PRACE FILOLOGICZNE

Warszawa 2012 TOM LXIII

DMITRI V. SITCHINAVA
Institute of the Russian Language
Russian Academy of Sciences

PARALLEL CORPORA WITHIN THE RUSSIAN NATIONAL CORPUS

SŁOWA KLUCZOWE: korpusy równoległe, wyrównywanie, tagowanie morfosyntaktyczne, tagset

KEYWORDS: parallel corpora, alignment, morpho-syntactic tagging, tagset

1. Introduction

The parallel corpora within the Russian National Corpus (Национальный корпус русского языка; search available at: http://ruscorpora.ru/search--para.html for a Russian interface) have been under development since 2005. Some publications in Russian have already been published by the RNC team (Добровольский, Кретов, Шаров 2005a, b, Андреева, Касевич 2005, Добровольский 2009).

The RNC includes bilingual corpora with Russian as one of the languages, being either the original or translated text. Particular attention is given to the Slavic languages; as of March 2011, only the Ukrainian corpora are searchable (the other two non-Russian languages available are English and German), but at present also Belorussian and Polish parallel corpora are under preparation. The XML markup currently in use allows for building tri- or polylingual corpora with multiple translations of original texts (including the option of multiple translations into the same language), but presently this kind of text is not yet available for search within the RNC.

2. Alignment

The sentences in the RNC are aligned sentence-by-sentence. The texts kindly offered for the use in the RNC by Adrian Barentsen and included into the Amsterdam Slavic Parallel Aligned Corpus multilingual corpus are already aligned paragraph-by-paragraph. This segmentation has been additionally refined by us semi-automatically, introducing boundaries between the sentences within these pre-defined paragraphs. For the majority of the texts the alignment is undertaken completely by the RNC team. The original text's sentence segmentation overrides the segmentation in the translation. So, each single element of parallel tagging corresponds to the original sentence, whilst its translated counterpart(s) may well be either a part of a sentence starting with a space or a comma, or more than one sentence.

Multiple alignment tools have been used for the RNC parallel corpora. It is evident that the procedure of alignment consists of two stages: introducing sentence boundaries into texts and the alignment in the narrow sense of the word. There exist programs that do not have an embedded sentence-splitting alogithm (HunAlign by Andras Farkas, http://mokk.bme.hu/resources/hunalign; LeoBilingua by Leonid Brodsky, www.hot.ee/bclogic/) and those who enable sentence-boundaries markup like TextAlign (http://www.englishelp.ru/soft/soft--for-translator/151-textalign.html) or Parallelnye Texty (Параллельные тексты), a program developed for the RNC by A.A. Kretov's team in Voronezh University and used for markup of some English-Russian and German-Russian texts. The algorithm of breaking the text down into sentences is straightforward in both programs; it uses punctuation marks, e.g. exclamation marks, quotation marks and full stops, without taking into consideration initial letters, abbreviations, quotation and parenthesis marks, or the rules of direct speech (for problems in using TextAlign in a Ukrainian-Polish parallel corpus see also Kotsyba 2009). The segmentation, in both programs, can be corrected manually, although the algorithm itself cannot be corrected, and some general mistakes are to be treated each time they occur. In the TextAlign program, additionally, the automatic sentence-breaking is obligatory, and one cannot escape it by creating a standalone program for this purpose. For LeoBilingua and HunAlign this is the only option, and it is possible to elaborate rules of sentence-breaking and change them as the new texts bring new challenges.

The alignment proper for all the four tools is automatic with possible manual verification. This is further divided, however, into two possible modes: step-by-step (with corrections possible in the middle of the consequent alignment) and total alignment with post-correction. The first approach is embraced by LeoBilingua and Parallelnye teksty, while the other is chosen by TextAlign and HunAlign. The last two programs, therefore, call for a re-reading of an already aligned text with correction of the wrongly aligned sentences. While in TextAlign a GUI interface is provided for this (however a single correction calls for re-aligning the whole text), in HunAlign only a manual editing of the output file is possible.

The choice of parallel sentences may be additionally verified from the point of view of sentence length and/or lexical contents; this feature is supported by LeoBilingua (one sentence should not be twice or more longer than another) and HunAlign uses a statistical mechanism evaluating the probability of a good alignment using sentence length, and, optionally, a bilingual dictionary. If the evaluation count in HunAlign is below zero, the alignment is usually mismatched; these places should be corrected manually.

