Parallel Corpora within the Russian National Corpus

«Parallel Corpora within the Russian National Corpus»

by Dmitri V. Sitchinava

Source:
PARALLEL CORPORA WITHIN THE RUSSIAN NATIONAL CORPUS

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 2005а, 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; the Ukrainian, Belarusian and Polish parallel corpora have been included. The XML markup currently in use allows for building tri- or polylilingual corpora with multiple translations of original texts (including the option of multiple translations into the same language), presently a multilingual corpus featuring different other Slavic languages is 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 algorithm (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 Teksty (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 Zielenaska (the translation has the title “Marie” and omits some sentences, as in this example).
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 Belarusian-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 the project has bilingual (with Russian) corpora for the following languages: English, German, French, Italian, Polish, Ukrainian, and Belarusian, as well as the multilingual corpus mentioned above. The parallel corpora within the RNC are supported by the Russian Academy project “Corpus linguistics” (Программа Президиума РАН «Корпусная лингвистика» № 36-П). Additionally, the Ukrainian corpora are developed by a joint Russian-Ukrainian team with the help of the Institute of the Ukrainian language in Kiev. The Belarusian corpora are developed within a joint project together with the National Belarusian Academy (РГНФ-БРФФИ № 11-24-01004a/Bel). This is a continuation of a project which
was created by Alexander Zubov’s group in Minsk [Зубов 2010]. The Polish–Russian is developed together with Marek Łaziński’s team at Warsaw University. As of September 2012 the size of the searchable corpora was as follows:

<table>
<thead>
<tr>
<th>Languages</th>
<th>Texts</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Russian</td>
<td>146</td>
<td>13,468,367</td>
</tr>
<tr>
<td>German-Russian</td>
<td>23</td>
<td>2,211,615</td>
</tr>
<tr>
<td>Italian-Russian</td>
<td>6</td>
<td>701,835</td>
</tr>
<tr>
<td>Polish-Russian</td>
<td>22</td>
<td>2,241,453</td>
</tr>
<tr>
<td>Ukrainian-Russian</td>
<td>421</td>
<td>6,047,859</td>
</tr>
<tr>
<td>Belarusian-Russian</td>
<td>99</td>
<td>1,693,282</td>
</tr>
<tr>
<td>Multilingual</td>
<td>12</td>
<td>5,012,566</td>
</tr>
</tbody>
</table>

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, довірно поділилася 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 держалась 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 some 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

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 be 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 interoperability, common bank of resources, and perhaps a common search system for all Slavic parallel corpora available online.

References


Korpusy równoległe w Narodowym Korpusie Języka Rosyjskiego

Streszczenie

Artykuł przedstawia korpusy równoległe 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.