Kristīne LEVĀNE-PETROVA, Ilze AUZIŅA, Kristīne POKRATNIECE

Latviešu valodas apguvēju korpusa datu ieguves un apstrādes metodoloģijas izstrāde

Development of methodology for Learner Corpus of Latvian data acquisition and processing

Popularity of learning Latvian as a foreign language is increasing. Latvian as a foreign language is being taught not only in the higher educational institutions of Latvia, but also in more than 20 universities outside of Latvia (Šalme 2008; Šalme 2011; Laizāne 2019). Therefore, corpus-based and corpus-driven teaching materials are crucial for the international students that acquire Latvian both in Latvia and abroad.

Since September 2018 the project *Development of Learner corpus of Latvian: methods, tools and applications*[1] has been carried out. During the project, based on the already existing experience of designing Learner Corpus of Latvian (LaVA), a corpus of students' essays with different language backgrounds will be created. The newly developed corpus will be publicly available. Although the corpus creation pipeline includes text collection, digitization, and morphological and error annotation; this article will cover just the first phases of the creation of the corpus – the development of a methodology for data collection and digitization.

The agreement form with data subject about data inclusion into the corpus and the metadata (gender, age, mother tongue, language proficiency level) collection form was developed. Guidelines for teachers on preferred topics have been prepared.

The corpus is built on an integrated multifunctional platform that provides a single interface for uploading, annotating and search. Moreover, the web platform can also be used for storing scanned copies of essays, comparing texts entered by two independent digitizers, correcting texts, and error-annotated texts and making inter-annotator agreement.

At least 1000 essays on different topics from students with different language backgrounds are planned to be included in the LaVA corpus. For data collection, multiple universities and language teachers have been contacted and have agreed to support the corpus creation process by providing their students the assignments developed previously.

Collected essays with metadata are handwritten; therefore, they need to be digitized for further data processing steps. The digitization is being carried out in three steps: (1) scanning of the assignments and essays; (2) metadata input; (3) text rewriting in digital format. Scanned images of the assignments help to validate data correctness if such concern arises. Metadata is entered manually.

Keywords: learner corpus, language acquisition, metadata, digitization, permission, essays

[1] Latvian Council of Science Grant *Development of Learner corpus of Latvian: methods, tools and applications* (No. lzp-2018/1-0527)

Latviešu valodas kā svešvalodas mācīšana kļūst aizvien populārāka Latvijas un arī ārvalstu augstākajās mācību iestādēs, tāpēc jaunu, korpusā balstītu mācību materiālu izstrāde ir ļoti nozīmīga.

Kopš 2018. gada septembra LU MII tiek īstenots LZP projekts ”Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums”, kura laikā tiks izveidots “Latviešu valodas apguvēju korpuss” (LaVa), kas būs jaunu mācību materiālu pamats. Tajā varēs arī pētīt ārvalstnieku latviešu valodas apguves īpatnības. Šajā rakstā tiks aplūkoti korpusa LaVa izstrādes sākotnējie posmi – korpusa datu atlases kritēriji, datu izmantošanas atļaujas ieguve, datu apstrādes vietne un datu digitalizēšana –, kā arī problēmas un risinājumi, kas ar to saistīti.

Atslēgvārdi: apguvēju korpuss, latviešu valodas apguve, metadati, tekstu atlases kritēriji, atļauja, digitalizēšana, tekstu apstrāde

Ievads

Latviešu valoda kā svešvaloda tiek mācīta gan Latvijas, gan vairāk nekā 20 ārvalstu augstskolās (Šalme 2008; Šalme 2011; Laizāne 2019), tāpēc aizvien aktuālāka ir jaunu, korpusā balstītu mācību materiālu izstāde. Lai pētītu ārvalstnieku latviešu valodas apguves īpatnības un nodrošinātu uz datiem balstītu latviešu valodas mācību un metodisko materiālu izstrādi, kopš 2018. gada septembra LU MII tiek īstenots LZP projekts ”Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums”[2], kura laikā tiks izveidots “Latviešu valodas apguvēju korpuss” (turpmāk – korpuss LaVA), kas būs pētnieciska bāze latviešu valodas apguves īpatnību izpētei, kvantitatīvai un kvalitatīvai valodas apguvēju pieļauto kļūdu analīzei un metodisko materiālu izstrādei.