Therefore, LeoBilingua and HunAlign seem to be the best choice for the RNC and both are currently used, both allowing for user-defined sentence splitting and using statistical mechanisms of alignment. Both have their advantages. While LeoBilingua allows for a slow well-controlled process, with the possibility to split sentences manually in all tricky places and correct possible text misprints in a GUI, sending the results directly into a Unicode XML file, HunAlign aligns the whole text quickly with very few mismatches, marked and easily discerned. The latter is currently used by Ruprecht von Waldenfels in ParaSol, a project which has tasks similar in scope to the RNC parallel corpora (earlier aka Regensburg Parallel Corpus, http://www-korpus.uni-r.de/ParaSol/, see also Waldenfels 2006). However, Slavic corpora offers some challenges, including dictionary problems; as languages with rich inflection including most forms of the paradigm into dictionaries used in alignment.

3. Format and morphological tagging

The parallel texts in the RNC are presented in XML format where sentences are paired using the <para></para> tag. Each sentence has an attribute indicating the language (this may be changed when tri- or polylingual texts are inaugurated). If a sentence is not translated, an omission is marked by three dashes.

The texts are further automatically annotated using the morphological analyzers designed by the Yandex search engine. Lexical and grammatical annotation is included in the <ana></ana> tag. The tags are not currently disambiguated: however, some Russian texts selected for the parallel corpora are already manually disambiguated for the monolingual corpus and may be may be later also included in the parallel corpus.

Here is an example of a non-disambiguated aligned text. This is a 19-century translation of Pushkin's *Kapitanskaya Dochka* into English by M. de Zielenska (the translation has the title "Marie" and omits some sentences, as in this example).

```
<para>
<se lang=”rus”><w><ana lex=”между” gr=”PR”/>Между</w> <w><ana lex=”то”
gr=”SPRO,inan,ins,n,sg”/><ana lex=”тот” gr=”APRO,dat,pl”/><ana lex=”тот”
gr=”APRO,ins,m,sg”/><ana lex=”тот” gr=”APRO,ins,n,sg”/><ana lex=”тема”
gr=”S,f,gen,inan,pl”/><ana lex=”тема” gr=”S,anim,gen,m,persn,pl”/><ana
lex=”тема” gr=”S,acc,anim,m,persn,pl”/><ana lex=”тем” gr=”CONJ”/>тем</w>
<w><ana lex=”минуть” gr=”V,act,indic,n,pf,praet,sg”/>минуло</w> <w><ana
lex=”я” gr=”SPRO,1p,anim,dat,sg”/><ana lex=”я” gr=”SPRO,1p,loc,sg”/>мне</
w> <w><ana lex=”шестнадцать” gr=”NUM,nom”/><ana lex=”шестнадцать”
gr=”NUM,acc”/><ana lex=”шестнадцать” gr=”NUM”/>шестнадцать</
w> <w><ana lex=”лета” gr=”S,gen,inan,n,pl”/><ana lex=”год”
gr=”S,gen,inan,m,pl”/>лет</w>.</se>
<se lang=”eng”><w><ana lex=”in” gr=”ADV”/><ana lex=”in” gr=”PR”/>In</
w> <w><ana lex=”this” gr=”APRO,ireg,sg”/><ana lex=”this” gr=”SPRO”/>this</
w> <w><ana lex=”way” gr=”S,nom,sg”/><ana lex=”way” gr=”ADV”/>way</
w> <w><ana lex=”i” gr=”SPRO,ireg,nom”/>I</w> <w><ana lex=”reach”
gr=”V,praet,partcp”/><ana lex=”reach” gr=”V,praet,indic”/>reached</w> <w><ana
lex=”my” gr=”APRO”/>my</w> <w><ana lex=”sixteenth” gr=”S,nom,sg”/><ana
lex=”sixteenth” gr=”ANUM”/>sixteenth</w> <w><ana lex=”year”
gr=”S,nom,sg”/><ana lex=”year” gr=”A”/>year</w>.</se>
</para>
<para>
<se lang=”rus”><w><ana lex=”тут” gr=”S,inan,m,nom,sg”/><ana lex=”тут”
gr=”S,acc,inan,m,sg”/><ana lex=”тут” gr=”ADV”/><ana lex=”тута”
gr=”S,f,gen,inan,pl”/><ana lex=”тут” gr=”ADVPRO”/>Тут</w> <w><ana
lex=”судьба” gr=”S,f,inan,nom,sg”/>судьба</w> <w><ana lex=”мой”
gr=”APRO,f,nom,sg”/>моя</w> <w><ana lex=”переменяться” gr=”V,act,f,indic,p
f,praet,sg”/><ana lex=”переменяться” gr=”V,act,f,indic,pf,praet,sg”/>переменила
сь</w>.</se>
<se lang=”eng”>-----.</se>
</para>
```

The texts are made available for search online at the www.ruscorpora.ru website. Due to copyright reasons, no text is available for full view, search results are always presented in the form of separate sentences with minimal context (so--called "snippets"). The following parameters are searchable:

- any combination of lemmata, exact word forms, and morphological tags within a 10-word combination (e. g. "had" + Past Participle yields the English Pluperfect);

- names of the author and the translator, language of the original text, language of the translation text. These parameters are available by selecting a subcorpus for further textual or grammatical search. It is interesting to note than in the Ukrainian-Russian parallel corpora some writers, like Taras Shevchenko, Marko Vovchok or Lesya Ukrainka, appear as authors in both languages as well as translators (either of their own work, as Shevchenko or Vovchok, or of others' work, like Lesya), in both directions of translation.

