Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 727–731
Marseille, 20-25 June 2022

LaVA – Latvian Language Learner corpus

Roberts Darģis[1], Ilze Auziņa[1], Inga Kaija[1,2], Kristīne Levāne-Petrova[1], Kristīne Pokratniece[1]
Institute of Mathematics and Computer Science, University of Latvia[1], Rīga Stradiņš University[2]
29 Raiņa boulevard, Riga, LV-1459, Latvia, 16 Dzirciema Street, Rīga, LV-1007, Latvia
{roberts.dargis, ilze.auzina, kristine.levane-petrova, kristine.pokratniece}@lumii.lv, inga.kaija@rsu.lv

Abstract

This paper presents the Latvian Language Learner Corpus (LaVA) developed at the Institute of Mathematics and Computer Science, University of Latvia. LaVA corpus contains 1015 essays (190k tokens and 790k characters excluding whitespaces) from foreigners studying at Latvian higher education institutions and who are learning Latvian as a foreign language in the first or second semester, reaching the A1 (possibly A2) Latvian language proficiency level. The corpus has morphological and error annotations. Error analysis and the statistics of the LaVA corpus are also provided in the paper. The corpus is publicly available at: http://www.korpuss.lv/id/LaVA.

Keywords: learner corpus, acquisition, Latvian, annotated

1. Introduction

Learner corpora are computerized textual databases of the language produced by foreign language learners (Leech, 1998). Such corpora collect the language produced by people learning their first, second or foreign language, mostly the latter two (Granger, 2002); the distinction between those was explored by (Laizane, 2018).

Nowadays, a learner corpus is useful for teachers who teach a foreign language, as well as by researchers analyzing the language of the learners and developing teaching and methodological materials. Another use case for a learner corpus is training natural language processing (NLP) tools on the corpus to be later used in automatic analyses of errors occurring in the input text and in intelligent computer-assisted language learning systems (Meurers, 2015).

The 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, 2011), (Laizāne, 2017). Therefore, corpus-based and corpus-driven teaching materials are crucial for the international students that acquire Latvian both in Latvia and abroad. Learner corpora are not only used to study the language of learners, but they also have a strong connection to the development of educational applications.

This paper presents a newly created *Learner corpus of Latvian (LaVA)* (Auziņa et al., 2021) of students' essays with different language backgrounds. The development was carried out at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL), from 2018 to 2021 as a part of a research project *Development of Learner corpus of Latvian: methods, tools and applications*. The newly developed corpus is publicly available at: http://www.korpuss.lv/id/LaVA.

2. Related Work

Learner corpora have been collected and analyzed for more than 25 years now. There are many learner corpora for English (Granger et al., 2009), (Gilquin et al., 2010), and other languages are also gaining popularity. This includes French (Granfeldt et al., 2006), Swedish (Volodina et al., 2019), Norwegian (Tenfjord et al., 2006), Dutch (Lemmens and Perrez, 2010), Japanese (Gries and Adelman, 2014), Arabic (Alfaifi et al., 2014), Chinese (Wang et al., 2015), Portuguese (Mendes et al., 2016), Russian (Rakhilina et al., 2016), and others.

There have also been some noticeable developments in creating learner corpora of the Latvian language before:

- During the Latvian Language agency's research project *Quality of the Latvian language: results of the state language proficiency test*, researchers created a corpus of the texts collected from the successfully passed tests of the State Language Proficiency Testing which is used to evaluate a person's state language proficiency level. For every language proficiency level (A1, A2, B1, B2, C1, C2), 150 tests have been used. That makes in total 900 tests (Dargis et al., 2018). However, the data set is not publicly available.
- A learner Corpus of the second Baltic language (Baltic languages being Latvian and Lithuanian) was developed (www.esamkorpuss.lv) by Inga Znotiņa. The corpus Esam is a learner corpus that consists of the texts that have been written by university students, Lithuanian learners of Latvian, and Latvian learners of Lithuanian (Znotiņa, 2015), (Znotiņa, 2017).