Korpusā paredzēts iekļaut to Latvijas augstākajās mācību iestādes studējošo ārvalstnieku darbus, kas latviešu valodu apgūst kā svešvalodu pirmo vai otro semestri. Korpusā plānots iekļaut aptuveni 1000 darbu (apmēram 100 000 vārdlietojumu).

Šāda korpusa, tāpat kā citu valodas korpusu (Levāne-Petrova 2012; Levāne-Petrova 2019), izstrāde notiek vairākos posmos (korpusa koncepcijas izstrāde, datu atlases kritēriju noteikšana, datu atlase, digitalizēšana, ja nepieciešams, normalizēšana, morfoloģiskā marķēšana un kļūdu marķēšana, infrastruktūras izveide korpusa datu vākšanai un apstrādei u.c.). Šajā rakstā tiks aplūkoti sākotnējie korpusa LaVA izveides posmi, proti, korpusa datu atlases kritēriji, datu izmantošanas atļaujas ieguve, datu apstrādes vietne un datu digitalizēšana.

LZP projekta mērķi un uzdevumi ir arī saistāmi ar VPP projekta “Latviešu valoda”[3] mērķiem un uzdevumiem.

[2] Latvijas Zinātnes padomes Fudamentālo un lietišķo pētījumu projekts ”Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums” (lzp-2018/1-0527).

[3] Valsts pētījumu programmas projekta “Latviešu valoda” Nr. VPP-IZM-2018/2-0002 apakšprojekts “Latviešu valodas apguve”.

Korpusa datu atlases kritēriji

Pirms katra korpusa izveides tiek noteikti, kādi datu avoti korpusā būtu iekļaujami jeb tekstu atlases kritēriji. Arī korpuss LaVA nav izņēmums. Tādējādi korpusa galvenie datu atlases kritēriji ir:

1) tekstus ir rakstījuši ārvalstu studenti, kas latviešu valodu apgūst kā svešvalodu kādā no Latvijas augstskolām;

2) studenti latviešu valodu apgūst 1. vai 2. semestri, sasniedzot latviešu valodas prasmes pamatlīmeni (A1 vai A2);

3) teksti ir tapuši mācību procesa laikā;

4) vēlamais tekstu apjoms – vismaz 100 vārdlietojumu.

Lai iegūtu šādus darbus, par projekta ieceri tika informēti to augstskolu mācībspēki, kurās studenti atpgūst latviešu valodu kā svešvalodu (Šalme 2008; Šalme 2011; Laizāne 2019), un ir aicināti iesaistīties korpusa datu vākšanā. Pašlaik (2019. gada jūlijā) studentu darbi ir iegūti no Rīgas Stradiņa universitātes (RSU) (84%), Latvijas Universitātes (LU) (6%), Liepājas Universitātes (LiePu) (5%), Rēzeknes Tehnoloģiju akadēmijas (RTA) (3%), Latvijas Kultūras akadēmijas (LKA) (2%). Tika apzinātas arī citas Latvijas augstākās mācību iestādes, kur ārvalstu studentiem māca latviešu valoda, tomēr ne visur pārbaudes darbi notiek rakstiski, tāpēc datus no šo augstskolu pasniedzējiem iegūt nebija iespējams, vai arī augstskolu docētāji nepiekrita iesaistīties datu vākšanā.

Plānotais korpusa LaVA apjoms – 1000 studentu darbi. Pieņemot, ka katra korpusā iekļaujamā teksta noteiktais garums būs vismaz 100 vārdlietojumu, ir paredzams, ka korpusu veidos aptuveni 100 tūkstoši vārdlietojumu. Tomēr arī tad, ja teksta apjoms ir mazāks, bet atbilst citiem tekstu atlases kritērijiem, tas tiek iekļauts korpusā.

Rakstudarbu tematu izvēle ir docētāju ziņā, tomēr parasti tā ir saistīta ar apgūstamo vielu un tēmām, ar kurām parasti sāk kādas svešvalodas apguvi, t. i., personas raksturojums, ģimene un draugi, ikdiena, mācības, vaļasprieki ("Es un mana ģimene", "Mani hobiji", "Manas studijas" u. tml.). Lai nodrošinātu datu aizsardzību, studentiem tiek lūgts rakstīt par izdomātu ģimeni, notikumiem u. tml.

