Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5123–5129
Marseille, 20-25 June 2022

Latvian National Corpora Collection – Korpuss.lv

Baiba Saulīte*, Roberts Darģis*, Normunds Grūzītis*, Ilze Auziņa*, Kristīne Levāne-Petrova*, Lauma Pretkalniņa*, Laura Rituma*, Pēteris Paikens*, Artūrs Znotiņš*, Laine Strankale*, Kristīne Pokratniece*, Ilmārs Poikāns*, Guntis Bārzdiņš*, Inguna Skadiņa*, Anda Baklāne†, Valdis Saulespurēns†, Jānis Ziediņš‡

*Institute of Mathematics and Computer Science, University of Latvia (IMCS UL)
Raina bulv. 29, LV-1459, Riga, Latvia
{baiba.valkovska, roberts.dargis, normunds.gruzitis}@lumii.lv

†National Library of Latvia (NLL)
Mukusalas iela 3, LV-1048, Riga, Latvia
{anda.baklane, valdis.saulespurens}@lnb.lv

‡Culture Information Systems Centre (CISC)
Terbatas iela 53-2, LV-1011, Riga, Latvia
janis.ziedins@kis.gov.lv

Abstract

LNCC is a diverse collection of Latvian language corpora representing both written and spoken language and is useful for both linguistic research and language modelling. The collection is intended to cover diverse Latvian language use cases and all the important text types and genres (e.g. news, social media, blogs, books, scientific texts, debates, essays, etc.), taking into account both quality and size aspects. To reach this objective, LNCC is a continuous multi-institutional and multi-project effort, supported by the Digital Humanities and Language Technology communities in Latvia. LNCC includes a broad range of Latvian texts from the Latvian National Library, Culture Information Systems Centre, Latvian National News Agency, Latvian Parliament, Latvian web crawl, various Latvian publishers, and from the Latvian language corpora created by Institute of Mathematics and Computer Science and its partners, including spoken language corpora. All corpora of LNCC are re-annotated with a uniform morpho-syntactic annotation scheme which enables federated search and consistent linguistics analysis in all the LNCC corpora, as well as facilitates to select and mix various corpora for pre-training large Latvian language models like BERT and GPT.

Keywords: Latvian language, text corpora, spoken language, Universal Dependencies, federated search

1. Introduction

Latvian corpora are increasingly used for large language model pre-training such as LVBERT (Znotins and Barzdins, 2020), LitLatBERT (Ulcar et al., 2021) and GPT2-LV (Plenert, 2021). The crucial zero-shot learning capability of the large language models depends not only on the model and data size but also on the quality and the encyclopaedic knowledge coverage of the training corpora, giving rise to the term 'GoodData' (Press, 2021). The groundbreaking GPT-3 (Brown et al., 2020) language model was trained on 750GB of mostly English GoodData, while the GPT-SW3 model for the relatively "small" Swedish language – on a 100GB text collection (Ekgren et al., 2022). Latvian National Corpora Collection (LNCC) with its current aggregated size of nearly 10GB and broad coverage is a step towards Latvian GoodData suitable for training high quality Latvian large language models essential for various downstream tasks (particularly zero-shot NLU and NLG – the future of NLP).

Since Latvian is a relatively less-resourced language and even smaller than Swedish, there will probably never be a single corpus available, in terms of size or quality, to pre-training equally large language models if compared, for instance, to the GPT-3 model for English. Also, no single balanced Latvian text corpus is sufficient for the modern lexicographic needs and grammar studies, since new text types and sources (user-generated content, spoken language, etc.) and specialised domains are insufficiently covered.

Therefore we see the Latvian National Corpus as a diverse and open-ended corpus collection which continues to evolve and enlarge through multi-institutional and multi-project efforts. Moreover, there is no single endpoint of LNCC in terms of institutional corpus platform instances: each member of the LNCC consortium runs its own instance or cooperates with other members, and the consortium decides which corpora are appropriate (w.r.t. type, quality, size, status) to be included in the collection. Nevertheless, users of LNCC can choose to work only with a subcollection based on the common set of high-level LNCC metadata tags. Apart from the meta-tags, all LNCC corpora are uniformly (re-)processed – tokenized, morphologically and syntactically tagged – to ensure consistent querying results and language modelling across all selected corpora of all LNCC endpoints.

