Радован Гарабик, Виктор Захаров

ПАРАЛЛЕЛЬНЫЙ РУССКО-СЛОВАЦКИЙ КОРПУС

1. Введение

Одной из интенсивно развивающихся областей современной корпусной лингвистики является развитие многоязычных ресурсов, в том числе особенно параллельных корпусов, которые позволяют исследовать проблемы перевода текстов, сравнивать (при соответствующей разметке) лексические, грамматические и синтаксические структуры разных языков, а также являются базой для отладки систем автоматического перевода и для создания словарей.

2. Формат и обработка текстов

Тексты, входящие в состав корпуса, подвергаются обработке и конверсии на нескольких уровнях, причём на каждом уровне проводится специфический тип обработки. Эта модульная система при необходимости внесения изменений позволяет заменить только нужную часть без переработки целой системы. Сначала тексты переводятся с входных форматов (HTML, MS Word, Open Document Text и др.) в общий текстовый формат в кодировке UTF-8[1] с абзацами, разделёнными пустой строкой. Такой формат удобно редактировать вручную, чтобы сравнивать начало и конец параллельных текстов, или удалять части, которые отсутствуют в одном из текстов (как, например, предисловие переводчика). Этот файл копируется в неизменном виде на следующий уровень (что позволяет проверить редактирование на предыдущем шаге и исправить ошибки или вернуть неправильно удалённые части текста). После этого текст лемматизируется,

1 *The Unicode Consortium*. The Unicode Standard, Version 4.0. Boston, MA: Addison-Wesley Developers Press, 2003.

морфологически размечается и записывается в формате TEI XML[2]. Этот формат конвертируется в следующий файл в формате, удовлетворяющем требованиям программы выравнивания (каждое предложение отдельной строкой, абзацы определены специальным символом ¶). После сравнения этого файла с соответствующим файлом на параллельном языке результаты выравнивания включаются в TEI XML файл, где каждое предложение снабжается ссылками в параллельный файл, которые записываются как атрибуты предложений (напр., `<s link="20+21+22">` значит, что этому предложению во втором языке соответствуют предложения с номерами 20, 21 и 22). После того размеченный таким образом текст конвертируется в формат корпусного менеджера.

3. Морфологическая разметка

Тексты в Словацком национальном корпусе автоматически лемматизированы и морфологически размечены[3]. Система морфологических тегов описывает все грамматические категории слова и основывается на позиционном кодировании. Каждой отдельной грамматической категории соответствует один символ и определенная позиция. Первую позицию занимает код части речи, включая сюда и коды для сокращений, знаков препинания, цифр, иностранных слов и неопределенных элементов текста.

Морфологическая разметка русских текстов базируется на программе морфологической разметки, разработанной А. Сокирко, на основе морфологического анализа системы

---

2 *Ide, N., Bonhome, P., Romary, L.* XCES: An XML-based Encoding Standard for Linguistic Corpora. In: Proceedings of the Second International Language Resources and Evaluation conference. Paris: European Language Resources Association, 2000.

3 *Garabík, R., Gianitsová, L., Horák, A., Šimková, M.*: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. In: **http://korpus.juls.savba.sk/publications/block2/**

«Диалинг». В этой программе граммемы записываются в виде ключевых слов с их значениями[4]. Далее эта форма записи приводится к формату, принятому в корпусном менеджере.

4. Библиографическая разметка

Библиографическая разметка в принципе следует систему аннотации Словацкого национального корпуса[5], где аннотация каждого документа включает библиографическое описание источника, стиль и жанр текста, дату издания оригинала, дату издания перевода, оригинальное название, имя и пол переводчика и автора.

5. Выравнивание

Для выравнивания использована программа hunalign[6], которая автоматически сравнивает тексты на основе совпадения относительных длин предложений, разделения текста на абзацы и внешнего словаря. Тексты могут поступать на вход программы выравнивания либо без всякой лингвистической обработки (только с сегментацией на предложения), либо в лемматизированном виде. Присутствие лемм является необходимым условием для использования словаря (так как в словаре содержатся только основные формы слов). В нашем корпусе сначала выравнивание производилось без словаря, затем на основе результатов выравнивания автоматически возник предварительный словарь совпадающих слов, из которого вручную были удалены неверные пары (около 75% пар) и добавлены переводы самых частотных слов. Далее этот словарь был использован для повторного выравнивания тех же текстов.