The learner corpus LaVA is the largest learner corpus in Latvian. It is also the only corpus with manually annotated morphology. The annotation schema used in the corpus allows to perform detailed quantitative error analysis that is not common among other learner corpora.

3. Data Source

LaVA corpus contains 1015 essays (190k tokens and 790k characters excluding whitespaces) from foreigners studying at Latvian higher education institutions who are learning Latvian as a foreign language in the first or second semester, reaching the A1 (possibly A2) Latvian language proficiency level.

The most common topics of essays are: *Me and my family*, *My routine* and *My studies*. Students are asked to use fictional information instead of truth due to personal data protection.

Data comes from five universities: Riga Stradiņš University (87%), Rezekne Academy of Technologies (4%), University of Latvia (3%), Liepaja University (3%), Latvian Academy of Culture (3%).

An agreement / questionnaire form was created for corpus data collection. The form is printed on one side of an A4 size paper sheet and includes three parts – an information letter, a permission form, and a metadata collection questionnaire (information about the author). The metadata includes gender, age, mother tongue, other language knowledge and how long they have been studying Latvian. The other side of the form is blank, and authors are requested to hand-write an essay there. Copyright and personal data protection are two of the most important legal aspects that should be resolved before data collection for the learner corpus is started. The purpose of the agreement form is to inform the authors of the ways their texts are used and to receive the authors' permission to use them in the stated way (Kaija and Auzina, 2020).

4. Corpus Creation Pipeline

The corpus creation pipeline has been developed based on research teams experience of designing learner corpora of Latvian (Znotiņa, 2017), (Dargis et al., 2018), (Levane-Petrova et al., 2020). After the data is uploaded to the corpus platform, the corpus creation pipeline consists of four steps:

1. data digitization;
2. text correction;
3. morphological annotation;
4. error annotation.

Each step is done independently by two annotators and inconsistencies are finalized by a third independent annotator. Error types are automatically determined based on morphological annotations and alignment between original and corrected text. More on error analysis can be found in Section 6.

Most of the essays are handwritten, so they need to be digitized. Only last few essays are typed on a computer by the students due to distance studies during COVID-19 restrictions. Character level agreement for text digitization between the two annotators is 97.4%.

In text correction step the original text is edited to a literally correct version of the text based on assumed target hypothesis (Auzina et al., 2020). Character level agreement for text correction between the two annotators is 96.8%.

The original and corrected text is morphosyntactically annotated. The initial annotation version is generated by the IMCS morphological tagger (Paikens, 2016) and then it is manually verified by two annotators. Morphosyntactic annotations contain part-of-speech tag, lemma and other Latvian specific morphological and syntactic information. In addition, token with corrected typos is added to the original token. The alignment between original token and token with corrected typos is used in error analysis step to calculate exactly what kind of typos leaner has made. Without the token with corrected typos, it would not be possible to differentiate which differences between original and corrected token are due to typos and which ones are due to inflectional and word formation errors. Annotator agreement for all layers is 92.5% and for each layer separately the agreement ranges from 95.5% for original tag to 99.3% for corrected lemma.

The last step of corpus creation is automatic error annotation. Currently errors are classified into 6 categories, but since this step is automatic, it could be changed later on. The 6 categories are: spelling errors, inflectional and word formation errors, lexical errors, punctuation errors, syntactic errors, complex errors.

Time for each activity is automatically measured in the corpus platform. The timer automatically stops if the annotator is inactive for more than 15 seconds and the timer automatically resumes on any activity (mouse movement, keyboard input). The time it takes to complete each step is shown in Table 1. The total time required to process one essay on average is 45 minutes (the initial steps needs to be done twice).

Step	Rate (characters per minute)	Average time per step (minutes)
Initial digitization	128.9	6.1
Final digitization	370.6	2.1
Initial text correction	177.8	4.4
Final text correction	274.1	2.8
Initial annotation	95.1	8.4
Final annotation	117.1	6.6

Table 1: Time consumption per corpus creation step

5. Statistics

The metadata added shows that 63% of the essays were written by female, and 37% by male students. The authors of the texts are young people: 88% of authors are between the ages of 17 and 25, and 12 % are between the ages of 26 and 46.

