Inga KAIJA

Jaunu burtu veidošana ar diakritiskajām zīmēm latviešu valodas kā svešvalodas apguvēju tekstos

Anotācija
Veidojot latviešu valodas apguvēju korpusu LaVA, tajā tiek iekļauti iesācēju latviešu valodas kā svešvalodas apguvēju rakstīti teksti, kuru viena no īpatnībām ir arī nereta neatbilstība pareizrakstības normām. Šāda neatbilstība ir arī jaunu burtu veidošana ar diakritiskajām zīmēm, proti, tādu diakritisko zīmju pievienošana burtiem, kādas šiem burtiem latviešu literārās valodas alfabētā netiek lietotas. Tā kā dažādas šādas kombinācijas datorrakstā nav ierakstāmas, korpusa datus digitalizējot, šādas neatbilstības tajā ne vienmēr tiek iekļautas un nebūs automātiski analizējamas.
Pētījumā ekscerpēti šādi piemēri no pirmajiem korpusā iekļautajiem tekstiem un sniegti primārie kvantitatīvie dati, izvirzot hipotēzes par lietojuma tendencēm, kā arī piedāvājot tālāko pētījumu virzienus pilnīgāka priekšstata gūšanai.

Atslēgvārdi: valodas apguvēju korpuss, diakritiskās zīmes, burti, latviešu valoda, svešvaloda, pieaugušie valodas apguvēji

Ievads

Valodas apguves pētījumos nu jau stabilu vietu ir ieņēmuši valodas apguvēju korpusi, kuru popularitāte un izplatība turpina strauji pieaugt (McEnery, Brezina u. c. 2019, 78). Tie dažādu valodu apguves pētījumos ir izmantoti arī Latvijā (vairāk sk. Znotiņa 2015, Znotiņa 2018), un pašlaik Latvijas Universitātes Matemātikas un informātikas institūtā Latvijas Zinātnes padomes finansētā pētījuma projektā "Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums" (projekta Nr. lzp-2018/1-0527) tiek veidots jauns latviešu valodas apguvēju korpuss LaVA. Tajā tiek apkopoti teksti, kurus rakstījuši ārvalsti studenti, kas Latvijas augstākās izglītības iestādēs apgūst latviešu valodu kā svešvalodu. Tie ir gan apmaiņas studenti, gan pilna laika studijās Latvijā studējošie, un visu korpusā iekļauto tekstu autori ir iesācēji, kas latviešu valodu mācās pirmo vai otro semestri. Līdz ar to, lai arī studentu apmeklētie kursi ne vienmēr paredz noteikta līmeņa sasniegšanu, teksti aptuveni atbilst A1–A2 līmenim saskaņā ar Eiropas kopīgajām pamatnostādnēm valodu apguvei (EKP 2006). Tekstus rakstot, to autoriem ir atļauts izmantot palīglīdzekļus pēc saviem ieskatiem, izņemot automātiskās tulkošanas rīkus. Taču A1 un A2 līmenī no valodas apguvēja vēl nav sagaidāma prasme vārdus no sava vārdu krājuma uzrakstīt pilnībā atbilstoši pareizrakstības normām (Šalme, Auziņa 2016, 216). To atspoguļo arī korpusa teksti, un, anotējot tajos kļūdas, uzmanība tiek pievērsta arī pareizrakstības kļūdām, piemēram, diakritisko zīmju neatbilstošam lietojumam vai īpašvārdu atveides jautājumiem (Auziņa, Kaija, Levāne-Petrova 2019).

Korpusa izveides metodoloģija paredz, ka teksti tiek rakstīti ar roku, tādējādi neprasot latviešu datorraksta apguvi, kas svešvalodā var sagādāt

grūtības (Lally 2000, 900). Pēc tam projekta dalībnieki tekstus digitalizē – pārraksta datorrakstā, pēc iespējas ciešāk pieturoties pie oriģinālteksta un saglabājot tā īpatnības. Tomēr ne visu ir iespējams digitalizēt pilnīgi precīzi: dažkārt tiek veidoti jauni burti ar diakritiskajām zīmēm, kādi latviešu valodas literārajā alfabētā (un datorraksta simbolu klāstā) nav atrodami (sk. 1. attēlā). Šādi gadījumi datorrakstā netiek atveidoti, tāpēc tie nebūs pieejami datorizētai kļūdu analīzei, kad korpuss būs izveidots.

1. attēls. Datorrakstā neatveidojamu burtu piemēri.

