А.Б. Кутузов

Тюменский государственный университет

akutuzov72@gmail.com

Опубликовано: Сборник научных трудов "Проблемы теории, практики и дидактики перевода", вып.14, т.1. - Нижний Новгород, 2011, сс. 51-54

Корпус несовершенных переводов: необходимость проекта

Прошло уже почти 20 лет с тех пор, как корпусная лингвистика и переводоведение осознали свой взаимный интерес друг к другу. Отсчёт использованию корпусной лингвистики в переводоведении, видимо, следует вести с 1993 года, когда Мона Бейкер опубликовала свою работу «Corpus Linguistics and Translation Studies: Implications and Applications». [1]

Сегодня уже окончательно ясно, что научная парадигма в лингвистике сдвигается в сторону изучения именно корпусных данных и корпусными методами. Чисто качественные методы уступают место синтезу количественных и качественных исследований, проводимых над огромными массивами текстов.

Переводоведение как ветвь лингвистики не отстаёт от этого процесса. Уже упомянутая Мона Бейкер (Центр переводоведческих и межкультурных исследований университета Манчестера) создала Translational English Corpus (TEC)[1] — массив письменных текстов, переведённых на английский с различных языков, насчитывающий 10 миллионов словоупотреблений. В него включены тексты, относящиеся к художественным и биографическим, а также новости и журналы. TEC подробно размечен, позволяет производить поиск по самым разным характеристикам переводчиков и тем самым решать исследовательские задачи, связанные с контекстуализацией переводов.

Российские корпусные лингвисты также не забывают переводоведов. В Национальном корпусе русского языка выделены параллельные подкорпусы[2], содержащие переводы и оригиналы художественных текстов. В них можно найти все переводы для определённого слова или словосочетания. Таким образом, исследователь получает возможность увидеть в большом масштабе, как реальные переводчики работают с реальными текстами и какой продукт (переводной текст) получается в результате.

Итак, параллельные корпусы англо-русских и русско-английских письменных переводов существуют. И тем не менее, мы (кафедра перевода и переводоведения ТюмГУ) предлагаем проект ещё одного корпуса, а именно — корпуса «несовершенных» или ошибочных переводов.

Основное отличие от перечисленных аналогов заключается в том, что этот корпус предполагается использовать для исследований, связанных с языковыми и

1 http://www.llc.manchester.ac.uk/ctis/research/english-corpus

2 http://ruscorpora.ru/search-para.html

переводческими ошибками. Для этого необходимо разработать формальную лингвистическую систему разметки ошибок на основе языка XML. XML позволит этой системе быть достаточно гибкой, чтобы реализовать применение различных классификаций ошибок и удобный поиск по любым критериям, включая серьёзность и характер ошибки, а так же характеристики переводчика.

К сожалению публикации, посвящённые собственно переводческой разметке, в отечественном и зарубежном переводоведении редки, хотя тема представляется весьма актуальной. Разметка переводных устных корпусов разрабатывается в университете Тампере, Финляндия [2], но в ней не учитываются переводческие ошибки. В данном проекте мы планируем основываться на работе «Применение дескриптивной разметки для формализации оценки качества перевода» [3].

Вообще, корпусные исследования ошибок перевода важны как в общетеоретическом плане (дают возможность глубже понять процессы межъязыкового переноса и взаимодействие между разными языковыми системами), так и в практическом (появляется прочная база для улучшения существующих курсов обучения переводу и для совершенствования систем автоматизированной оценки переводов). Отметим также, что концепция нашего проекта хорошо укладывается в общемировую тенденцию к автоматизации и стандартизации всей цепочки процессов порождения, перевода и публикации текстов [4].

Тем не менее, насколько нам известно, в России подобные проекты пока до конца не осуществлялись, хотя слова о необходимости «корпуса переводческих ошибок» уже звучат (см., например, [5]). Известно лишь, что с 2004 года в Ульяновском государственном техническом университете на базе кафедры «Прикладная лингвистика» осуществлялся проект по созданию и анализу электронного учебного корпуса переводов RuTLC (Russian Translation Learner Corpus) [6]. Но, к сожалению, этот корпус недоступен через Интернет, что делает невозможным его полноценное использование. Кроме того, объём в 1 миллион словоупотреблений, заявленный создателями этого корпуса, мы считаем недостаточным.