The rest of this paper is structured as follows. After briefly mentioning related work in Section 2., Section 3. serves as a description and index of 21 Latvian language corpora (developed by 11 institutions) currently included in LNCC. Section 4. introduces the common morpho-syntactic and spoken language tagsets used to (re-)annotate the corpora. Section 5. outlines the simple but efficient implementation of federated search within LNCC, and Section 6. concludes the paper highlighting the future perspective of LNCC.

2. Related Work

National corpora have been created for many languages. A national corpus can be a single corpus like BNC (Consortium, 2007), or it may be a collection of different types of

corpora like the Czech National Corpus (Křen, 2020), Bulgarian National Corpus (Koeva et al., 2012). LNCC follows the latter approach.

A national corpus (also, a reference corpus) may consist of different types of texts that are not necessarily balanced. Such national or reference corpora have been created for American English (Ide and Suderman, 2006), Croatian (Tadić, 2002), Czech (Křen et al., 2016), German (Kupietz et al., 2018), Hungarian (Oravecz et al., 2014), Polish (Przepiórkowski et al., 2011), Romanian (Mititelu et al., 2018), Turkish (Aksan et al., 2012), and other languages. LNCC as a whole collection is a reference corpus.

Conceptually, the idea of LNCC is somewhat similar to the Leipzig Corpus Collection, particularly its Deutscher Wortschatz sub-collection which focuses on the German language (Goldhahn et al., 2012). The main differences are that LNCC includes not only web-crawled corpora in the collection but also other previously or recently created Latvian language corpora, both general and specialised, covering various time periods. Also the aim of LNCC is more general, supporting various use cases apart from the lexicographic use case.

3. Corpus Collection

LNCC consists of wide variety of corpora.[1] Currently, 21 text and spoken corpora (total size 1.3B tokens) representing different types and genres are available.

Text corpora are widely represented in this collection (see Table 1): LVK2018, UDLV-LVTB, Hugo.lv, Tīmeklis2007, Tīmeklis2020, Vikipēdija, Emuāri, Barometrs, Saeima, Likumi, LiLa, MuLa, LaVA, Pārspriedumi, Disertācijas, LatSenRom, Rainis, Senie. LNCC also contains three corpora of spoken language: LRK2013 (Pinnis et al., 2014), LVMED (Dargis et al., 2020b), and Subtitri. These corpora currently include only transcriptions of speech – the aligned audio recordings are not available via the current LNCC user interface. Note that the Corpus of Saeima (transcriptions of parliamentary debates, years 1993–2017) can be considered also as a corpus of edited spoken language.

LVK2018 is designed as a general language, representative and balanced 10 million word corpus of contemporary Latvian that aims to cover the variety of existing texts in certain estimated proportions (Levane-Petrova, 2019). It is used as a data source for the continuous development of a balanced multilayer (UD, FrameNet, PropBank, AMR, as well as named entity and coreference layers) corpus of Latvian (Gruzitis et al., 2018).The general spoken language corpus LRK2013 has been designed to be phonetically balanced and representative in term of speakers and types of speech acts (Pinnis et al., 2014). The balanced corpus of contemporary Latgalian texts (MuLa) consists of certain proportions of texts published in Latgalian. Domain, genre etc. specific corpora are: a learner corpus LaVA, parliamentary corpus Saeima (Dargis et al., 2018), literary corpora (Rainis, LatSenRom), and other specialised corpora – Pārspriedumi (Levāne-Petrova and Pokratniece, 2021), Saeima Disertācijas, Likumi, Vikipēdija, Barometrs, Emuāri, Subtitri, LVMED (Dargis et al., 2020b). Two comprehensive web corpora are also available for the Latvian language: Tīmeklis2007 (Dzerins and Dzonsons, 2007) and Tīmeklis2020.