4. Choice of texts

Linguistic corpora need to be representative and include texts of different genres, styles, and topics. This is problematic for parallel corpora because within a given pair of languages different kinds of texts are translated with different intensity. For example, few texts in mathematics have ever been translated from Ukrainian into Russian or few newspaper articles from Russian into English, whereas renowned works of fiction are typically translated from one language into another without considerable restrictions. The existing large parallel Slavic corpora (ASPAC, the Czech National Corpus or Waldenfels' ParaSol) consist almost exclusively of fiction. Nevertheless non-fiction texts are worth including in parallel corpora, a good example is the multilingual corpus of the *acquis communautaire* of the European Union or the English-German parliamentary proceedings corpus (http://corpus.leeds.ac.uk/paraquery.html). Russia, Ukraine and Belarus share a common Slavic heritage and common imperial and Soviet past; bilingualism with Russian is widespread in both Ukraine and Belarus, and Russian is a state official language in the latter. A great deal of official and semi-official documents exist in both language versions, bilingual media also are important; translated non-fiction books belong mainly to the Soviet era but are also considerable in number. So these kinds of texts are also to be included in the Ukrainian-Russian and Belorussian-Russian parallel corpora.

Texts are also provided by authors or publishers with other parallel corpora (we would like to thank Adrian Barentsen and Ruprecht von Waldenfels for their help), and can be found online (including those with expired copyright) or are scanned from printed material.

Currently five parallel corpora are available for search: English-Russian, Russian-English, German-Russian, Ukrainian-Russian and Russian-Ukrainian. The latter two are developed by a joint Russian-Ukrainian team with the help of the Institute of the Ukrainian language in Kiev; other corpora are developed together by the teams of Dmitry Dobrovolsky in Moscow and Vienna and Alexei

Kretov in Voronezh. As of September 2010 the size of the searchable corpora was as follows:

	Texts	Sentences	Tokens
English-Russian	57	436,431	6 229,619
Russian-English	24	61,259	963,507
German-Russian	15	72,553	1,270,128
Ukrainian-Russian	64	32,496	316,649
Russian-Ukrainian	16	16,330	168,125

The Russian-Belorussian and Belorussian-Russian parallel corpora are prepared for online search. This is a continuation of a project which was created and developed by Alexander Zubov's group in Minsk [Зубов 2010]. To these will be added a Polish-Russian corpus developed by Boris Orekhov's team in Ufa (Russia) and Norway and by Marek Łaziński's team at Warsaw University.

5. Corpora-based research

Notwithstanding the relatively small current size of the RNC parallel corpora, important typological, grammatical and lexical studies on its core are possible. For example, the Ukrainian Pluperfect (*прийшов був*-type) is only partly equivalent to its traditional Russian counterpart (*пришел было*). In many contexts, the cancelled result or unfinished situation expressed by the Ukrainian Pluperfect can be yielded by simple context in Russian (*Ляпнула була з маху, так би мовити, довірчо поділилась цікавим спостереженням* [O. Zabuzhko] – *Ляпнула с пылу с жару, так сказать, доверительно поделилась интересным наблюдением*). There exist also examples with "irreal modality" semantics, which are seldom described in Ukrainian grammars: *Ледве держалася на ногах, і нікого не було коло неї, о кого могла була обпертися* [O. Kobylyans'ka]. – *Едва держалась на ногах, и никого не было около нее, на кого могла бы опереться*. What is more, if we explore the Russian *пришёл было* construction where it appears as a translated equivalent of English and German sentences (the so-called translation pattern, compare: Santos 1999), we see that it is used in particular with adverbials signaling a short period of time (*for a while, einen Augenblick lang*) or with characteristic German particles as *eben,* both say a great deal about its semantics.

