Liepājas Universitāte

Inga Znotiņa

OTRĀS BALTU VALODAS APGUVĒJU KORPUSS: IZVEIDES METODOLOĢIJA UN LIETOJUMA IESPĒJAS

Promocijas darba kopsavilkums
filoloģijas doktora grāda iegūšanai
valodniecības zinātņu nozares lietišķās valodniecības apakšnozarē

LEARNER CORPUS OF THE SECOND BALTIC LANGUAGE: METHODOLOGY OF CREATION AND USAGE POSSIBILITIES

Summary of Doctoral Thesis
Submitted for the Conferment of the Doctoral Degree in Philology
Linguistics Subfield: Applied Linguistics

Liepāja 2018

Znotiņa, Inga. *Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas // Learner Corpus of the Second Baltic Language: Methodology of Creation and Usage Possibilities* : promocijas darba kopsavilkums. Zin. vad. Ilze Auziņa. Liepāja : Liepājas Universitāte, LiePA, 2018. 65 lpp.

Promocijas darba zinātniskā vadītāja:
Dr. philol. **Ilze Auziņa**, Latvijas Universitātes Matemātikas un informātikas institūta vadošā pētniece

Promocijas darba recenzentes:
Dr.paed. **Diāna Laiveniece**, Liepājas Universitātes profesore
Dr.philol. **Anta Trumpa**, Latvijas Universitātes Latviešu valodas institūta vadošā pētniece
Dr.hum. **Egle Žilinskaite-Šinkūniene** (Eglė Žilinskaitė-Šinkūnienė), Viļņas Universitātes Filoloģijas fakultātes lektore

Promocijas darbs izstrādāts ESF projektā „Doktora studiju attīstība Liepājas Universitātē". Vienošanās Nr. 2009/0127/1DP/1.1.2.1.2./09/IPIA/VIAA/018

Promocijas darba aizstāvēšana notiks Liepājas Universitātes valodniecības nozares promocijas padomes atklātajā sēdē 2018. gada 8. februārī plkst. 13.00 Liepājā, Kūrmājas prospektā 13, 118. auditorijā.

Ar promocijas darbu un tā kopsavilkumu var iepazīties Liepājas Universitātes bibliotēkā (Lielajā ielā 14, Liepājā) un Liepājas Universitātes mājaslapā (*www.liepu.lv*).

ISBN 978-9934-569-30-2

Saturs

Promocijas darba vispārīgs apraksts 4

Promocijas darba satura īss raksturojums 8
I daļa. Valodas apguvēju korpusi, to izveide un lietošana 8
II daļa. Otrās baltu valodas apguvēju korpuss 10

Nobeigums 19

Aizstāvēšanai izvirzītās tēzes 22

Promocijas darba aprobācija 24

General Overview of Doctoral Thesis 28

Summary of the content of Doctoral Thesis 32
I part. Learner Corpora, Their Creation and Use 32
II part. Learner Corpus of the Second Baltic Language 34

Conclusion 43

Theses for Defence 46

Approbation of the Doctoral Thesis 48

Promocijas darba kopsavilkumā izmantotā bibliogrāfija /
Bibliography Used in the Summary of Doctoral Thesis 52

PROMOCIJAS DARBA VISPĀRĪGS APRAKSTS

Līdz ar datortehnoloģiju uzplaukumu valodniecībā arvien plašāku popularitāti gūst korpusi. 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). 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 2010, Rutenberga 2012, Juknevičienė 2013, 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. 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

[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.

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 deskriptīvā 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, esošo valodas apguvēju korpusu izstrādes informācija un tehniskais apraksts; literatūra par dažādu korpusu izveidi, marķēšanu un anotēšanu, sevišķi attiecībā uz baltu valodām, baltu valodu apguves pētījumi un lingvodidaktikas terminoloģijas avoti, korpusa lingvistikas terminoloģijas avoti latviešu, lietuviešu un angļu valodā, kā arī spēkā esošie normatīvie akti, kas regulē autortiesību un personas datu aizsardzības jautājumus Latvijas Republikā un Lietuvas Republikā:

- pētījumi par valodas apguvēju korpusiem (Barlow 2005, Cigankova, Vinčela 2012, 2013, Dagneaux u. c. 1998, De Cock, Granger 2005, De Mönnink 1999, Díaz-Negrillo 2012, Gilquin 2007, Granger 1994, 1997, 1998, 2002, 2003a, 2004, 2007, 2008a,b, 2009, 2013, Grigaliūnienė, Juknevičienė 2012, Hana u. c. 2012, Hardy, Römer 2011, Jantunen 2011, Lüdeling u. c. 2005, Meunier 2007 , Myles 2005, Nesselhauf 2004, Rosen u. c. 2013, Tono 2002, Камшилова 2013 u. c.);
- valodas apguvēju korpusu izstrādes informācija (Bohát u. c. 2015, EIC-e, Granger 1993, 2003b, Grigaliūnienė u. c. 2008, Hana u. c. 2010, Juknevičienė 2013, Lüdeling et al. 2008, Vinčela 2010, Zinsmeister, Breckle 2012 u. c.);
- dažādu korpusu izveides, marķēšanas, anotēšanas raksturojums (Bārzdiņš u. c. 2007, Deksne, Skadiņa 2014, EAGLES 1996, Levāne-Petrova 2011, 2012, LVKK 2005, LVMPK 2009, Paikens 2007, 2016, Rimkutė 2006, Rimkutė u. c. 2013, Römer, O'Donnell, 2011, Rūtenberga, Kalnbērziņa 2013, Vinčela 2014, Zinkevičius 2000 u. c.);
- pētījumi par baltu valodu apguves jautājumiem un lingvodidaktikas terminoloģijas avoti (Bussmann 1996, Dabašinskienė, Čubajevaitė 2009, EKP 2006, Grīnberga 2004a,b, Kalnbērziņa u. c. 2011, Laizāne

2014a,b,c, LDLTAL 2010, LTSV 2011, LTŽ 2012, Savickienė 2006, Šalme, Auziņa 2013, VPSV 2007, Žīgure 1999 u. c.);

- korpusa lingvistikas terminoloģijas avoti (Akadterm_e, Baker u. c. 2006, Biber u. c. 2006, Bussmann 1996, Crystal 1992, 2008, ELE 2008, Grūzītis 2012, Helviga 2012 u. c., Marcinkevičienė 2000, Skadiņa u. c. 2014, Skadiņa, Vasiļjevs 2013, Spektors 2000, VPSV 2007 u. c.);
- normatīvie akti par autortiesībām un personas datu aizsardzību (ADTAI, AL, ATGTI, FPDAL).

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)[2]. 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 docētāji četrās universitātēs: Latvijas Universitātē, Liepājas Universitātē, Viļņas Universitātē un Vītauta Dižā universitātē.

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

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

PROMOCIJAS DARBA SATURA ĪSS RAKSTUROJUMS

I daļa. Valodas apguvēju korpusi, to izveide un lietošana

Darba pirmajā daļā raksturoti dažādi esošie valodas apguvēju korpusi, to klasifikācija, izveides un lietošanas īpatnības. Šajā daļā ir trīs nodaļas. **1. nodaļā *Valodas apguvēju korpusa jēdziens un raksturojums*** skaidrots valodas apguvēju korpusa termins, jēdziens un definīcija, kā arī sniegta informācija par valodas apguvēju korpusu veidiem.

1. apakšnodaļā *Valodas apguvēju korpuss: termins un definīcija* raksturoti termini un definīcijas, kādas dažādās valodās izmanto nozares pētnieki. Aplūkojot piedāvātos variantus, nolemts promocijas darbā izmantot šādus terminus: latviešu val. *valodas apguvēju korpuss*, lietuviešu val. *besimokančiųjų tekstynas*, angļu val. *learner corpus*. Aprakstīti arī daži problēmjautājumi attiecībā uz dažu korpusu paveidu piederību valodas apguvēju korpusiem.

Valodas apguvēju korpusā iekļautos tekstus var uzskatīt par starpvalodas paraugiem, un tieši starpvalodas izpētē īpaši nozīmīgs ir valodas apguvēju korpusu izmantojums. Tā kā šajā pētījumā tiek runāts par otrās baltu valodas apguvi, proti, lietuviešu valodas apguvi latviešiem vai latviešu valodas apguvi lietuviešiem, šeit tiek piedāvāts jauns termins *baltu starpvaloda*, ar to saprotot starpvalodu, kāda veidojas otrās baltu valodas apguves gadījumā.

Balstoties 1. apakšnodaļā aplūkotajās definīcijās, 2. apakšnodaļā *Valodas apguvēju korpusa raksturīgās īpašības* sīkāk aprakstīti būtiskākie valodas apguvēju korpusa parametri (autori, teksti, apjoms, mašīnlasāmība) un pētnieku izvirzītās prasības tiem. Šo prasību pamatojums salīdzināts ar korpusa izveides reālajām iespējām un to ierobežojumiem. Runājot par mašīnlasāmību, pamatots šajā darbā lietotais terminu *marķēšana* un *anotēšana* šķirums pretēji Latvijā ierastajai praksei.

3. apakšnodaļā *Valodas apguvēju korpusu veidi*, analizējot un apkopojot līdzšinējos nozares attīstības aprakstus, raksturotas valodas apguvēju korpusu klasifikācijas iespējas atbilstoši četrām dimensijām:

1) parametri, kas ir saistīti ar korpusā iekļaujamajiem tekstiem un to īpašībām (mērķvaloda, valodu skaits, producēšanas veids, tekstu tips, tematika un oriģinalitāte, specializācija);
2) parametri, kas ir saistīti ar korpusā iekļaujamo tekstu tapšanas apstākļiem, autoriem un ieguvi (dzimtā valoda, dzimto valodu skaits korpusā, valodas apguvēju vecums, valodas prasmju līmenis un to skaits korpusā, valodu apguves secība, tekstu tapšanas secīgums un to rakstīšanas veids);
3) parametri, kas ir saistīti ar korpusā iekļaujamo tekstu apstrādi un tehnisko noformējumu, korpusu sagatavojot pētniecības darbam (tekstu nedalāmība, kopējais apjoms, anotējums, izmantotā programmatūra, valodu pāru virzieni);

4) parametri, kas ir saistīti ar korpusa lietošanu, tās iespējām un līdzšinējo darbu ar attiecīgajiem korpusiem (korpusa pedagoģiskais lietojums, korpusa komerciāls vai akadēmisks lietojums un pieejamība).

Klasifikācijai ir divas funkcijas: no vienas puses, tā parāda valodas apguvēju korpusu daudzveidību un iespējamās variācijas, bet, no otras puses, noskaidrojot konkrēta korpusa vietu klasifikācijā, kļūst skaidrāka tajā iekļauto datu būtība un to sniegtās pētniecības iespējas.