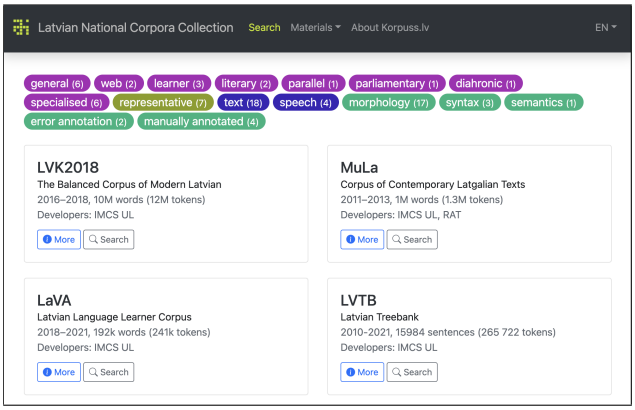

Figure 1: Screenshot of the Korpuss.lv website: the most popular corpora are listed first (see Table 1 for details).

Almost all corpora included in LNCC are monolingual; only one parallel corpus (its Latvian counterpart) has currently been added to LNCC – the LiLa corpus (Utka et al., 2012); however, more parallel corpora will be included in collaboration with Culture Information Systems Centre.

LNCC is mainly intended for use in synchronic research as most of the corpora are synchronic corpora of contemporary language – texts included in them cover the period from the 1990s to the 20s of the 21st century. An earlier period (1873–1940) is covered by literary corpora (Rainis, LatSenRom). In addition, LNCC includes also one diachronic corpus – Senie (Andronova, 2007). The corpus of the early written Latvian covers both printed texts and some manuscript transcripts of the 16th–18th century. All the texts were transliterated from the Gothic to the Latin script. All the texts were later converted into Unicode, and transliteration of the old spelling into the modern spelling has been started in 2021. This will allow for experimental automatic morpho-syntactic annotation of the diachronic corpus as well.

During the development of the LNCC morphological annotations were uniformly annotated across all the corpora. Corpora with no annotations were morphologically annotated for the first time and corpora with older annotations were re-annotated with the newest tagger to have the same tagset across all corpora.

Almost all corpora included in LNCC are automatically morphologically annotated and UD-parsed (Nivre et al., 2020). Two of them have been manually validated: the UDLV-LVTB treebank (Pretkalniņa et al., 2018) and the learner corpus of non-native Latvian speakers LaVA with manual error and morphological annotation (Dargis et al., 2020a). The common morpho-syntactic annotation scheme does not cover morphological features present in the Latgalian language (dialectal), early written Latvian (historical) and news portal comments (ungrammatical), therefore the MuLa, Senie and Barometrs corpora are not annotated and parsed.

The LNCC corpus platform (Figure 1) lists all the corpora and provides filtering by meta-tags: type of data included in