Viens no aspektiem, kam ir īpaša nozīme rakstīšanā un tās apguvē, ir tieši atbilstošu zīmju lietošana (LTSV 2011, 70), t. sk. "ievērojot noteiktus zīmju atveides likumus, raksta tehnikas paņēmienus" (Šalme 2011, 16), tāpēc neatbilstošu diakritisko zīmju lietojuma izpēte palīdz izprast latviešu starpvalodas rakstības sākotnējās īpatnības.

Šī pētījuma **mērķis** ir atklāt, kādi burti tiek veidoti ar diakritiskajām zīmēm latviešu valodas kā svešvalodas apguvēju tekstos un izvirzīt hipotēzes par šāda lietojuma motivāciju. Lai šo mērķi sasniegtu, izvirzīti šādi **uzdevumi**:

- definēt ar diakritiskajām zīmēm veidotus jaunus burtus;
- ekscerpēt piemērus;
- iegūt primāros kvantitatīvos datus;
- izvirzīt hipotēzes par iespējamām lietojuma tendencēm;
- piedāvāt tālāko pētījumu virzienus.

Metode

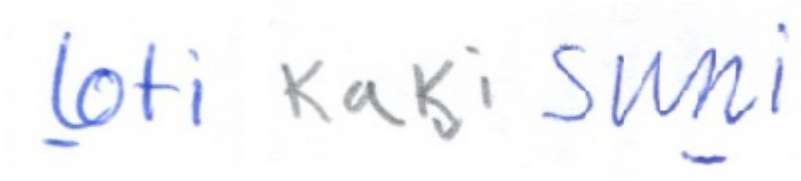
Ar diakritiskajām zīmēm veidoti jauni burti ir tādi burti, kuriem pievienotās diakritiskās zīmes neveido latviešu literārās valodas alfabētā sastopamu burtu. Tādi var būt:

- burti ar latviešu literārajā valodā nelietojamu diakritisko zīmi (piem., jumtiņu, sk. 2. attēlā);

2. attēls. Burtu piemēri ar latviešu literārajā valodā nelietojamām diakritiskajām zīmēm.

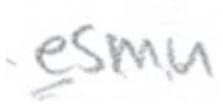
- burti, kuriem pievienota latviešu literārajā valodā citiem (bet ne šim) burtiem lietojama diakritiskā zīme (piemēram, garumzīme līdzskanim) – tā var būt novietota ierastā (piem., garumzīme virs burta, sk. 3. attēlā) vai neierastā pozīcijā[1] (piem., karons zem burta, sk. 4. attēlā);

3. attēls. Burtu piemēri ar latviešu literārajā valodā citiem burtiem lietojamām diakritiskajām zīmēm ierastā pozīcijā.



4. attēls. Burtu piemēri ar latviešu literārajā valodā citiem burtiem lietojamām diakritiskajām zīmēm neierastā pozīcijā.

- burti, kuros diakritiskā zīme ir novietota nepareizā pozīcijā (līdz šim atrasts viens šāds piemērs – garumzīme zem burta, sk. 5. attēlā).



5. attēls. Diakritiskās zīmes nepareiza novietojuma piemērs.

Šajā kopā neietilpst jebkādi citi latviešu literārās valodas alfabētā nelietoti burti (piem., *x* vai *y*), kā arī rakstzīmes no citiem, ne latīņu, alfabētiem (piem., kirilicas). Tāpat netiek ieskaitīti gadījumi, kuros teksta autors, rakstīdams garo *ī*, to ir uzrakstījis gan ar punktu, gan garumzīmi, jo tā nav jauna burta veidošana, bet gan esoša burta pareizrakstības problēma. Netiek ņemti vērā arī neatveidoti citvalodu īpašvārdi, jo šādos gadījumos autora mērķis nav bijis izmantot latviešu valodas, bet gan oriģinālvalodas rakstzīmes.

[1] Runājot par diakritisko zīmju pozīciju pie burtiem, kuriem šāda diakritiskā zīme latviešu literārajā valodā nav lietojama, ar ierastu pozīciju saprotama garumzīmes un karona (apgrieztā jumtiņa) rakstība virs attiecīgā burta, kā arī komata (mīkstinājuma zīmes) rakstība zem attiecīgā burta, kā tas tiek darīts arī latviešu literārajā valodā lietotos burtos. Ar neierastu pozīciju saprotams jebkāds cits novietojums.

