Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200614

Lessons Learned from Creating a Balanced Corpus from Online Data

Roberts DARĢIS [1], Kristīne LEVĀNE-PETROVA, and Ilmārs POIKĀNS
Institute of Mathematics and Computer Science, University of Latvia, Latvia

Abstract. This paper describes lessons learned from developing the most recent Balanced Corpus of Modern Latvian (LVK2018) from various online sources. Most of the new corpora are created from data obtained from various text holders, which requires cooperation agreements with each of the text holders. Reaching these cooperation agreements is a difficult and time consuming task and may not be necessary if the resource to be created is not of hundred millions of size. Although there are many different resources available on the Internet today for a particular language, finding viable online resources to create a balanced corpus is still a challenging task. Developing a balanced corpus from various online sources does not require agreements with text holders, but it presents many more technical challenges, including text extraction, cleaning and validation.

Keywords. Balanced corpus, general corpus, corpus development, metadata

1. Introduction

Nowadays, the research of different scientific disciplines would not be possible without the use of corpora, especially a reference corpus, that is designed to provide comprehensive information about a language [1].

A corpus is used in linguistics to conduct language research, create dictionaries and grammars; in sociology to analyze mass opinion and behavior and in computer science to develop natural language processing components, such as machine translation, speech recognition and various text taggers.

Most of the new corpora are created from data obtained from various text holders that makes corpus creation much more easier, because the texts are of high quality, in easily parsable formats with structured metadata [2], [3]. Obtaining cooperation agreements from different text holders is a difficult and time consuming task and may not be necessary if the resource to be created is not of hundred millions or even billions of size.

This paper describes the development of the latest corpus in the LVK series. The Balanced Corpus of Modern Latvian (LVK2018) [4] is a new 10 million representative corpus of contemporary Latvian, created mostly from various online sources. Although there are many different resources available on the Internet today for a particular language, finding viable online resources to create a balanced corpus is still a challenging

[1]Corresponding Author: Roberts Darģis, Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Riga, LV-1459, Latvia; E-mail: roberts.dargis@lumii.lv.

task. Extracting text and metadata is a much more complicated task, and much more effort is required to clean and validate the data. This paper describes lessons learned from developing the most recent balanced corpus of LVK series from various online sources, which do not require agreements with text holders, but it presents many more technical challenges.

2. Background of LVK Series

The Balanced Corpus of Modern Latvian (LVK) has been developed in multiple rounds. The history of the LVK series goes back to 2007 when the first 1 million corpus was created. The experience from the designing of other general corpora was taken into account as well. The reviewed list of corpora includes British National Corpus [5], [6], Czech National Corpus [3], [7], [8], Corpus of the Contemporary Lithuanian Language [9], [10], and others. The same corpus design criteria were also used for the subsequent LVK series. The previous corpus from this series (LVK2013) was released on 2013 with 4.5 million words [11]. All corpora are morphologically annotated [12], [13], [14] and with metadata descriptions.

Previous corpora in LVK series were manually created. The main innovation in LVK2018 is autmatization process in all corpus development steps.

3. Design Principles of LVK

LVK2018 is designed as a general-language, representative and publicly available corpus. It is a monolingual, fully morphologically and partly syntactically and semantically annotated corpus. Presently, it consists of 10 million tokens. Characteristics of LVK2018:

- **General** – the corpus includes sources from different domains, styles, genres, etc.
- **Balanced** – the corpus that aims to cover the variety of existing texts in estimated proportions.
- The corpus represents the **synchronic** state of the language. It covers sources as from the end of the last century until the present.
- **Originality** – the corpus should only contain texts originally written in Latvian. The obvious translations of the different texts into Latvian will not be included in LVK2018.
- The corpus is **representative**, it contains texts from all language styles, major domains and many subdomains.

The corpus contains five different sections – *journalism, fiction, scientific, legal and parliamentary transcripts* (figure 1).

To cover different magazines and newspapers, subsequently the Journalism section also has been divided into the following subsections: *nationwide media, regional media, leisure media, and popular science media* (figure 2). The previously defined corpus sections were not enough to achieve fully balanced and representative corpus, so multiple additional text selection criteria were set.

- **Time** – the corpus should contain texts created and published after 1991.