Korpusa datu izmantošanas atļauja un korpusa datu ieguve

Lai aizsargātu izglītojamo tiesības, vācot viņu uzrakstītos tekstus, tika izstrādāta veidlapa un tekstu ieguves kārtība / principi. Veidlapa ir angļu valodā, jo

tā ir studentu mācību valoda, kurā viņi apgūst specialitāti. Veidlapas otrā pusē (tukša lapa) studenti ar roku raksta tekstu.

Dati tiek iegūti, izmantojot veidlapu, kurā ir:
1) dota informācija par projektu un raksturoti datu vākšanas un izmantošanas mērķi; 2) atļauja datu iekļaušanai korpusā; 3) anketa, kurā studenti sniedz informāciju par sevi (metadati) (sk. 1. attēlu).

Sastādot atļauju, kuru studenti paraksta un tā sniedz atļauju iekļaut savus tekstus korpusā LaVA, ir ņemti vērā noteikumi par personas datu aizsardzību un autortiesību jautājumiem, kas saistoši Latvijā. Dokuments, kas regulē autortiesību aizsardzību Latvijas Republikā, ir Autortiesību likums[4], savukārt personas datu aizsardzību regulē Personas datu apstrādes likums un Eiropas Savienības jaunā Vispārīgā datu aizsardzības regula[5]. (Kaija, Auziņa 2019) Veidojot atļauju, tika ņemta vērā otrās baltu valodas apguvēju korpusa "Esam" veidotājas pieredze šādas atļaujas izstrādē (Znotiņa, 2017).

Lai ievērotu noteikumus, ir izstrādāta standartizēta veidlapa visiem korpusa tekstu autoriem. Tam tika izveidota datu izmantošanas atļauja, kuru studentiem ir jāparaksta un tādējādi jāsniedz sava piekrišana datu iekļaušanai korpusā.

Ar savu parakstu izglītojamie:

1) apliecina, ka piekrīt nesaņemt nekādu finansiālu atlīdzību par to, ka viņu teksti tiek iekļauti korpusā, turklāt apzinās, ka korpuss būs pieejams bez maksas zinātniskiem un mācību mērķiem;

2) apstiprina, ka tekstā minētie dati neļauj identificēt konkrētu personu;

3) piekrīt, ka teksts ir anonīms un izglītojamā vārds neparādās nekur korpusa mājaslapā vai tā publiskajā dokumentācijā; turklāt katram autora tekstam tiek piešķirts unikāls kods, kas ļauj atpazīt tā paša autora rakstītus tekstus, bet neatklāt autora identitāti;

4) piekrīt, ka korpusā iekļautie dati / teksti var tikt citēti mācību materiālos, pētījumos un citos dažāda veida darbos;