Tā kā korpusā LaVA nav iespējams automātiski atlasīt visus gadījumus, kuros autors ir lietojis tādu diakritisko zīmi, kas latviešu valodā ar attiecīgo burtu netiek lietota, piemēri tiek ekscerpēti manuāli, lasot tekstus, un katrs atbilstošs gadījums tiek reģistrēts izklājlapā, fiksējot teksta identifikācijas numuru korpusā, datnes nosaukumu, vārdu, kurā jaunais burts konstatēts, latīņu alfabēta burtu, kuram pievienotā diakritiskā zīme veido jaunu burtu, diakritiskās zīmes veidu un novietojumu (pozīciju), kā arī attiecīgā teksta autora latviešu valodas apguves secīgo semestri (1. vai 2.). Pēc šiem datiem, tos skatot kvantitatīvi, izdarītie secinājumi izmantoti, izvirzot hipotēzes par šāda lietojuma cēloņiem. Atsevišķā datnē saglabāts arī vārda attēls ar kodu, pēc kura to var atrast minētajā izklājlapā.

Dažkārt diakritiskā zīme nav īsti skaidra, piemēram, slīpa svītra var būt iecerēta kā garumzīme, karons, akūts vai gravis (atkarībā no slīpuma virziena). Tāpat daļai tekstu autoru rokraksta īpatnību dēļ punkts uz i mēdz drīzāk atgādināt garumzīmi. Šādos un citos neskaidros gadījumos lēmums par to, kāda garumzīme izmantota, pieņemts, vērojot attiecīgās diakritiskās zīmes rakstību citos tā paša teksta vārdos. Ja diakritiskā zīme ir novietota starp diviem burtiem, no kuriem vienam tā ir iespējama, tad tiek uzskatīts, ka lietojums ir pareizs, resp., diakritiskā zīme lietota ar to burtu, kuram tā ir iespējama.

Rezultāti

Kopumā izskatīti pirmie 322 korpusā iekļautie teksti. To sadalījums pēc semestra, kopējais vārdu skaits un vidējais vārdu skaits vienā tekstā sniegts 1. tabulā.

	Tekstu skaits	Vārdu skaits	Vidējais vārdu skaits tekstā
1. semestris	176	24 427	138,79
2. semestris	146	22 505	154,14
KOPĀ	322	46 932	145,75

1. tabula. Tekstu kopskaits, apjoms un vidējais apjoms.

Kā redzams, 2. semestra tekstu ir par ~17% mazāk nekā 1. semestra tekstu, taču vidējais vārdu skaits tajos ir par ~11% lielāks, tāpēc kopējais vārdu skaits tajā ir tikai par ~8% mazāks nekā 1. semestrī. Līdz ar to nav sagaidāmas būtisku apjoma atšķirību noteiktas īpatnības kādā no semestriem.

Ekscerpējami piemēri nav konstatēti visos tekstos, piemēru kopskaits pat nesasniedz tekstu kopskaitu – kopā ekscerpēti 175 piemēri. Piemēru skaita sadalījums pēc semestriem sniegts 2. tabulā.

	Piemēru skaits	Vidējais piemēru skaits tekstā[2]	Vidējais piemēru skaits 1000 vārdos[3]
1. semestris	126	0,716	5,158
2. semestris	49	0,336	2,177
KOPĀ	175	0,543	3,729

2. tabula. Tekstu kopskaits, vidējais skaits tekstā un 1000 vārdos.

2. semestra tekstos atrasto piemēru skaits ir vairāk nekā divreiz mazāks nekā 1. semestra tekstos atrasto piemēru skaits – gan absolūtos skaitļos, gan attiecībā pret tekstu skaitu un vārdu skaitu tekstā. Tomēr vienā tekstā var būt vairāki šādi jaunie burti, un šāds skats neļauj to izvērtēt. Tāpēc 3. tabulā sniegti dati par kļūdaino tekstu (šeit ar to jāsaprot teksti, kuros ir veidoti jauni burti ar diakritiskajām zīmēm) skaitu un no tā izrietošie aprēķini.

	Tekstu skaits	Kļūdaino tekstu skaits	Kļūdaino tekstu īpatsvars	Vidējais piemēru skaits kļūdainā tekstā
1. semestris	176	59	33,52%	2,136
2. semestris	146	29	19,86%	1,690
KOPĀ	322	88	27,33%	1,989

3. tabula. Tekstu kopskaits, apjoms un vidējais apjoms.

