Bilingual Parallel Corpora Featuring the Circum-Baltic Languages within the Russian National Corpus

Dmitri Sitchinava[1] and Natalia Perkova[2]

[1]Higher School of Economics, 21/4 Staraya Basmannaya 105066 Moscow, Russia / Institute of the Russian language, 18/2 Volkhonka 119019 Moscow, Russia
mitrius@gmail.com
[2]Stockholm University, SE-106 91 Stockholm, Sweden / Uppsala University, Box 256, 751 05 Uppsala, Sweden
nofpernat@gmail.com

Abstract. The paper presents parallel corpora within the Russian National Corpus (RNC) featuring Circum-Baltic/Russian language pairs and describes the choice of texts, morphological annotation and possible applications. The following languages of the Circum-Baltic linguistic area are included into the bilingual pairs of the corpus: Estonian, Finnish, Latvian, Lithuanian, Polish, and Swedish. The corpus includes both fiction and non-fiction texts and has a diachronic dimension. The morphological annotation of different languages is sensitive for language-specific categories and features. For each language an expanded RNC tagset is constructed which provides cross-linguistic comparison but at the same time takes into consideration differences in grammatical systems. The corpora can be used for exploring some grammatical and lexical features for the Circum-Baltic region that have no straightforward correspondence in Russian and are often rendered by other means. Further expansion of the corpus by non-fiction genres is particularly important for the study of lexicon and syntax specific for legalese, media or academic style.

Keywords: parallel corpora, Circum-Baltic area, grammatical typology, contrastive linguistics

1 General Overview of the Project

Parallel corpora are linguistic corpora in multiple languages consisting of original and translated texts with corresponding alignment, most typically sentence-by-sentence. The translations included into a parallel corpus are never made for this purpose; they are already available for ordinary readers and are supposed to convey the original meaning of the text as accurately as possible (however, some problems and challenges are inevitable, in particular when fiction or religious texts are involved). It is also generally assumed that the translation is a naturally sounding text in the target

language (which is not always the case, cf. the notion of *translationese* [6] that means a style more or less heavily influenced by, and transparent of, the source language).

Parallel corpora, including so-called *massive parallel corpora* [2] have been actively used as the sources of typological and contrastive linguistic information for the analysis of different lexical and grammatical phenomena. The semantics of lexical and grammatical items can be analyzed via natural translations of their occurrences to another language or a group of languages. The meaning that is conveyed in the process of translation and is (ideally) shared by the original and its translations, can be used as *tertium comparationis* for the comparison of linguistic phenomena, that is the ways in which particular meanings, or contexts, are expressed (cf. the discussion of comparable concepts in linguistics, see [7]). Translations also have the property (for some linguists, also the advantage) of not being specially elicited for linguistic purposes, unlike translational questionnaries widely used by typologists that aim to collect the data which in an indirect way can be treated as parallel texts, too (see the seminal study [3] which discusses this type of data). Among the phenomena which have been investigated in multilingual parallel corpora, one can mention motion verbs [17], aspect in Slavic imperative forms [18], the perfect gram in European languages [4].

Our purpose is to incorporate parallel bilingual corpora with Russian and the Circum-Baltic languages ([5], [14]) into the Russian National corpus. These languages, belonging to different language groups (Finnic, Baltic, Slavic, and Germanic), share some common typological traits and exhibit mutual influences within smaller areas. The parallel corpora with Russian can be seen as a tool for researching some of these phenomena (and of course many other grammatical and lexical items), including those for which Russian has no direct structural correspondence, rendering them by other means (for example, Perfect tense or marking of reported information, the so-called evidentiality). For both geographical and historical reasons, Russia being a close neighbour of the areal in question (and the Baltic States and Poland previously being part of the Russian Empire and later of the USSR resp. the Soviet bloc), many texts written in the Circum-Baltic languages are available in Russian translations and vice versa, and new translations of modern texts in both directions appear. This makes it possible to rely on the existence on translations from and into Russian, which is especially useful, considering that Russian can be seen as a rather high-resource language, compared to many languages of the area.

