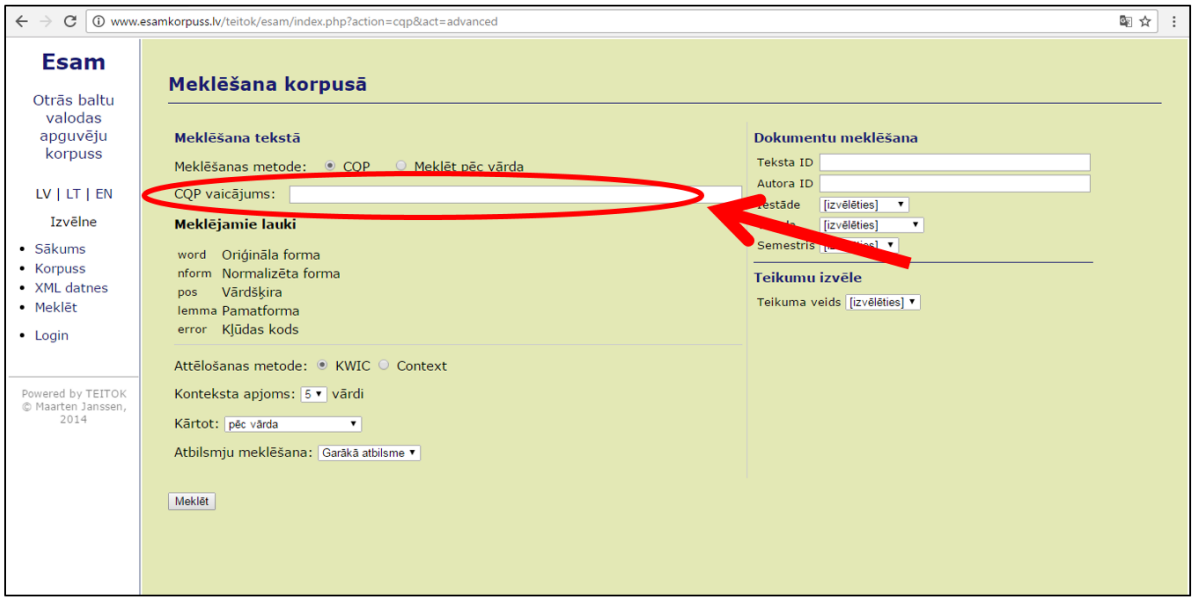
Logā, kas atveras, ir redzams vaicājuma ievades lauks un mazliet papildinformācijas par vaicājuma formulēšanu. Korpusā darbojas CQP vaicājumu sintakse (par to vairāk informācijas ir dažādos avotos tiešsaistē, piemēram, *cwb.sourceforge.net/temp/CQPTutorial.pdf*). Pēc vaicājuma ierakstīšanas laukā ir jāspiež “Search” pa labi no lauka.



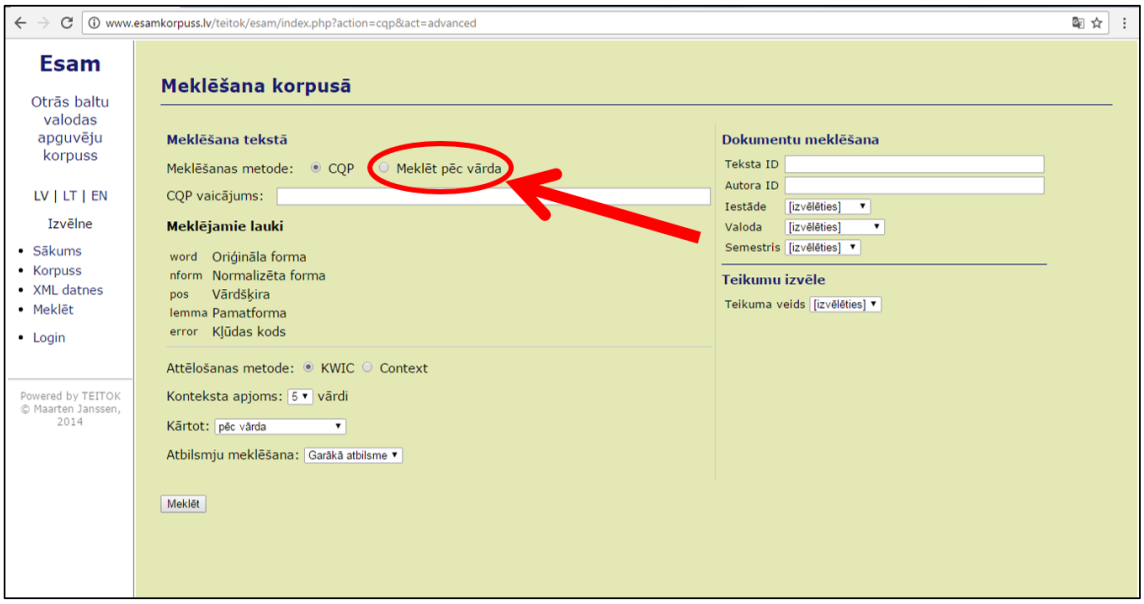
Papildu meklēšanas iespējas ir pieejamas izvērstajā meklēšanā. Lai to atvērtu, jāspiež uz saites pa labi no pogas “Search”.