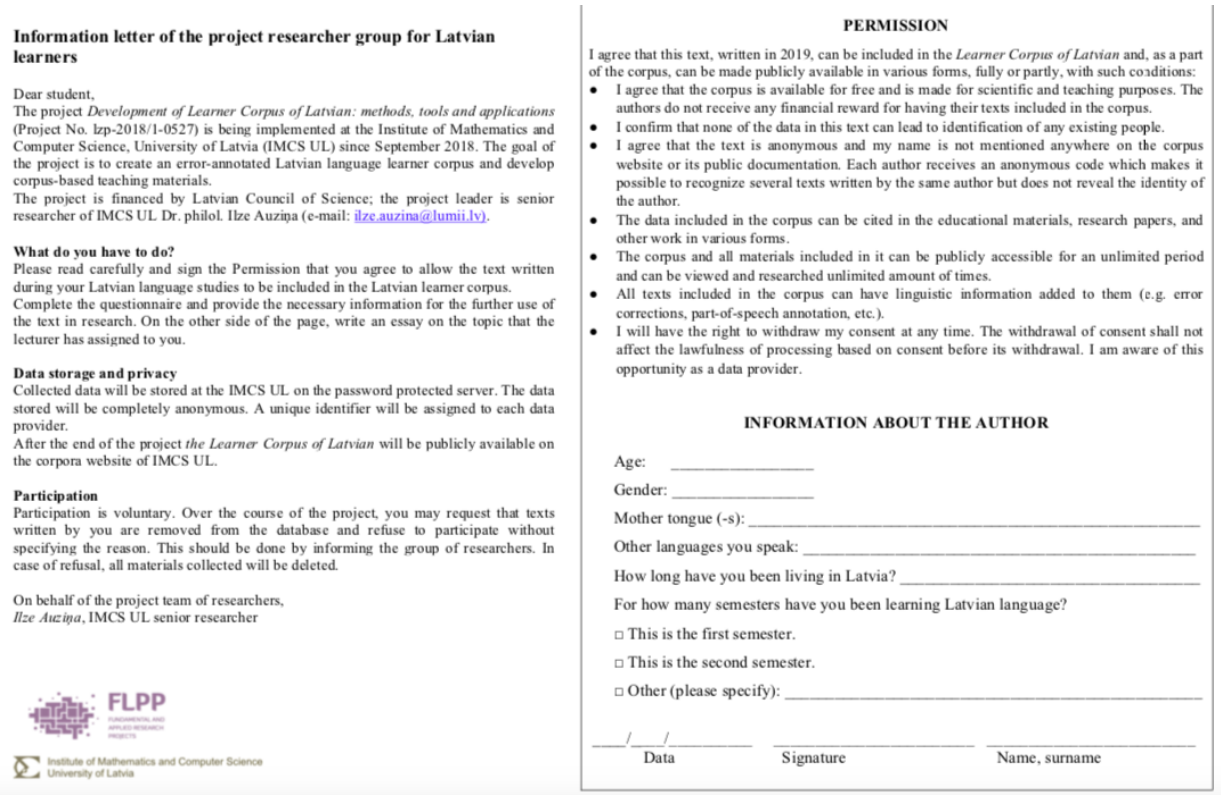
5) piekrīt, ka korpusā iekļautie dati var būt publiski pieejami neierobežotu laiku, un tos var skatīt un petīt neierobežoti daudz reižu;

6) piekrīt, ka korpusā iekļautajiem tekstiem tiek pievienota papildu lingvistiskā informācija, piemēram, marķētas kļūdas, morfoloģiskās pazīmes;

7) patur tiesības jebkurā laikā atsaukt piekrišanu izmantot datus, tomēr vienošanās atsaukšana neietekmēs datu apstrādes likumību līdz vienošanās atsaukšanai.

[4] https://likumi.lv/doc.php?id=5138

[5] https://eur-lex.europa.eu/legal-content/LV/TXT/HTML/?uri=CELEX:32016R0679&from=LV



Information letter of the project researcher group for Latvian learners

Dear student,
The project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The goal of the project is to create an error-annotated Latvian language learner corpus and develop corpus-based teaching materials.
The project is financed by Latvian Council of Science; the project leader is senior researcher of IMCS UL Dr. philol. Ilze Auziņa (e-mail: ilze.auzina@lumii.lv).

What do you have to do?
Please read carefully and sign the Permission that you agree to allow the text written during your Latvian language studies to be included in the Latvian learner corpus.
Complete the questionnaire and provide the necessary information for the further use of the text in research. On the other side of the page, write an essay on the topic that the lecturer has assigned to you.

Data storage and privacy
Collected data will be stored at the IMCS UL on the password protected server. The data stored will be completely anonymous. A unique identifier will be assigned to each data provider.
After the end of the project *the Learner Corpus of Latvian* will be publicly available on the corpora website of IMCS UL.

Participation
Participation is voluntary. Over the course of the project, you may request that texts written by you are removed from the database and refuse to participate without specifying the reason. This should be done by informing the group of researchers. In case of refusal, all materials collected will be deleted.

On behalf of the project team of researchers,
Ilze Auziņa, IMCS UL senior researcher

FLPP
Institute of Mathematics and Computer Science
University of Latvia

PERMISSION

I agree that this text, written in 2019, can be included in the *Learner Corpus of Latvian* and, as a part of the corpus, can be made publicly available in various forms, fully or partly, with such conditions:

- I agree that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.
- I confirm that none of the data in this text can lead to identification of any existing people.
- I agree that the text is anonymous and my name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched unlimited amount of times.
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).
- I will have the right to withdraw my consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. I am aware of this opportunity as a data provider.