[1] http://korpuss.lv

Code name	Full name	Size	Type	Release
Written language text corpora				
LVK2018	Balanced Corpus of Modern Latvian (Levāne-Petrova and Darģis, 2018)	12M tokens	text, general, representative	2016–2018
UDLV-LVTB	Latvian UD Treebank, subset of LVK2018, part of UD v2.9 (Zeman et al., 2021)	266k tokens (16k sent.)	text, general, representative, manually annotated	2015–2021
Hugo.lv	Hugo.lv Parallel Corpora	10.5M tokens	text, general, culture	2018
Tīmeklis2007	Latvian Web Corpus (Džeriņš and Džonsons, 2007)	123.5M tokens	text, web	2006–2007
Tīmeklis2020	Latvian Web Corpus	492.6M tokens	text, web	2020–2022
Vikipēdija	Latvian Wikipedia	27.7M tokens	text, specialised	2022
Emuāri	Latvian Blog Corpus	8M tokens	text, specialised	2014–2015
Barometrs	Corpus of News Portal Comments	447.3M tokens	text, specialised	2011–2021
Saeima	Corpus of Latvian Parliament Debates (Auziņa et al., 2018)	24M tokens	text, specialised:parliamentary	2013–2018
Likumi	Corpus of Legal Acts of the Republic of Latvia	116.2M tokens	text, specialised	2022
LiLa	Lithuanian-Latvian-Lithuanian Parallel Text Corpus (Utka et al., 2013)	5.7M tokens	text, parallel, representative	2011–2013
MuLa	Corpus of Contemporary Latgalian Texts (Sperga et al., 2013)	1.3M tokens	text, specialised:dialect, representative	2011–2013
LaVA	Latvian Language Learner Corpus (Auziņa et al., 2021)	241k tokens	text, specialised:learner, manually annotated, error annotation	2018–2021
Pārspriedumi	Corpus of Students Essays (Levāne-Petrova et al., 2021)	226k tokens	text, specialised	2018–2021
Disertācijas	Corpus of Latvian PhD Theses	23.4M tokens	text, specialised	2022
LatSenRom	Corpus of Latvian Early Novels	3.3M tokens	text, specialised:literary	2019–2021
Rainis	Corpus of Texts Written by Rainis (Spektors et al., 2018)	2.3M tokens	text, specialised:literary	2018
Senie	Corpus of Early Written Latvian Texts (Andronova et al., 2002)	2.7M tokens	text, specialised:diachronic	2002–2021
Spoken language text corpora				
LRK2013	Latvian Speech Recognition Corpus	975k tokens (100 hours)	spoken, general, representative	2013
Subtitri	Latvian Subtitles of Public Broadcasting	10.8M tokens (1200 hours)	spoken, specialised	2020–2022
LVMED	Latvian Medical Speech Corpus	157k tokens (35 hours)	spoken, specialised	2022
LNCC:	**21 corpora**	**1.3B tokens**		**Jun 2022**

Table 1: The current compilation of LNCC.

the corpus (written vs. spoken); type of corpus (general vs. specialised; some specialised corpora are divided in more detail – learner, literary, dialectal, parliamentary); annotation levels (morphology, syntax, manually annotated, error annotation).

4. Common Tagsets

All LNCC datasets are automatically tokenized and morphologically tagged (with few exceptions as mentioned in Section 3.). In general, tokens are separated from each other by whitespace with some language and domain specific exceptions, e.g. ordinal numerals, like abbreviations, are written in Latvian together with the full-stop mark (e.g. '1.' – 'first') therefore an ordinal numeral together with the full-stop mark is tokenized as one token. Systematic tokenization phenomena of the written language – numbers, dates, URLs, email addresses, initials, etc. – are identified using regular expressions.

In spoken language corpora, orthographic transcription is used, taking into account the basic principles of ortho-

graphic annotation, e.g., the norms of the standard orthography of the Latvian language are followed (in the case of significant deviation from the norm, both the correct and incorrect forms are given); capital letters are used in proper names and acronyms only; the numerals and abbreviations are expanded; non-verbal elements, unclear speech and physiological noise (e.g., snuffling, smacking, coughing, etc.) are annotated.

Morpho-syntactic annotation is done by the open-source IMCS UL morphological tagger which ensures 92.7% full morphological tag accuracy and 97.6% lemmatization / POS (part-of-speech) accuracy (Paikens et al., 2013; Paikens, 2016).

Latvian has a classic Indo-European (Baltic) system with diverse grammatical inflection and extensive word formation (Vanags, 2021). Word order is relatively free, i.e. pragmatically governed, however, the basic word order is subject-verb-object. Due to this, morpho-syntactic annotation contains not only POS and lemma, but also case, number, tense and various language-specific attributes.

The common tagset of Latvian has been developed and fine-tuned at IMCS UL over the years. It is a positional tagset, generally compliant with the MULTEXT-EAST standard (Erjavec, 2012), adapted to the Latvian specifics. The tagset contains 13 POS classes: 10 POS classes correspond to the word classes defined in Latvian Grammar (Kalnaca and Lokmane, 2021) – 5 for inflected word classes (nouns, adjectives, verbs, pronouns, numerals) and 5 for non-inflected word classes (adverbs, prepositions, particles, conjunctions, interjections); the tagset also contains 3 POS classes for abbreviations, punctuation and residuals.