---

4 См. **http://www.aot.ru**

5 Garabík, R. Словацкий национальный корпус. In: Труды международной конференции Корпусная лингвистика, Санкт-Петербург: Издательство С.-Петербургского университета, 2004, p. 99 – 121.

6 **http://mokk.bme.hu/resources/hunalign**

6. Проблемы

Избыточные фрагменты в начале или в конце текстов (как, например, предисловие переводчика или информация об авторе) необходимо удалять вручную. Также необходимо сравнивать начало оригинального и переводного текстов, в том числе название, имя автора, и их концовки, особенно последние строки, которые часто содержат разные данные (напр., год или место возникновения книги, подпись автора), так как для хорошего выравнивания требуется, чтобы оригинал и перевод максимально совпадали.

Значительная часть переводов с русского на словацкий – это русская классика 19-го века, следовательно, это достаточно важная часть корпуса. Тексты русской классики отличаются несколькими специфическими свойствами. Например, присутствие французских фраз в русском тексте, которые иногда переведены в сносках, а иногда и нет. Надо иметь в виду, что оригинальная версия русского текста часто включала французский текст без перевода как отражение языковой ситуации своего времени. Также следует учитывать, что мы работаем с электронными версиями книг, в процессе создания которых такие сноски могли потеряться. В словацком тексте эти фразы либо поясняются (в сносках), либо переводятся на словацкий язык без примечаний, что в оригинале они представлены не на русском языке. Таким образом возникают разные ситуации:, лишний текст (сноски) в словацкой части корпуса либо в русской (в словацкой сноски отсутствуют), или же французскому тексту в русской части соответствует фрагмент на словацком языке. Всё это портит качество выравнивания.

Вторая важная проблема связана с тем, что выравнивание происходит на основе предложений, а членение текста на предложения в переводном тексте часто отличается от оригинала. Чаще всего это происходит в прямой речи, как мы это эмпирически обнаружили. Кроме того, запись прямой речи в

переводе часто использует другие типографские знаки и другие правила оформления по сравнению с оригиналом. Эту проблему, по-видимому, можно решить автоматически путём тщательной настройки алгоритма сегментации текста на предложения.

7. Поиск в корпусе

Для поиска в корпусе пользуется система Manatee/Bonito[7], которая состоит из сервера (Manatee) и клиента (Bonito), но в нашем параллельном корпусе клиент не используется. Вместе него был создан пользовательский веб-интерфейс с использованием программного шаблона для веб-приложений Karrigell[8] в языке программирования Python[9]. Пользовательский интерфейс включает виртуальную клавиатуру с буквами русского алфавита, словацкими буквами с диакритикой и несколькими другими полезными буквами и символами. Сервер позволяет осуществлять простой поиск одного слова, или фразы (несколько слов в определенном порядке), или произвольных регулярных выражений из слов, лемм и морфологических тегов. Веб-интерфейс (рис. 1) доступен в открытом доступе на страничке Словацкого национального корпуса[10].

---

7 Rychlý, P. PhD Thesis: Korpusové manažery a jejich efektivní implementace. Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2000.

8 **http://karrigell.sf.net**

9 **http://www.python.org**

10 **http://korpus.juls.savba.sk/parus/**

Рис. 1. Пользовательский интерфейс корпуса

8. Состояние дел и направления дальнейшего развития

В настоящее время корпус содержит в словацкой части 818 097 слов, 43 381 предложений, и в русской части 819 009 слов и 46 832 предложений. Разница в количестве предложений, скорее всего, происходит от несовершенства алгоритма сегментации и не имеет других важных причин. Из предварительного исследования в корпусе мы получили, что выравнивание совсем отсутствует в 2.4% предложений и 0.6% предложений содержат лишние сноски в словацком тексте, объясняющие французские фразы. Также 24.1% пар предложений таковы, что одному предложению в одном языке соответствуют два или больше предложений во втором (но при этом они выровнены таким образом, что в этом предложении всегда есть правильная ссылка хотя бы к одному из соответствующих ему предложений второго языка). Из этого следует, что для усовершенствования выравнивания самым полезным будет доработка алгоритмов автоматической сегментации по предложениям.

В дальнейшем развитии корпуса мы собираемся, прежде всего, включить в корпус как можно большее число текстов, главным образом, текстов русской классики, но также и тексты других жанров, расширить возможности отображения результатов поиска (в частности, дать возможность просмотра дополнительного контекста конкордансов) и сделать исправления и усовершенствования в пользовательском интерфейсе.