2. nodaļā ***Vēsture un izplatība*** īsi raksturota valodas apguvēju korpusu pētniecības virziena attīstība. 1. apakšnodaļā *Valodas apguvēju korpusi pasaules kontekstā* atklāta valodas apguvēju korpusu izveidošanās no kļūdu kartotēkām un pašreizējā izaugsme dažādās valstīs, īpaši Eiropā, Amerikā un Āzijā.

2. apakšnodaļa *Valodas apguvēju korpusi latviešu un lietuviešu valodniecībā* koncentrējas uz situāciju Latvijā un Lietuvā. Kā liecina šajā nodaļā raksturotais, lai arī sākotnēji vairums korpusu veidoti angļu valodai, to straujā izplatība dažādās valstīs veicina arī citu šajās valstīs lietotu valodu apguvēju korpusu izveidi, un tas attiecināms arī uz Latviju un Lietuvu. Abās valstīs tiek strādāts ar svešvalodu (galvenokārt angļu, franču) apguvēju korpusiem, taču ir bijuši arī centieni veidot baltu valodu apguvēju korpusus. Neviens no tiem gan nav bijis plašākam pētnieku lokam pieejams, savukārt iepriekš Viļņas Universitātē latviešu valodas apguvēju korpusam savāktie dati tika nodoti otrās baltu valodas korpusa izveidei, veicinot šī pētījuma tapšanu.

3. nodaļā ***Darbs ar valodas apguvēju korpusiem*** raksturots darbs ar valodas apguvēju korpusiem un galvenās metodes to izmantojumā. Vispirms 1. apakšnodaļā *Korpusa lingvistikas procedūras darbā ar valodas apguvēju korpusiem* raksturotas galvenās korpusa lingvistikas procedūras (konkordanču rindu meklēšana, biežuma sarakstu sastādīšana u. c.) un to lietojuma īpatnības valodas apguvēju korpusa gadījumā. Aprakstīts to lietojums kvalitatīvos un kvantitatīvos, kā arī jaukta tipa pieejas pētījumos.

Pēc tam skaidrots, kādos pētījumos un ar kādām metodēm valodas apguvēju korpusi lielākoties tiek lietoti. Visbiežāk sastopamās metodes darbā ar valodas apguvēju korpusiem ir kontrastīvā starpvalodas analīze un datorizētā kļūdu analīze, tāpēc atsevišķi aprakstītas šo divu metožu lietojuma tendences.

Visbeidzot aprakstīts arī valodas apguvēju korpusu lietojums valodas apguvē un valodas apguvei paredzētu materiālu izstrādē. Ir nosaukti vairāki korpusu izmantojumu virzieni (programmu, materiālu, pedagoģisko vārdnīcu izstrāde, tiešs pedagoģiskais lietojums klasē u. c.), taču daļa no tiem nav sevišķi izplatīti. Secināts, ka lielākā daļa valodas apguvēju korpusu pedagoģiskā izmantojuma veidu ir saistīti ar izpēti, gatavojoties nodarbībām un izstrādājot tām paredzētos materiālus, tāpēc docētājam, kurš pats šajā procesā nav iesaistīts un galvenokārt nodarbojas tikai ar tiešo mācīšanu, valodas apguvēju korpusi tiešā veidā nav noteikti nepieciešami. Nav arī sagaidāms, ka korpusi ievērojami mainītu līdzšinējo mācību praksi kopumā. Tomēr arī šie docētāji gūst labumu no korpusu izpētes, izmantojot citu valodnieku

izstrādātos materiālus, turklāt tieši šādiem docētājiem nereti ir plaša pieeja valodas apguvēju producētajam materiālam, kas, nodots pētnieku rokās, veicinātu šādu materiālu izstrādi. Līdz ar to arī gadījumos, kad docētājs apzināti izvēlas pats valodu apguvēju korpusus neizmantot, sadarbība ar pētniekiem, kas to dara, var sniegt vērtīgu ieguldījumu nozares attīstībā.

II daļa. Otrās baltu valodas apguvēju korpuss

Darba otrajā daļā raksturota otrās baltu valodas apguvēju korpusa izveide. Tajā raksturots jaunveidojamais korpuss un tam izvēlētie risinājumi.

Promocijas darba otrajā daļā ir 4 nodaļas. **1. nodaļā *Avotu atlase*** runāts par avotu atlasi, ieskaitot ne vien tekstu ieguvi, bet arī personas datu aizsardzības un autortiesību jautājumus. 1. apakšnodaļā *Tekstu ieguve un atlases kritēriji* stāstīts, kā docētāji ar kontrolēta vērojuma metodi ieguvuši tekstus, kādi nosacījumi to autoriem izvirzīti un kā no iegūtā materiāla atlasīts korpusā iekļaujamais materiāls.

Korpusā iekļautie teksti ir Lietuvas un Latvijas augstskolu studentu rakstītie sacerējumi, kas tikuši uzdoti kā patstāvīgi veicami mājasdarbi otrās baltu valodas kursā iesācējiem. Visi korpusā iekļautie teksti ir tapuši, to autoriem otro baltu valodu mācoties pirmo vai otro semestri bez priekšzināšanām. 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 saskaņā ar ES vadlīnijām (EKP 2006). Tas gan lielā mērā ir atkarīgs no nodarbību skaita semestrī. 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. Visiem korpusa tekstu autoriem otrā baltu valoda ir svešvaloda.

Šobrīd korpusā ir iekļauti dati no Latvijas Universitātes (LU; Rīga, Latvija), Liepājas Universitātes (LiepU; Liepāja, Latvija), Vītauta Dižā universitātes (VDU; Kauņa, Lietuva) un Viļņas Universitātes (VU; Viļņa, Lietuva). Latviešu valoda Lietuvā un lietuviešu valoda Latvijā tiek mācīta vēl vairākās augstākās izglītības iestādēs, taču pagaidām teksti no tām nav iegūti. Sadarbība ar tām iespējama, korpusu tālāk papildinot.

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. Daļa tekstu docētājiem iesniegta digitāli, bet gadījumos, kad teksts iesniegts rokrakstā, tas digitalizēts, visu tekstu manuāli pārrakstot datorrakstā. Teksti ir studentu patstāvīgi rakstīti, docētāji iespēju robežās novērsuši plaģiātu.

2. apakšnodaļā *Korpusa „Esam" vieta valodas apguvēju korpusu klasifikācijā* otrās baltu valodas apguvēju korpuss ir raksturots atbilstoši promocijas darba pirmajā daļā sniegtajai klasifikācijai:

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

- Korpusa mērķvaloda ir otrā baltu valoda.
- Valodas producēšanas veids – rakstveida.
- Tekstu tips – apraksts, dažos gadījumos robežojoties ar eseju.
- Pēc tekstu tematikas šis būtu uzskatāms par netematisku korpusu.
- Tekstu oriģinalitāte – korpusā ir oriģināli teksti, nevis tulkojumi.
- Pēc valodu skaita korpuss „Esam" ir divvalodīgs.

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

- Teksta autora dzimtā valoda – viena no baltu valodām.
- Valodas apguvēju vecums – korpusā iekļauto tekstu autori visi ir pieaugušie, tie ir augstskolu studenti. Vecums lielākajai daļai ir līdzīgs – ap divdesmit gadiem, taču ir arī nedaudzi studenti, kas ir vecāki par kursabiedriem.
- Valodas prasmju līmenis – iesācēji bez priekšzināšanām.
- Pēc valodu apguves secības visiem autoriem otrā baltu valoda ir svešvaloda.
- Pēc tekstu tapšanas secīguma korpuss ir nosacīti sinhronisks. Tajā iekļautie teksti ir tapuši 8 gadu laikā no 2007. gada līdz 2014. gadam.
- Pēc tekstu rakstīšanas veida korpuss ir neviendabīgs. Tajā ir gan datorrakstā tapuši teksti, gan arī tādi, kas ir rakstīti ar roku.

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.
- Korpusam nav noteiktas specializācijas, tas ir vispārīga rakstura valodas apguvēju korpuss.
- Dzimtās valodas šajā korpusā ir divas: latviešu un lietuviešu (ar iespējamām papildu dzimtajām valodām).
- Valodas prasmes līmenis korpusā „Esam" ir A.
- Pēc kopējā apjoma korpuss ir mazs – apm. 52 000 vārdlietojumu.
- Anotējums – korpusā ir anotētas pamatformas, vārdšķiras, sintaktiskie teikumu veidi un valodas lietojuma kļūdas.
- Izmantotā programmatūra – programma *TEITOK*, kas darbojas, balstoties uz CQP/CWB platformas.
- Valodu pāru virzieni šajā korpusā ir divi – tas ir divvirzienu korpuss.

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.

- Pēc korpusa izmantojuma sfēras šis ir akadēmisks korpuss.
- Pēc pieejamības „Esam" ir publiski pieejams korpuss.

3. apakšnodaļā *Personas datu aizsardzība un autortiesības* īsumā raksturots šo jautājumu regulējums Latvijas Republikas un Lietuvas Republikas tiesību aktos, ciktāl tas skar otrās baltu valodas apguvēju korpusa izveidi. Personas datu aizsardzības nolūkos daļas no tekstiem ir anonimizētas, korpusā iekļaujamajos tekstos reālos personas datus aizstājot ar izdomātiem, taču pēc iespējas cenšoties saglabāt katra teksta īpatnības. Korpusā iekļautajos tekstos aizstāti teksta autora un teksta autoram pazīstamu personu vārdi un uzvārdi; pilsētu un apdzīvotu vietu nosaukumi, mājvārdi, adreses; profesijas; mācību iestāžu un/vai to struktūrvienību nosaukumi. Ikviena no šīm informācijas vienībām tiek individuāli izvērtēta. Lai nezaudētu pārlieku daudz no sākotnējā teksta, katrā gadījumā tiek lemts atsevišķi. 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.

Autortiesību aizsardzības nolūkos ar visu korpusā iekļaujamo tekstu autoriem slēgta vienošanās par atļauju darbus izmantot valodas apguvēju korpusā, kā arī par veidu, kādā tiek vai netiek norādīts katra atsevišķa teksta autora vārds un uzvārds. Apstrādāti un korpusā iekļauti tika tikai tie teksti, par kuriem ir saņemta to autoru rakstveida piekrišana.

2. nodaļā *Tekstu apstrāde, marķēšana un anotēšana* aprakstīta tekstu mašīnlasāmības nodrošināšanas gaita. 1. apakšnodaļā *Digitalizēšana* raksturoti principi, pēc kādiem digitalizēti rokrakstā vai drukā iesniegtie teksti: kļūdas un īpatnības iespēju robežās tiek saglabātas, autora paša labojumi tiek ņemti vērā, taču docētāja labojumi tiek atmesti, lai nodrošinātu teksta autentiskumu.

2. apakšnodaļā *Marķēšana* raksturota ekstralingvistiskas papildinformācijas, kura nav atkarīga no lingvistiskas analīzes, jeb metadatu iekļaušana teksta datnēs. Korpusā ir divu veidu marķējums: strukturālais marķējums (marķēts teksta, rindkopas un izteikuma sākums un beigas, kā arī anonimizētās vietas tekstā) un metadatu marķējums (teksta autors, augstskola, semestris, mērķvaloda un nosaukums).