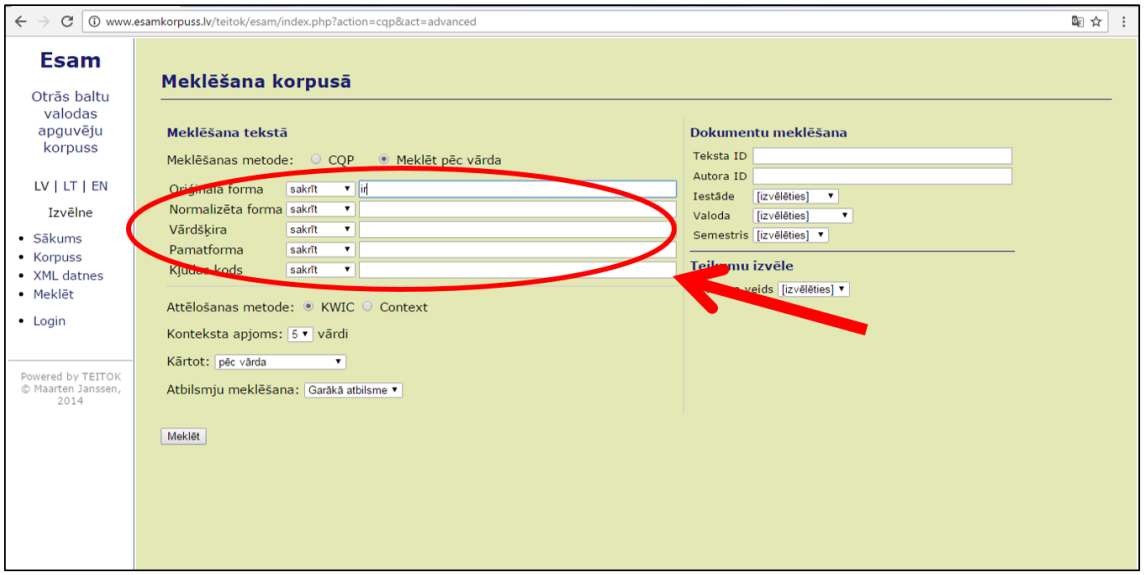
Izvērstajā meklēšanā tāpat ir iespējams ievadīt vaicājumu, kas atbilst CQP sintaksei.



Ja lietotājs vēlas, tad var izvēlēties arī iespēju “Meklēt pēc vārda”.

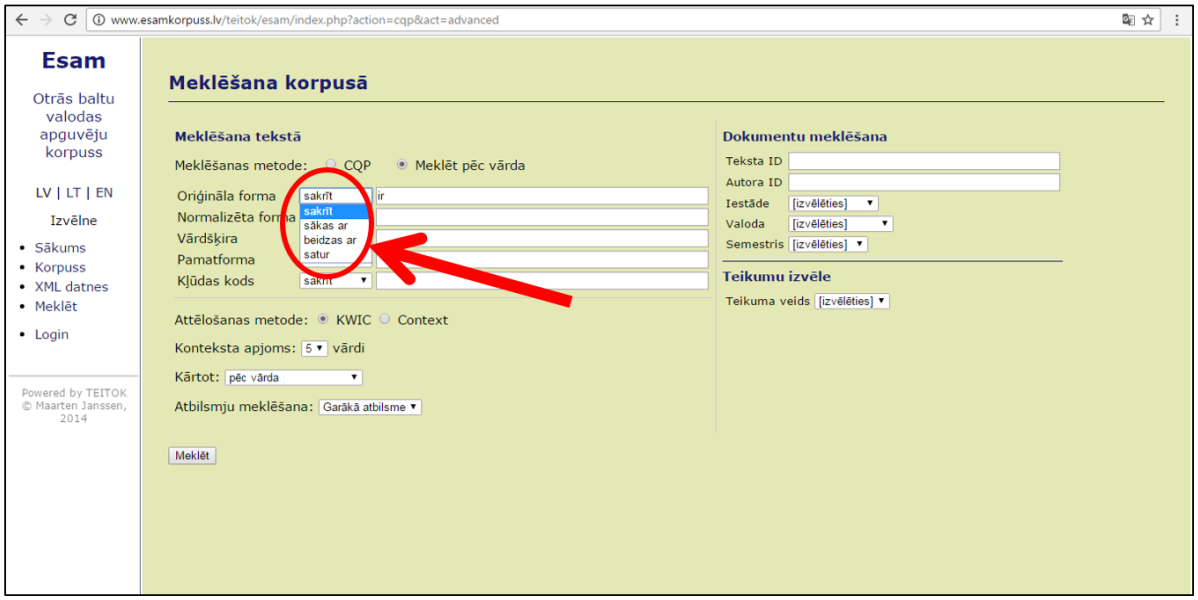


To izvēloties, tiek parādīti vairāki lauki, kuros ievadot vairākas vērtības, programma CQP sintaksei atbilstošu vaicājumu sastāda pati.



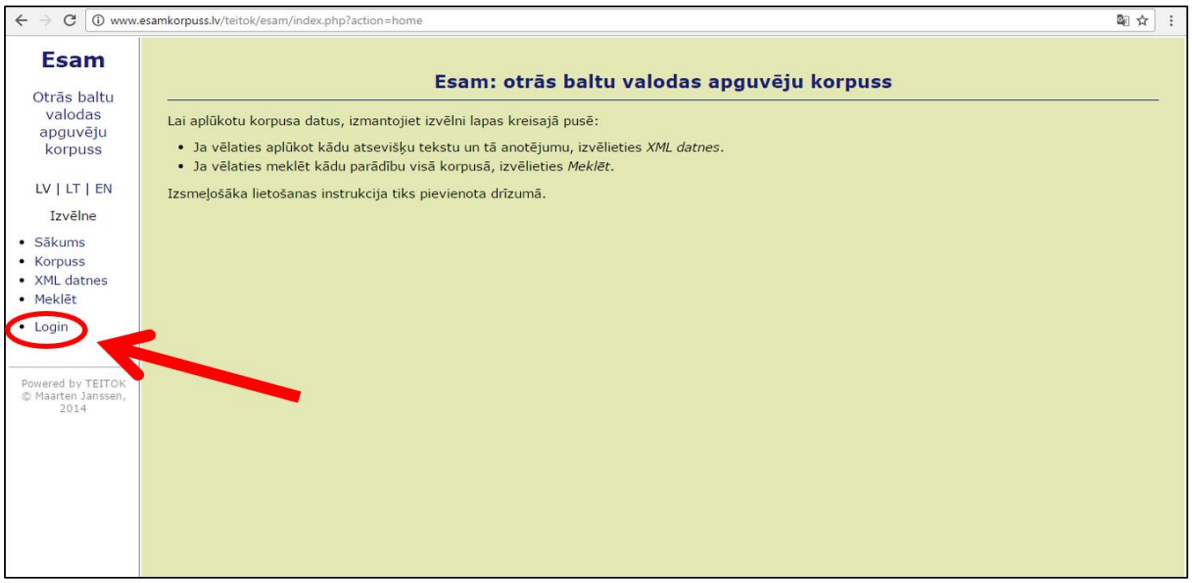
Katram no šiem laukiem ir iespējams norādīt, vai tajā ievadītajai vērtībai ir jāsakrīt ar vārdu tekstā, jābūt tā sākumā vai beigās, vai jābūt jebkurā vārda daļā:

- sakrīt – tiks atrasti vārdi, kas precīzi atbilst vaicājumam;
- sākas ar – tiks atrasti vārdi, kas sākas ar vaicājumā norādīto simbolu virkni;
- beidzas ar – tiks atrasti vārdi, kas beidzas ar vaicājumā norādīto simbolu virkni;
- satur – tiks atrasti vārdi, kuru jebkura daļa sakrīt ar vaicājumā norādīto simbolu virkni.

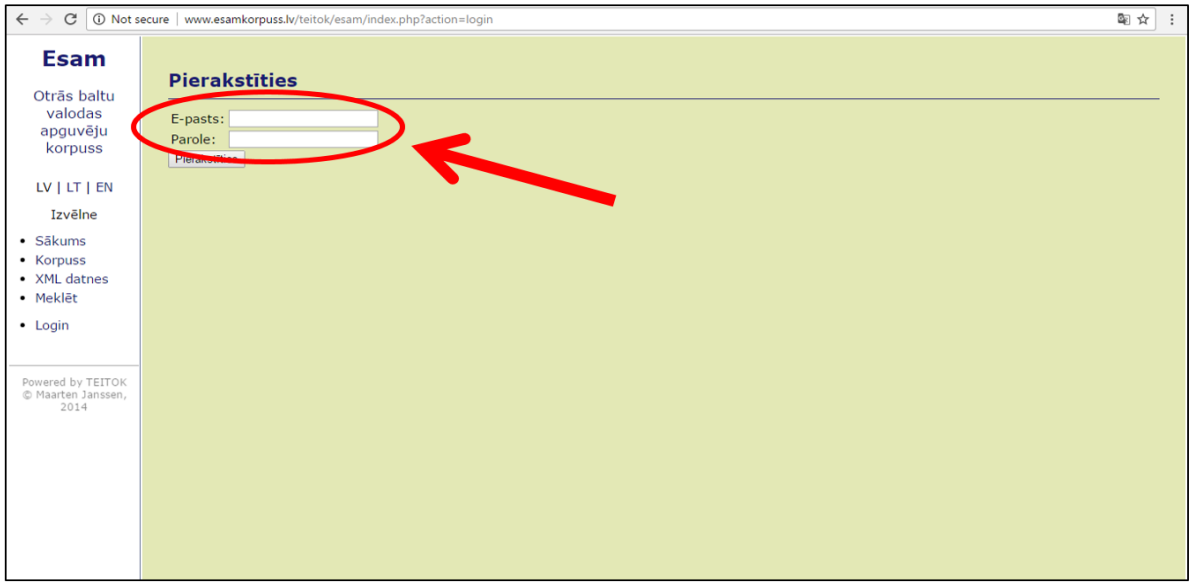


Pierakstīšanās sistēmā

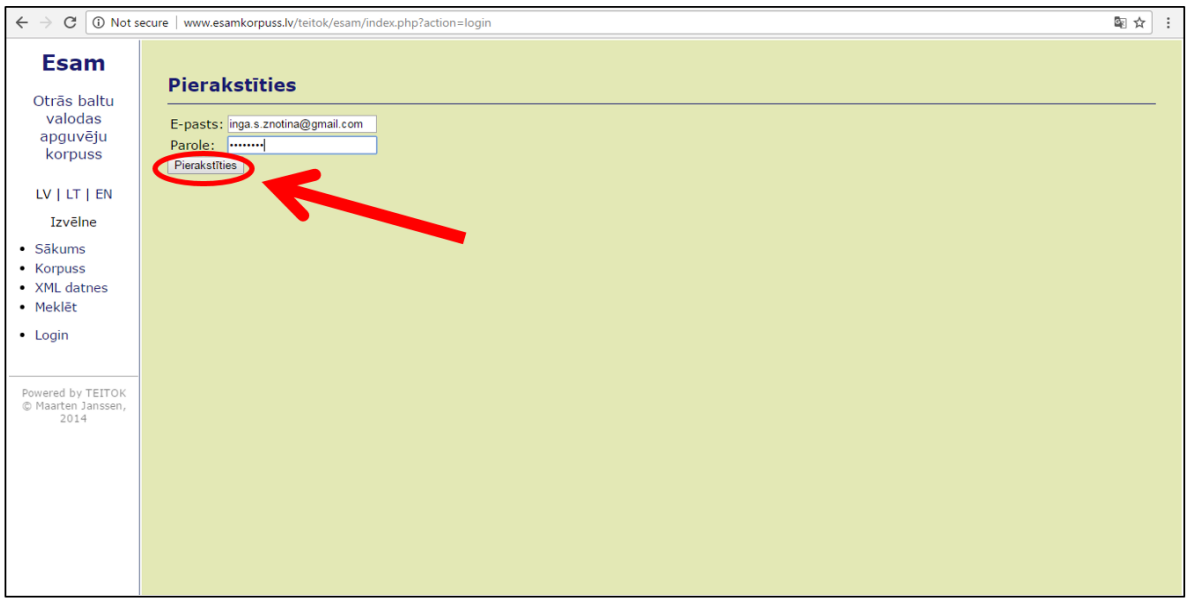
Pierakstīšanās sistēmā ir paredzēta tikai tiem lietotājiem, kuri piedalās korpusa anotēšanā un/vai uzturēšanā. Lai pierakstītos sistēmā, izvēlnē pa kreisi ir jāizvēlas “Login”.