INFORMATION ABOUT THE AUTHOR

Age: _____
Gender: _____
Mother tongue (-s): _____
Other languages you speak: _____
How long have you been living in Latvia? _____
For how many semesters have you been learning Latvian language?
□ This is the first semester.
□ This is the second semester.
□ Other (please specify): _____

____/____/_____ _____ _____
Data Signature Name, surname

1. attēls. Studentu apstiprinājuma veidlapa

Studenti anketā ieraksta ziņas par sevi, kas tālāk kā metadati tiek izmantoti, analizējot datus: studenta vecums, dzimums, dzimtā valoda, svešvalodu zināšanas, ziņas par to, cik ilgi uzturas Latvijā un mācās latviešu valodu. Studenti norāda arī savu vārdu un uzvārdu, taču šī informācija tālāk netiek izmantota, jo katram darbam, to ieskenējot, tiek piešķirts identifikators, nekādā veidā nenorādot uz autora personību. Vārds, uzvārds un paraksts ir nepieciešams kā apliecinājums tam, ka students ir izlasījis informāciju par projektu, sapratis un piekritis teksta izmantošanas nosacījumiem. Turklāt tas ļauj gadījumā, ja students, laikam ejot, pārdomā un nevēlas, lai viņa rakstudarbs ir iekļauts korpusā, atrast un izņemt no datu kopas atbilstošo darbu. Tā kā studenti ir norādījuši savu vārdu un uzvārdu, pasniedzēji pēc tam, kad veidlapa un rakstudarbs ir ieskenēti, var atdot oriģinālu studentam.

Pasniedzējiem, kas piekrita sadarboties un nodot studentu datus korpusa izveidei, tika sniegti norādījumi, kā korpusam nepieciešamie dati būtu iegūstami un kādi tieši dati ir nepieciešami.

Rakstudarbi top mācību procesa laikā (ne kā pārbaudes darbi) – vai nu nodarbību laikā, vai arī kā mājasdarbs. Studenti drīkst izmantot savus latviešu valodas nodarbību pierakstus, vārdnīcas un citus materiālus, bet ir lūgums neizmantot mašīntulku, piemēram, *Google tulkotāju, Hugo.lv.* Jāatzīmē, ka ne vienmēr ir iespējams identificēt, vai, rakstot latviešu valodas tekstus, ir izmantots

mašīntulks, tomēr ir gadījumi, kad tas ir nosakāms, jo tajā parādās sarežģītas konstrukcijas, kuras students vēl nav mācījies u.tml.

Gan aizpildītās anketas un studenta parakstītā atļauja par datu izmantošanu, gan studentu radītais teksts tiek ieskenētas un digitālā veidā nodots korpusa veidotājiem tālākai datu apstrādei. Pirms datu ieskenēšanas darbi netiek laboti.

Ja students ir uzrakstījis tekstu par vienu no docētāja dotajiem tematiem, bet veidlapu nav aizpildījis un nav parakstījis atļauju par teksta iekļaušanu korpusā, dati netiek izmantoti.

Korpusa datu apstrādes vietne

Kad ieskenētās studentu veidlapas un darbi attēlu formātā ir saņemti, tie tiek ievietoti speciāli projekta vejadzībām izveidotā korpusa vietnē to tālākai apstrādei (sk. 2. attēlu). Korpusa vietnei ir ierobežota piekļuve – tikai korpusa lietotāji var tajā pieteikties ar lietotājvārdu un paroli. Korpusa vietnē redzams katra darba unikālais identifikators (Id), metadati par katru tekstu un tekstu apstrādes posmi: oriģinālā teksta digitalizēšana (*Original text*), normalizēšana (*Corrected text*) un kļūdu marķēšana (*Error annotation*).