3. apakšnodaļā *Anotēšanas veidu izvēle* raksturoti kritēriji, pēc kādiem vadoties, pieņemts lēmums par korpusā izmantojamajiem anotējuma veidiem. Lai to izdarītu, jāsalīdzina ieguldāmie resursi (anotētāja zināšanas, pieejamie tehnoloģiskie risinājumi, laika patēriņš, anotēšanas sistēmas pieejamība vai nepieciešamība tādu radīt u. tml.) ar ieguvumiem (ieguvēju pētījumu potenciālais skaits, apmērs, nozīmīgums u. tml.). 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 raksturotas valodas apguvēju korpusu izveides un pētniecības tendences Latvijā, Lietuvā un citur. Vadoties pēc līdzšinējām tendencēm, nolemts otrās baltu valodas apguvēju korpusā anotēt vārdšķiras, pamatformas, teikuma veidus un kļūdas. Zemas

popularitātes un ierobežotas lietderības dēļ nav izmantots morfosintaktisks un problēmorientēts anotējums.

Tekstos ir sastopamas gan kļūdas, gan neskaidras gramatiskās konstrukcijas, tāpēc automātisko anotēšanas rīku iespējas ir ierobežotas, un rezultātu pārskatīšana būtu darbietilpīga. Ņemot to vērā, teksti ir anotēti manuāli vai pusautomātiski.

4. apakšnodaļā *Pamatformu anotēšana* skaidrots pamatformu anotēšanas process un formāts otrās baltu valodas apguvēju korpusā. Nolemts, anotējot pamatformas, vadīties pēc līdzšinējiem darbiem attiecīgās valodas pamatformu noteikšanā, par pamatu ņemot abu valodu morfoloģiskās anotēšanas rīku sniegtos variantus un nepieciešamības gadījumā tos labojot ar pārbaudi vārdnīcās. Anotēšanas procesā sastopoties ar sarežģītākiem piemēriem, radīti tiem īpaši noteikumi, piem., deminutīvi tiek anotēti kā atsevišķi vārdi, nevis vārdformas.

5. apakšnodaļā *Morfoloģiskā anotēšana* raksturots vārdšķiru anotējums otrās baltu valodas apguvēju korpusā. Tajā izmantota projekta *Semti-Kamols* latviešu valodas morfoloģisko pazīmju kopa (LVMPK 2009) ar papildinājuma kodu *z* pieturzīmēm. 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.

6. apakšnodaļā *Sintaktiskā anotēšana* raksturots izteikumu anotējums pēc teikuma veida. Korpusā iekļautie teksti ir sadalīti izteikumos un izsacījumos, un izteikumi ir iedalīti vairākās pamatgrupās pēc tā, kādam teikuma modelim tie atbilst. Izmantota šāda klasifikācija:

- vienkāršs teikums (nepaplašināts/paplašināts);
- salikts teikums (sakārtots/pakārtots/jaukts);
- izsacījums;
- neskaidri veidots teikums.

Gadījumos, kad tekstā ir tiešā runa, tā tiek uzskatīta par tā paša izteikuma neatkarīgu(-ām) daļu(-ām).

7. apakšnodaļā runāts par kļūdu anotējumu. Vispirms raksturotas dažādas izpratnes par kļūdas jēdzienu. Balstoties dažādu pētnieku definīcijās, otrās baltu valodas apguvēju korpusa anotēšanas nolūkā kļūda tiek saprasta kā neatbilstība priekšstatam par to, kā attiecīgajai valodas struktūrai pareizi jābūt veidotai, jeb mērķa hipotēzei, kādu izvirza teksta labotājs.

Tālāk raksturots anotēšanas process un lomu sadalījums labotāju/anotētāju pārī. Labošana tiek uzticēta cilvēkam, kuram teksta mērķvaloda ir dzimtā valoda, savukārt kļūdu tipus nosaka un anotē cilvēks, kuram teksta mērķvaloda nav dzimtā valoda, taču ir ļoti labas šīs valodas prasmes, un dzimtā valoda sakrīt ar teksta autora dzimto valodu.

Tā kā līdz šim nav tikusi izveidota iesācēju tekstu kļūdu anotēšanai piemērota pazīmju kopa latviešu un lietuviešu valodai, otrās baltu valodas apguvēju korpuss anotēts pēc S. Greindžeres (Granger 2003a) izstrādātās klasifikācijas, 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 klasifikācijā ir pieci kļūdu tipi ar apakštipiem (sk. 1. tabulu):

1. 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)[3] ‘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’
Forma	F	Citas pareiz-rakstības kļū-das (ieskaitot pārrakstīša-nos)	FP	*man patīk tiktis (tikties) ar draugiem*	*kikvieną dieną* ‘katru dienu’
Morfoloģija un vārddarināšana	M	Atvasināšana	MA	*patīk futbols, basketbols, vazinātes*[4] *ar ritēni (riteni)*	*todėl užmiegojome anksti* ‘tāpēc aizmigām agri’
		Salikteņdari-nāšana	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) atvaļi-nājumiem*	*įvariuose gyvenimo valandą* ‘dažādās dzīves stundās’

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

		(Ne)noteiktā galotne	MG	*Fotoaparātā bija manas … skaistas fotogrāfijas*	*žmonių kamšatis ir ilgoji (ilgos) valandos viešajame transporte* ‘cilvēku saspiestība un ilgās stundas sabiedriskajā transportā’
Morfoloģija un vārddarināšana	M	Salīdzināmās pakāpes	MQ		*Aš esu jaunesnioji (jaunausia).* ‘es esmu visjaunākā’
		Persona	MP	*Tėvs interesėjies par automobiliem (automobi-ļiem)*	*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’
Morfoloģija un vārddarināšana	M	Divdabis	MV		*vairuotojas, matytint, kad bėgtu (bėgu), (..) pristabdau (pristabdo)* ‘vadītājs, redzot, ka skrienu, piebremzē’

Kļūdas tips	**Apzīmējums**	**Kļūdas apakštips**	**Apzīmējums**	**Piemērs latviešu valodā**	**Piemērs lietuviešu valodā**
		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?[5]*)* ‘Bioloģijas fakultātē ir ļoti daudz (kā?)’
Sintakse	S	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’

[5] Šajā piemērā tekstā nav kontekstuāla saistījuma.

		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’

Katrai kļūdai ir izvēlēta tikai viena, pēc labotāja domām, iederīgākā mērķa hipotēze un viens, pēc anotētāja domām, iederīgākais kļūdas tips.

Tā kā korpuss ir elektronisks, tā veiktspēja ir atkarīga no izmantotās programmatūras. **3. nodaļā *Programmatūra*** aprakstīts, kādas programmas ir izmantotas korpusa izveidē un kādas ir paredzēts izmantot, lietojot korpusu. 1. apakšnodaļā *Korpusa izveidē izmantotā programmatūra* nosaukti teksta redaktori, kas izmantoti, tekstus digitalizējot, un raksturota tekstu un autoru kodu ģenerēšana ar nejaušo skaitļu virkņu ģeneratoru. Norādīts, kā teksti sagatavoti aplūkošanai izmēģinājuma korpusā ar programmu *AntConc*. Tālāk parādīts anotēšanas process korpusa gala versijā ar programmas *TEITOK* saskarnes starpniecību un sniegts anotējuma tehniskais raksturojums: marķēšanas un anotēšanas iezīmes. Korpusa pilnās versijas datnes atbilst *TEI* standartam.

2. apakšnodaļā *Korpusa lietošanai nepieciešamā programmatūra* aprakstīta programmatūra, kāda tiek izmantota, korpusu lietojot. Tā kā izmēģinājuma korpuss ir pieejams kā atsevišķas datnes lejupielādei, to aplūkošanai var izmantot jebkuru programmu pēc lietotāja izvēles, taču ieteiktā korpusa izpētes programma ir *AntConc*, un visu datņu saderība ar šo programmu ir pārbaudīta. Tomēr šī programma nav izmantota korpusa gala versijā, jo programmu nepieciešams lejupielādēt katrā datorā, kurā iecerēts strādāt ar korpusu, turklāt atkarībā no drošības uzstādījumiem datorā to var neizdoties palaist. Tāpēc korpusa gala versija ir pieejama tiešsaistē, un to nav nepieciešams lejupielādēt. 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.

Lietotāja saskarne ir redzama interneta pārlūkā, vietnē *www.esamkorpuss.lv*. Saskarne ir Mārtena Jansena (*Maarten Janssen*) izstrādātā *TEITOK*. Promocijas darbā sniegts īss tās lietošanas apraksts. Tā kā *TEITOK* darbojas uz *CQP* bāzes, tajā var izmantot regulārās izteiksmes.

4. nodaļā *Pētījumi otrās baltu valodas apguvēju korpusā* sniegts īss apraksts ar piemēriem par šī korpusa izmantošanu konkrētu metožu kontekstā. 1. apakšnodaļā *Kontrastīvā starpvalodas analīze* raksturotas iespējas salīdzināt korpusa materiālu lietuviešu valodā ar materiālu latviešu valodā, lai atklātu iespējamās atšķirības latviešu valodas kā otrās baltu valodas un lietuviešu valodas kā otrās baltu valodas

apguvē. Savukārt 2. apakšnodaļā *Datorizēta kļūdu analīze* raksturotas datorizētas kļūdu analīzes iespējas otrās baltu valodas apguvēju korpusā.

3. apakšnodaļā *Baltu starpvaloda* runāts par iespējām izmantot divvirzienu materiālu, to nedalot pēc mērķvalodas. Tādējādi iespējams analizēt visai baltu starpvalodai raksturīgas iezīmes. 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.

Baltu starpvalodas pētīšana nav atsevišķa metode, drīzāk gan aspekts, kādā ir iespējams aplūkot korpusa datus, izmantojot citas metodes. 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.

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 docētāju 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 (tajā ir 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 docētājus 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 docētājs 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. 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 docētājiem 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 docētāji šajā procesā var sniegt būtisku ieguldījumu, turpinot vākt tekstus ievietošanai korpusā. Tātad ļoti liela nozīme ir pētnieku un docētāju savstarpējai sadarbībai – ieguvējas būtu abas puses.

Korpusa izveidē ņemtas vērā citu valodas apguvēju korpusus veidojušu pētnieku atziņas, taču ir ieviesti arī jauninājumi. 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 mācīta arvien retāk. Cerams, ka otrās baltu valodas apguvēju korpusa izveide varētu veicināt otrās baltu valodas popularitātes atgriešanos gan studentu, gan pētnieku lokā.

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 Silviānes 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).