2. semestrī kļūdaino tekstu ir daudz mazāk gan absolūtos skaitļos, gan procentuāli. Vidējais kļūdu skaits tekstā gan nevienā no semestriem nav augsts – 1. semestrī tās ir mazliet vairāk nekā divas kļūdas tekstā, savukārt 2. semestrī – nedaudz vairāk par pusotru kļūdu tekstā. Vērtējot šos skaitļus, jāņem vērā, ka zemāka vērtība par 1 šeit nav iespējama vispār, jo aprēķinos ir iekļauti tikai tie teksti, kuros ir vismaz viens ar diakritiskajām zīmēm veidots jauns burts.

Secinājumi

Praktiski strādājot ar ārvalstu studentiem, kas mācās latviešu valodu, empīriskie novērojumi var likt uzskatīt, ka jaunu burtu veidošana ar diakritiskajām zīmēm ir samērā reta parādība. Jo sevišķi tā šķistu tādēļ, ka, tekstus rakstot, tiek izmantoti palīglīdzekļi. Tomēr tas, ka pirmajos divos semestros šī īpatnība sastopama vairāk nekā ceturtdaļā korpusa tekstu, bet pirmajā semestrī – pat trešdaļā tekstu, apstiprina pretēju pieņēmumu – ka tā drīzāk ir raksturīga starpvalodas īpatnība valodas apguves sākumposmā, taču ar

[2] Aprēķināts pēc kopējā aplūkoto tekstu skaita.

[3] Aprēķināts pēc aplūkoto tekstu kopējā vārdu skaita.

tendenci samazināties, proti, turpinot apgūt valodu (nākamajā semestrī), mazāk kļūst apguvēju, kas šādus jaunus burtus veido.

Ar iegūtajiem datiem gan nepietiek, lai izdarītu pieņēmumus par to, kurā semestrī šai parādībai vajadzētu kļūt reti sastopamai. Korpusā ir iekļauti teksti tikai no pirmajiem diviem mācību semestriem, par tālāku latviešu apguves procesu datu nav. Jāņem vērā arī, ka dati būtu citādi, ja teksti būtu rakstīti, neizmantojot nekādus palīglīdzekļus. Šādos tekstos nav iespējams nošķirt apguvēja prasmi *pārrakstīt* vārdus un tekstu no apguvēja prasmes tos *uzrakstīt*.

Iesācēju rakstītajos tekstos ir daudz atkārtojošos konstrukciju un vārdu – to nosaka viņu valodas prasmes līmenis. Līdz ar to varētu sagaidīt, ka, ja kļūda ir pieļauta vienreiz, tas tā notiks arī visos pārējos līdzīgos vārdos vai konstrukcijās. Dažos tekstos tas ir novērojams, tomēr tas, ka vidējais piemēru skaits kļūdainā tekstā kopumā nav augsts, liek domāt, ka jaunu burtu veidošanai ar diakritiskām zīmēm drīzāk ir nevis sistēmisks, bet gan gadījuma raksturs. No otras puses, tas neizslēdz pieņēmumu, ka ir faktori, kas veicina jaunu burtu veidošanu noteiktās pozīcijās, burtos un/vai vārdos. Jau šobrīd ir redzams, ka tādos vārdos kā *garšo*, *viņš*, *ļoti*, *četri* jaunus burtus veidojuši vairāki apguvēji. Tāpat nereti šādi piemēri sastopami blakus burtiem, kuriem attiecīgā diakritiskā zīme būtu lietojama, taču nav lietota. Par šāda lietojuma iemesliem vairāk varētu uzzināt kvalitatīvā pētījumā.

Tekstu skaits korpusā turpina pieaugt. Raksta iesniegšanas brīdī korpusā ir iekļauti 586 teksti, un kopskaitā ir paredzēts savākt 1000 tekstu. Kad korpusā iekļaujamie teksti būs apkopoti, būtu ieteicams atkārtot šo pētījumu, lai pārliecinātos, vai jaunie dati apstiprina līdzšinējos pieņēmumus. Tad arī būtu vēlama datu sīkāka kvantitatīva apstrāde, grupējot piemērus pēc burta, pēc diakritiskās zīmes, pēc diakritiskās zīmes novietojuma, kā arī korpusā savāktajiem metadatiem, piemēram, tekstu autoru dzimuma, dzimtās valodas un citu valodu prasmēm. Līdz šim ekscerpēto piemēru vēl ir pārāk maz šādai kvantitatīvai analīzei.