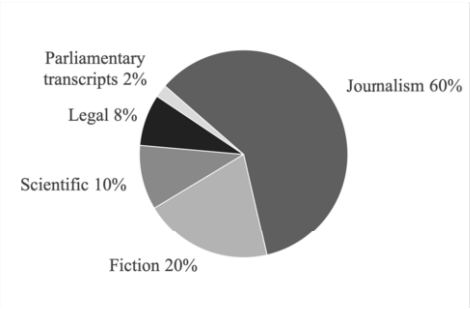

Figure 1. Composition of LVK2018

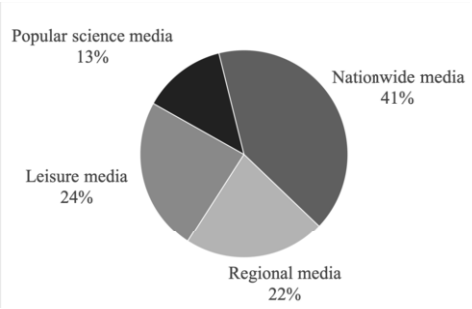

Figure 2. Composition of Journalism section

- The corpus should contain **full-text**.
- **Diversity** – texts should cover as wide range of topics as possible. The sample cannot exceed more than 5 % and 50,000 words of the particular section of the corpus to not dominate one author or domain in the corpus. If the text is longer than the limit, it should be cut from the end of the sample only.
- **Uniqueness** – the corpus sample should be represented in corpus just once.
- **Quality** – samples should only contain clean text written in literary language with appropriate usage of diacritics and punctuation in Latvian. Tables and other non-text parts should be removed.

4. Corpus Development Process

All the criteria set previously were implemented in the corpus development process which was divided in three steps.

- **Data collection** – for each section of the corpus, the most suitable data sources were identified, taking into consideration time and originality selection criteria. Text with all of the available metadata were extracted from each of the data sources.
- **Data processing** – data quality and uniqueness assurance processes were used in data processing. Texts were filtered through multi-level system to ensure that only unique and qualitative texts are considered as candidates for final data selection.
- **Final data selection** – a subset was selected from the remaining texts to be included in the corpus, taking into account corpus size limitation and diversity criteria.

Each corpus section (journalism, fiction, scientific, legal, and parliamentary transcripts) required multiple exceptions in different corpus development steps due to the diverse nature of data sources and structure. Next sections of the article will go into more details on how each step was implemented.

5. Data Collection

The most challenging tasks in the whole corpus development process were finding the most suitable freely available online data sources for each of the corpus sections and gathering the data in a structured format with metadata. Texts that are available online are considered to be freely available for the corpus creation purposes. Although the texts themselves are subject to copyright, the concordances that are available through the corpus query interface and used for educational or research purposes do not infringe copyrights [2].

Data for the Journalism section was gathered from various online media sites through a media monitoring system. The data also included forum posts and articles' comments that needed to be removed. The only available metadata for each article were the source URL and the date of publication. From the source URLs, a list of domains were obtained and manually categorized into the one of the four subcategories according to the most common content type.

Fiction is the only section with almost no freely available online data sources. Some sites publish freely available works of new authors and hobby writers. Book samples are also perfect for corpus creation. Unfortunately, there was no enough easily available data for Latvian, so the data was taken from books digitized in previous projects. The data also included metadata (title, author, publisher and year).

Data for the scientific section was collected from online freely available doctoral theses in PDF format. Thematic domain and publishing year were extracted as metadata.

Legal documents were crawled from the database of legal acts of the Republic of Latvia. They contain different types of documents (such as law, regulations, protocols, rulings and orders) from multiple institutions.

Parliamentary transcripts are extracted from The Corpus of the Saeima, which consists of the transcription of Latvian parliamentary debates crawled from the official website [15].

6. Data Processing

Gathering data from online sources can introduce many errors. A quality filter must be put in place to ensure only literary correct texts are included in the corpus. A multi-level system was used to filter out all the texts that did not meet the quality standards. Each of the quality filters were implemented due to a specific problem.

Some documents were written in poor language without the correct use of diacritical signs. Some documents besides the Latvian version also contained versions in a different language. A dictionary approach was used to filter out these kinds of documents. The document was considered to be of poor quality if fewer than 85 % of the words were found in a dictionary. This percentage was empirically calculated from LVK2013, which was manually created and validated.