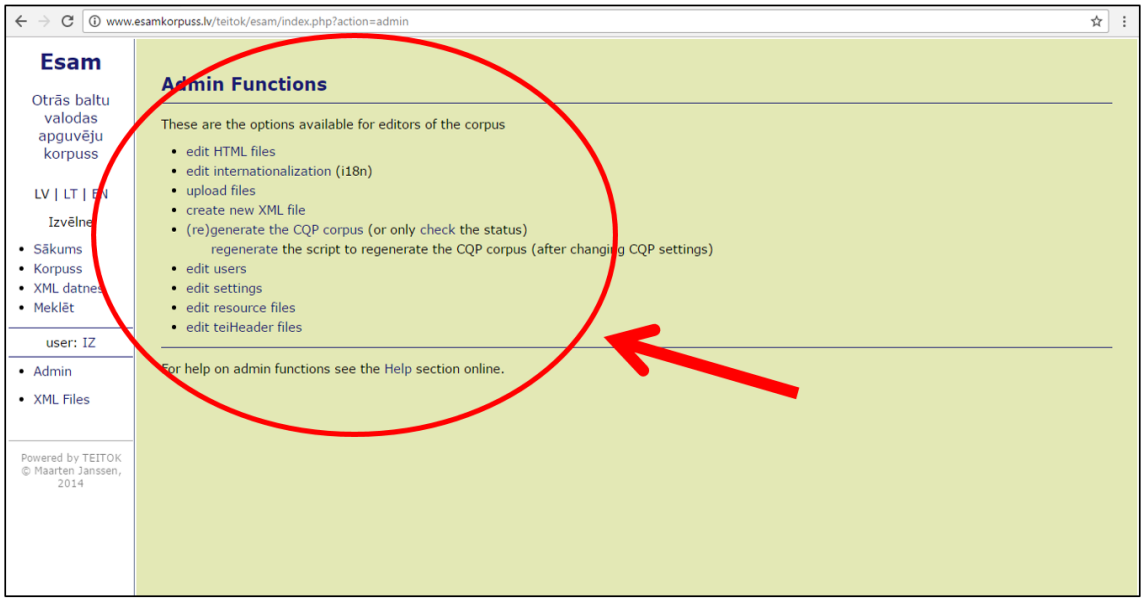
Pēc tam ir jānorāda savs e-pasts un parole, ko reģistrētajam lietotājam ir nosūtījusi korpusa uzturētāja.



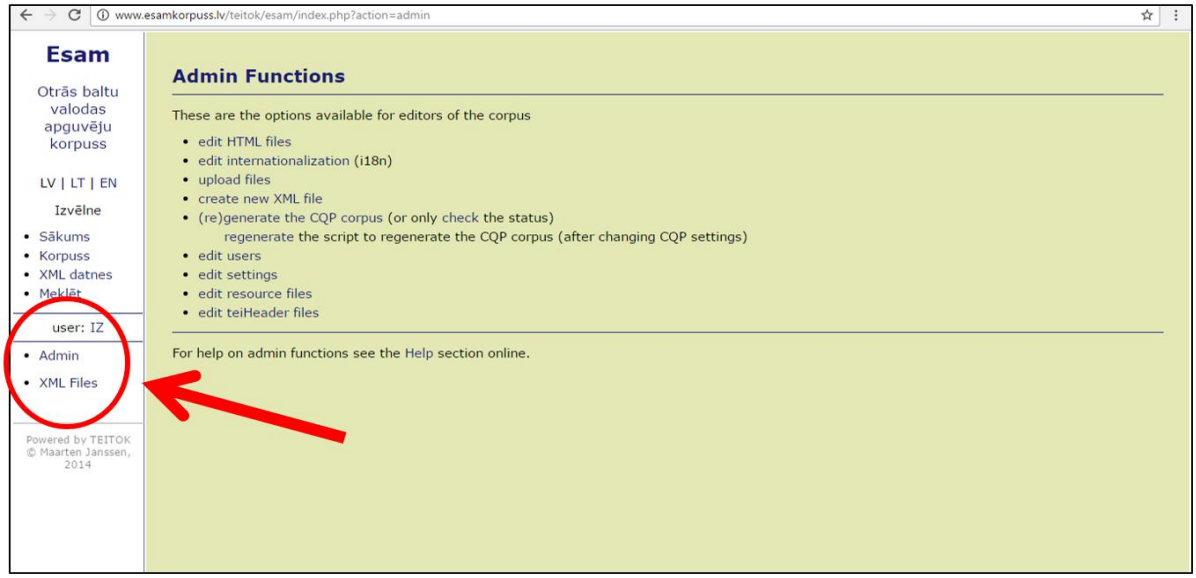
Kad tas ir izdarīts, jāspiež “Pierakstīties”.



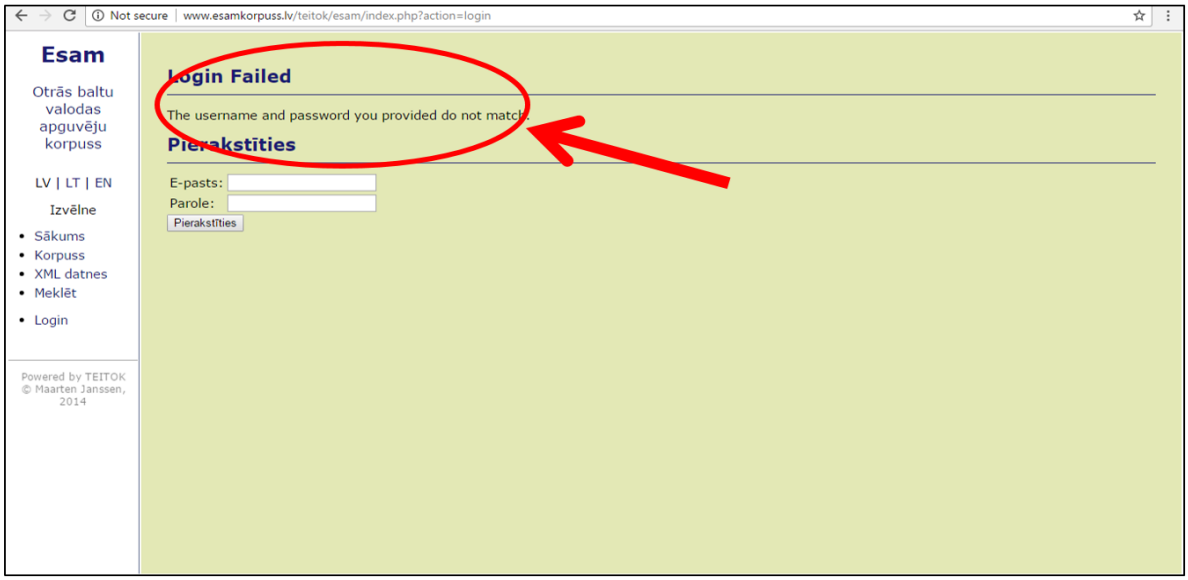
Ja dati ir ievadīti pareizi, tad parādās administratoram pieejamo funkciju saraksts:



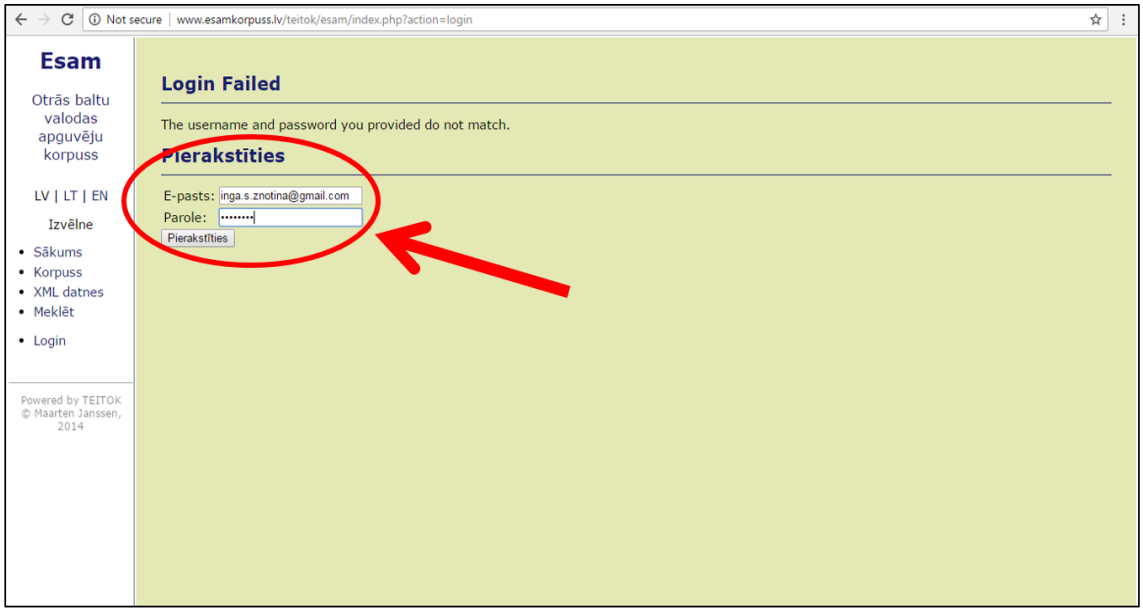
Tāpat parādās izmaiņas arī izvēlnē kreisajā pusē: lietotājvārda saīsinājums un administratora izvēlne.



Ja dati nav ievadīti pareizi, tad parādās paziņojums, ka lietotājvārds un parole nesakrīt.



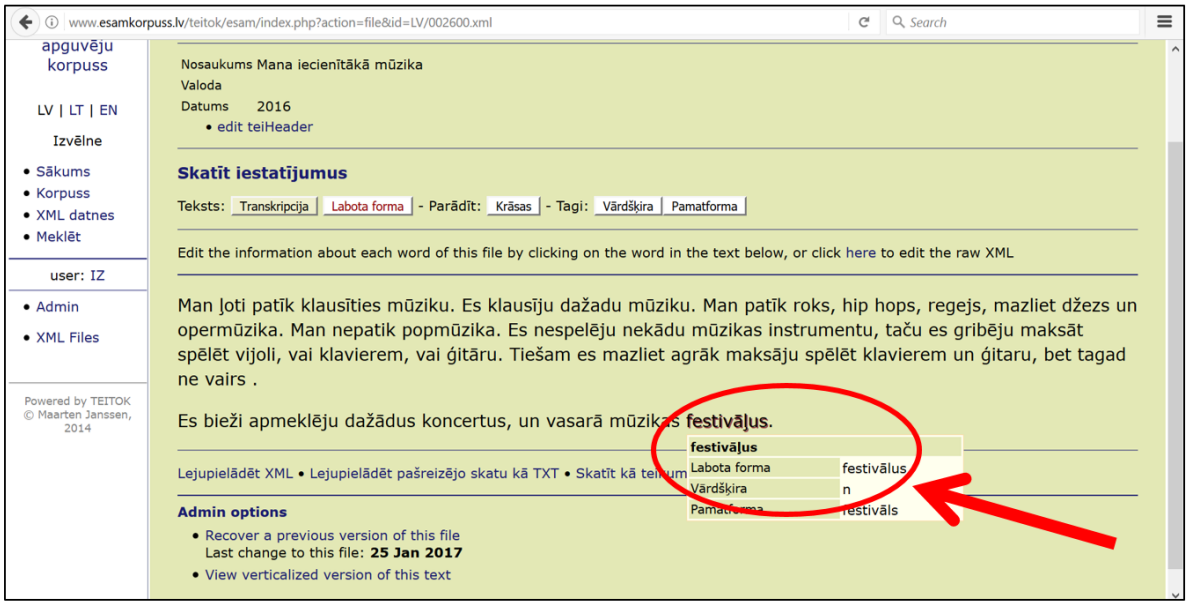
Tādā gadījumā var mēģināt vēlreiz.