Language learners have indicated 35 different mother tongues of which 22 are mentioned seven or more times. Majority of language learners (86%) have indicated only one mother tongue, 13% have indicated two and 1% have indicated three or four mother tongues. The most common mother tongue in the corpus is German (37%), followed by Swedish (11%) and Finnish (9%). Other languages indicated at least seven times are: Norwegian, Italian, Arabic, Turkish, Portuguese, Russian, Persian, Urdu, Spanish, Sinhala, French, Tamil, Hindi, Punjabi, Chinese and Flemish, Hebrew.

The corpus of LaVA contains more than 18k word forms of more than 8k words. The relative word frequencies and ranks were compared to a reference corpus to see what are the main language differences. General corpus *The Balanced Corpus of Modern Latvian (LVK2018)* (Levāne-Petrova and Darģis, 2018) were used as a reference corpus. LVK2018 contains 12M tokens and 396k word forms of 175k words. Words *būt* 'to be', *un* 'and', *bet* 'but', *ar* 'with', *uz* 'to', *no* 'from' are among top 20 words in both corpora with similar rank. As expected, words related to topics of the essays ranks much higher with significant increase in relative frequency. These are personal and possessive pronouns (*es* 'I', *mans/mana* 'my' (masculine/feminine), *mēs* 'we', *viņš* 'he', *viņa* 'she'), family members (*māsa* 'sister', *brālis* 'brother', *tēvs* 'father', *māte* 'mother') and words related to daily routine (*patikt* 'to like', *dzīvot* 'to live', *studēt* 'to study', *ēst* 'to eat', *draugs* 'friend').

6. Error Analysis

The corpus contains about 190k tokens out of which 26.3% contains errors. The relative error type breakdown of tokens with errors are shown in Table 2. Note, that the total percentage is more than 100%, because 6% of tokens had more than one error type.

Error Type	Relative frequency
Inflectional and Word Formation	45.9%
Spelling	44.6%
Lexical	16.9%
Punctuation	12.5%
Syntactic	1.4%
Complex	0.8%

Table 2: Error frequency by error type relative to tokens with errors

Further breakdown of inflectional and word formation (IWF) errors by part-of-speech type is shown in Table 3. In 44% of IWF errors word base form is used instead of the form required by the context. The base form is infinitive for verbs and nominative for nouns, pronouns, adjectives and numerals.

Part-of-speech	Relative to WF	Relative to PoS
Noun	61.7%	28.5%
Pronoun	17.7%	13.4%
Verb	9.7%	6.7%
Adjective	4.2%	31.9%
Numeral	3.9%	13.1%
Other	2.8%	2.1%

Table 3: Inflectional and word formation (IWF) error frequency by part-of-speech (PoS) relative to tokens with IWF errors and relative to all tokens with the corresponding PoS

Table 4 shows the breakdown of spelling errors. If more than one consecutive letter contained an error, each consecutive letter was counted as complex error.

Error Type	Relative frequency
Diacritical marks	78.1%
Capital letters	7.8%
Missing letter	4.1%
Redundant letter	3.6%
Complex	6.4%

Table 4: Frequency of spelling errors by error type

Standard Latvian orthography uses 22 unmodified letters of the Latin alphabet (*q*, *w*, *x*, *y* are not used) extended with 10 modified letters. Some Latvian letters are written with diacritical marks: macron indicates vowel length (*ā*, *ē*, *ī*, *ū*), palatal consonants are marked with a cedilla or a small comma placed below a letter (*ģ*, *ķ*, *ļ*, *ņ*), and some sibilants and africates are marked with corona (*š*, *ž*, *č*, *dž*).

Misuse or absence of the additional diacritical marks is the most common spelling error (78.1%). Incorrect use of short vowels *i, a, e* instead of the corresponding long vowels makes up to majority of the spelling errors (51.73%). These vowels on average are incorrectly used in 6% of cases relative to the corresponding letter frequency in the original text.

