Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-130

On the Development of a Latvian-Russian Parallel Corpus

Natalia PERKOVA [a,1] and Dmitri SITCHINAVA [b]
[a] *Stockholm University*
[b] *Vinogradov Russian Language Institute, Russian Academy of Sciences*

Abstract. This paper presents the current status of the Latvian-Russian parallel corpus, which is an ongoing project within the Russian National Corpus. It discusses the existing parallel corpora including Latvian texts, availability of sources and the main principles and tools of alignment and morphological annotation, as well as further plans for developing the corpus.

Keywords. Corpus linguistics, parallel corpus, Latvian, Russian

1. Introduction

Parallel corpora are generally considered to be a valuable source of established correspondences between languages, the information which can be widely used in various translation-related projects, as well as in corpus-based contrastive studies. Parallel texts have also recently gained a certain popularity among typologically oriented linguists who treat them as a very useful tool for getting natural contextualised examples allowing for language comparison, see [1], [2].

The corpus under discussion is the first bidirectional parallel corpus of Russian and Latvian texts, available through an online search interface[2]. It has been built within the *Russian National Corpus (RNC)*[3], and therefore some of the general principles of the RNC have been applied to its development and design. The corpus is aimed to be the first dynamic representative parallel corpus of Latvian with morphological annotation: new texts will be gradually added to make it more representative and balanced. The first version of the corpus was launched in 2014. The current version (as for August 2016) is slightly larger, and it consists of eight Latvian and four Russian texts with corresponding translations, all lemmatised and morphologically annotated. Its size is about 800 thousand words.

For many years, Russian has been one of the languages of particular relevance for translations from and into Latvian, and a representative collection of Latvian-Russian aligned texts can be used for many purposes. It can become an important tool for translation studies and language teaching and learning, grammar-focused and lexicographic research. To a certain extent, the corpus parts can also be used as

[1] Corresponding Author: Natalia Perkova, Stockholm University, Department of Linguistics, Universitetsvägen 10 C, SE-10691, Stockholm, Sweden; E-mail:Natalia@ling.su.se
[2] http://ruscorpora.ru/search-para-lv.html
[3] http://ruscorpora.ru

monolingual corpora, as authentic examples can be easily drawn from the search results.

2. Background

During the development of the corpus under discussion, the language resources existing for Latvian were thoroughly taken into account. It was done with the primary goal to build a user-friendly corpus which includes the texts underpresented elsewhere, as well as to look for possible solutions of the problems of the currently available Latvian corpora. Latvian became the major focus in this project because of its relatively small number of speakers and therefore of a particular importance of developing language resources for it.

The biggest parallel corpora with Latvian texts (*EuroParl, JRC-Acquis* and *EU Bookshop*, see [3], [4]) tend to be very genre-specific, consisting of official texts. Among other large multilingual corpora including Latvian texts, *ParaSol*[4] (currently with only one text in Latvian) and *InterCorp*[5] deserve mention. Latvian translations are also available in the massive *Parallel Bible Corpus*[6] [5]. The recently launched Lithuanian-Latvian parallel corpus *LiLa* [6, 7] is probably the most representative parallel corpus, as considers contemporary Latvian fiction (written after 1991) included in the sample. Film subtitles in Latvian aligned to other languages are available in the *Opus* project[7].

Russian National Corpus, similarly to *Intercorp*, the corpus developed within the Czech National Corpus, has a significant parallel part with the texts translated from and into Russian [8]. At present, a dozen of languages, including Latvian, are represented in RNC by corresponding bilingual corpora. In addition, the RNC includes a multilingual segment with several texts provided by Adrian Barentsen, the compiler of *ASPAC* (Amsterdam Slavic Parallel Aligned Corpus)[8]. Latvian is represented by only one text in this subcorpus.

3. Principles of Text Selection and the Corpus Composition

For the primary purposes, only fiction is currently considered as the data used for sampling. Even though the corpus lacks some genre representativity, the included texts combine segments imitating natural speech with narrative structures, which makes the data valuable for linguistic research. Such a corpus can also be seen as a supplementary collection of parallel texts in addition to the existing corpora of a more formal character commonly used for the purposes of machine translation and related problems.