Мы планируем набирать параллельные тексты для корпуса в основном из студенческих переводов в связи с тем, что этот материал широко доступен и содержит достаточное количество переводческих ошибок. Часто оригиналы и студенческие переводы сразу существуют в электронном виде, что исключает необходимость их сканирования и распознавания. В целях репрезентативности переводы необходимо будет собирать не только в Тюменском государственном университете, но и в других высших учебных заведениях, на постоянной основе занимающихся подготовкой переводчиков. Целевой объём корпуса установлен в 10 миллионов словоупотреблений, по аналогии с существующими параллельными и переводоведческими корпусами: соответствующим подкорпусом НКРЯ и The Translational English Corpus.

Работа над проектом включает в себя как лингвистическую так и дополнительную программистскую составляющую (техническое обеспечение корпуса). Задача лингвистов-переводоведов будет заключаться в разработке критериев отбора текстов в корпус, самой процедуре отбора, создании схемы дескриптивной разметки ошибок и её применении к корпусу. Конечно же, основной проблемой станет разметка корпуса по ошибкам: если небольшой корпус ещё можно разметить вручную, то огромные объёмы текстов потребуют автоматизации этого процесса (с последующей проверкой человеком).

Круг потенциальных пользователей проекта довольно широк: это и лингвисты, специализирующиеся в области переводоведения, и психолингвисты, и преподаватели перевода, а также организации, заинтересованные в оптимизации своих переводческих процессов. Вероятнее всего, поиск в корпусе будет организован через свободно доступный веб-сайт, подобно перечисленным выше корпусам.

Реализация проекта логически распадается на несколько этапов:

1. Набор пилотного параллельного корпуса в 1 миллион словоупотреблений.
2. Организация публичного доступа к пилотному корпусу через интернет.
3. Разметка переводческих ошибок в пилотном корпусе.
4. Совершенствование методики.
5. Набор основного корпуса (до 10 миллионов словоупотреблений).
6. Разметка основного корпуса.

Таким образом, на первом этапе нами будет создан сравнительно небольшой неглубоко аннотированный параллельный корпус «ошибочных» переводов, который будет полностью доступен для исследователей через Интернет. Учтя замечания пользователей, мы проведём в нём разметку ошибок, а затем приступим к формированию основного массива корпуса.

Подводя итоги, отметим, что одна из основных задач корпусной лингвистики — предоставление другим ветвям науки о языке полноценного и адекватного материала для исследований. Наш корпус несовершенных переводов нацелен именно на это: дать лингвистам (и другим заинтересованным субъектам) ту базу, на основе которой они смогут проверить свои предположения о различных аспектах сложнейшей человеческой деятельности под названием «перевод».

Библиография

1. **Baker, M.** Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology: In Honour of John Sinclair. - Amsterdam & Philadelphia: John Benjamins, 1993. - pp. 233-250.
2. **Михайлов, М.Н., Исолахти, Н.Б**. Корпус устных переводов как новый тип

корпуса текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14).– М.: РГГУ, 2008. с. 376-381

3. **Кутузов, А.Б.** Применение дескриптивной разметки для формализации оценки качества перевода // Индустрия перевода и информационное обеспечение внешнеэкономической деятельности предприятий: материалы Международной научно-практической конференции. - Пермь: Изд-во ПГТУ, 2008 г.
4. **M.T. Carrasco Benitez.** Open architecture for multilingual parallel texts // Электронный сборник препринтов arxiv.org, URL: http://arxiv.org/abs/0808.3889, 2008
5. **Степанова М.М.** Анализ переводческих ошибок в подготовке преподавателей перевода // Дидактика перевода: Материалы научной конференции / Под ред. проф. В.Н. Базылева. - М.: Гос.ИРЯ им. А.С. Пушкина, 2010. - 96 с.
6. **Ekaterina Sosnina.** Russian Translation Learner Corpus: The First Insights // The proceedings of the 6 international scientific conference «Interactive systems: problems of human-computer interaction» - Ulyanovsk: UlSTU, 2005