Most spelling, inflectional and word formation, and lexical errors are related to a single unit of text – usually a single word or word form – but there are analytical grammatical forms that use a standalone word to express grammatical meaning along with an auxiliary word or a word used in an auxiliary function, e.g., perfect tense forms (the indicative mood) *es esmu lasījis* 'I have read', *ir bijis* 'there have been', etc.), prepositional phrases (*ar draugiem* 'with friends', *uz skolu* 'to school', etc.). Multi-token errors in the use of analytical grammatical forms are manually marked as syntax errors. Overall these errors are rare, only 0.2% of tokens are marked in multi-token errors.

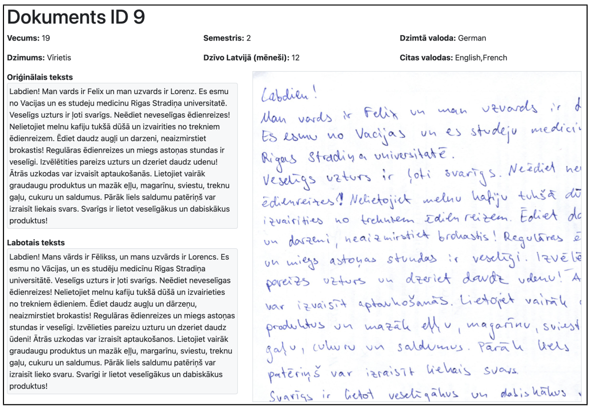
In Latvia the use of punctuation in a sentence is very strict. Punctuation is based on grammatical principles, and different use of punctuation often completely changes the meaning of a sentence. Although punctuation knowledge is not required for beginners whose essays are included in the corpus, punctuation is corrected in accordance with the Latvian language punctuation rules. This, in turn, allows automatically detect punctuation errors.

7. Published formats and interfaces

The corpus is published in the corpus homepage for easy browsing[1] (Figure 1). The homepage also provides concordancer for simple queries. More advanced queries can be constructed in the *noSketchEngine* (Rychlỳ, 2007) instance[2].

Unfortunately, some error analysis functionality is only available in the full, paid version of *SketchEngine*[3]. The files necessary for a researcher to upload the LaVA corpus in the *SketchEngine* are available in the *Download* section of the corpus homepage.

The full LaVA corpus in a comma separated format (CSV) can also be found in the *Download* section. The *essays.csv* file contains all the metadata (including URL to a image of the handwritten essay), original text and corrected text. The *annotations.csv* contains the token aligned version of original and corrected text with all the manual morphosyntactical annotations and automatically computed error type annotations.



Dokuments ID 9

Vecums: 19
Semestris: 2
Dzimtā valoda: German
Dzimums: Vīrietis
Dzīvo Latvijā (mēneši): 12
Citas valodas: English,French

Oriģinālais teksts

Labdien! Man vards ir Felix un man uzvards ir Lorenz. Es esmu no Vacijas un es studeju medicinu Rigas Stradiņa universitatē. Veselīgs uzturs ir ļoti svarīgs. Neēdiet neveselīgas ēdienreizes! Nelietojiet melnu kafiju tukšā dūšā un izvairities no trekniem ēdienreizem. Ēdiet daudz augļi un darzeni, neaizmirstiet brokastis! Regulāras ēdienreizes un miegs astoņas stundas ir veselīgi. Izvēlētities pareizs uzturs un dzeriet daudz udenu! Ātrās uzkodas var izvaisīt aptaukošanās. Lietojiet vairāk graudaugu produktus un mazāk eļļu, magarīnu, sviestu, treknu gaļu, cukuru un saldumus. Pārāk liels saldumu patēriņš var izraisīt liekais svars. Svarīgs ir lietot veselīgākus un dabiskākus produktus!

