Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums

Ilze Auziņa

Ilze.auzina@lumii.lv

LATVIJAS UNIVERSITĀTE ANNO 1919

FLPP FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

Mākslīgā intelekta laboratorija LU MII

Projekts “Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums” Nr. lzp-2018/1-0527

Projekta mērķi

Vispārīgais mērķis ir sagatavot pētniecisku bāzi latviešu valodas apguves īpatnību izpētei, balstoties uz jaunizveidotā *Latviešu valodas apguvēju korpusa* datiem.

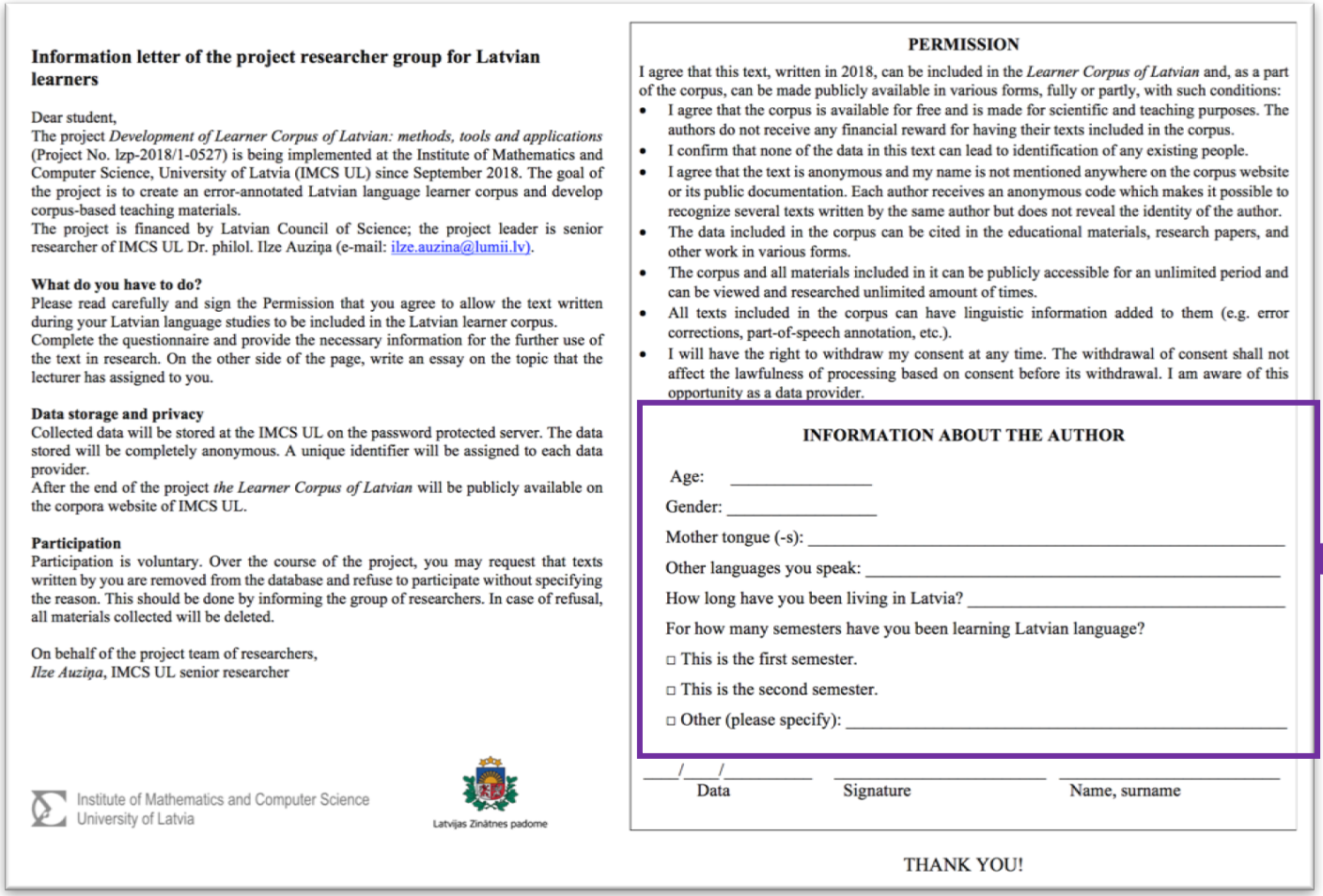
Specifiskie mērķi:

- korpusa datu vākšanai un apstrādei nepieciešamās vietnes izveide,
- tekstu digitalizēšanas, normalizēšanas un kļūdu marķēšanas metodoloģijas izstrāde,
- *Latviešu valodas apguvēju korpusa* (LaVA) izveide,
- kvantitatīva un kvalitatīva valodas apguvēju pieļauto kļūdu analīze,
- korpusā balstītu mācību materiālu un pašnovērtējuma tīmekļa platformas izstrāde.

Datu atlases kritēriji

- Latvijas augstskolās studējošo to ārvalstu studentu darbi, kas papildus savām pamatstudijām apgūst latviešu valodu.
- Latviešu valodas prasmes līmenis – A1, A2.
- Plānotais apjoms – 1000 studentu darbi (esejas), vēlamais katra teksta garums – vismaz 100 vārdi.
- Temati: *Es un mana ģimene, Mana ikdiena, Manas studijas* u.tml.

Datu izmantošanas atļauja



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

Institute of Mathematics and Computer Science
University of Latvia

Latvijas Zinātnes padome

PERMISSION

I agree that this text, written in 2018, 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): _____

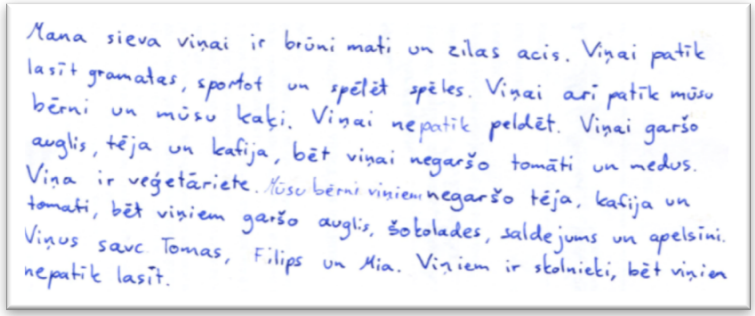
____/____/_____ _____ _____
Data Signature Name, surname

THANK YOU!