There has been no intention to restrict the selected texts by contemporary literature. On the contrary, the aim was to include in the corpus texts of as many different time spans as possible, considering the period from the end of the 19th century. This is most important, considering that the existing corpora of Latvian tend to include the texts written not earlier than 1991. Old Latvian texts are available in the

4 http://www.slavist.de/
5 http://www.korpus.cz/intercorp/
6 http://paralleltext.info/data/
7 http://opus.lingfil.uu.se/index.php
8 http://home.medewerker.uva.nl/a.a.barentsen/page3.html

Corpus of Early Written Latvian texts[9] ; however, this corpus covers only the period of the 16-18th centuries. As for the texts written in the 19th century and before 1991, there is a reasonable gap in the existing language resources. Latvian classical literature is available online[10], but this cannot be considered as a corpus proper. There are also text collections available via the National Library of Latvia[11] which cover different time spans (including the period of the first independence and Soviet texts); however, they mostly consist of newspapers and cannot be counted as a very representative and balanced corpus.

At present, the Latvian texts included in the corpus represent mostly classical literature (short prose by Rainis and Rūdolfs Blaumanis) and late Soviet literature (texts by Zenta Ērgle, Regīna Ezera and Andris Puriņš). The novel *"Dzīves svinēšana"* by Nora Ikstena has also been processed, and at the current stage this is the only work by a contemporary Latvian author in the corpus. The choice of youth literature (Z. Ērgle and A. Puriņš) was primarily motivated by the character of its language, which is very close to colloquial speech.

As for the texts written in other periods, it is planned to extend the sample, more particularly by including more classical works of Rūdolfs Blaumanis, Rainis, Jānis Jaunsudrabiņš, Kārlis Skalbe, Anna Brigadere, as well as texts written by such well-known Latvian authors as Vilis Lācis, Andrejs Upīts, Alberts Bels, Imants Ziedonis, Vizma Belševica. It is also planned to include some relatively recent translations from Latvian, which are mostly various short stories by modern authors.

Russian originals are currently represented by four texts: the novel *"Belaja gvardija"* by Mikhail Bulgakov and three short stories (*"Čelovek v futljare"*, *"Hameleon"*, *"Tolstyj i tonkij"*) by Anton Chekhov. Needless to say, Russian texts have been extensively translated into Latvian, so the problem of the representativity disbalance is rather characteristic for the current stage of the project, while it is planned to extend the sample considerably, adding both most important classical works and texts of other periods.

Table 1. Structure of the Latvian-Russian corpus.

Author	Original language	Year	Title	Size
Rūdolfs Blaumanis	LV	1897	*Nezāle*	7574
Rūdolfs Blaumanis	LV	1898	*Purva bridējs*	28108
Rainis	LV	1880-1920	*Ideāla disciplina, un kas no tās iznāca*	1920
Zenta Ērgle	LV	1976	*Starp mums, meitenēm, runājot...*	65714
Andris Puriņš	LV	1977	*Nevaicājiet man neko*	115094
Regīna Ezera	LV	1977	*Zemdegas*	260586
Zenta Ērgle	LV	1983	*Bez piecām minūtēm pieauguši*	94874
Nora Ikstena	LV	1998	*Dzīves svinēšana*	52747
Anton Chekhov	RU	1883	*Tolstyj i tonkij*	1111
Anton Chekhov	RU	1884	*Hameleon*	1799

[9] http://www.korpuss.lv/senie/

[10] http://www.korpuss.lv/klasika/saturs.htm and http://www.letonika.lv/literatura/Section.aspx?f=1&id=3112615

[11] http://bonito.korpuss.lv/lnb

Anton Chekhov	RU	1898	*Čelovek v futljare*	52747
Mikhail Bulgakov	RU	1924	*Belaja gvardija*	143955

The corpus includes both public domain texts (for example, Chekhov or Blaumanis) and copyright protected texts; the rate of free texts is naturally higher among the original texts as they are created earlier than the translations. However the texts cannot be accessed online in full. According to the general policy of the Russian National Corpus, the maximum size of the extended context retrieved by the search engine is seven sentences. This policy helps to avoid the copyright restriction, as all the texts can be freely quoted for scholarly purposes.