The study of lexical semantic highlights multiple translation patterns for come culture-significant lexemes, like Russian *тоска*, extensively commented by Anna Wierzbicka as a typical Russian concept, can be translated as *gloom, dismay, grief,* *agony, excruciating feelings, anguish, ache, despair, loneliness, yearning, longing, distress, frustration, wistfully, desperation, nostalgia, misery*... A vivid example of the Russian-Ukrainian lexical mismatch: *То ж любити, а то – кохати... Любиш – батька, матір, людей; а кохаєш – милого* [П. Мирный]. – *Любовь любви рознь. Одно – любить отца, мать, людей; а другое – милого.*

A corpus featuring non-fiction texts may also be used as a tool for exploring terminological subtleties. For example, *rule of law*, a term from Anglo-Saxon common law, can by translated into the German civil law tradition as *Rechtstaat*, *Rechstaatlichkeit* or even *Demokratie*.

Some tricky problems are unveiled about translation, translators, cultures, and ideologies. For example, Victorian translators of Tolstoy erased any allusion to prostitution; Soviet translators of Ukrainian pre-revolutionary authors changed the non-deferential references to Jews or "Muscovites" and religion-related issues. A linguist must take into consideration these possible non-linguistic distortions, while a scholar in culture studies may analyze them in their own right.

6. Cooperation

The RNC collaborates with existing Slavic parallel corpora, built on the basis of close methodology and ideas. These include ParaSOL (Waldenfels 2006), ASPAC (Amsterdam Slavic Parallel Aligned Corpus, a project by A. Barentsen http://home.medewerker.uva.nl/a.a.barentsen/), InterCorp within the Czech National Corpus (http://www.korpus.cz/intercorp/, see also Vavřín, Rosen 2009). A large, separate project *Slovo o Polku Igoreve* is also being developed with the support of the RNC (http://nevmenandr.net/slovo/, see also: Orekhov 2009).

Currently the teams of all these corpora are discussing the problem of interoperablity, common bank of resources, and perhaps a common search system for all Slavic parallel corpora available online.

References

Kotsyba N., 2009, *The Current State of Work on the Polish-Ukrainian Parallel Corpus (PolUKR)*, [in:] *Problems of Slavic Lexicography. Proceedings of the international workshop within MONDILEX project, Kyiv, 2–4 February 2009*.

Santos D., 1999, *The Pluperfect in English and Portuguese: What Translations Patterns Show*, [in:] *Out of Corpora: Studies in Honour of Stig Johansson*, eds. H. Hasselgård, S. Oksefjell, Amsterdam–Atlanta, pp. 283–299.

Vavřín M., Rosen A., 2009, *Korpus InterCorp*, available at: http://korpus.cz/intercorp--info.php

v. Waldenfels R., 2006, *Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment,* [in:] *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, eds. B. Brehmer, V. Zdanova, R. Zimny, München, pp. 123–138 (available at: http://www-nw.uni-regensburg.de/%7E.war05297.slavistik.sprachlit.uni-regensburg.de/pub/WaldenfelsParallelCorpora2006.pdf)

Андреева Е.Г., Касевич В.Б., 2005, *Грамматика и лексика (на материале англо--русского корпуса параллельных текстов)*, [in:] *Национальный корпус русского языка: 2003–2005*. Москва, с. 297–307.

Добровольский Д.О., 2009, *Корпус параллельных текстов в исследовании культурно-специфичной лексики*, [in:] *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы*, Санкт-Петербург, с. 383–401.

Добровольский Д.О, Кретов А.А., Шаров С.А., 2005a, *Корпус параллельных текстов*, [in:] *Научная и техническая информация*, сер. 2. *Информационные процессы и системы*, 2005, № 6.

Добровольский Д.О., Кретов А.А., Шаров С.А., 2005b, *Корпус параллельных текстов: архитектура и возможности использования*, [in:] *Национальный корпус русского языка: 2003–2005*, Москва, с. 263–296.

Зубов А.В., 2010, Лингво-методические возможности русско-белорусского параллельного корпуса текстов, in: Русский язык: Исторические судьбы и современность. IV Международный конгресс исследователей русского языка. Труды и материалы. Москва, с. 516–517

Орехов Б.В., 2009, *Параллельный корпус переводов «Слова о полку Игореве»: итоги и перспективы*, [in:] *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы*, Санкт-Петербург, с. 462–473.

Korpusy równoległe w Narodowym Korpusie Języka Rosyjskiego

Streszczenie

Artykuł przedstawia korpusy równolegle w ramach Narodowego Korpusu Języka Rosyjskiego. Szczególny nacisk położono na zasady wyrównywania tekstu zdaniami (alignment), zaprezentowano różne dostępne programy wyrównujące, jak LeoBilingua czy HunAlign, omówiono ich wady i zalety. Poruszono problemy tagowania gramatycznego tekstów z języków, których kategorie gramatyczne się nie pokrywają. Przedstawiono też dotychczasowy stan korpusów równoległych w ramach NKJR oraz plany na przyszłość (projekt jest daleki od ukończenia). Zaprezentowano również przykłady analiz językoznawczych opartych na korpusach równoległych.