МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

ПАРАЛЛЕЛЬНЫЕ КОРПУСА ВОСТОЧНОСЛАВЯНСКИХ ЯЗЫКОВ: ОТРАЖЕНИЕ ИСТОРИЧЕСКОЙ СПЕЦИФИКИ ТЕКСТА И ПЕРЕВОДА[1]

Д. В. Сичинава

Институт русского языка им. В. В. Виноградова Российской академии наук, Москва

Two East Slavic languages, Ukrainian and Belorussian, remain less-resourced as far as corpora are concerned. The paper deals with two topics concerning the experience of building parallel corpora for East Slavic: the availability of quality translations and handling free (loose) translations of fiction. A system of markup for loose translations is proposed and illustrated by corpora examples.

Введение

Восточнославянские языки — украинский и белорусский — сами по себе не могут считаться недостаточно документированными (less-resourced). Это официальные письменные языки в соответствующих государствах, а также литературные языки, представленные большим количеством опубликованных и оцифрованных текстов. Однако ни один из этих языков на начало 2011 г. не имеет доступного национального корпуса, что для славянских языков после корпусного прорыва 1990–2000-х годов уже необычно. Существует Corpus Albaruthenicum — собрание белорусских научных текстов[2] и корпуса украинского языка, разработанные в Лаборатории компьютерной лингвистики Киевского университета[3], но оба эти корпуса невелики; корпус же Языкового информационного фонда Национальной академии наук Украины не доступен для исследователей. Оба языка также сравнительно скудно представлены в коллекциях выровненных «массивов параллельных текстов» (термин [Cysouw and Wälchli, 2007]; ср. проект PARASOL [Waldenfels, 2006]).

В докладе мы обсудим разработку параллельных аннотированных украинско-русских и белорусско-русских корпусов и специфику этого процесса, связанную с близкородственностью и социолингвистической историей восточнославянских языков. Данный опыт небесполезен и для создания параллельных корпусов других языков постсоветского пространства.

1. Параллельные корпуса НКРЯ.

В настоящее время НКРЯ (Национальный корпус русского языка) совместно с украинскими и белорусскими исследователями работает над созданием украинско-русского и белорусско-русских корпусов. Эти корпуса

---

[1] Работа подготовлена при поддержке совместного российско-белорусского гранта РГНФ-БРФФИ № 11–24–01004а/Bel (руководители А.М.Молдован и В.А.Кощенко), а также программы Президиума РАН «Корпусная лингвистика».

[2] http://grid.bntu.by/corpus/

[3] http://www.mova.info/corpus.aspx?l1=209

входят в состав более широкого проекта параллельных корпусов НКРЯ. Данные языковые пары в настоящее время доступны для онлайн-поиска (соответственно http://ruscorpora.ru/search-para-uk.html and http://ruscorpora.ru/search-para-be.html). Оба корпуса растут: украинско-русский корпус достиг 6 миллионов словоупотреблений, белорусско-русский — 2 миллионов. Планируемый объём — 10 миллионов словоупотреблений. Предполагается, что в оба корпуса войдут тексты различных типов: художественная литература, научные тексты, публицистика, правовые документы.

Тексты выровнены при помощи свободно распространяемой программы HunAlign[1], разработанной в Венгрии. Соответствия вычисляются в основном по длине предложений. Первичные результаты выравнивания корректируются вручную. Сотрудниками НКРЯ (Т.А.Архангельский) разработан графический интерфейс пользователя (GUI) для постредактирования выровненных текстов и приписывания метаинформации (сведения об авторстве, названии, дате создания и т. п.). Выровненные тексты сохраняются в формате XML и кодировке UTF-8. Тексты проходят морфологическую разметку на базе анализатора Mystem[2], после чего доступны для поиска онлайн по грамматическим тегам, лексемам и словоформам.

Для восточнославянских языков особую роль играет отбор текстов. Как отмечено в [Cysouw and Wälchli, 2007], перед составителями параллельных корпусов стоит проблема репрезентативности доступных текстов: например, обычно широко доступны переводы достаточно специфических текстов, как, например, юридических сочинений с их особым языком или Библии, для которой миссионеры обычно создавали специальный подъязык-«агиолект». Схожие проблемы со своей спецификой стоят и перед разработчиками параллельного корпуса.

2. От сотворчества к машинному переводу: специфика доступных текстов.

Казалось бы, украинские и белорусские тексты и их переводы на русский (и обратно) доступны в большом количестве. Однако это множество текстов достаточно далеко отстоит от «идеального» параллельного корпуса — репрезентативного набора текстов, представляющих все жанры функционирования и переведённых на другой язык с большой точностью.