Some documents did not contain correct Latvian: insufficient use of punctuation, too many punctuation marks or garbage symbols due to some parsing errors or embedded content with many hashtags. Documents having an unreasonable percentage of punctuation marks calculated against the number of tokens were filtered. Reasonable proportion

[2] Copyright Law, 48/150 (2059/2061), 27.04.2000. https://likumi.lv/ta/en/en/id/5138-copyright-law

was considered to be between 12 % and 36 %. These percentage were also empirically calculated from LVK2013.

In rare cases, the conversion from PDF corrupted some letters with diacritics. Documents of a reasonable length are expected to contain every letter of the alphabet. Documents that did not meet the requirement were filtered out.

After the quality filters, the next step was uniqueness filters. In the media industry, it is a common practice to republish the same in multiple sites with little or no modifications at all. Uniqueness issues are also common in the parliamentary section due to standard phrases and in the legal section due to templates, especially in rulings. To filter out too similar articles Bray-Curtis similarity over bag of words models [16] was used. If the similarity between any two documents was greater than 0.8, only the longest document was kept. This threshold was also calculated from LVK2013. Vocabulary in legal documents tends to be much more limited than in journalism, so the threshold in legal documents was set to 0.65 instead of 0.80.

7. Data Selection

The last step of the corpus development process is the final document selection, taking into account the corpus size limitation and diversity criteria. The main challenge is finding the right balance between diverse and representative subsets. Each section of the corpus was balanced according to the available metadata, such as date, author, industry, and others. The way how each section of the corpus was balanced was different due to different metadata properties.

Documents from journalism section were chosen based on the date of publication. To keep the original balance between article categories (local news, global news, sports, finance, etc.), articles were grouped by date of publication and the whole day was included or excluded from the corpus. To obtain the most diversified subset, documents were chosen evenly across the available timespan.

To choose the most representative subset of fiction, documents were chosen so that each author is included in the corpus. If some author had less data available, than the remaining quota was evenly distributed to other authors.

Subset of documents for scientific section was chosen the same way as the subset for fiction section were chosen only instead of choosing by author the documents were chosen based on scientific discipline of the document. In total, there were 30 scientific disciplines.

Data for legal section was already as balanced as it could be because the threshold for the longest common word string was chosen in such a way that the remaining amount of data was only a bit bigger than the required amount for the corpus. In final selection, a few random documents were removed to obtain the target word count.

In final data selection for parliamentary section, documents were grouped by date to cover as wide time span as possible. To achieve higher diversity in the selected subset, iteratively the shortest document from each date was selected until the total word count reached the goal.

8. Annotation of LVK2018

LVK2018 has three publicly visible metadata fields – unique identifier (id), section and reference. A different reference template was designed for each of the five sections to incorporate all the relevant metadata fields for that sections.

- for legal texts – *{title}*, adoption *{date of adoption}*, published on *{date of publication}*;
- for Parliamentary transcripts:
    - for the samples of LVK2013 – from Parliamentary transcripts *{date}*, *{speaker}* (*{parliamentary group}*);
    - for the samples of LVK2018 – from Parliamentary transcripts theme: "*{theme}*", *{sample URL}*;
- for science – *{author}*, *{title}* (*{the branch of science}*), *{year}*;
- for fiction – *{author}*, *{title}* (*{chapter}*). *{publishing place}*, *{publisher}*, *{year}*;
- for journalism – *{naming}*, *{source}*, *{published}*, *{subsection}*.

LVK2018 contains morphosyntactic annotation by the IMCS morphological tagger [12][13][14]. Morphosyntactic annotations contain PoS tag, lemma and other Latvian specific morphological and syntactic information.

A balanced subcorpus of LVK2018 (10,000 sentences), containing samples of texts from the different styles, domains and subdomains existent in the corpus, is also syntactically manually annotated [17], using hybrid dependency-constituency grammar formalism developed in the previous Latvian Treebank pilot project [18]. Afterwards, the hybrid annotation is automatically converted to Universal Dependencies to achieve the cross-lingual compatibility, as well as to provide training data for efficient and robust parsers [19]. The same subset of 10,000 sentences is also manually named-entity, coreference and FrameNet annotated [20].

9. Availability

LVK2018 has been released in the framework of Latvian National Corpus. LVK2018 is freely available via the corpus query interface NoSketch Engine [21].