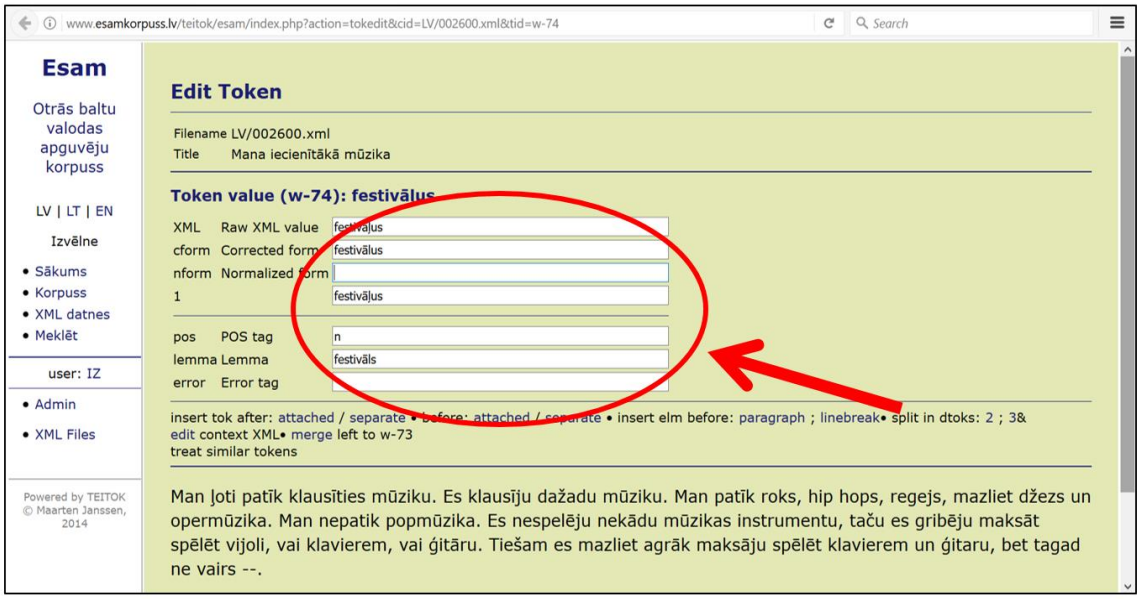
Ja atkārtoti neizdodas pieslēgties, tad ir jāsazinās ar korpusa uzturētāju un jālūdz izveidot jaunu paroli.

Anotēšana

Lai anotētu korpusu, ir jāpierakstās sistēmā. Anotēt tekstu var, atverot šo tekstu atsevišķai aplūkošanai (skat. iepriekš: Atsevišķu tekstu aplūkošana). Pēc tam, kad anotētājs ir pierakstījies sistēmā, šis skats piedāvā arī anotēšanu. Lai to darītu, tekstā jāuzklikšķina uz vārda, kuru nepieciešams anotēt.

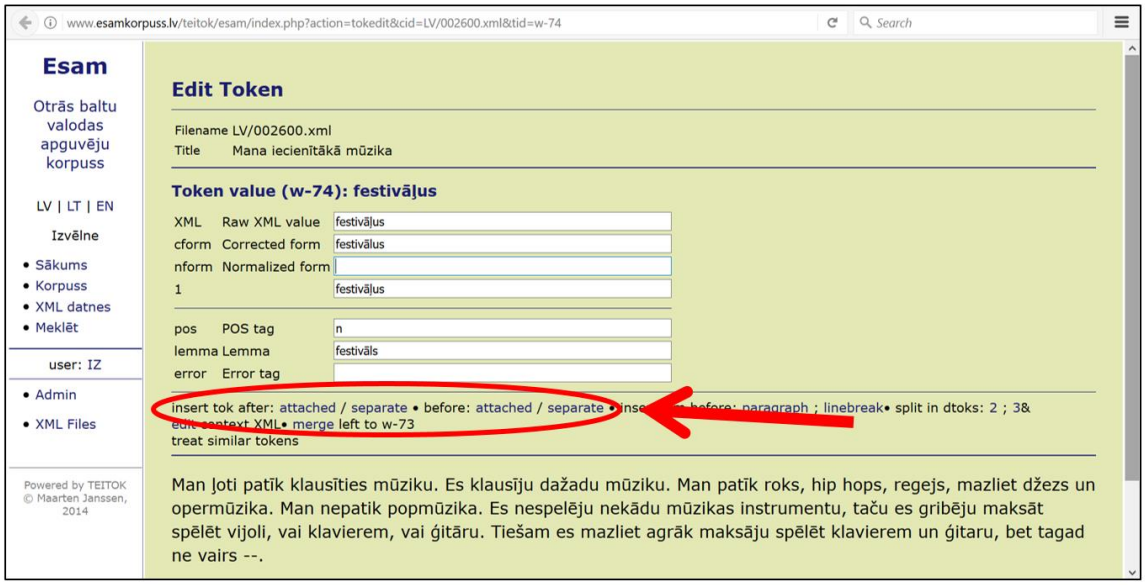


Noklikšķinot tiek atvērta vārda labošanas forma. Tajā ir septiņi lauki, kuros var labot informāciju.



- Laukā “Raw XML value” ir redzama attiecīgā vārda sākotnējā forma, kāda tā ir redzama XML datnē. Ja vien nav pārliecības, ka, digitalizējot tekstu, ir pieļauta kāda kļūda (lai par to pārliecinātos, jāsazinās ar korpusa sastādītāju), šajā laukā esošo informāciju mainīt nedrīkst.

- Lauks “Corrected form” ir paredzēts labojumiem. Pēc šiem labojumiem tiek veikta kļūdu anotēšana. Ja teksta oriģinālā ir vārds vai pieturzīme, kuras vietā labojumā nebūtu jābūt nekam (piem., lieks komats), tad laukā “Corrected form” ieraksta divas defises bez atstarpes starp tām: --. Ja teksta oriģinālā nav nekādas vienības, bet pēc labojuma principiem tā ir nepieciešama, tad labošanai izvēlas blakus esošu vārdu un tieši zem laukiem sadaļā “insert tok after/before” klikšķina uz “attached” vai “separate”. Tā tekstā tiek ievietots jauns vārds. Izvēloties “attached”, starp jauno vārdu un esošo vārdu tekstā nebūs atstarpes (ja, piemēram, nepieciešams ievietot pieturzīmi). Izvēloties “separate”, atstarpe būs. Jaunizveidotajam vārdam ir jāaizpilda tikai lauks “Corrected form”.