Известно, что украинско-русский и белорусско-русский перевод, появившийся в XIX веке, изначально был сильно ограничен тематически (прежде всего художественной литературой и в основном стихотворными текстами), что было связано с функциональным статусом соответствующих идиомов (до 1917 года фактически не признававшихся отдельными от русского и до 1905 года сильно ограниченных с точки зрения цензурной возможности публикаций). Для начального периода украинской литературы характерно существование двуязычных авторов и авторских переводов, в том числе весьма вольных. Вольность этих переводов диктовалась, среди прочего, со-

---

[1] http://mokk.bme.hu/resources/hunalign/

[2] http://company.yandex.ru/technologies/mystem/

циологическим контекстом функционирования украинского языка и ориентацией авторов на простонародный колорит (и на соответствующую аудиторию — что могло не соответствовать реальному кругу читателей). Ср. пример из романа П.А.Кулиша «Чёрная рада» (русская и украинская версия создавались в 1846–1857 гг. и опубликованы примерно одновременно):

И он достал с полки большую серебряную кружку с барельефами, представлявшими греческих вакханок. Крышка была украшена литою статуйкою Фауна.

І дістав із полички жбан, прехимерно з срібла вилитий і що то вже заприукрашений! Не жалували пани грошей для своєї пихи і потіхи. Побоках бігли босоніж дівчата — інша і в бубон б'є, а зверху сидів, мовживий, божок гречеський, Бахус.

В советское время с языков народов СССР на русский переводилась национальная художественная литература, а с русского на эти языки — помимо русской литературы (а часто и иностранной, с языка-посредника), также официальные документы, пропаганда, марксистско-ленинский канон. Нехудожественная литература переводилась очень ограниченно: с 1930-х годов эта область в СССР была почти повсеместно русифицирована. Сохранялась в украинской и белорусской советской литературе и традиция авторского и авторизованного перевода, часто весьма сильно отклоняющегося от оригинала (фактически — другая авторская редакция на другом языке). Что касается дореволюционной классики, она нередко подвергалась в переводе стилистическому искажению и/или идеологической цензуре. В постсоветский период сохраняется (особенно в Белоруссии, но во многом и на Украине) функциональное неравноправие языков и преобладание русского, при доступности национальных литератур в оригинале, поэтому переводы на русский и с русского вызывают ограниченный читательский интерес. На этом фоне есть отдельные изменения в лучшую сторону, прежде всего публикация современной украинской литературы в России, но полной репрезентативной картины пока не создаётся. С другой стороны, существует огромное количество двуязычных текстов (прежде всего новостных, например, украинско-русский параллельный корпус [Ланде и Жигало, 2010]), создаваемых при помощи постредактируемого машинного перевода, что достаточно удобно для близкородственных языков. Разумеется, эти тексты не свободны от недовыправленных ошибок машинного перевода, что затрудняет их массовое использование в корпусе.

Наконец, важной особенностью массовых двуязычных текстов является неопределенность направления перевода (важного параметра для изучения параллельных корпусов).

3. Разметка неточного перевода.

Неточности перевода ранее не размечались в параллельных корпусах, по крайней мере, систематически (нередко из автоматически создаваемых корпусов для целей машинного перевода такие пары предложений просто выбрасываются), однако эта информация важна для контрастивных и типологических исследований. Используются следующие теги:

Loose=add = “в переводе добавлена часть предложения или новое предложение”

Loose=omit = “в переводе опущена часть предложения или новое предложение”

Loose=change = “в переводе изменена часть предложения

Пример — перевод с белорусского на русский (И. Шамякин, «Торговка и поэт», авторизованный перевод Т. Шамякиной)

<se lang="be">Вольга ўскочыла ў магазін, не вельмі задумаўшыся, чаму людзі так паводзяць сябе, і здзівілася, калі ўбачыла міліцыянераў, разгубілася і спалохалася.</se> <se lang="ru" loose="add">Ольга вбежала в магазин, не очень задумываясь, почему люди так странно ведут себя, и... увидела милиционеров. Испугалась. **Вот те и нет власти!**</se>

Перевод с украинского на русский (А. Довженко, «Ночь перед боем», перевод автора).

<se lang="uk">— Ідіть собі **під три чорти... Чорт вас носить**, — сказав байдужим голосом дід Савка.</se>

<se lang="ru" loose="omit">— Идите себе... — равнодушно сказал дед Савка.</se>

Возможна разработка более тонкого механизма разметки неточного перевода, а также применение существующей разметки к другим парам языков.

Список литературы

Cysouw and Wälchli, 2007 — Cysouw, M., Wälchli, B. Parallel texts: using translational equivalents in linguistic typology // STUF — Language Typology and Universals. 2007. Vol. 60, № 2, Pp. 95–99.

Ландэ и Жигало, 2010 — Ландэ Д.В., Жигало В.В. О создании параллельного двуязычного корпуса веб-публикаций // http://infostream.ua/ling/ml-small-end.pdf

Waldenfels, 2006 — Von Waldenfels, R. Compiling a parallel corpus of Slavic languages: text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B., Zdanova, V., Zimny, R. (Hrsg.); Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 2006. Pp. 123–138.