Latvian nouns inflect for number and case, adjectives inflect for case, number, gender and definiteness, and verbs inflect for tense, mood, voice and person (Nau, 1998). By representing this information in a positional tagset, the length of the tag can vary greatly – from one (e.g. for particles) to 11 for verbs.

Latvian Treebank (LVTB) is manually syntactically annotated using a hybrid dependency-constituency grammar model (Barzdins et al., 2007; Nespore et al., 2010; Pretkalnina et al., 2011) and then transformed to the UD model (Pretkalnina et al., 2018). Other corpora are automatically UD-parsed using a BERT-based parser for Latvian, trained on the UDLV-LVTB treebank, that gives the labelled attachment score (LAS) of 89.9% (Znotins and Barzdins, 2020). The parser is periodically re-trained on the latest UDLV-LVTB data that follows the global UD release cycle.

5. Federated Search

Open-access federated search facility is available through the LNCC website.[2] It gives an overview about the absolute and relative (per million) frequency of a given search term across all the LNCC corpora. Each 'tile' in the result set (as illustrated in Figure 2) leads to a detailed search result in the form of a corpus concordance.

The federated search combines multiple corpora from multiple corpus indexer instances (endpoints) maintained by

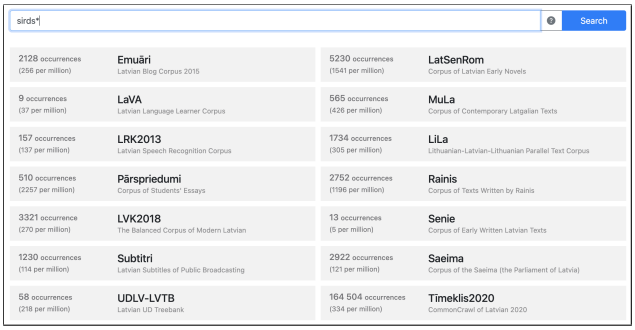

Figure 2: Screenshot of a federated search result for the search term *sirds** ('heart*').

different organisations (members of the informal LNCC consortium). Currently, all the three endpoints of LNCC – the IMCS UL endpoint, the NLL endpoint and the CISC endpoint – use the NoSketch Engine platform (Rychlỳ, 2007), and the current implementation of the federated search relies on the NoSketch Engine API, but the functionality can be easily extended to support other corpus indexers if necessary.

The list of corpora included in LNCC and, thus, in the federated search is carefully curated to maintain the representativity and quality of LNCC. The minimum requirement (apart from general content quality requirements) to include a Latvian language corpus in LNCC is open access to the corpus, even if it is not available as open data. Compliance to the common morpho-syntactic tagset is preferred but not mandatory for basic federated search.

6. Conclusion

We have presented the conception and the current content of the Latvian National Corpus Collection (LNCC). As a whole, it can be seen as an open-ended national reference corpus; open-ended in a sense that more corpora developed through different projects can be added any time by the LNCC partners.

Compliance to the common tagset and the simple yet efficient federated search facility has instantly paid off. The otherwise separate and diverse corpora have become easily accessible through a simple and unified search interface. Because of the much lower entry barrier and much better discoverability, the number of queries and sessions have significantly increased for all LNCC corpora.

We have also organised a CLARIN-LV (Skadiņa et al., 2020) K-centre on-line tutorial and user study sessions for linguists (researchers and teachers). Although most of the attendees had already used some Latvian language corpora, they were surprised that many more corpora are available. Even if they primarily use a certain corpus or few corpora, the federated search allows for quick quantitative comparison across the whole corpora collection. Of course, there has been already available the pan-European federated content search via CLARIN,[3] which has also been disseminated though the CLARIN-LV K-centre seminars, however,

[2] http://korpuss.lv/search

[3] https://clarin.eu/content/content-search

it is less convenient and less efficient for the users when it comes to everyday use.

In this paper and in LNCC so far, we have focused on text corpora, including spoken language text corpora. Inclusion of full-fledged Latvian speech corpora in LNCC is a near future task. We will also add a new, extended (100M tokens) version of the Balanced Corpus of Modern Latvian will be released by the end of 2022.