- Laukā “Normalized form” ir dati, ar kuriem tiek aizstātas anonimizētās vietas teksta lasāmības saglabāšanai. Ja šajā laukā kaut kas ir ierakstīts, tad laukā “Raw XML value” ir kods *xxx*.

- Lauks “1” ir nepieciešams sistēmas tehniskās darbības nodrošināšanai. Normālā gadījumā vērtībai tajā būtu jāsakrīt ar laukā “Raw XML value” redzamo. Ja tas tā nav, lūdzu sazināties ar korpusa veidotāju.

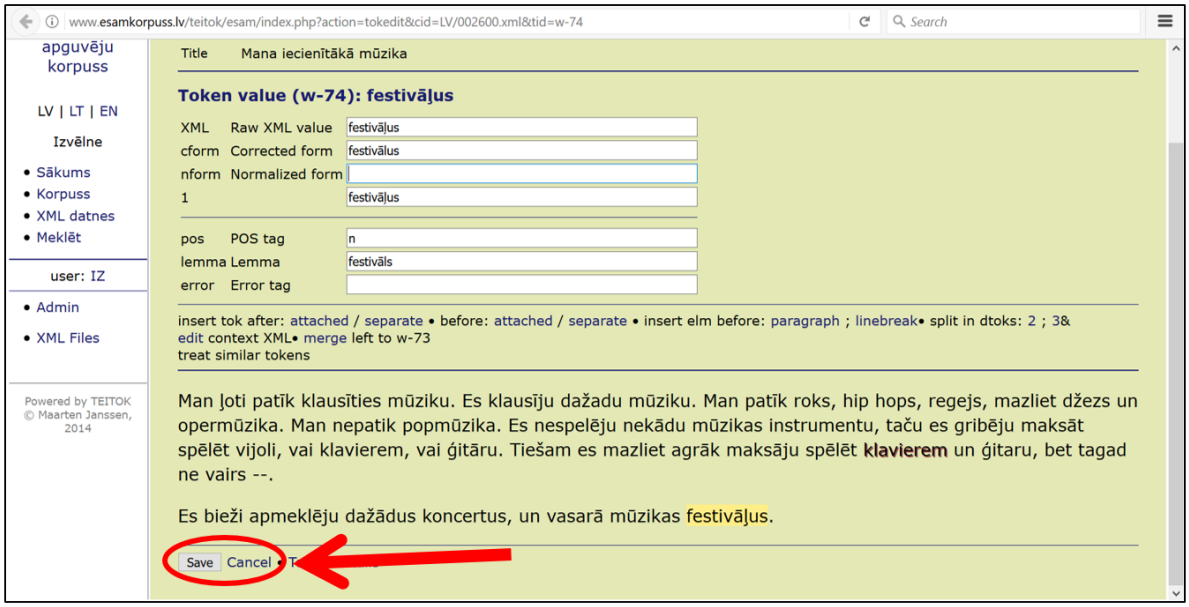
- Lauks “POS tag” ir paredzēts vārdšķiru anotējumam. Tajā ieraksta attiecīgajam vārdam atbilstošās vārdšķiras kodu (sk. 1. pielikumā).

- Lauks “Lemma” ir paredzēts pamatformu anotējumam. Tajā ieraksta attiecīgajam vārdam atbilstošo pamatformu.

- Lauks “Error tag” ir paredzēts kļūdu anotējumam. Pēc oriģināla un labotās formas salīdzināšanas šajā laukā ieraksta atbilstošās kļūdas kodu (sk. 2. pielikumā).

Aizpildīt un/vai labot var vienu vai vairākus laukus. Lai saglabātu jauno anotējumu, jānospiež “Save” lapas apakšā. Lai atceltu labojumus un atgrieztos iepriekšējā skatā, jānospiež “Cancel” lapas apakšā.

Teikumu tipi tiek anotēti, jaunu tekstu pievienojot korpusam. Lai anotētu teikumu tipus un/vai labotu to anotējumu, jāsazinās ar korpusa veidotāju.



Apstiprināt un saglabāt labojumus var arī, nospiežot *Enter* pogu uz datora klaviatūras, kamēr kursors atrodas vienā no septiņiem aprakstītajiem laukiem.

Ja darbā ar korpusu rodas kādas neskaidrības vai šaubas par darītā pareizību, lūdzu sazināties ar korpusa veidotāju pa e-pastu inga.s.znotina@gmail.com.

Pielikums

1. pielikums. Vārdšķiru anotējuma kodi

Vārdšķira	**Piezīmes**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Lietvārds		n	mājas	stebuklą ‘brīnumu’
Darbības vārds	Ieskaitot divdabjus	v	gribu	supratau ‘sapratu’
Īpašības vārds		a	brūnas	lėtais ‘lēniem’
Vietniekvārds		p	man	aš ‘es’
Apstākļa vārds		r	ļoti	labai ‘ļoti’
Prievārds		s	pie	į ‘uz’
Saiklis	Ieskaitot saliktos saikļus	c	un	jei ‘ja’
Skaitļa vārds	Ieskaitot daļskaitļus un vairākvārdu skaitļa vārdus; vairākvārdu skaitļa vārda gadījumā katram no vārdiem vārdšķira ir norādīta atsevišķi	m	viens	dvi ‘divas’
Izsauksmes vārds		i	labdien	laba diena ‘labdien’
Partikula		q	nē	ne ‘nē’
Bezmorfoloģijas elements	Simbols vai simbolu virkne, kam nav mērķvalodas morfoloģiskās struktūras: cipari, abreviatūras, saīsinājumi, vārdi citās valodās, formulas u.tml.[53]	x	Varniukų	07:07

[53] Formulējums pielāgots no LVMPK 2009.