4. Text preparation, Alignment and Annotation

The prepared texts (scanned, recognised, and proofread, if needed) are saved in the Unicode UTF-8 format and then get sentence-to-sentence aligned with the help of the Euclid Parallel Text Aligner, a GUI interface based on HunAlign [9], developed by T. Arkhangelsky. Aligned texts are then manually corrected in order to get rid of possible alignment mistakes. For the input texts which are well-prepared (having no major scanning errors, etc.) the quality of alignment is very high. However, sometimes errors occur because of some non-standard correspondences between two texts, which can occur due to some translation-related factors.

The resulting files are kept in the XML format. In addition, the necessary meta-information is kept in separate CSV files: it includes the date of publication, the author and translator names, the text title, etc. These data can be used while setting a desired subcorpus: one can choose necessary parameters to delimit the search. An example of metadata information is provided in Figure 1.

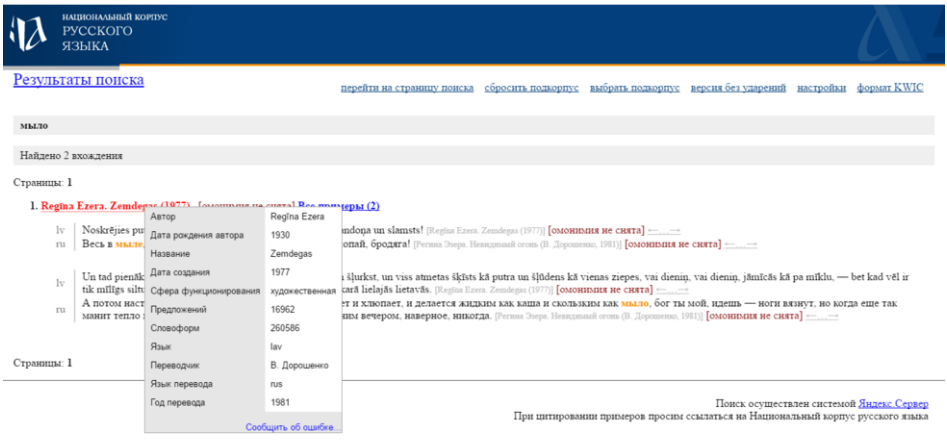

Figure 1. Metadata structure in the Latvian-Russian corpus.

For the current version of the corpus, the system of morphological annotation has been developed by Danko Aleksejevs and Natalia Perkova[12]. It is based on the publicly available morphological tagger developed by the team from the Institute of Mathematics and Computer science at the University of Latvia [10][13]. The morphological tagset of the corpus follows the general lines chosen for all the corpora within the RNC. The tagset for Russian texts is the one already used for the Russian-language RNC texts. The POS and grammatical tags are uniform codes separated by commas or the = sign, and they are abbreviations of Latin or English terms (e. g. *praet* for past or *time* for time adverbials). The Latvian tagset includes not only inflectional categories but also classifiers (time adverbials, personal pronouns, etc.). The homonymy is not disambiguated in the Russian part; as for the Latvian part, only one variant is chosen on the basis of the algorithms of the morphological tagger, which can be seen as a shortcoming. Still, the quality of the generated morphological analysis is quite high for the present state of the corpus. The online search interface allows to easily choose necessary categories, as well as to build longer searches.