The Russian National Corpus (henceforth RNC, http://ruscorpora.ru) already has a set of bilingual corpora and a multilingual parallel corpus [16], searchable online and aligned sentence-by-sentence (represented in a XML format, with the metainformation represented in a CSV-format table). Of the languages included in bilingual parallel corpora with Russian, the following ones belong to the linguistic area in question: Polish (the Polish-Russian corpus available since 2010, 6 million tokens), Estonian (łaunched in 2015, 600 thousand tokens), Latvian (since 2016, see Perkova, Sitchinava 2016 for more detail; 2.5 million tokens), Swedish (since 2017, 3.6 million tokens), Lithuanian (since 2018, 560 thousand tokens), and Finnish (yet to be published online, counting 2 million tokens). The corpora are searchable online,

and contexts up to seven sentences can be extracted and downloaded. The full texts are not downoadable due to copyright restrictions.

The architecture of the Latvian, Lithuanian and Swedish parallel corpora with Russian is planned (and the texts themselves aligned) by one of the present authors, Natalia Perkova. She also participates, together with Elizaveta Fomina, in the Estonian subproject. The Polish-Russian Parallel Corpus was compiled by the RNC team together with the Polish Academy of Sciences, and the Finnish-Russian corpus is being prepared by Karina Mishchenkova in collaboration with the ParRus and ParFin projects, headed by Mikhail Mikhailov at the University of Tampere.

2 Metadata markup and architecture

All texts in the corpus get metatextual markup that specifies their genre, authorship (including information concerning translators), date (both of the source and the target text), and the direction of translation. Using these parameters subcorpora can be customized, e.g. some linguistic phenomena can be searched within different time periods to show possible diachronic changes. A subcorpus of a given author or translator can be built to track the author-specific patterns that are more prominent in the so-called "translationese" style.

All the bilingual parallel corpora are planned to include both fiction and non-fiction texts (all the texts are included in full). However, the primary focus has been on fiction as the most obvious choice and the source of rather varied texts potentially representative of a wide range of stylistic phenomena. Ideally, the corpus should be representative, comprising dialogical, narrative, scientific, official and other types of texts with their specific characteristics reflected in linguistic structures and lexicon.

The bilingual corpora also feature a diachronic dimension, including texts (and translations, though the latter naturally tend to be more recent) from different historical periods from the 19th to the 21st century. In a multilingual ("massive parallel") corpus only the internationally renowned texts (e.g. Russian or Swedish 19th-century classics) that are translated to many languages can be included; they cannot be very numerous and in any case there cannot be many recent original texts available. In a situation like with the Circum-Baltic/Russian language pairs where many texts have been translated in both directions, bilingual corpora can feature a more representative collection of culturally significant fiction of different periods, both the "cultural canon" and contemporary texts of the 2000s and 2010s (the latter include the works by authors like Danny Wattin or Carl-Johan Vallgren in Swedish, Ljudmila Petrushevskaja or Marina Stepnova in Russian etc.). The Swedish texts included in the corpus feature also the "Finnish Swedish" variant (for example, the works by Tove Jansson). This diachronical representativeness is the major innovative feature of the corpus as compared to the existing parallel corpora featuring Circum-Baltic languages.

Some fiction texts are included in more than one translated version, thus allowing for polyvariant texts. For example, in the Latvian-Russian corpus this is the case of

some of Chekhov's stories represented in the translations by Anna Grēviņa, Oskars Kalnciems, Paulis Kalva and Regīna Ezera, or "Four Rides" by Vilis Lācis translated to Russian by G. Ceitlin and V. Rugais. Alternative translations are provided for the Swedish-language children's books by Tove Jansson and Astrid Lindgren. These texts can be used to explore variation in translation and to study the contextual synonymy of different grammatical or lexical items.