10. How to quote LVK

The corpus material is to be quoted in the bibliography in the following way: The Balanced Corpus of Modern Latvian – LVK2018. The Institute of Mathematics and Computer Science, University of Latvia. Riga, 2018. Available at: www.korpuss.lv

11. Conclusions

LVK2018 was developed in the FullStack project framework and served as a basis for multilayered syntactically and semantically annotated text corpus for Latvian [19]. Creating LVK2018 from online data allowed successfully complete the project due to time constraints, because the online approach is much faster for developing a corpus of this size compared to the typical approach via signing the cooperation agreements with text holders.

Although nowadays a 10 million balanced corpus is not considered as a large corpus, it is useful for many Latvian language studies [22]. The success of LVK2018 has helped to secure a new project which is fully dedicated to the development of a new 100 million balanced corpus of contemporary Latvian. The new corpus will be created in cooperation with text holders, because there is not that much freely available online data for Latvian. The LVK2018 serves as a great example in the conversations with the text holders.

Acknowledgements

This work has received financial support from the Latvian Language Agency through the grant agreement No. 4.6/2019-029.

References

[1] Sinclair J. EAGLES. Preliminary recommendations on corpus typology. EAGLES Document EAG TCWG-CTYP/P. 1996.

[2] Mititelu VB, Irimia E, Tufis D. CoRoLa – The Reference Corpus of Contemporary Romanian Language. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 2014. p. 1235–1239.

[3] Křen M, Cvrček V, Čapka T, Čermáková A, Hnátková M, Chlumská L, et al. SYN2015: Representative corpus of contemporary written Czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); 2016. p. 2522–2528.

[4] Levāne-Petrova K. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. Valoda: nozīme un forma. 2019;(10):131–146.

[5] Aston G, Burnard L. The BNC handbook: exploring the British National Corpus with SARA. Capstone; 1998.

[6] Burnard L. Reference Guide for the British National Corpus (xml edition), 2007. URL http://www natcorp ox ac uk/XMLedition/URG. 2007.

[7] Čermák F. Today's corpus linguistics: Some open questions. International Journal of Corpus Linguistics. 2002;7(2):265–282.

[8] Hnátková M, Kren M, Procházka P, Skoumalová H. The SYN-series corpora of written Czech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 2014. p. 160–164.

[9] Kovalevskaitė J. Dabartinės lietuvių kalbos tekstynas – 10 metų kaupimo ir naudojimo patirtis. Prace Baltystyczne. 2006;3:231–241.

[10] Rimkutė E, Kovalevskaitė J, Melninkaitė V, Utka A, Vitkutė-Adžgauskienė D. Corpus of contemporary Lithuanian language–the standardised way. In: Human Language Technologies–The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT; 2010. p. 154–160.

[11] Levāne-Petrova K. Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. Baltistica. 2012;(8):89–98.

[12] Paikens P. Lexicon-based morphological analysis of Latvian language. In: Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT); 2007. p. 235–240.

[13] Paikens P, Rituma L, Pretkalniņa L. Morphological analysis with limited resources: Latvian example. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); 2013. p. 267–277.
[14] Paikens P. Deep Neural Learning Approaches for Latvian Morphological Tagging. In: Human Language Technologies–The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016. vol. 289. IOS Press; 2016. p. 160–166.
[15] Darģis R, Auziņa I, Bojārs U, Paikens P, Znotņš A. Annotation of the corpus of the Saeima with multilingual standards. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); 2018. .
[16] Bray J, Curtis J. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs (27). PRIMER-E Plymouth; 1957.
[17] Rituma L, Saulīte B, Nešpore-Bērzkalne G. Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis. Valoda: nozīme un forma. 2019;(10):200–216.
[18] Pretkalniņa L, Nešpore G, Levāne-Petrova K, Saulīte B. A Prague Markup Language profile for the SemTi-Kamols grammar model. In: Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011); 2011. p. 303–306.
[19] Gruzitis N, Pretkalnina L, Saulite B, Rituma L, Nespore-Berzkalne G, Znotins A, et al. Creation of a balanced state-of-the-art multilayer corpus for NLU. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); 2018. p. 4506–4513.
[20] Gruzitis N, Nespore-Berzkalne G, Saulite B. Creation of Latvian FrameNet based on Universal Dependencies. In: Proceedings of the International FrameNet Workshop (IFNW); 2018. p. 23–27.
[21] Rychlỳ P. Manatee/Bonito-A Modular Corpus Manager. In: RASLAN; 2007. p. 65–70.
[22] Holvoet A. The Middle Voice in Baltic. vol. 5. John Benjamins Publishing Company; 2020.