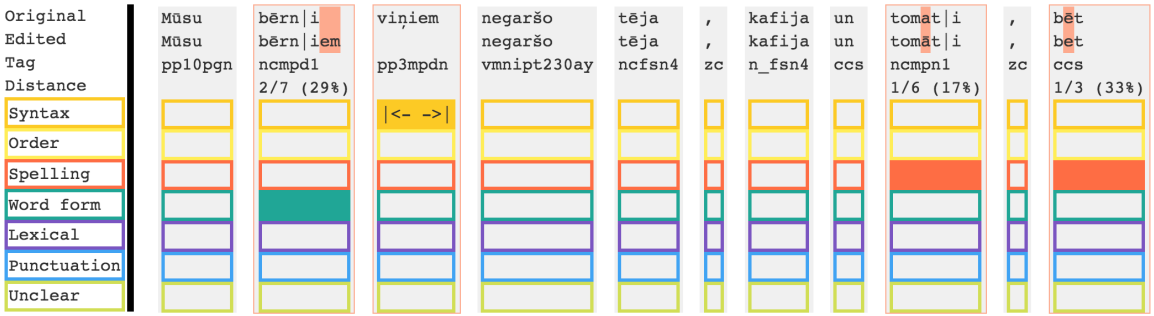
→ Metadati

Korpusa izveides posmi

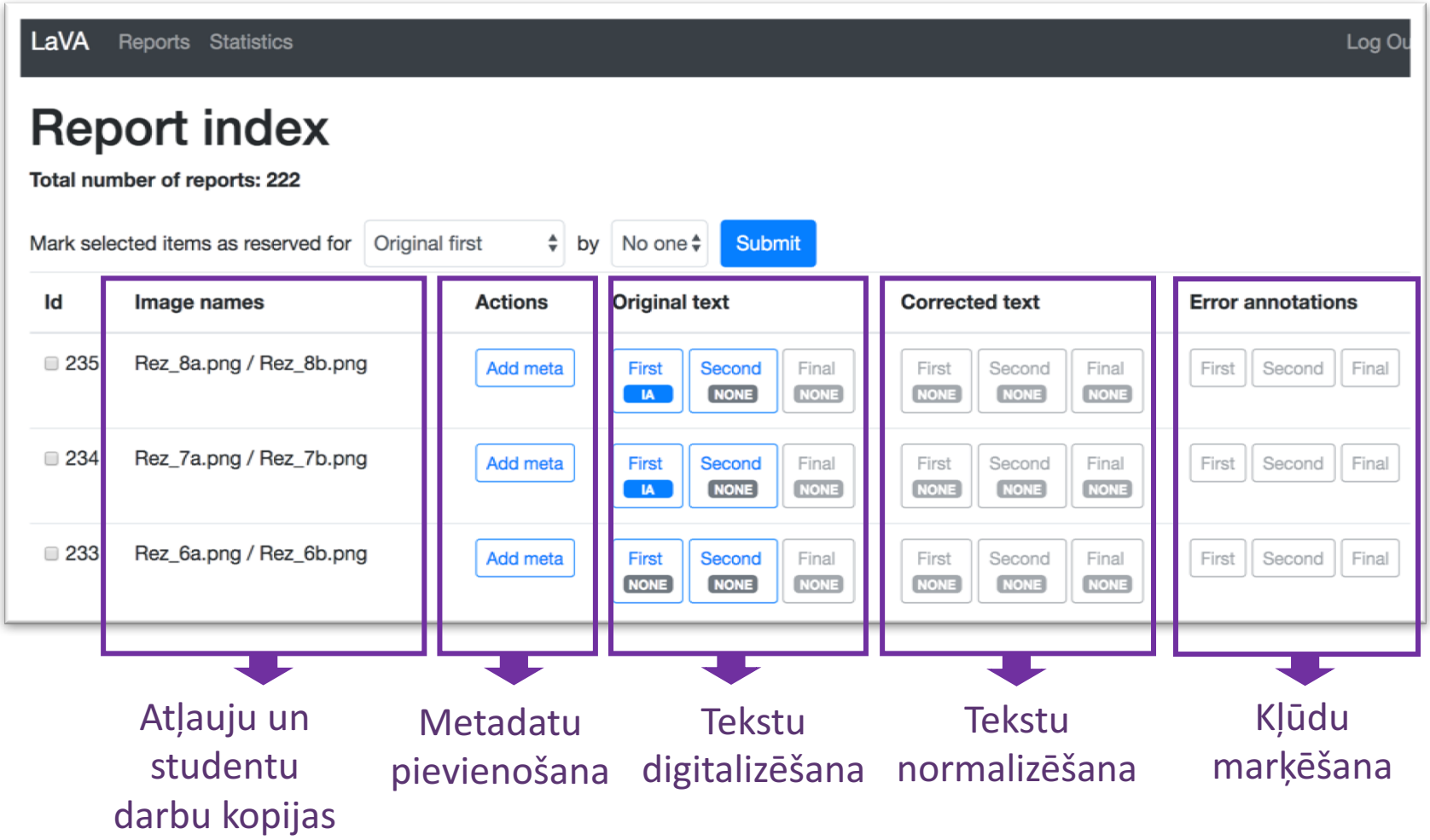
- Datu atlase un digitalizācija
- Tekstu normalizēšana (mērķa hipotēzes izvirzīšana)
- Automātiska morfoloģiskā marķēšana, tostarp vārdšķiru noteikšana, lemmatizēšana
- Oriģinālā un normalizētā teksta sastatīšana
- Automatizēta kļūdu anotēšana un manuāla pārskatīšana