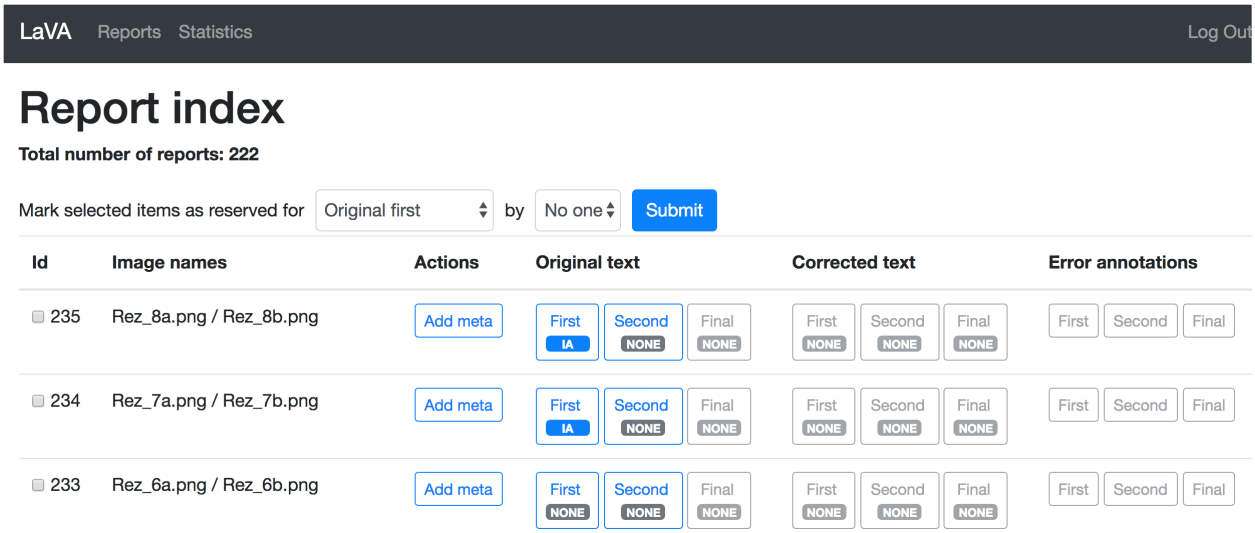
Korpusa vietnē korpusa veidotāji var izvēlēties apstrādājamo tekstu un sekot līdzi darba progresam. Tajā ir arī pieejama statistika par korpusā ievietoto tekstu skaitu, jau apstrādātajiem (digitalizētajiem un normalizētajiem) tekstiem un katra korpusa veidotāja (A, B, C, D, E) padarīto. (sk. 3. attēlu)

Lai vieglāk būtu sekot līdzi apstrādājamo datu plūsmai, ir dots krāsu marķējums/informācija par tekstu statusu – iesāktie teksti zilā krāsā, pabeigtie – zaļā krāsā. Tikai tad, kad teksta digitalizēšanas process ir pabeigts, var doties pie nākamā posma – tekstu normalizēšanas.

Galvenajā korpusa datu apstrādes vietnes logā var redzēt, kurš korpusa veidotājs ir konkrētu tekstu ir digitalizējis, normalizējis vai marķējis tajā kļūdas. Tas ļauj pārvaldīt katru korpusa izveides posmu, izsekot, kurš korpusa veidotājs ir apstrādājis konkrēto darbu un kādu korpusa izstrādes posmu veicis, pārrunāt problēmas un tādējādi sekmēt korpusa izveidi.

Datu digitalizēšana

Korpusa datu digitalizēšana notiek korpusa vietnē: 1) no ieskenētās anketas tiek ievadīti metadati, t.i., informācija par tekstu autoru dzimumu, vecumu, dzimto valodu, svešvalodu prasmi; personas vārds un uzvārds netiek saglabāts; 2) tiek manuāli pārrakstīts (digitalizēts) izglītojamā uzrakstītais teksts.



2. attēls. Korpusa datu apstrādes vietne

Report statistics

Total number of reports: 586

Action	A	B	C	D	E	None
Original text input	0	366	185	281	277	0
Original text finalization	0	202	140	48	196	0
Corrected text input	0	450	232	71	293	126
Corrected text finalization	0	197	170	0	201	18