The next example is taken from the short story "*Ideāla disciplina, un kas no tās iznāca*" by Rainis. It shows the annotation for the Latvian wordform *krietnais* and its Russian correspondence *усердный*. The Russian tag has more information, as it gets rich semantic annotation, as well as provides two possible variants of grammatical analysis (nominative or accusative):

<w><ana lex="krietns" gr="A,qual=pos,m,sg,nom,def"/>Krietnais</w>
<w><ana lex="усердный" sem="t:humq der:s dt:behav r:qual" disamb="yes" gr="A,acc,inan,m,norm,plen,sg" sem2=""/><ana lex="усердный" sem="t:humq der:s dt:behav r:qual" disamb="yes" gr="A,m,nom,norm,plen,sg" sem2=""/>Усердный</w>

5. Plans for Future Developments

The current state of the corpus is yet to be developed in terms of its size, representativity and balance. The further work on the corpus will be focused primarily on expanding it by adding more new texts. There is a special list with the titles of relevant texts which will serve as the basis for that.

The system of morphological annotation will be fully implemented, as well as further improved. To put it more precisely, better quality of lemmatisation and morphological tagging is needed in some cases. In addition, a big problem arises because of the non-availability of multiple annotation variants, which leads to the wrong analysis of certain wordforms. Such valuable information as established derivational relations between wordforms can also be implemented from the Latvian morphological tagger.

There is also strong need for developing non-Russian versions of the search interface and other available relevant information so that more people could use the corpus.

[12] https://bitbucket.org/virtulis/lvtagger2ruscorpora
[13] https://github.com/PeterisP/LVTagger

Acknowledgements

The work on the Latvian-Russian parallel corpus has been done within the project 15-04-12018 by the Russian Scientific Fund of Humanities. We would like to thank Danko Aleksejevs for helping us with building the system of morphological annotation for Latvian.

References

[1] B. Wälchli. Advantages and disadvantages of using parallel texts in typological investigations, *Sprachtypologie und Universalienforschung* **60(2)** (2007), 118-134

[2] R. von Waldenfels. Polish tea is Czech coffee : advantages and pitfalls in using a parallel corpus in linguistic research. In: Ender, Andrea; Leemann, Adrian; Wälchli, Bernhard (eds.) *Methods in contemporary Linguistics. Trends in Linguistics*: Vol. 247 (2012), 263-281. Berlin: Mouton De Gruyter.

[3] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. Schlüter, M. Przybyszewski, and S. Gilbro. An overview of the European Union's highly multilingual parallel corpora. Lang. Resour. Eval. 48, 4 (December 2014), 679-707.

[4] R. Skadiņš, J. Tiedemann, R. Rozis, & D. Deksne. Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (2014), 1850–1855.

[5] T. Mayer and M. Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (2014), 3158-3163.

[6] E. Rimkutė, A. Utka, K. Levāne-Petrova. Lietuvių-latvių ir latvių-lietuvių kalbų lygiagretusis tekstynas LILA. *Kalbų studijos* **23** (2012): 70-77.

[7] A. Utka, K. Levāne-Petrova, A. Bielinskienė, J. Kovalevskaitė, E. Rimkutė, D. Vēvere. Lithuanian-Latvian-Lithuanian Parallel Corpus. In: A. Tavast, K. Muischnek, M. Koit (eds.). *Human Language Technologies. The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012* (2013), 260–264. Amsterdam, Berlin, Tokyo, Washington, DC: IOS Press.

[8] D. Sitchinava. Parallel'nye teksty v sostave Nacional'nogo korpusa russkogo jazyka: novye napravlenija razvitija i rezul'taty [PARALLEL TEXTS WITHIN THE RUSSIAN NATIONAL CoRPUS: NEW DIRECTIONS AND RESULTS]. In: *Trudy Instituta Russkogo Iazyka imeni V.V. Vinogradova [Proceedings of the V.V. Vinogradov Russian Language Institute]* **6** (2015), 194-234.

[9] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón. Parallel corpora for medium density languages. In: *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*. Edited by Nicolas Nicolov, Kalina Bontcheva, Galia Angelova and Ruslan Mitkov. [Current Issues in Linguistic Theory 292] (2007), 247-258.

[10] P. Paikens, L. Rituma, L. Pretkalniņa. Morphological analysis with limited resources: Latvian example. In: *Proceedings of 19th Nordic Conference of Computational Linguistics* (2013), 267-277.