PROMOCIJAS DARBA APROBĀCIJA

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), Lielbritānijā (Šefīldā, Lenkasterā) un Zviedrijā (Gēteborgā):

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. „Valodas apguvēju korpusu izmantojums svešvalodas mācību procesā." Rīgas Stradiņa universitātes ikgadējā zinātniskā konference. Rīgā 2016. gada 17. –18. martā.
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 4. 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." 11. starptautiskā zinātniskā konference *Rural Environment, Education, Personality 2018*. Jelgavā 2017. gada 11.–12. maijā.
23. Inga Znotiņa. „Learner corpus of the second Baltic language: annotation and data comparability." Seminārs par otrās valodas resursu un rīku savietojamību (*Workshop on interoperability of Second Language Resources and Tools*). Gēteborgā 2017. gada 6.–8. decembrī.

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. 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.

GENERAL OVERVIEW OF DOCTORAL THESIS

The rise of computer technologies has facilitated the growth of popularity of language corpora. A corpus in linguistics is a "large body of texts available for computerized analysis" (VPSV 2007, 196)[6], and the respective subfield of applied linguistics is called corpus linguistics (VPSV 2007, 196). The texts included in a corpus are usually chosen so that they reflect a certain type of communication or variety of language as precisely as possible (ELL 2005, 234). For that, various specialized corpora are often created. Those are corpora that consist of one certain kind of texts (more on this see Flowerdew 2004), as opposed to general corpora which are made with the goal to reflect the language as a whole (McEnery u. c. 2006, 15). One subtype of specialized corpora are learner corpora (they are described as such in, e.g., Koester 2010). Such corpora consist of target language samples produced by learners of that language. Learner corpora are being used increasingly extensively for researching language learning, errors, and the influence native language may have on them. However, most learner corpora have been created for the most widespread languages in the world: English, Spanish, Chinese etc. Such corpora have also been created in Latvia and Lithuania among other countries (Vinčela 2010, Rutenberga 2012, Juknevičienė 2013, Grigaliūnienė u. c. 2008 etc.).

There are not a lot of corpora in less widespread "native language : target language" pairs, and while creating such corpora, one needs to solve various issues of methodological nature in such domains as text collection, error annotation, copyright. et al. Therefore, the conception of the doctoral thesis is to create a publicly accessible learner corpus of the second Baltic language. The second Baltic language in Baltistics is usually understood as Latvian for people whose mother tongue is Lithuanian, and Lithuanian for those whose mother tongue is Latvian[7] (Butkus 2008, 57). Thus, the corpus is intended for linguistic research of the beginnings of learning Latvian as the second Baltic language and Lithuanian as the second Baltic language.

The **novelty** of this research is that even though the development, researching and using language technologies is expanding, it is still a rather fragmented field, and a part of potential areas of work are not being paid enough attention (Skadiņa u. c. 2014, 232). That can also be said about learner corpora. This is the first publicly accessible learner corpus for Baltic languages, as well as the first publicly accessible learner corpus in Latvia.

The process of creating the corpus of the second Baltic language is atypical in various aspects. One of those – this is a bi-directional corpus: it includes the texts written by learners of not one but two languages with the condition that both language pairs include the same two languages where each language is the learner's native language in one of the pairs, and the target language in the other one. It is not known whether any researchers anywhere in the world have worked with a bi-

[6] Here and further – my translation into English – I. Z.

[7] Although learning Prussian has become more popular lately, it is still generally considered an extinct language (Blažek 2007, 100), and more attention is being paid to the issues of acquisition of the modern Baltic languages.

directional learner corpus in any language pair, so it can be assumed that this is, if not the first, then at least one of the first such corpora. It is, however, not a complete novelty in corpus linguistics in general – researchers all over the world, including the Baltic States, are working with bi-directional parallel corpora, namely, translation corpora that consist of translations from language A into language B as well as translations from language B into language A (Johansson 2007, 302; Rimkutė u. c. 2013, 73). In translations, just like in language acquisition, the source language has an important impact on the target language (Sīlis 2009, 29–35), and researchers often discuss issues of interference (Zauberga 2001). Bi-directional research can not only reveal tendencies of native language's influence, but also aid contrastive analysis of the systems of two languages (e.g., see a comparison of demonstrative pronouns in Latvian and Lithuanian for translation in Znotiņa 2012). Similar gains could likely be expected from learner corpora consisting of bi-directional material.

Topicality of the paper is determined by the fact that learner corpora are popular worldwide, and creating new corpora for languages that do not yet have such corpora would allow for perfecting researchers' knowledge on language acquisition in the global context. It is also topical locally: the learner corpus of the Second Baltic language helps one investigate the main difficulties in the second Baltic language acquisition, thus inducing improvements in study materials as well as comparative studies of both languages.

The **problems** addressed in the research include the necessity to solve various theoretical and practical, including legal, methodological, and technical, issues in order to create a publicly accessible learner corpus of the second Baltic language.

Based on the aforementioned, the **subject** of the doctoral thesis is "Learner corpus of the second Baltic language: methodology of creation and usage possibilities", and the object of the research – the actions necessary to create a learner corpus of the second Baltic language, and methods for researching the material of such a corpus.

The **research question** is – how to create and use a learner corpus of the second Baltic language?

The **goal** of the research is double-sided. On one hand, it is to create a learner corpus that is based on the independently written texts by learners of the second Baltic language (Latvians who study Lithuanian language, and Lithuanians who study Latvian language) in the respective target language, and to develop the methodology required for reaching this goal. On the other hand, it is to describe the creation of a publicly accessible learner corpus of the second Baltic language, to explain the more important choices and characterize the usage possibilities of this corpus in research.

In order to achieve this, the following **objectives** have been set:

1) to describe the concept of a learner corpus, their types and steps of creation process;
2) to collect texts for the corpus and to ensure protection of copyrights and personal data;
3) to mark and describe the texts included in the corpus;

4) to ensure technical support for the performance and public accessibility of the corpus;
5) to create or adapt existing corpus annotation tagsets;
6) to annotate the corpus;
7) to describe the usage of the most important features of the corpus.

Several **methods** have been used in the present research. The description of history and prevalence of learner corpora is mainly based on descriptive analysis of academic literature. Referative analysis of academic literature is used to explain the notion and usage of learner corpora.

Comparative analysis of methods is used to describe the most popular methods in work with learner corpora. It helps deciding which ones should be considered essential in the case of the learner corpus of the second Baltic language. After the choice is made, the analytically synthetic method is used to create the corpus of the second Baltic language. The corpus is given the name "Esam", and the basis of the name is the similarity between the 1st person plural forms of Latvian verb *būt* and Lithuanian verb *būti* 'to be': Latv. *esam*, Lith. *esame* (there is also a shortened form *esam* in colloquial Lithuanian). The lexical and grammatical meanings of this word emphasize the unity of Baltic languages as well as the fact that the authors of the texts, the creators of the corpus, and – to some extent – also the users of the corpus belong to the Baltic cultural environment.

The final part of the thesis is dedicated to showing how the newly created corpus can be used in research. It especially emphasizes two methods: contrastive interlanguage analysis and computerized error analysis.

The **theoretical** background of the research consists of various publications in learner corpus research; the information on development of existing learner corpora and their technical description; papers on building, markup and annotation of various corpora, especially for Baltic languages; research on acquisition of Baltic languages; sources of terminology for language pedagogy and corpus linguistics in Latvian, Lithuanian and English, as well as the current regulations of copyrights and personal data protection in Republic of Latvia and Republic of Lithuania:

- research on learner corpora (Barlow 2005, Cigankova, Vinčela 2012, 2013, Dagneaux u. c. 1998, De Cock, Granger 2005, De Mönnink 1999, Díaz-Negrillo 2012, Gilquin 2007, Granger 1994, 1997, 1998, 2002, 2003a, 2004, 2007, 2008a,b, 2009, 2013, Grigaliūnienė, Juknevičienė 2012, Hana u. c. 2012, Hardy, Römer 2011, Jantunen 2011, Lüdeling u. c. 2005, Meunier 2007 , Myles 2005, Nesselhauf 2004, Rosen u. c. 2013, Tono 2002, Камшилова 2013 etc.);
- descriptions of creating learner corpora (Bohát u. c. 2015, EIC-e, Granger 1993, 2003b, Grigaliūnienė u. c. 2008, Hana u. c. 2010, Juknevičienė 2013, Lüdeling et al. 2008, Vinčela 2010, Zinsmeister, Breckle 2012 etc.);
- papers on building, markup and annotation of various corpora (Bārzdiņš u. c. 2007, Deksne, Skadiņa 2014, EAGLES 1996, Levāne-

Petrova 2011, 2012, LVKK 2005, LVMPK 2009, Paikens 2007, 2016, Rimkutė 2006, Rimkutė u. c. 2013, Römer, O'Donnell, 2011, Rūtenberga, Kalnbērziņa 2013, Vinčela 2014, Zinkevičius 2000 etc.);

- research on issues of acquisition of Baltic languages and terminology of language pedagogy (Bussmann 1996, Dabašinskienė, Čubajevaitė 2009, EKP 2006, Grīnberga 2004a,b, Kalnbērziņa u. c. 2011, Laizāne 2014a,b,c, LDLTAL 2010, LTSV 2011, LTŽ 2012, Savickienė 2006, Šalme, Auziņa 2013, VPSV 2007, Žīgure 1999 etc.);
- sources of corpus linguistics terminology (Akadterm$_{-e}$, Baker u. c. 2006, Biber u. c. 2006, Bussmann 1996, Crystal 1992, 2008, ELE 2008, Grūzītis 2012, Helviga 2012 u. c., Marcinkevičienė 2000, Skadiņa u. c. 2014, Skadiņa, Vasiļjevs 2013, Spektors 2000, VPSV 2007 u. c.);
- regulations on copyrights and protection of personal data (ADTAI, AL, ATGTI, FPDAL).

It is a truism that corpora cannot be divided into "useful corpora" and "useless corpora" because the way a corpus is built depends on the type and expected usage of said corpus (Hunston 2008, 155). Therefore, it is important to not only create the corpus but also to understand what can and what cannot be expected from it, what conclusions can and what conclusions cannot be drawn from it. Every research that uses a corpus must account for the type of the corpus and the material included in it, the size of the corpus etc. The material and sources of the research are second Baltic language learners' independently written texts in the respective target languages. At the time of writing the texts, authors are students in higher education institutions. The texts have been collected and offered for including in the corpus by the teachers of the second Baltic language in four universities: University of Latvia, Liepaja University, Vilnius University, and Vytautas Magnus University.

The **structure** of the doctoral thesis is formed by an introduction, two parts with chapters and subchapters, conclusion, the theses for defence, bibliography (342 units) and 6 appendices. The paper includes 9 tables and 16 pictures.

The research was facilitated by a practice in the Centre of Computational Linguistics of Vytautas Magnus University (Kaunas, Lithuania) from May 1 2013 till July 31 2013. During the practice, the theoretical basis of the research was developed, and legal issues connected with copyrights and personal data protection were addressed.

The skills necessary for the research were improved in the Corpus Linguistics Summer School organized by the University Centre for Computer Corpus Research on Language (UCREL), Lancaster University, United Kingdom from July 15 2014 till July 18 2014.