Labotais teksts

Labdien! Mans vārds ir Fēlikss, un mans uzvārds ir Lorencs. Es esmu no Vācijas, un es studēju medicīnu Rīgas Stradiņa universitātē. Veselīgs uzturs ir ļoti svarīgs. Neēdiet neveselīgas ēdienreizes! Nelietojiet melnu kafiju tukšā dūšā un izvairieties no trekniem ēdieniem. Ēdiet daudz augļu un dārzeņu, neaizmirstiet brokastis! Regulāras ēdienreizes un miegs astoņas stundas ir veselīgi. Izvēlieties pareizu uzturu un dzeriet daudz ūdeni! Ātrās uzkodas var izraisīt aptaukošanos. Lietojiet vairāk graudaugu produktus un mazāk eļļu, margarīnu, sviestu, treknu gaļu, cukuru un saldumus. Pārāk liels saldumu patēriņš var izraisīt lieko svaru. Svarīgi ir lietot veselīgākus un dabiskākus produktus!

Labdien!
Man vards ir Felix un man uzvards ir
Es esmu no Vacijas un es studeju medici
Rigas Stradiņa universitatē.
Veselīgs uzturs ir ļoti svarīgs. Neēdiet ne
ēdienreizes! Nelietojiet melnu kafiju tukšā dū
izvairities no trekniem ēdienreizem. Ēdiet da
un darzeni, neaizmirstiet brokastis! Regulāras ē
un miegs astoņas stundas ir veselīgi. Izvēlē
pareizs uzturs un dzeriet daudz udenu! Ā
var izvaisīt aptaukošanās. Lietojiet vairāk
produktus un mazāk eļļu, magarīnu, sviest
gaļu, cukuru un saldumus. Pārāk liels
patēriņš var izraisīt liekais svars.
Svarīgs ir lietot veselīgākus un dabiskākus

Figure 1: Screenshot of a corpus document

8. Conclusion

The comprehensive error analysis carried out on the corpus showed the most common mistakes made by learners from many aspect. Based on error analysis of the Latvian Language Learner corpus the self-assessment platform is developed. It contains three types of exercises (typing, inflection and gap filling). The self-assessment platform is freely available online (http://uzdevumi.riks.korpuss.lv/en/) and the interface is translated in two language – Latvian and English.

The corpus presented in this paper is freely available and will hopefully lead to many more language acquisition research both quantitative and qualitative. We believe that this analysis will be useful for teachers in language pedagogy and for authors in learning aids.

The statistics about time necessary for each actions can be used in research proposals to estimate the person-month necessary to create similar corpora.

The corpus size is sufficient to accurately represent learners' knowledge at the beginners level. The main aim for the future is to expanded the corpora with error annotated texts from more advanced language proficiency levels.

9. Acknowledgements

The work reported in this paper is a part of the project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) that is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The project is financed by Latvian Council of Science.

This work is also a part of the Latvian State Research Programme *Letonika — Fostering a Latvian and European Society* project *Research on Modern Latvian Language and Development of Language Technology* (No. VPP-LETONIKA-2021/1-0006) and has received financial support from the Latvian Language Agency through the grant agreement No. 4.6/2019-029.

10. Bibliographical References

Alfaifi, A., Atwell, E., and Hedaya, I. (2014). Arabic learner corpus (alc) v2: a new written and spoken corpus of arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014*, volume 2, pages 77–89. Kobe International Communication Center.

Auzina, I., Kaija, I., and Levane-Petrova, K. (2020). Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā. *Valoda: nozīme un forma*, 11:7–26.

Dargis, R., Auzina, I., and Levane-Petrova, K. (2018). The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 4111–4115.

Gilquin, G., De Cock, S., and Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Presses universitaires de Louvain (Louvain-La-Neuve).