2. pielikums. Kļūdu anotējuma kodi

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
Forma	F	Kopā vai šķirti rakstāmi vārdi	FK		*širdyje kaž kas suvirpėja* (suvirpa)[54] ‘sirdī kaut kas ietrīsas’
		Lielie/mazie burti	FL	*...un Viņai patīk...*	*Olimpinėje* (Olympinėse) *Žaidynėse* ‘Olimpiskajās spēlēs’
		Diakritiskās zīmes	FD	*Viņas (Viņš) ir uzņemējs*	*dažnai nera pakankamai laiko* ‘bieži nav pietiekami daudz laika’
		Citas pareizrakstības kļūdas (ieskaitot pārrakstīšanos)	FP	*man patīk tiktis (tikties) ar draugiem*	*kikvieną dieną* ‘katru dienu’
Morfoloģija un vārddarināšana	M	Atvasināšana	MA	*patīk futbols, basketbols, vazinātes*[55] *ar ritēni (riteni)*	*todėl užmiegojome anksti* ‘tāpēc aizmigām agri’
		Salikteņdarinā-šana	MS		*aerouostas (oro uostas)* ‘lidosta’
		Locījums	ML	*Es gribu pastāstīt par mana ģimeni*	*didelė dalis drabužiai yra tokios spalvos* ‘liela daļa apģērbu ir tādā krāsā’
		Dzimte	MD	*Mans acis ir brūnas.*	*Jos visi yra šalia* ‘viņi visi ir līdzās’
		Skaitlis	MN	*es biju ļoti skumīga šoreiz pār (par) atvaļinājumiem*	*įvariuose gyvenimo valandą* ‘dažādās dzīves stundās’
		(Ne)noteiktā galotne	MG	*Fotoaparātā bija manas skaistas fotogrāfijas*	*žmonių kamšatis ir ilgoji (ilgos) valandos viešajame*

[54] Piemēros, kur nepieciešams, iekavās sniegts labojums; pasvītrota tā kļūda, kas atbilst attiecīgajam kļūdu apakštipam.

[55] Šķiet, šis vārds darināts no diviem vārdiem: lie. *važinėti* ‘braukāt’ un la. *vizināties.*

					transporte ‘cilvēku saspiestība un ilgās stundas sabiedriskajā transportā’
		Salīdzināmās pakāpes	MQ		*Aš esu <u>jaunesnioji</u> (jaunausia).* ‘es esmu visjaunākā’
		Persona	MP	*Tēvs <u>interesējies</u> par automobiliem (automobiļiem)*	*aš <u>nebuvo</u> name* ‘es nebiju mājās’
		Laiks	MT	*viņai patīk ceļot(,) un māte <u>apmeklēja</u> (ir apmeklējusi) Krieviju, Franciju...*	*aš <u>pasibundu</u> (pasibudau), nes buvau labai alkana* ‘es pamodos, jo biju ļoti izsalkusi’
		Izteiksme	MI		*Aš esu dėkinguma (dėkinga), ka (kad) <u>sutikčiau</u> (sutikau) jai (ją)* ‘es esmu pateicīga, ka satiku viņu’
		Kārta	MK		*Į ją <u>galėtų</u> įeiti ir iš lauko* ‘tajā varētu ieiet arī no āra’
		Refleksivitāte	MR		*netrukdėme ir <u>neriejomės</u>* ‘netraucējām un nebārāmies’
		Divdabis	MV		*vairuotojas, <u>matytint</u>, kad bėgtu (bėgu), (..) pristabdau (pristabdo)* ‘vadītājs, redzot, ka skrienu, piebremzē’
		Pabeigtība	MB		*Kada ji <u>ėjo</u> (atėjo) iš darbo...* ‘kad viņa atnāca no darba’
		Iterativitāte	MX		*mama man (mane) <u>išmokydavo</u> nekada nepasiduoti* ‘mamma man iemācīja nekad nepadoties’
Sintakse	S	Vārdu secība	SV	*..radoša tik (*tikai*) dėļ naudas (*naudas dēļ*)*	*Vieta, kur <u>visada aš</u> galiu grįžti yra...* ‘vieta, kur es vienmēr varu atgriezties’

		Izlaists vārds	SI	*tā vārds (*vārds ir*) Džekis*	*Biologijos fakultete yra labai daug (daug* ko?[56]*)* ‘Bioloģijas fakultātē ir ļoti daudz (kā?)’
		Lieks vārds	SL	*..ceļiauju (ceļoju) uz Klaipėdu būt brivdienās (brīvdienās)*	*ji pasiūlė man kartu su ja reikėjo ruošti pjesę* ‘viņa piedāvāja man kopā ar viņu vajadzēja gatavot lugu’
		Saistījums	SS	*Mans mātes vārds ir...*	*aš nešioju kepures, irgi pirštines* ‘es nēsāju cepures, arī cimdus’
Leksika	L	Nozīme	LN	*pelēks, biezs (resns) un ļoti labs kaķis Benas*	*Ne tik katris* (kiekvienas) *latvis* ‘ne tikai katrs latvietis’
		Saderība	LV	*zils paklājs, kurš der (piestāv) pie sienu (sienām)*	*nes esame tiek įvairios* ‘jo esam tik dažādas’
		Stabili vārdu savienojumi	LS	*..braukšu uz ciemus (ciemos)*	*Aš tą* (tai) *labai įvertinu* ‘es to ļoti novērtēju’
Interpunkcija	I	Nepiemērota pieturzīme	IN		*dar kartą užmigau..* ‘vēlreiz aizmigu’
		Lieka pieturzīme	IL	*Tāpēc, es biju ļoti skumīga*	*Trečią valandą naktį* (nakties)*,* *aš...* ‘trijos naktī es...’
		Pieturzīmes trūkums	IT	*Viņai patīk ceļot(,) un māte apmeklēja...*	*Viskas būtu(,) kaip aš norėčiau.* ‘viss būtu, kā es gribētu’

[56] Šajā piemērā tekstā nav kontekstuāla saistījuma.