The main source of non-fiction translated to Russian is currently the site inosmi.ru that features newspaper articles from foreign press translated to Russian from many languages, including all the Circum-Baltic languages involved. Alongside with this, also legal texts and international treaties are currently included into the Finnish-Russian part.

3 Morphological annotation

All the texts within the bilingual parallel corpora are morphologically annotated; more precisely, the POS and grammatical features of wordforms are specified. Thus, any combination of grammatical features, words and/or their parts is searchable within the corpus.

For Russian the Mystem analyzer developed by the Yandex company is used [15]; the Polish morphological analysis is based on the TaKiPi algorithm that predicts tags statistically [12]. The Latvian morphological analyzer was implemented in 2016 on the basis of LUMII morphological tagger [11]; this tagger, not unlike the Polish one, does not specify alternative morphological analyses and chooses only the one that is most probable statistically. The Lithuanian analyzer used in the corpus is based on the VDU (Kaunas) morphological annotator [13], for Estonian we use the Corpus analyzer developed in Tartu University [8]; for Swedish the open-source Stagger analyzer [9].

We owe the technical implementation of these taggers and their harmonization with the XML format used in the RNC parallel corpora to Danko Aleksejevs (Latvian and Lithuanian), Timofey Arkangelsky (Estonian and Swedish) and Boris Orekhov (Polish).

The morphological annotation of different languages is sensitive for language-specific categories and features. The originally obtained morphological information is preserved as much as possible, not affected by the conversion to RNC tags. It can be said that for each language the expanded RNC tagset is constructed, which provides cross-linguistic comparison, but at the same time takes into consideration differences in grammatical systems. For example, in Estonian texts, unlike elsewhere within the RNC, the compound nouns, productive in this language, are tagged as compound and different stems are separated by the plus sign. For Polish, the cliticized auxiliary 'be' in Conditional (*pojechał-by-m* go-BE.COND-1sg 'I would go') is marked as a separate word form (with an additional tag 'clitic'), orthographically attached to the main verb, and at the same time as a particle (the feaure *nwok* means a non-vocalized variant of the clitic as opposed to *-em*):

<w>
<ana lex="pojechać" gr="V,pf,indic,praet,sg,m,pers"/>
<ana lex="by" gr="PART"/>
<ana lex="być" gr="V,indic,praes,clit,sg,1p,ipf,nwok"/>
pojechałbym</w>

An example of aligned translations of a sentence with Russian, Latvian, and Lithuanian markup (*The Man in a Case* by Chekhov, translations resp. by Paulis Kalva and E. Viskanta; the phrase means 'It was already midnight'):

<se lang="ru"><w><ana lex="быть" sem="t:be:exist ca:noncaus d:root" disamb="yes" gr="V,act,f,indic,intr,ipf,norm,praet,sg" sem2="ca:noncaus d:root"/>Была</w> <w><ana lex="уже" sem="t:time" disamb="yes" gr="ADV,norm" sem2=""/>уже</w> <w><ana lex="полночь" sem="ev:posit r:abstr t:time" disamb="yes" gr="S,acc,f,inan,norm,sg" sem2="t:space r:concr r:abstr"/><ana lex="полночь" sem="ev:posit r:abstr t:time" disamb="yes" gr="S,f,inan,nom,norm,sg" sem2="t:space r:concr r:abstr"/>полночь</w>.</se>

<se lang="lv"><w><ana lex="būt" gr="V=indic,praet,act,3p"/>Bija</w> <w><ana lex="jau" gr="ADV,time="/>jau</w> <w><ana lex="pusnakts" gr="S,common,f=sg,nom"/>pusnakts</w> .</se>