Granfeldt, J., Nugues, P., Persson, E., Thulin, J., Ågren, M., and Schlyter, S. (2006). Cefle and direkt profil: A new computer learner corpus in french l2 and a system for grammatical profiling. In *LREC-2006,*

[1] LaVA corpus homepage: http://lava.korpuss.lv/

[2] LaVA in *noSketchEngine*: http://nosketch.korpuss.lv/#dashboard?corpname=lava

[3] SketchEngine: https://www.sketchengine.eu/

The fifth international conference on Language Resources and Evaluation, pages 565–570. ELRA.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., et al. (2009). *International Corpus of Learner English*. Presses universitaires de Louvain Louvain-la-Neuve.

Granger, S. (2002). A Bird's-Eye View of Learner Corpus Research. *Computer learner corpora, second language acquisition and foreign language teaching*, 6:3–33.

Gries, S. T. and Adelman, A. S. (2014). Subject Realization in Japanese Conversation by Native and Non-native Speakers: Exemplifying a New Paradigm for Learner Corpus Research. In *Yearbook of Corpus Linguistics and Pragmatics 2014*, pages 35–54. Springer.

Kaija, I. and Auzina, I. (2020). Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection. In *Selected papers from the CLARIN Annual Conference 2019*, pages 41–47.

Laizāne, I. (2017). Acquisition of Latvian as a Foreign Language in Latvia: Development and Trends. *Rural environment. Education. Personality*, pages 116–120.

Laizane, I. (2018). The Understanding of the Concepts of First Language, Second Language and Foreign Language Outside of Latvia. In *Rural Environment. Education. Personality.(REEP). Proceedings of the International Scientific Conference (Latvia)*, pages 81–87. Latvia University of Life Sciences and Technologies.

Leech, G. (1998). Learner corpora: what they are and what can be done with them. *Learner English on Computer. Addison Wesley Longman, London and New York*.

Lemmens, M. and Perrez, J. (2010). On the use of posture verbs by French-speaking learners of Dutch: A corpus-based study. *Cognitive Linguistics*, 21(2).

Levane-Petrova, K., Auzina, I., and Pokratniece, K. (2020). Latviešu valodas apguvēju korpusa datu ieguves un apstrādes metodoloģijas izstrāde. In *Valodu apguve: problēmas un perspektīva*, volume 16, pages 299–309. LiePA.

Mendes, A., Antunes, S., Jansseen, M., and Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. In *Proceedings of the Tenth Language Resources and Evaluation Conference–LREC'16*, pages 3207–3214. European Language Resources Association.

Meurers, D. (2015). Learner Corpora and Natural Language Processing. *The Cambridge handbook of learner corpus research*, pages 537–566.

Paikens, P. (2016). Deep Neural Learning Approaches for Latvian Morphological Tagging. In *Human Language Technologies – The Baltic Perspective*, volume 289. IOS Press.

Rakhilina, E. V., Vyrenkova, A., Mustakimova, E., Ladygina, A., and Smirnov, I. (2016). Building a Learner Corpus for Russian. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 66–75.

Rychlỳ, P. (2007). Manatee/Bonito-A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno: Masaryk University.

Šalme, A. (2011). *Latviešu valodas kā svešvalodas apguves pamatjautājumi*. Latviešu Valodas Aģentūra.

Tenfjord, K., Meurer, P., and Hofland, K. (2006). The ask corpus-a language learner corpus of norwegian as a second language. In *LREC*, volume 6, pages 1821–1824.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., et al. (2019). The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Wang, M., Malmasi, S., and Huang, M. (2015). The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Znotiņa, I. (2015). Learner Corpus Annotation in Latvia and Lithuania. *Darnioji daugiakalbystė, No. 7*, pages 145–159.

Znotiņa, I. (2017). Computer-Aided Error Analysis for Researching Baltic Interlanguage. In *Rural Environment, Education, Personality. Proceedings of the 10th International Scientific Conference*, pages 238–244.

11. Language Resource References

Auziņa, Ilze and Kaija, Inga and Levāne-Petrova, Kristīne and Pokratniece, Kristīne and Darģis, Roberts. (2021). *Latvian Learner Corpus (LaVa)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/42.

Levāne-Petrova, Kristīne and Darģis, Roberts. (2018). *Balanced Corpus of Modern Latvian (LVK2018)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/11.