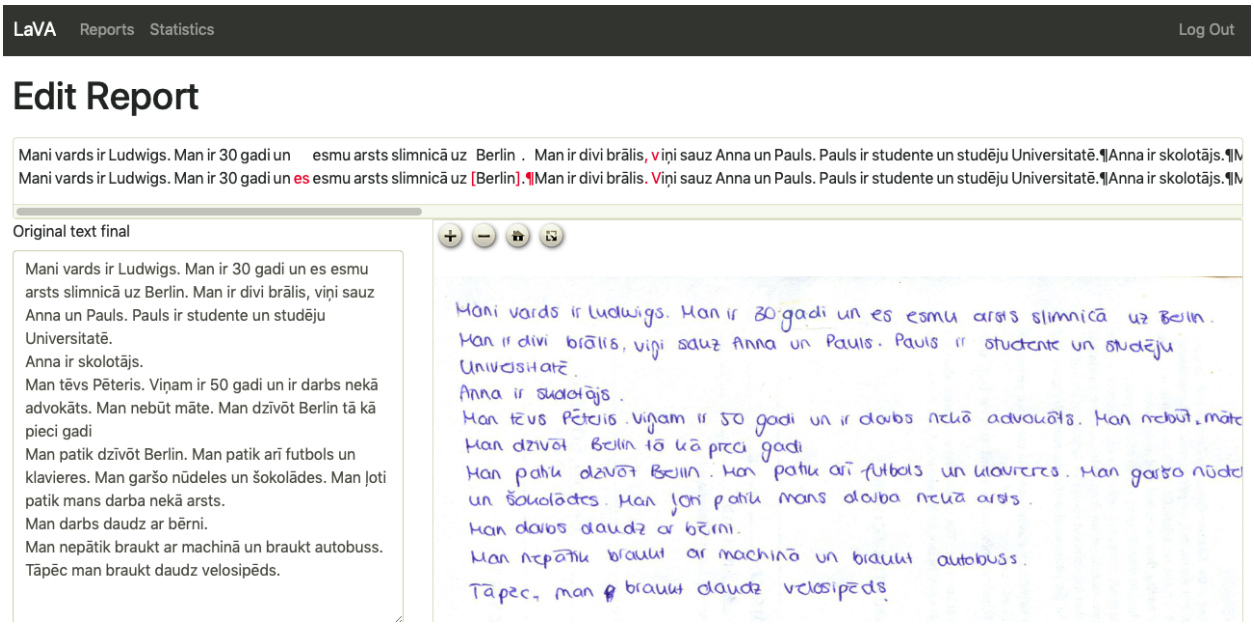
3. attēls. Korpusa datu apstrādes vietne: statistika par tekstu apstrādi

Izglītojamā rakstīto tekstu manuāli pārraksta divi korpusa veidotāji neatkarīgi viens no otra – proti, viens pārraksta (digitalizē) sacerējumu korpusa platformā un saglabā to (*Original text: First*), paturot visas studenta darbā pieļautās kļūdas un valodas īpatnības, tad to pašu dara otrs korpusa veidotājs (*Original text: Second*). Pēc tam trešais korpusa veidotājs izveido šī darba gala versiju (*Original text: Final*) (sk. 4. attēlu), panākot t.s. marķētāju vienotību *(inter-annotator agreement)*. Proti, 4. attēlā ir redzams, ka abu neatkarīgo sacerējumu ievadītāju teksti (*Original text: First; Original text: Second*) tiek automātiski sastatīti un nesakritības tiek parādītas sarkanā krāsā, tādējādi trešais korpusa veidotājs izlabo un vienādo darbu saturu atbilstoši oriģinālam. Tika izmēģināta arī automātiska rokraksta atpazīšana, bet tā kā

vairāk nekā 50% teksta tika atpazīta kļūdaini, korpusa veidotāji nolēma tekstus pārrakstīt manuāli.

Digitalizējot studentu darbus, ir sastopamas dažādas problēmas, piemēram, grūti salasāms vai neskaidrs rokraksts, teksta autors labojis rakstīto, vairākkārt pārvelkot vienu vai vairākus burtus un nav skaidrs, kas domāts, u. tml. Reizēm izgītojamais lietojis latviešu valodas grafētikai netipiskas rakstzīmes vai diakritiskās zīmes, piemēram, *w, ø,* umlautu, akūtu virs patskaņa burta u. c. Šādos gadījumos, digitalizējot tekstu, ja vien ir iespējams atrast atbilstošu simbolu, tiek saglabātas izglītojamā lietotās grafēmas un diakritiskās zīmes. (Kaija 2019) Ja teksts rakstīts ar drukātiem burtiem, digitalizējot tekstu, lielie sākumburti tiek lietoti tikai tur, kur tas pēc ortogrāfijas likumiem gaidāms. Ja studenta darbā teksts nav saprotams, digitalizējot neskaidro vārdu vai teksta fragmentu, tas tiek ievietots kvadrātiekavās, piemēram, *Man patīk iet vai [palaist]*.