Most of the current LNCC corpora have been developed and deployed by IMCS UL in close collaboration with other academic institutions and private companies: Riga Stradins University, Latvian Language Institute UL, Rezekne Academy of Technologies, Liepaja University, Vytautas Magnus University, LETA Ltd., Tilde Ltd., Riga East University Hospital. National Library of Latvia (NLL) and Culture Information Systems Centre (CISC) maintain their own corpus indexers, providing separate but integrated endpoints for the federated search of LNCC, while the NLL and CISC corpora are being processed and annotated using the common tagset and the NLP pipeline provided by IMCS UL. Other endpoints and selected corpora can be added to LNCC in the future.

Acknowledgements

This work has received financial support from the National Research Programmes "Digital Resources of the Humanities" (grant agreement No. VPP-IZM-DH-2020/1-0001: development of Korpuss.lv), "Letonika – Fostering a Latvian and European Society" (grant agreement No. VPP-LETONIKA-2021/1-0006: data-driven linguistic studies), from the Latvian Language Agency (grant agreement No. 4.6/2019-029: creation of LVK2022), and from the European Regional Development Fund (grant agreement No. 1.1.1.1/18/A/153: spoken language corpora).

7. Bibliographical References

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., Yılmazer, H., Atasoy, G., Öz, S., Yıldız, İ., et al. (2012). Construction of the Turkish National Corpus (TNC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3223–3227.

Andronova, E. (2007). The Corpus of Early Written Latvian: Current State and Future Tasks. In *Proceedings of the Corpus Linguistics Conference (CL)*. University of Birmingham, UK.

Barzdins, G., Gruzitis, N., Nespore, G., and Saulite, B. (2007). Dependency-based hybrid model of syntactic analysis for the languages with a rather free word order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, pages 13–20, Tartu, Estonia.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Consortium, B. (2007). British national corpus, XML edition. Oxford Text Archive.

Dargis, R., Auzina, I., Bojars, U., Paikens, P., and Znotins, A. (2018). Annotation of the Corpus of the Saeima with Multilingual Standards. In *Proceedings of the 2018 ParlaCLARIN Workshop*.

Dargis, R., Auzina, I., Levane-Petrova, K., and Kaija, I. (2020a). Quality Focused Approach to a Learner Corpus Development. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 392–396.

Dargis, R., Gruzitis, N., Auzina, I., and Stepanovs, K. (2020b). Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 135–141. IOS Press.

Dzerins, J. and Dzonsons, K. (2007). Harvesting National Language Text Corpora from the Web. In *Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT)*.

Ekgren, A., Gyllensten, A. C., Gogoulou, E., Heiman, A., Verlinden, S., Öhman, J., Carlsson, F., and Sahlgren, M. (2022). Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language resources and evaluation*, 46(1):131–142.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 759–765, Istanbul, Turkey.

Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., and Paikens, P. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 4506–4513, Miyazaki, Japan.

Ide, N. and Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Kalnaca, A. and Lokmane, I. (2021). *Latvian Grammar*. University of Latvia Press.

Koeva, S., Stoyanova, I., Leseva, S., Dekova, R., Dimitrova, T., and Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, (1):65–110.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petke-

vic, V., Procházka, P., et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528.

Křen, M. (2020). Czech National Corpus in 2020: Recent Developments and Future Outlook. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 52–57.

Kupietz, M., Lüngen, H., Kamocki, P., and Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Levane-Petrova, K. (2019). Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Language: Meaning and Form*, 10:131–146. The Balanced Corpus of Modern Latvian, its role in grammar studies.

Levāne-Petrova, K. and Pokratniece, K. (2021). Skolēnu pārspriedumu korpusa izveide. In *Latviešu valodas apguve: XIII Starptautiskais baltistu kongress*. LiePA.

Mititelu, V. B., Tufiş, D., and Irimia, E. (2018). The reference corpus of the contemporary romanian language (corola). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nau, N. (1998). *Latvian*, volume 217. Lincom Europa.