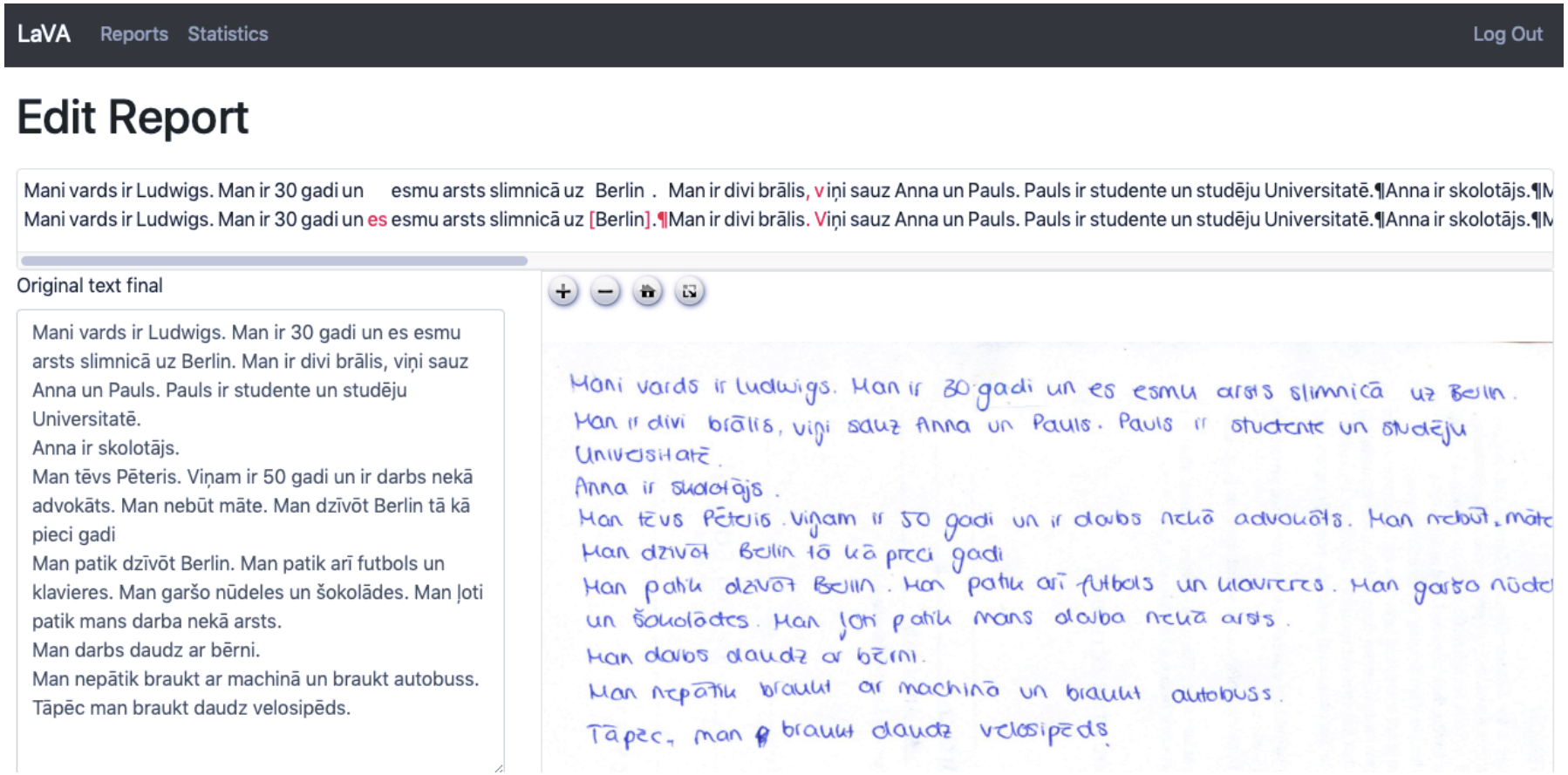
***Manai sievai** ir brūni mati un zilas acis. Viņai patīk lasīt **grāmatas**, sportot un spēlēt spēles. Viņai arī patīk mūsu bērni un mūsu kaķi. Viņai nepatīk peldēt. Viņai garšo **auglis**, tēja un kafija, **bet** viņai negaršo tomāti un medus. Viņa ir veģetāriete. Mūsu **bērniem** negaršo tēja, kafija un **tomāti**, **bet** viņiem garšo **auglis**, **šokolādes**, saldējums un apelsīni. Viņus sauc Tomass, Filips un Mia. **Viņi** ir **skolnieki**, **bet** viņiem nepatīk lasīt.*