Lai atvieglotu digitalizētāju darbu, korpusa vietnē tiek piedāvāti vairāki tehniski risinājumi, piemēram, ekrānlogā, kurā redzams ieskenētais teksts, ir iespējams pietuvināt teksta attēlu un samazināt to.



4. attēls. Korpusa datu digitalizēšana

Nobeigums

Rakstā aplūkoti korpusa LaVa izveides sākuma posmi – datu vākšanas metodoloģijas un atļaujas izstrāde, datu digitalizēšanas process, kā arī īsi raksturota korpusa datu apstrādes vietne. Kad izglītojamo darbi ir digitalizēti un metadati par

katru teksta autoru ir saglabāti, sākas darbs pie nākamajiem korpusa izveides posmiem. Tie ir:

- datu normalizēšana, proti, datu pārrakstīšana atbilstoši latviešu valodas pareizrakstības un gramatikas normām;
- automātiska morfoloģiskā marķēšana, tostarp vārdšķiru un pamatformu noteikšana;
- oriģinālā un normalizētā teksta sastatīšana,
- kļūdu tipu un kļūdu marķēšanas metodoloģijas izstrāde;
- automatizēta kļūdu anotēšana un manuāla pārskatīšana. (Darģis et al. 2018; Auziņa et al. 2019)

Digitalizēšanā izmantotie tehniskie risinājumi tiek izmantoti arī nākamajos datu apstrādes posmos.

Literatūra

1. Auziņa et al. 2019 = **Auziņa, Ilze, Levāne-Petrova, Kristīne, Darģis, Roberts.** Latviešu valodas apguvēju kļūdu analīze: pareizrakstības kļūdas. Raksts iesniegts publicēšanai krājumā "Vārds un tā pētīšanas aspekti".
2. Darģis et al. **2018** = **Darģis, Roberts, Auziņa, Ilze, Levāne-Petrova, Kristīne.** The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2018), 2018 Miyazaki, Japan: European Language Resources Association (ELRA).
3. Kaija 2019 = **Kaija, Inga**. Jaunu burtu veidošana ar diakritiskajām zīmēm latviešu valodas kā svešvalodas apguvēju tekstos. Raksts iesniegts publicēšanai krājumā "Valodu apguve: problēmas un perspektīva".
4. Kaija, Auziņa 2019 = Kaija, Inga, Auziņa, Ilze. Data collection for learner corpus of Latvian: copyright and personal data protection. Extended abstract accepted for publication in CLARIN Annual Conference 2019 proceedings.
5. Laizāne 2019 = **Laizāne, Inga.** Latviešu valoda kā svešvaloda: lingvodidaktikas virziena attīstība Latvijā un ārpus tās. Promocijas darbs filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares lietišķās valodniecības apakšnozarē. Liepāja : Liepājas Universitāte, 2019. Pieejams arī: https://www.liepu.lv/uploads/dokumenti/prom/Disertacija_Laizane_2019.pdf
6. Levāne-Petrova 2012 = **Levāne-Petrova, Kristīne.** Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. *Baltistica* VIII priedas, Vilnius : 2012. 89.–98. lpp.
7. Levāne-Petrova 2019 = **Levāne-Petrova, Kristīne.** Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos *Valoda: nozīme un forma 10. LU Humanitāro zinātņu fakultātes Latviešu un vispārīgās valodniecības katedras rakstu krājums.* Rīga LU Akadēmiskais apgāds. 2019. Iesniegts publicēšanai.

8. Šalme 2008 = **Šalme, Arvils.** *Latviešu valodas kā svešvalodas apguve Eiropas augstskolās.* Rīga: LVA.
9. Šalme 2011 = **Šalme, Arvils.** *Latviešu valodas kā svešvalodas apguves pamatjautājumi.* Rīga: LVA.
10. Znotiņa 2017 = **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, LiePA, 2017. Pieejams arī: https://www.liepu.lv/uploads/files/Otrās baltu valodas apguvēju korpusa izveide.pdf