<se lang="lt" variant_id="1"><w><ana lex="būti" gr="V,nrefl=indic,praet,sg,3p"/><ana lex="būti" gr="V,nrefl=indic,praet,pl,3p"/>Buvo</w> <w><ana lex="jau" gr="ADV=pos"/><ana lex="jau" gr="PART="/>jau</w> <w><ana lex="vidurnaktis" gr="S=m,sg,nom"/>vidurnaktis</w>.</se>

4 Directions of corpus-based research

The corpora can be used for exploring some grammatical and lexical features for the Circum-Baltic region that have no straightforward correspondence in 1Russian and are often rendered by other means. The Circum-Baltic linguistic area is characteristic for having the perfect aspectual gram and in addition the possessive perfect construction (see [1] on the Latvian and Lithuanian forms in a parallel corpus); even Polish has a new possessive perfect construction. This opposition is lost in Modern Russian, but, interestingly, reappeared in the Russian and Belarusian North-Western dialects under the influence of Finnic and Baltic languages (see [14] for more detail). Such perfect-based tenses like pluperfect or future perfect tend to get secondary meanings. More particularly, the future perfect forms in the Baltic languages have secondary semantics of hypothetic events (or inferential with past time reference, not unlike its use in other languages of Europe, see [10] on typological context). Overall, only 25% of the Lithuanian future prefect forms in the corpus have future time refer-

ence, whereas about a half of the examples signal hypotheses and/or inference concerning the past events and the remaining 25% fall to the category of precedence with regard to an irreal or habitual situation. For Latvian, the corresponding numbers are even more impressing in terms of semantic non-compositionality, resp. 27%, 66% and 2% (with some other marginal or ambiguous uses).

In Russian, a wide range of discourse markers corresponds to the hypothetic and inferential usages of Future Perfect in Baltic, e.g., *navernoe* 'perhaps', *dumaju* 'I think', *konečno* 'certainly' or even ordinary past tense forms without additional markers:

Latv. Zini, vecomāt, es laikam arī **būšu iemīlējusies** [fall.in.love.FUT.PERF]. [Zenta Ērgle. Starp mums, meitenēm, runājot... (1976)]

Rus. Znaeš, babuška, ja **navernoe** [probably] vljubilas' [fall.in.love.PST]. [(Ž. Ezit trans. 1979)]

'You know, Granny, I have [probably] fallen in love'

Rus. Ètogo nikogda ne bylo… serdce šalit… ja pereutomilsja [overworked.PST] [M. Bulgakov. Master and Margarita, 1925-1940].

Latv. Tas nu nekad nav bijis … sirds streiko … **būšu pārpūlējies** [overworked.FUT.PERF]. [Ojārs Vācietis trans.]

'This has never happened before. My heart's acting up… I'm [evidently] overworked…'

Lit. Čia, tose plynėse, tuose miškuose, ant šitų kelių ir takų viskas **bus prasidėję** [start.FUT.PERF], ėmę gauti prasmę… [Juozas Aputis. Lidija Skoblikova ir tėvo žingsniai (1980-1989)]

Rus. Imenno zdes', v ètix pustošax, v ètix lesax, na ètix dorogax i tropax, vsë, **požaluj** [perhaps], i načalos' [start.PST], stalo obretat' smysl… [Virgilijus Čepaitis trans., 1989]

'It was perhaps here, in these wastelands and woods, on these roads and paths, where all these things emerged and started to make sense'

Lexical correspondences can also be investigated on the basis of data from parallel texts. For instance, the Russian word *toska* '~yearning, anguish, misery' is rendered in most languages with a very high diversity and statistical entropy of different translations. More particularly, the Estonian-Russian parallel corpus (currently relatively small) already counts seven Estonian correspondences for *toska*, namely *koduigatsus* '~nostalgia/homesickness', *ahastama* '~anguish, depression', *masendus* '~depression', *mure* '~anxiety', *nukrus* '~sadness, grief', *tusk* '~chagrin' (interestingly, an old borrowing from Slavic and related to *toska*), *igatsus* 'yearning, longing'.