SUMMARY OF THE CONTENT OF DOCTORAL THESIS

I part. Learner Corpora, Their Creation and Use

The first part of the thesis describes various existing learner corpora, their classification, specifications of creation and use. This part consists of three chapters. The **1st chapter** ***Notion and characteristics of learner corpus*** explains the term *learner corpus*, its notion and definition. It also offers information on various kinds of learner corpora.

The 1st subchapter *Learner corpus: term and definition* describes the terms and definitions used by various researchers in various languages. After discussing the offered variations, it is decided to use terms *valodas apguvēju korpuss* in Latvian, *besimokančiųjų tekstynas* in Lithuanian, and *learner corpus* in English. The possibilities of considering some subtypes of corpora as learner corpora are also discussed.

The texts included in a learner corpus can be called samples of interlanguage, and it is exactly interlanguage research where using learner corpora is especially valuable. Since this research is aimed at acquisition of the second Baltic language, namely, acquisition of Lithuanian for Latvians, and acquisition of Latvian for Lithuanians, it offers a new term *Baltic interlanguage* to speak about the interlanguage that is formed in the case of learning the second Baltic language.

Based on the definitions described in the 1st subchapter, the 2nd supchapter *Characteristics of learner corpora* further elaborates on the most essential parameters of a learner corpus (authors, texts, computer-readability) as well as the requirements set by various researchers. The basis of those requirements is compared with the actual possibilities and limitations of creating a corpus. The description of machine-readability issues also explains the usage of two different terms *markup* and *annotation* which opposes the common practice in Latvia.

The 3rd subchapter *Types of learner corpora* analyzes and summarizes descriptions of the development of learner corpus research, thus emphasizing four dimensions of learner corpus classification:

1) parameters connected with the texts and their qualities (target language, number of languages, production type, text type, topic, originality, specialization);
2) parameters connected with the conditions in which the texts were created, their authors and collection (native language, number of native languages in corpus, age of learners, language skill level, number of language skill levels in the corpus, sequence in which languages were learned, writing medium);
3) parameters connected with the handling and technical processing of the texts while preparing the corpus for research (fullness of texts, size of the corpus, annotation, used software, language pair directions);

4) parameters connected with usage of the corpus, the possibilities it offers and the research done so far (pedagogical use of the corpus, commercial or academic use of the corpus, accessibility).

The classification serves two functions: on the one hand, it shows the diversity of learner corpora, and, on the other hand, it allows for specifying the place a corpus takes in the classification making it clearer what data it includes and what research options it offers.

The **2nd chapter** ***History and prevalence*** shortly describes the development of the field of researching learner corpora. The 1st subchapter *Learner corpora in global context* describes learner corpora's evolution from error collections and their current spread in various countries, especially in Europe, America, and Asia.

The 2nd subchapter *Learner corpora in Latvian and Lithuanian linguistics* is aimed at the situation in Latvia and Lithuania. Even though most corpora were initially made for English, their swift popularity gain facilitates the creation of learner corpora for other languages, and this is true for Latvian and Lithuanian as well. Both Latvian and Lithuanian researchers are working with foreign (mostly English, French) language learner copora, but there have also been attempts to create learner corpora for Baltic languages. None of those has been accessible for a wider scope of researchers. The material collected for a Latvian learner corpus in Vilnius University was eventually handed over for including in the learner corpus of the second Baltic language, thus helping in this research.

The **3rd chapter** ***Work with learner corpora*** explains how learner corpora are used. The 1st subchapter *Corpus linguistics procedures in learner corpora* describes the main procedures of corpus linguistics (concordancing, frequency lists etc.) and how they are used in the case of a learner corpus. Usage in qualitative, quantitative and mixed approach research is outlined.

Next, the most common learner corpus research types and methods are covered. The most common methods are contrastive analysis and computerized error analysis. Therefore, tendencies of using these methods are pointed out especially.

Finally, the use of learner corpora in classroom and learning materials has been discussed. Various fields of corpus use have been listed (study program creation, study material creation, pedagogical lexicography, direct use in classroom etc.), but a part of them are not very widespread. It is concluded that most of the pedagogical use of learner corpora is connected with research during the preparation for lessons and study material creation, so a teacher who is not involved in this process and mostly only teaches does not necessarily need learner corpora. These corpora are also not expected to significantly change the study process overall. However, such non-involved teachers also benefit from corpus research by using study materials created by other linguists. Besides, these teachers frequently have access to a wide range of learner produced texts which, given to researchers, would benefit the creation of such study materials. Therefore, even if a teacher chooses not to use learner corpora, collaboration with researchers who do it could greatly support the development of the field.

II part. Learner corpus of the second Baltic language

The 2nd part of the doctoral thesis elaborates on the creation of the creation of the learner corpus of the second Baltic language. It offers a characterization of the corpus as well as the solutions chosen for it.

There are four chapters. The **1st chapter *Source collection*** goes into obtaining texts as well as protection of personal data and copyrights. The 1st subchapter *Text collection and sampling criteria* tells how teachers have obtained the texts using the method of controlled experiment. It also points out the requirements set for the authors of the texts and the selection of corpus material from all the obtained texts.

The texts included in the corpus are compositions written by students of universities in Latvia and Lithuania. The task was assigned as an independently written homework in the course of the second Baltic language for beginners. All texts that were included in the corpus were written when their authors were studying the second Baltic language for the 1st or 2nd semester with no preliminary knowledge. Each text is marked for the respective semester. According to the content and programs of the courses of the second Baltic language, the 1st semester should lead one to roughly A1 language skill level which should grow into A2 level in the 2nd semester, while the 3rd and 4th semester should allow one to reach B1, perhaps B2 level according to the EU guidelines (EKP 2006). This highly depends on the number of lessons in each semester, though. Since the corpus currently only includes texts written during the 1st and 2nd semester of language learning, the represented language skill level can be assumed to be A level (in a broad sense). The second Baltic language is a foreign language to all of the authors.

As of now, the corpus includes texts from University of Latvia (Latvijas Universitāte; Riga, Latvia), Liepaja University (Liepājas Universitāte; Liepaja, Latvia), Vytautas Magnus University (Vytauto Didžiojo Universitetas; Kaunas, Lithuania) and Vilnius University (Vilniaus Universitetas; Vilnius, Lithuania). Latvian in Lithuania and Lithuanian in Latvia is taught in several more higher education institutions, but texts have not been obtained from them yet. Collaboration with them is possible in the further process of adding new material to the corpus.

The corpus currently consists of 257 texts by 83 authors. The token count is approx. 45 000 in the Lithuanian part of the corpus and approx. 7 000 in the Latvian part of the corpus. Some of the texts were handed in to the teacher digitally, but the ones in handwriting were digitized by typing manually. Texts were independently written by students, and teachers have attempted to prevent plagiarism.

In the 2nd subchapter *The place of corpus "Esam" in the learner corpus classification*, the learner corpus of the second Baltic language is characterized according to the classification offered in the 1st part of the doctoral thesis:

1. Parameters connected with texts and their characteristics.

- The target language is the second Baltic language.
- Production type – written.
- Text type – essay.

- By themes this is a general corpus.
- Originality – the corpus consists of original texts rather than translations.
- By the number of languages corpus "Esam" is bilingual.

2. Parameters connected with the conditions in which the texts were created, their authors and collection.

- Authors' native language – one of the Baltic languages (with possible additional native languages).
- There are two native languages.
- Age of learners – the authors of the texts included in the corpus are all adult students of universities. Their age is mostly similar – around 20, but there are a few students who are older than their peers.
- Language skill level – beginners, no premilinary knowledge.
- Sequence in which languages were learned – the second Baltic language is a foreign language for all authors.
- Writing time – the corpus is relatively synchronic. The texts included in it were written from 2007 till 2014.
- Writing medium is not uniform. The corpus includes handwritten as well as digitally typed texts.

3. Parameters connected with the handling and technical processing of the texts while preparing the corpus for research.

- Fullness of texts – the corpus consists of full texts.
- There is no specialization, this is a general learner corpus.
- By size this is a small corpus – approx. 52 000 tokens.
- Annotation – the corpus is annotated by lemmas, parts-of-speech, syntactic sentence types, and language errors.
- Software – *TEITOK*, based on the CQP/CWB platform.
- Language pair directions are two – it is a bi-directional corpus.

4. Parameters connected with usage of the corpus, the possibilities it offers and the research done so far.

- This is a corpus of indirect pedagogical use.
- The corpus is indended for academic use.
- By accessibility "Esam" is a publicly accessible corpus.

The 3rd subchapter *Protection of personal data and copyrights* shortly describes the regulation of these issues in Republic of Latvia and Republic of Lithuania as far as it matters for the creation of the learner corpus of the second Baltic language. In order to protect personal data, parts of texts have been anonymized by replacing the real personal data with imaginary data in the texts while attempting to maintain the characteristics of each text. The replaced information includes names and surnames of the author and people known to the author; names of towns and settlements, addresses; occupations; names of education institutions and/or their departments. Each item of information is individually evaluated. In order to not

lose too much of the original text, each case is separately considered. All places where replacement was done are marked in order to be able to find them if there are any doubts whether a part of the material is suitable for a certain research. If a specific data in a text cannot be replaced or are especially difficult to replace, they are omitted and the place is marked as such.

For protecting copyrights, all authors of the texts included in the text were asked to sign a permission for making their texts a part of the corpus and to indicate whether they want their names to be included in the authors list. Only the texts with a written author's permission were included in the corpus.

The **2nd chapter** ***Text handling, markup and annotation*** deals with ensuring the computer readability of the texts. The 1st subchapter *Digitizing* characterizes principles according to which the texts received in paper form were digitized: the errors and other characteristics are kept as much as possible, the corrections done by authors themselves are considered, but the corrections done by the teacher are not kept so as not to lose the authenticity of the texts.

The 2nd subchapter *Markup* describes how the text files are supplemented with extralinguistic additional information that does not depend on linguistic analysis, namely, metadata. The corpus has two kinds of markup: structural markup (marking the beginning and end of each text, paragraph, and utterance, as well as anonymized places in the text) and metadata markup (author of the text, university, semester, target language, and title).

The 3rd subchapter *Choice of annotation* discusses criteria that help decide what kind of annotation should be used in the corpus. In order to make the decision, one needs to compare the necessary resources (annotator's knowledge and skills, available technological solutions, time consumption, necessity to create a tagset or availability of a suitable one, etc.) with the gains (the potential amount, size and importance of benefiting research etc.). The learner corpus of the second Baltic language is not created for one or several specific research questions. It aims to aid as diverse analysis of the collected texts as possible, so one of the main criteria when deciding about annotation is its relevance in the existing research context. For that reason the tendencies of learner corpus creation and use in Latvia, Lithuania, and elsewhere are described. Based on these tendencies, it is decided to annotate parts-of-speech, lemmas, sentence types, and errors. Morphosyntactic and problem-oriented annotation is not done due to low popularity and limited usefulness.