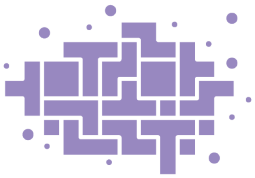
Vietne datu apstrādei



Datu digitalizēšana



Datu normalizēšana (1/3)



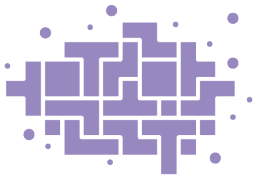
Mērķa hipotēze (vācu val. *Zielhypothese*, angļu val. *target hypothesis*)

- Priekšstats, kā ir pareizi veidota attiecīgā valodas struktūra, „valodas apguvēju izteikumu rekonstrukcija mērķvalodā" (Ellis 1994, 54; sk. arī Lüdeling et al. 2005; Siemen et al. 2006).
- LaVA korpusa uzbūve nepieļauj vairāku mērķa hipotēžu izvirzīšanu vienam un tam pašam teksta fragmentam, tāpēc par neviennozīmīgiem gadījumiem jāvienojas, sniedzot iespējami ticamāko mērķa hipotēzi.

*Vinš ir **ari** no italijas un vinš ir italis.* →
*Viņš ir **arī** no Itālijas, un viņš ir itālis.*

*Man garso **auglis**.* →
*Man garšo **auglis**.*

Datu normalizēšana (2/3)



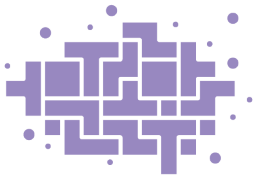
- Ne vienmēr ir iespējams izvirzīt mērķa hipotēzi, bieži ir iespējami vairāki varianti

*Es **dzivoju Essenā**, Vācijā ar manu ģimeni, bet no **pagājušāja** gada es **dzīvo** Rīgā, Latvijā **par studēt medicinu**.*

Es biju divdesmit cetri nodarbiba latviska valoda un cetridesmit stundas.

1. var. *Es biju divdesmit četrās latviešu valodas nodarbībās un četrdesmit stundās.*
2. var. *Man bija divdesmit četras latviešu valodas nodarbības un četrdesmit stundas.*

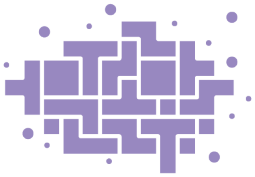
Datu normalizēšana (3/3)



Pamatprincipi

- Tiek ievērotas latviešu valodas pareizrakstības normas.
- Minimāla iejaukšanās – lai saglabātu apguvēju valodas īpatnības, tiek labots pēc iespējas mazāk.
- Netiek labots valodas stils – valodas apguves pamatlīmenī stila kļūdu ir pārāk daudz.
- Netiek labots netipiskais – ja lietojums ir netipisks, bet būtībā nav nepareizs, tas netiek mainīts. Arī netipiska vārdu secība netiek mainīta.

Kļūdu marķēšana



Kļūdu tipi

- Pareizrakstības kļūdas
  *t**ē**v – t**e**v, b**ē**t – b**e**t, bi**l**ete – bi**ļ**ete; **vis** lielākais – **vis**lielākais, l**aj**me – l**ai**me*
- Interpunkcijas kļūdas
  *Kā tev iet**.** – Kā tev iet**?** Sveiki Anna! – Sveiki**,** Anna!*
- Formveidošanas un vārddarināšanas kļūdas
  *Es dzīvoju Latvij**a**. – Es dzīvoju Latvij**ā**. **Es** gribas ēst. – **Man** gribas ēst.*
- Sintakses kļūdas
  *Apsveicu tevi **ar** svētku dienu. – Apsveicu Tevi svētku dienā!*
- Leksikas kļūdas
  *Aptauja piedalījās Latvijās **dzivotāji**. – Aptaujā piedalījās Latvijas **iedzīvotāji**.*

LATVIJAS UNIVERSITĀTE
ANNO 1919

FLPP
FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

AILab
Mākslīgā intelekta laboratorija
LU MII

Jautājumi? Komentāri?

Projekts “Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums” Nr. lzp-2018/1-0527