Future development of the Circum-Baltic parallel corpora with Russian includes expansion of corpora aimed to cover all the periods of fiction since the 19th century to the modern texts. Further expansion of the corpus by non-fiction genres is particularly

important for the study of lexicon and syntax specific for legalese, media or academic style.

Acknowledgements

The research is supported by the Russian Foundation for Basic Research, project 1734-01061-OGN "Slavic future anterior in a typological perspective"

References

1. Arkadiev, P., Daugavet, A.: The perfect in Lithuanian and Latvian: a contrastive investigation. Talk at Academia Grammaticorum Salensis Tertia Decima, 1–6 August 2016 (http://inslav.ru/sites/default/files/arkadievdaugavet2016_baltperf_salos.pdf)
2. Cysouw, M., Wälchli B. (eds.): Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in: Sprachtypologie & Universalienforschung STUF 60(2) (2016).
3. Dahl, Ö.: Tense and Aspect Systems. Blackwell, Oxford (1985).
4. Dahl, Ö.: The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus. In: Szmrecsanyi, B., Wälchli, B. (eds.). Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & litterae 28, pp. 268–289. Walter de Gruyter, Berlin (2014).
5. Dahl, Ö., Koptjevskaja-Tamm, M. (eds.): Circum-Baltic languages. Typology and contact. Vol. 1-2, Benjamins, Amsterdam—Philadelphia (2001).
6. Gellerstam, M.: Translationese in Swedish novels translated from English. In: Translation studies in Scandinavia, pp. 88–95. CWK Gleerup, Malmö (1986).
7. Haspelmath, M.: Comparative concepts and descriptive categories in crosslinguistic studies. Language 86(3), 663–687 (2010).
8. Kaalep, H.-J., Muischnek, K., Müürisep, K. . Rääbis, A., Habicht K.: Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. Keel ja Kirjandus 9, 623–633 (2000).
9. Östling, R.: Stagger: an Open-Source Part of Speech Tagger for Swedish. Northern European Journal of Language Technology 3, 1–18 (2013).
10. Pen'kova, J.: Ot retrospektivnosti k prospektivnosti: grammatikalizacija predbuduschego v jazykax Evropy [Russian: From Retrospectiveness to Prospectiveness: Grammaticalization of Antefuturum in the Languages of Europe]. Voprosy jazykoznanija 2, 53–70 (2018).
11. Perkova, N., Sitchinava, D.: On the Development of a Latvian-Russian Parallel Corpus. In: Skadiņa, I., Rozis, R. (eds.). Human Language Technologies – The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016, pp. 130–135. IOS Press, Amsterdam (2016).
12. Piasecki, P., Radziszewski, A., Godlewski G. et al.: TaKIPI, CLARIN-PL digital repository (2014), http://hdl.handle.net/11321/31, last accessed 2019/02/11.
13. Rimkutė, E., Daudaravičius, V., Utka, A.: Morphological Annotation of the Lithuanian Corpus. 45th Annual Meeting of the Association for Computational Linguistics. In: Workshop Balto-Slavonic Natural Language Processing, pp. 94–99 (2007).
14. Seržant, I., Wiemer, B. (eds.).: Contemporary approaches to dialectology: the area of North, North-West Russian and Belarusian dialects. University of Bergen, Bergen (2014).

15. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: Machine Learning; Models, Technologies and Applications, Las Vegas (2003)
16. Sitchinava, D.: Parallel corpora within the Russian National Corpus. Prace Filologiczne 63, 271–278 (2012).
17. Wälchl B., Cysouw M.: Lexical typology through similarity semantics: Toward a semantic map of motion verbs. Linguistics 50(3), 671–710 (2012).
18. Waldenfels R. von: Explorations into variation across Slavic: Taking a bottom-up approach. In: Szmrecsanyi, B., Wälchli, B.. (eds.): Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & Litterae 28, pp. 290-323. Walter de Gruyter, Berlin (2014).