The texts contain errors and unclear grammatical constructions, so the possibilities to use automatic annotation tools are limited, and reviewing the results would be time consuming. Because of that, the texts are annotated manually or semi-automatically.

The 4th subchapter *Lemmatizing* discusses the process and format of annotating lemmas in the learner corpus of the second Baltic language. It is decided to use the previous works on lemmatizing in the respective languages as a guide for determining lemmas. The basis are the results offered by morphological annotation tools for Latvian and Lithuanian which are corrected by checking dictionaries as

needed. For more complicated cases, special rules are created, e.g., diminutives are annotated as separate lemmas.

The 5th subchapter *Morphological annotation* describes how parts-of-speech are annotated in the corpus. The tagset created for the project *Semti-Kamols* (LVMPK 2009) is used, and one extra tag *z* is added for punctuation. In some cases, the Latvian linguistic tradition and the Lithuanian linguistic tradition may have different views on the part-of-speech of a specific word (this is especially true for particles), so it is attempted to annotate each language's texts according to that language's linguistic tradition rather than try to merge them.

The 6th subchapter *Syntactic annotation* deals with annotating utterances for sentence types. The texts are divided into utterances and quasi-sentences where utterances are further divided into several basic groups according to the model of sentence they match. The classification follows:

- simple sentence (unextended/extended);
- complex, compound, mixed compound sentence;
- quasi-sentence;
- sentence of unclear structure.

When direct speech is used in a text, it is treated as independent part(-s) of the same utterance.

The 7th subchapter focuses on error annotation. First, various understandings of the notion of error are reviewed. Based on the definitions offered by various researchers, a definition for the case of annotating the learner corpus of the second Baltic language is formulated. An error here is a deviation from the way the respective language structure should be made – namely, from the target hypothesis, as proposed by the corrector of the text.

The process of annotation and role casting in the corrector/annotator pair is then described. Correcting is delegated to a person whose mother tongue matches the target language of the text, while the error types are determined and annotated by a person whose native language does not match the target language but is the same as the author's native language; still, the annotator must have a very good command of the target language.

So far, no suitable tagset has been created for annotating texts in Latvian and Lithuanian written by beginner learners. The corpus of the second Baltic language is annotated according to the classification developed by Sylviane Granger (Granger 2003a) which is adapted to the systems of Baltic languages. Adaptation is based on grammar of Latvian (LVG 2013) and Lithuanian (DLKG 1994) as well as various language pedagogy publications. The adapted classification consists of five error types with subtypes (see Table 1):

Table 1. Error classification

Error type	Tag	Error subtype	Tag	Example in Latvian	Example in Lithuanian
Form	F	Together or separately written words	FK		*širdyje kaž kas suvirpėja* (suvirpa)[8] 'something trembles in the heart'
		Capitalization	FL	*...un Viņai patīk...* 'and she likes'	*Olimpinėje* (Olympinėse) *Žaidynėse* 'in the Olympic Games'
		Diacritics	FD	*Viņas (Viņš) ir uzņemējs* 'he is a businessman'	*dažnai nera pakankamai laiko* 'there is often not enough time'
		Other spelling errors (inclyding typos)	FP	*man patīk tiktis (tikties) ar draugiem* 'I like meeting friends'	*kikvieną dieną* 'every day'
Morphology and word formation	M	Derivation	MA	*patīk futbols, basketbols, vazinātes*[9] *ar ritēni (riteni)* 'likes football, basketball, riding a bike'	*todėl užmiegojome anksti* 'for that reason we fell asleep early'
		Compounding	MS		*aerouostas (oro uostas)* 'airport'
		Case	ML	*Es gribu pastāstīt par mana ģimeni* 'I want to tell about my family'	*didelė dalis drabužiai yra tokios spalvos* 'a great part of clothes are that color'
		Gender	MD	*Mans acis ir brūnas.* 'my eyes are brown'	*Jos visi yra šalia* 'they all are near'

[8] In the examples, where necessary, correction is given in brackets; underlined is the mistake that matches the error subtype.

[9] This word has apparently been made out of two words: Lith. *važinėti* 'to ride around' and Latv. *vizināties* 'to go for a ride'.

		Number	MN	*es biju ļoti skumīga šoreiz pār (par) atvaļinājumiem* ‘I was very sad this time about the vacation’	*įvariuose gyvenimo valandą* ‘in various hours of life’
		(In)definite ending	MG	*Fotoaparātā bija manas … skaistas fotogrāfijas*	*žmonių kamšatis ir ilgoji (ilgos) valandos viešajame transporte* ‘crowd of people and long hours in public transportation’
Morphology and word formation	M	Degree of comparison	MQ		*Aš esu jaunesnioji (jaunausia).* ‘I am the youngest’
		Person	MP	*Tēvs interesējies par automo-biliem (automo-biļiem)* ‘father is interested in cars’	*aš nebuvo name* ‘I was not home’
		Tense	MT	*viņai patīk ceļot(,) un māte apmeklēja (ir apmeklējusi) Krieviju, Franciju...* ‘she likes traveling, and mother has visited Russia, France…’	*aš pasibundu (pasibudau), nes buvau labai alkana* ‘I woke up because I was very hungry’
		Voice	MI		*Aš esu dėkinguma (dėkinga), ka (kad) sutikčiau (sutikau) jai (ją)* ‘I am grateful I met her’
		Euphony	MK		*Į ją galėtų įeiti ir iš lauko* ‘in it one could go from outside’

Error type	Tag	Error subtype	Tag	Example in Latvian	Example in Lithuanian
		Reflexivity	MR		*netrukdėme ir neriejomės* ‘we did not disturb or quarrel’
		Participle confusion	MV		*vairuotojas, matytint, kad bėgtu (bėgu), (..) pristabdau (pristabdo)* ‘driver stops as he sees me running’
Morphology and word formation	M	Perfective	MB		*Kada ji ėjo (atėjo) iš darbo...* ‘when she came from work’
		Iterativity	MX		*mama man (mane) išmokydavo nekada nepasiduoti* ‘mom taught me to never give up’
Syntax	S	Word order	SV	*..radoša tik (*tikai*) dēļ naudas (*naudas dēļ*)* ‘creative only for money’	*Vieta, kur visada aš galiu grįžti yra...* ‘place where I can always return’
		Word missing	SI	*tā vārds (*vārds ir*) Džekis* ‘its name (is) Džekis’	*Biologijos fakultete yra labai daug (daug ko?[10]) ‘in the Faculty of Biology there’s a lot (of what?)’*
		Word redundant	SL	*..ceļiauju (ceļoju) uz Klaipēdu būt brivdienās (brīvdienās)* ‘I travel to Klaipeda to be for holidays’	*ji pasiūlė man kartu su ja reikėjo ruošti pjesę* ‘she offered me together with her needed to prepare a play’

[10] In this example, there are no contextual ties.

		Cohesion	SS	*Mans mātes vārds ir...* ‘my name of mother is…’	*aš nešioju kepures, irgi pirštines* ‘I wear hats, also gloves’
Lexis	L	Meaning	LN	*pelēks, biezs (resns) un ļoti labs kaķis Benas* ‘grey, thick (fat) and very good cat Benas’	*Ne tik katris* (kiekvienas) *latvis* ‘not only every Latvian’
		Matching	LV	*zils paklājs, kurš der (piestāv) pie sienu (sienām)* ‘blue carpet that fits (matches) the walls’	*nes esame tiek įvairios* ‘because we are so much different’
		Prefab	LS	*..braukšu uz ciemus (ciemos)* ‘I will go to visit’	*Aš tą* (tai) *labai įvertinu* ‘I evaluate (value) that very much’
Punctuation	I	Punctuation confusion	IN		*dar kartą užmigau..* ‘I fell asleep again’
		Punctuation redundant	IL	*Tāpēc, es biju ļoti skumīga* ‘so I was very sad’	*Trečią valandą naktį* (nakties)*, aš...* ‘at three a.m. I…’
		Punctuation missing	IT	*Viņai patīk ceļot(,) un māte apmeklēja...* ‘she likes traveling, and mother visited…’	*Viskas būtu(,) kaip aš norėčiau.* ‘everything would be as I want’

Each error is assigned only one target hypothesis that seems most likely for the corrector, and only one error code that seems most likely for the annotator.

Since the corpus is digital, its performance depends on the software used. The **3rd chapter *Software*** describes what programs are used creating the corpus and what programs are expected to be used by the users of the corpus. The 1st subchapter *Software used in corpus creation* lists text editors that were used for digitizing texts, and describes the generation of text codes and author codes using a random number generator. It also explains how the texts were prepared for using as the sample corpus with the *AntConc* concordancer. Further, it elaborates on the process of annotating the final version of the corpus using the interface of *TEITOK* software developed by

Maarten Janssen. The technical description showing the markup and annotation tags is provided. The files of the full version of the corpus match the *TEI* standard.

The 2nd subchapter *Software necessary for using the corpus* deals with the software that is being utilized for working with the ready corpus. Since the sample corpus is available for downloading as separate files, it can be used with any program according to the user's choice, but the recommended corpus research software is *AntConc*, and the sample corpus's compatibility with this program has been tested. This program is not used in the final version of the corpus because it must be downloaded to every computer that needs to access the corpus. Besides, depending on the safety settings on the specific computer it may not be possible to launch *AntConc*. For that reason, the final version of the corpus is available online, and it doesn't have to be downloaded. Accessibility of the corpus is ensured by creating a special website which is run by a private *Linux Ubuntu* server. The server is located in Riga and is physically accessible only for the creator of the corpus.

The *TEITOK* user interface is visible in the user's internet browser, on the website *www.esamkorpuss.lv*. The paper offers a short description of using it. Since *TEITOK* is based on *CQP*, it supports regular expressions.

The **4th chapter** ***Researching the learner corpus of the second Baltic language*** provides a short outline on using the corpus in the context of specific methods, along with examples. The 1st subchapter *Contrastive interlanguage analysis* outlines the possibilities to compare the Latvian subcorpus with the Lithuanian subcorpus in order to reveal any possible differences between acquisition of Latvian as the second Baltic language and acquisition of Lithuanian as the second Baltic language. The 2nd subchapter *Computerized error analysis* deals with the possibilities of analyzing errors in the learner corpus of the second Baltic language.

The 3rd subchapter *Baltic interlanguage* discusses the possibilities to use the bi-directional material as a whole, without specifying a subcorpus. This allows for analyzing features that are common for the whole Baltic interlanguage. It can be useful in two aspects:

- for investigating how a beginner learner writes texts in a language which he/she knows is similar to a language he/she already knows well;
- for revealing and investigating similarities and differences between both Baltic languages.

Researching Baltic interlanguage is not a separate method, rather an aspect in which the corpus data can be viewed while using any other methods. Similarly, the notions of e.g., Germanic interlanguage, Slavic interlanguage etc. could be introduced in order to describe the interlanguage that forms when a learner is learning a new language that is closely related to his/her native language or, possibly, a language which the specific learner knows in a rather high level and which is being used as the medium in the language learning process.