Nespore, G., Saulite, B., Barzdins, G., and Gruzitis, N. (2010). Comparison of the SemTi-Kamols and Tesniere's dependency grammars. In *Human Language Technologies – The Baltic Perspective*, volume 219 of *Frontiers in Artificial Intelligence and Applications*, pages 233–240. IOS Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Oravecz, C., Váradi, T., and Sass, B. (2014). The Hungarian Gigaword corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Paikens, P., Rituma, L., and Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 267–277, Oslo, Norway.

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

Pinnis, M., Auzina, I., and Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*.

Plenert, A. (2021). gpt2-lv. In *Hugging Face*. https://huggingface.co/aidan-plenert-macdonald/gpt2-lv edition.

Press, G. (2021). Andrew Ng Launches A Campaign For Data-Centric AI. *Forbes*, Jun 16, 2021.

Pretkalnina, L., Nespore, G., Levane-Petrova, K., and Saulite, B. (2011). A Prague Markup Language profile for the SemTi-Kamols grammar model. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 303–306, Riga, Latvia.

Pretkalnina, L., Rituma, L., and Saulite, B. (2018). Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. In *Text, Speech, and Dialogue*, volume 11107, pages 95–105. Springer.

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., Łaziński, M., and Pezik, P. (2011). National Corpus of Polish. In *Proceedings of the 5th Language & Technology Conference: Human language technologies as a challenge for computer science and linguistics*, pages 259–263.

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.

Skadiņa, I., Auziņa, I., Grūzītis, N., and Znotiņš, A. (2020). Clarin in Latvia: From the preparatory phase to the construction phase and operation. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pages 342–350.

Tadić, M. (2002). Building the Croatian National Corpus. In *LREC2002 Proceedings, Las Palmas, ELRA, Pariz-Las Palmas*, volume 2, pages 441–446.

Ulcar, M., Zagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., and Robnik-Sikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *CoRR*, abs/2107.10614.

Utka, A., Levane-Petrova, K., Bielinskiene, A., Kovalevskaite, J., Rimkute, E., and Vevere, D. (2012). Lithuanian-Latvian-Lithuanian Parallel Corpus. In *Human Language Technologies - The Baltic Perspective*, volume 247. IOS Press.

Vanags, P. (2021). Latviešu valoda. In *Nacionālā Enciklopēdija*. Latvijas Nacionālā enciklopēdija.

Znotins, A. and Barzdins, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.

8. Language Resource References

Andronova, Everita and Spektors, Andrejs and Vanags, Pēteris and Baltiņa, Maija and Trumpa, Anta and Trumpa, Edmunds and Grūzītis, Normunds and Siliņa-Piņķe, Renāte and Frīdenberga, Anna and Skrūzmane, Elga and Ķauķīte, Sintija. (2002). *The Corpus of Early Written Latvian (Senie)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/12.

Auziņa, Ilze and Darģis, Roberts and Bojārs, Uldis and Paikens, Pēteris and Znotiņš, Artūrs and Rābante-Buša, Guna. (2018). *Corpus of the Saeima (the Parliament of Latvia) (Saeima)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/50.

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.

Džeriņš, Jānis and Džonsons, Kristaps. (2007). *Latvian Web Corpus 2007 (Timeklis2007)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/46.

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.

Levāne-Petrova, Kristīne and Pokratniece, Kristīne and Darģis, Roberts. (2021). *Corpus of Students' Essays (Pārspriedumi)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/51.

Spektors, Andrejs and Grūzītis, Normunds and Darģis, Roberts and Auziņa, Ilze and Saulīte, Baiba and Levāne-Petrova, Kristīne. (2018). *Rainis*. CLARIN-LV digital library at IMCS, University of Latvia.

Sperga, Ilze and Pokratniece, Kristīne and Briška, Anna. (2013). *Latgalian Corpus (MuLa)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/8.

Utka, Andrius and Levāne-Petrova, Kristīne and Vēvere, Daira and Rābante-Buša, Guna and Kovalevskaitė, Jolanta and Rimkutė, Erika. (2013). *Lithuanian-Latvian-Lithuanian Parallel Corpus (LiLa)*. CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/6.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia. (2021). *Universal Dependencies 2.9*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-4611.