Literatūra

Auziņa, Kaija, Levāne-Petrova 2019 = **Auziņa, Ilze, Kaija, Inga, Levāne-Petrova, Kristīne.** *Mērķa hipotēžu izvirzīšana latviešu valodas apguvēju korpusā*. Stenda referāts Latvijas Universitātes 77. konferences latviešu valodniecības sekcijā "Gramatika un valodas normēšana" 14.02.2019. Rīga : Latvijas Universitāte, 2019.

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

Lally 2000 = **Lally, Carolyn Gascoigne.** The Effect of Keyboarding on the Acquisition of Diacritical Marks in the Foreign Language Classroom. *The French Review*, 73(5), 2000, 899–907.

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

McEnery, Brezina u. c. = **McEnery, Tony, Brezina, Vaclav, Gablasova, Dana, Banerjee, Jayanti.** Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, 39, 2019, 74–92.

Šalme, Auziņa 2016 = **Šalme, Arvils, Auziņa, Ilze.** *Latviešu valodas prasmes līmeņi: pamatlīmenis A1, A2, vidējais līmenis B1, B2.* Rīga : Latviešu valodas aģentūra, 2016.

Šalme 2011 = **Šalme, Arvils.** *Latviešu valodas kā svešvalodas apguves pamatjautājumi.* Rīga : LVA, 2011.

Znotiņa 2015 = **Znotiņa, Inga.** Learner corpus annotation in Latvia and Lithuania. *Sustainable Multilingualism*, 7, 2015, 145–159.

Znotiņa 2018 = **Znotiņa, Inga.** *Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas.* Promocijas darbs filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares lietišķās valodniecības apakšnozarē. Liepāja : Liepājas Universitāte, 2018.

Using diacritical marks to make new letters in texts written by learners of Latvian as a foreign language

Summary

A Latvian learner corpus LaVA is being built in the Institute of Mathematics and Computer Science, University of Latvia. The corpus includes texts written by beginner learners in the first two semesters of learning Latvian as a foreign language. The texts are written by hand and digitized afterwards in order to reduce the issues that could be caused by the necessity to learn not only writing itself but also using a foreign keyboard.

One of the features that cannot be digitized is the new letters created by adding diacritical marks which are not used that way in the standard Latvian alphabet. Since one of the essential steps in learning to write in a language is learning the letters and diacritical marks of that language, this study aims to find instances of such newly made letters and to discuss the basic quantitative measures in order to define hypotheses and areas of interest for further research of such usage.

Altogether 322 texts were searched and 175 examples were found. The amount of examples found in 2nd semester texts was less than half the amount of examples found in the 1st semester texts, but the percentage of texts containing examples was higher than expected – more than 33% in the 1st semester and almost 20% in the 2nd semester. It leads to a conclusion that this is quite a common occurrence but also prone to reduction in the second semester. The corpus does not provide any data on later semesters so it cannot be predicted when such instances should become a rare, individual feature rather than a common one.

The average amount of examples in a text is not high, though. Counting only the texts where at least one example was found, the average amount of examples per text is 2,136 in the 1st semester and 1,690 in the 2nd semester. Considering that the absolute lowest possible value here is 1, it should not be seen as a high value. Therefore, using diacritical marks to make new letters, while a common feature of the Latvian interlanguage, could be characterized as casual rather than systemic.

However, that does not exclude the possibility of certain patterns in usage. The currently collected data already shows that there are some words – such as *garšo, viņš, ļoti, četri* – where examples were found in more than one author's text. Examples of using unsuitable diacritical marks are also sometimes found next to letters for which said diacritical marks would be suitable. This should be explored more thoroughly using qualitative methods.

The size of the corpus keeps growing, the expected size upon completion is 1000 texts. When it is reached, it would be useful to repeat the study and check whether the larger amount of data still confirms the same assumptions. The larger sample size would also allow for more detailed quantitative analysis discussing each letter, diacritical mark, placement of the diacritical mark, and metadata collected for the corpus, such as gender, native language and other spoken languages by the authors of the texts.

Keywords: learner corpus, diacritical marks, letters, Latvian, foreign language, adult language learners

Autore:
Inga Kaija – Dr. philol., docente Rīgas Stradiņa universitātē, pētniece Latvijas Universitātes Matemātikas un informātikas institūtā, 26176142, Inga.Kaija@rsu.lv, kaijainga@gmail.com