CONCLUSION

The goal of the research – to create a learner corpus of the second Baltic language and to describe the creation process – has been reached by carrying out the predefined tasks.

The first part of the doctoral thesis describes the notion of a learner corpus, types of such corpora, and the information provided in academic literature on the creation and use of such corpora. Although learner corpora are just one subtype of corpora, they have been gaining popularity rapidly in the later years. However, the creation and research of such corpora is still fragmented. This can also be observed in Latvia and Lithuania – in both countries, work on learner corpora is mainly done individually without any collaboration between neighbouring countries.

Various well-known corpus linguistics procedures can be used in the work with learner corpora as long as they are chosen accordingly to the method of research. There are two especially popular methods in learner corpus research: contrastive interlanguage analysis and computerized error analysis. Both methods can be used in the learner corpus of the second Baltic language.

In the process of creating a learner corpus of the second Baltic language, two teachers from Latvian universities and two teachers of Lithuanian universities have helped to collect 1st and 2nd semester students' learner texts in their second Baltic language. The technical support has been ensured to offer online access to the corpus for everyone who agrees to the terms and conditions. These parts of the work have been described in the 2nd part's 1st and 3rd chapter of the doctoral thesis.

The newly created corpus is small (slightly more than 50 000 tokens), but it is considered representative of the language variety it represents because the overall number of second Baltic language learners is also not high. In order to increase the size of the corpus in the future, the teachers and students from other higher education institutions where the second Baltic language is being learned should be involved in text collection.

The size of the corpus is also affected by issues of copyrights and personal data protection. They make the greatest challenge if texts are included in the corpus when longer time has passed since the texts were written: one needs to find a way to contact the author and obtain the permission. The contact information of the author is often not easily found, or authors do not trust a person they do not know who contacts them via a social network. If permission is asked by the teacher immediately after the text was written, the process is simpler and more convenient for both sides. It is also possible to note that the text is going to be included in the corpus and inform the authors about the conditions before the text is written – when assigning the task. However, in such a case, one could wonder if the creation of the text (choice of words, constructions etc.) could be affected by realization that the text is going to be publicly accessible.

The corpus was annotated, partly by using and adapting existing tagsets that have been used in other corpora of Baltic languages. An already existing tagset was used for morphological annotation, while the tagset for syntactic annotation has been

adapted based on the current understanding about sentences and their types. This was done to be able to successfully annotate quasi-sentences, parcelates, direct speech. The error annotation tagset was created based on the existing error classifications created for other languages. Annotation process is described in the 2nd part's 2nd chapter of the doctoral thesis.

The notion of error is not unequivocal. During the annotation of the corpus, the corrector of texts proposes a target hypothesis in each specific case, and disagreements with the target hypothesis are considered errors. The classification created for error annotation is made as effective and comprehensive as possible so that it could be used not only in this corpus, but also in various other corpora of Baltic language learners.

The annotation process of beginner learner texts can be challenging due to the errors – in many occasions, it is difficult to understand what the author has intended to say. In order to not make annotation too complicated, the seemingly most likely version is chosen in each case, but it can be controversial and, should it be decided that the initial choice is not suitable, it can be corrected later.

Program *Teitok* was chosen for the technical support of the corpus. It has been especially developed for small annotated corpora. One of the main factors in choosing the software and creating the interface was user convenience, trying to ensure that high level computer user skills or any specific knowledge is not required for using the corpus.

The 2nd part's 4th chapter of the doctoral thesis tells about the most essential usage possibilities of the corpus. It is not a whole limiting description because the corpus may be used for academic purposes with any methods. The rather small size, however, means that one can not always expect to have enough data for any kind of research question to reach reliable conclusions. The decision whether the corpus is sufficient for a research question has to be made in each case separately.

Discussing the use of the corpus, one must bear in mind that it is primarily intended as research material rather than a pedagogical tool, and the direct pedagogical use of the corpus may not seem suitable or necessary for some teachers of the second Baltic language. However, the results of researching this material can give additional insight which in turn could help improve study materials and therefore indirectly affect the language learning process. Besides, the teachers can make a great contribution to the research by continuing to collect texts for including in the corpus. Thus, mutual collaboration between researchers and teachers is highly valued, as both sides benefit from it.

While creating the corpus, various ideas expressed by other researchers building learner corpora have been taken into the account, but some novelties have also been introduced. The doctoral thesis offers two new notions which would be worth especially noting. One of them is *bi-directional learner corpus*. Direction as a parameter of describing a corpus is already well-known in parallel (translation) corpus research. Since not only the target language but also the learner's native language affect the learning process considerably, the opposition of the native language and the target language is considered a direction in this research.

Accordingly, a bi-directional learner corpus is a corpus where the represented target languages are also represented as native languages, and vice versa.

From this emerges the other offered new notion – *Baltic interlanguage*. Work on a bi-directional corpus leads to an etiological question – what chould its material be called? Therefore, it is offered to use the term *Baltic interlanguage* when talking about the interlanguage that forms when a speaker of one Baltic language is learning the other Baltic language (Lithuanian for a Latvian, and Latvian for a Lithuanian). Since both languages are closely related, this interlanguage is likely to have different qualities from the qualities of the interlanguage that forms when a Baltic language is learned by someone whose native language belongs to a different language group.

In the continuing research, primarily the characteristics of target language production in the second Baltic language acquisition should be investigated. This is the aim for which the corpus was created. As the more (or less) challenging language elements for learners are discovered, changes and improvements can be made in existing or new study materials, dictionaries, learning/teaching methods etc. It would also be valuable to investigate possible direct use of the corpus in learning and teaching languages.

Finally, it should be emphasized that the corpus is only useful while the academic environment is interested in the second Baltic language and its acquisition process. The current tendencies in choosing to study the second Baltic language in the higher education institutions are not too optimistic – several institutions have been dropping it ever so often. Hopefully, the creation of the learner corpus of the second Baltic language could facilitate the return of popularity of the second Baltic language among students and researchers alike.

THESES FOR DEFENCE

1. Learner corpora are gaining popularity in many countries. However, in Latvia and Lithuania the use of learner corpora is not widespread and is rather fragmented. The corpora created here are mainly only accessible to their creators, or elsewhere created learner corpora are used. It would be advisable to collaborate more in creating and researching learner corpora of various languages with Latvian or Lithuanian as the native language, especially when it comes to comparable corpora. Considering the relationship and similarity between the Baltic languages, similarities or differences in the interlanguage data can help understand the impact each Baltic language has on acquisition of another language.
2. In this research, a publicly accessible online learner corpus of the second Baltic language was created and its use in research was described. It is the first publicly accessible learner corpus in Latvia and Lithuania, the first publicly accessible learner corpus of Baltic languages, and the first bi-directional learner corpus. The notion of a bi-directional corpus is introduced in this paper meaning a corpus that includes texts in two language pairs where each of the languages is a native language in one of the pairs and target language in the other pair.
3. The corpus was annotated in four levels: syntactic, morphological, lexical, and error annotation was done. Pre-existing classifications were used for syntactic, morphological, and lexical annotation. New classification was developed for error annotation, adapting S. Granger's classification of errors developed for annotating learner texts in French. Error here is a deviation from the target hypothesis proposed by a corrector whose native language matches the target language of the text. Errors are also annotated in aspects that are not learned or are learned partly in the respective level (e.g., in A level – punctuation, perfect tenses of verbs etc.).
4. When speaking of the interlanguage that forms when a speaker of one Baltic language learnes the other Baltic language, it is offered to use the term *Baltic interlanguage*; it can be researched using a bi-directional corpus. If research reveals any features that considerably differ from the interlanguage that forms when a Baltic language is learnt by someone who does not speak any other Baltic language, it could help characterize Baltic languages in comparison with other languages.
5. When continuing this research, it would be valuable to add new texts to the corpus, including texts that represent other language levels, to help investigate the development of language skills. It would also be desirable to create a comparable corpus consisting of matching level texts written by authors who do not speak another Baltic language. The further research could be expected in two directions:

a. researching the learner corpus of the second Baltic language – by using this corpus, features of the Baltic interlanguage could be revealed; The results of such research could then be used for improving study materials of the second Baltic language, dictionaries of Baltic languages etc.;

b. adding more data to the corpus (e.g., texts that reflect higher skill level, or speech samples) or creation of new comparable corpora (e.g., a corpus of Baltic languages where informants' native languages are not Baltic languages).

APPROBATION OF DOCTORAL THESIS

The results of different aspects of the research have been discussed in 23 conference talks in Latvia (Liepāja, Ventspils, Rīga, Daugavpils, Jelgava), Lithuania (Vilnius, Kaunas), Estonia (Tartu), the United Kingdom (Sheffield, Lancaster), and Sweden (Gothenburg):

1. Inga Znotiņa. "Valodas apguvēju korpuss: lietuviešu un latviešu termins un definīcija." 18th International Scientific Conference *Vārds un tā pētīšanas aspekti*. Liepaja, 28–29 November 2013.
2. Inga Znotiņa. "Learner corpora in Latvia and Lithuania." 8th International Scientific Conference *Native Language and Other Languages*. Tartu, 28–29 November 2013.
3. Inga Znotiņa. "Learner corpus research methods and requirements for the corpora used." International Scientific Conference organized by Tartu University and Tallinn University *Mapping Methods: Approaches to Language Studies*. Tartu, 8–10 May 2014.
4. Inga Znotiņa. "Valodas apguvēju korpusi Latvijā un Lietuvā." Scientific Workshop for Doctoral Students of Linguistics in University of Latvia, Liepaja University and Ventspils University College. Riga, 16–17 May 2014.
5. Inga Znotiņa. "Pētniecības iespējas nemarķētā baltu valodu apguvēju korpusā." 19th International Scientific Conference *Vārds un tā pētīšanas aspekti*. Liepaja, 27–28 November 2014.
6. Inga Znotiņa. "Error-tagging a learner corpus of Baltic languages." Sheffield University's Annual Postgraduate Conference in Linguistics *ShefLing PGC*. Sheffield, 5–6 March 2015.
7. Inga Znotiņa. "Valodas apguvēju korpusa anotēšanas veidi." 3rd International Scientific Conference *Via scientiarum*. Ventspils, 12–13 March 2015.
8. Inga Znotiņa. "Semantiski tuvu leksēmu apguves izpēte neanotētā valodas apguvēju korpusā." Annual Scientific Conference of Riga Stradiņš University. Riga, 26–27 March 2015.
9. Inga Znotiņa. "Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība." 20th International Scientific Conference *Vārds un tā pētīšanas aspekti*. Liepaja, 3–4 December 2015.
10. Inga Znotiņa, Daiva Puškorjute-Riduliene. "Kļūdu anotēšana otrās baltu valodas apguvēju korpusā." 20th International Scientific Conference *Vārds un tā pētīšanas aspekti*. Liepaja, 3–4 December 2015.
11. Inga Znotiņa. "Besimokančiųjų tekstynų anotavimas Lietuvoje ir Latvijoje." 3rd International Scientific Conference *Sustainable Multilingualism: Language, Culture and Society*. Kaunas, 29–30 May 2015.
12. Inga Znotiņa. "Learner corpus *Esam*: a new corpus for researching Baltic interlanguage." 8th International Scientific Conference *Corpus Linguistics*. Lancaster, 21–24 July 2015.

13. Inga Znotiņa. "Lemmatization in a beginner learner corpus." 3rd International Scientific Students' Conference *Bridges in the Baltics*. Vilnius, 2–3 October 2015.
14. Inga Znotiņa. "Kļūdu klasifikācija otrās baltu valodas apguvēju tekstos." XII International Congress of Baltistics. Vilnius, 28–31 October 2015.
15. Inga Znotiņa, Daiva Puškorjute-Riduliene. "*Mans* un *savs* baltu valodās: vietniekvārdu kļūdu klasifikācijas problēmas". 26th International Scientific Conference *Scientific Readings*. Daugavpils, 28–29 January 2016.
16. Inga Znotiņa. "Valodas apguvēju korpusu izmantojums svešvalodas mācību procesā." Annual Scientific Conference of Riga Stradiņš University. Riga, 17–18 March 2016.
17. Inga Znotiņa. "Publiski pieejama valodas apguvēju korpusa izveide: programmatūras meklējumi." Scientific Workshop for Doctoral Students of Linguistics in University of Latvia, Liepaja University and Ventspils University College. Riga, 20 May 2016.
18. Inga Znotiņa. "Vārdšķiru un pamatformu noteikšana otrās baltu valodas korpusā: problemātiskie gadījumi latviešu valodā." 27th International Scientific Conference *Scientific Readings*. Daugavpils, 26–27 January 2017.
19. Inga Znotiņa. "Otrās baltu valodas korpusa sintaktiska anotēšana." Annual Scientific Conference of Riga Stradiņš University. Riga, 6–7 April 2017.
20. Inga Znotiņa, Inga Laizāne. "Otrās baltu valodas apguve Latvijas un Lietuvas augstākās izglītības iestādēs." 4th International Scientific Conference *Via scientiarum*. Liepaja, 7–8 April 2017.
21. Inga Znotiņa, Inga Laizāne. "Digitālie resursi baltu valodu apguvē." 14th International Scientific Conference *Language Acquisition: Problems and Perspective*. Liepaja, 21 April 2017.
22. Inga Znotiņa. "Computer-aided error analysis for researching Baltic interlanguage." 11th International Scientific Conference *Rural Environment, Education, Personality 2018*. Jelgava, 11–12 May 2017.
23. Inga Znotiņa. „Learner corpus of the second Baltic language: annotation and data comparability." Workshop on interoperability of Second Language Resources and Tools. Gothenburg, 6–8 December 2017.

Certain parts and conclusions of the research have been published in 6 articles and 11 collections of conference theses and abstracts.

Articles:

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.

Theses and abstracts:

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. 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. Available: 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 KOPSAVILKUMĀ IZMANTOTĀ BIBLIOGRĀFIJA

ADTAI – *Asmens duomenų teisinės apsaugos įstatymas* [skatīts 2014. gada 5. oktobrī]. Pieejams: *http://www3.lrs.lt/pls/inter3/dokpaieska.showdoc_l?p_id=400103*

Akadterm$_{-e}$ – Akadēmiskā terminu datubāze AkadTerm [skatīts 2016. gada 19. jūnijā]. Pieejams: *http://termini.lza.lv/term.php*

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*

Baker u. c. 2006 – Baker, Paul, Hardie, Andrew, McEnery, Tony. A Glossary of Corpus Linguistics. Edinburgh : Edinburgh University Press, 2006.

Barlow 2005 – Barlow, Michael. Computer-based analyses of learner language. *Analysing Learner Language*. Rod Ellis, Gary Barkhuizen. Oxford : Oxford University Press, 2005.

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.

Biber u. c. 2006 – Biber, Douglas, Conrad, Susan, Reppen, Randi. *Corpus Linguistics.* Investigating Language Structure and Use. Cambridge : Cambridge University Press, 2006.

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.

Bussmann 1996 – Bussmann, Hadumond. *Routledge Dictionary of Language and Linguistics*. London, New York : Routledge, 1996.

Butkus 2008 – Butkus, Alvydas. *Baltiškos impresijos.* Kaunas : Aesti, 2008.

Cigankova, Vinčela 2012 – Cigankova, Nataļja, Vinčela, Zigrīda. Specializēd 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.

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.

Crystal 1992 – Crystal, David. *An Encyclopedic Dictionary of Language and Languages*. Oxford, Cambridge, Mass. : Blackwell, 1992.

Crystal 2008 – Crystal, David. *A Dictionary of Linguistics and Phonetics*. 6th edition. Malden, MA, Oxford : Blackwell Pub., 2008.

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.

Dagneaux u. c. 1998 – Dagneaux, Estelle, Denness, Sharon, Granger, Sylviane. Computer-aided error analysis. *System*, Vol. 26, No. 2, 1998, pp. 163–174.

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.

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.

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.

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*

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*

EIC$_{-e}$ – *What is Estonian Interlanguage Corpus (EIC)?* [skatīts 2014. gada 2. septembrī]. Pieejams: *http://evkk.tlu.ee/wwwdata/what_is_evk?language=en*

EKP 2006 – **Eiropas Padome. Valodas politikas nodaļa.** *Eiropas kopīgās pamatnostādnes valodu apguvei: mācīšanās, mācīšana, vērtēšana*. Rīga : Madonas poligrāfists, 2006.

ELE 2008 – *Encyclopedia of Language and Education*. Nancy H. Hornberger (ed.). New York : Springer, 2008

FPDAL – *Fizisko personu datu aizsardzības likums* [skatīts 2014. gada 5. oktobrī]. Pieejams: *http://likumi.lv/doc.php?id=4042*

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.

Granger 1993 – Granger, Sylviane. The International Corpus of Learner English. *The European English Messenger*, Vol. 2(1), 1993, p. 34.

Granger 1994 – Granger, Sylviane. The learner corpus: a revolution in applied linguistics. *English Today*, Vol. 39, No. 10/3, 1994, pp. 25–29.

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.

Granger 1998 – Granger, Sylviane. *Learner English on Computer*. Sylviane Granger (ed.). London, New York : Longman, 1998.

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.

Granger 2003a – Granger, Sylviane. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, Vol. 20, No. 3, 2003, pp. 465–480.

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.

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.*

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.

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.

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.

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.

Granger 2013 – Granger, Sylviane. *Learner English on Computer*. Sylviane Granger (ed.). London, New York : Routledge, 2013.

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.

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.

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.

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.

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*

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.

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.

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.

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*

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.

Jantunen 2011 – Jantunen, Jarmo Harri. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivőrdlusi. Lähivertailuja,* No. 21, 2011, pp. 86–105.

Johansson 2007 – Johansson, Stig. *Seeing through Multilingual Corpora*. On the use of corpora in contrastive studies. Amsterdam, Philadelphia : John Benjamins, 2007.

Juknevičienė 2013 – 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.

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*

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.

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.

Laizāne 2014c – Laizāne Inga. Lietvārda locījumu nozīmes latviešu valodas kā svešvalodas apguvē. *Zinātniski metodisks izdevums „Tagad",* LVAVA, 2014, 25.–31. lpp.

LDLTAL 2010 – Richards, Jack C., Schmidt, Richard. *Longman Dictionary of Language Teaching and Applied Linguistics.* 4th ed. Harlow : Longman, 2010.

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.

Levāne-Petrova 2012 – 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.

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*

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.

LTSV 2011 – Skujiņa, Valentīna, Anspoka, Zenta, Kalnbērziņa, Vita, Šalme, Arvils. *Lingvodidaktikas terminu skaidrojošā vārdnīca*. Rīga : Latviešu valodas aģentūra, Latviešu valodas institūts, 2011.

LTŽ 2012 – Ramonienė, Meilutė, Brazauskienė, Jelena, Burneikaitė, Nida, Daugmaudytė, Jurga, Kontutytė, Eglė, Pribušauskaitė, Joana. *Lingvodidaktikos terminų žodynas*. Vilnius : Vilniaus universiteto leidykla, 2012.

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*

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*

Marcinkevičienė 2000 – Marcinkevičienė, Rūta. Tekstynų lingvistika: teorija ir praktika. *Darbai ir dienos*, Nr.. 24, 2000, 7–64 psl.

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.

Myles 2005 – Myles, Florence. Interlanguage corpora and SLA research. *Second Language Research*, Vol. 21, No. 4, pp. 373–391.

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.

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.

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.

Rimkutė 2006 – Rimkutė, Erika. *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne.* Daktaro disertacija. Kaunas : Vytauto Didžiojo universitetas, 2006.

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.

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.

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.

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.

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

Savickienė 2006 – Savickienė, Ineta. Linksnio kategorijos įsisavinimas: lietuvių kalba kaip gimtoji ir svetimoji. *Kalbotyra*, Vol. 56, No. 3, 2006, 122–129 psl.

Sīlis 2009 – Sīlis, Jānis. *Tulkojumzinātnes jautājumi. Teorija un prakse*. Ventspils : Ventspils Augstskola, 2009.

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.

Skadiņa, Vasiļjevs 2013 – Skadiņa, Inguna, Vasiļjevs, Andrejs. Valodas tehnoloģijas. *Latviešu valoda.* Andrejs Veisbergs (red.). Rīga : Latvijas Universitāte, 2014, 453.–475. lpp.

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.

Šalme, Auziņa 2013 – Šalme, Arvils, Auziņa, Ilze. Latviešu valodas prasmes līmeņi: Pamatlīmenis A1, A2. Rīga : Latviešu valodas aģentūra, 2013.

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.

Vinčela 2010 – Vinčela, Zigrīda. *Student-Composed Electronic Discourse as a Result of Applied Linguistic Research.* Promocijas darbs. Rīga : Latvijas Universitāte, 2010.

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.

VPSV 2007 – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Red. V. Skujiņa. Rīga : LU Latviešu valodas institūts, 2007.

Zauberga 2001 – Zauberga, Ieva. Discourse interference in translation. *Across Languages and Cultures* 2 (2), 2001, pp. 265–276.

Zinkevičius 2000 – Zinkevičius, Vytautas. *Lemuoklis* – morfologinei analīzei. *Darbai ir dienos,* 24 (2000), 245–274 psl.

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.

Znotiņa 2012 – Znotiņa, Inga. *Parodomieji įvardžiai lietuvių–latvių lygiagrečiajame tekstyne.* Magistro darbas. Kaunas : Vytauto Didžiojo universitetas, 2012.

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.

Камшилова 2013 – Камшилова, Ольга Николаевна. 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.