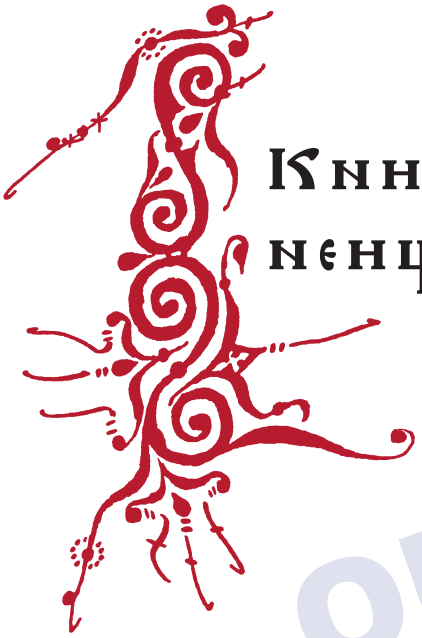
UPPSALA
UNIVERSITET



Книгамъ бо есть
ненщетнаѧ глубина

Essays in Honour of Irina Lysén

Edited by
Hanne Martine Eckhoff and Thomas Rosén

Параллельный корпус в системе университетских лингвистических курсов

Атле Грённ, Университет Осло

Корпус RuN-Euro, первоначально разработанный как русско-норвежско-английский параллельный корпус в университете Осло, возник не благодаря нашему интересу к корпусной лингвистике, а скорее как побочный продукт при решении исследовательских и учебно-методических проблем в теоретической лингвистике, контрастивной лингвистике и преподавании русского языка как иностранного. По этой причине работа над параллельным корпусом была неразрывно связана с конкретными теоретическими и практическими задачами. Некоторые из них представлены в настоящей статье, в частности описывается онлайн-программа RuN Interactive Translation (RuN-IT).

1. Все ли мы являемся корпусными лингвистами?

Все лингвисты ценят хорошие аутентичные данные. В этом смысле мы все рады приходу эры корпусной лингвистики, например, в изучении связи формы и содержания (взаимодействия синтаксиса и семантики), что является моей специализацией. Однако, прежде чем я начну традиционную презентацию нашего корпуса, мне бы хотелось обратить внимание читателя на одну проблему. Тогда как теоретические лингвисты имеют четкое представление о том, что значит оценка носителей языка для их теорий (теория должна предсказать, будет ли то или иное предложение признано правильным носителями языка) – пока остается неясным, как проблемы частотности связаны с грамматикой.

Корпусная лингвистика предполагает квантитативные методы, однако значение частотности для синтактико-семантического взаимодействия остается пока неопределенным. Мой собственный подход является весьма традиционным. Я пытаюсь выстроить лингвистический анализ таким образом, чтобы он смог объяснить найденные примеры – неважно, встречаются ли они в корпусе десять, сто или тысячу раз. В действительности единичные (нечастотные) случаи часто оказываются наиболее интересными для лингвистической теории. С этой точки зрения, я занимаюсь квалитативной лингвистикой, и читатель вправе счи-

тать, что подобный подход не относится к собственно корпусной лингвистике.

2. Проект RuN и корпус RuN-Euro

Проект RuN (2008–2010 гг.), «Встреча русского с норвежским – языки в пересечении», спонсировался Норвежским центром международного сотрудничества в сфере высшего образования (SIU) в рамках программы сотрудничества с Россией, инициированной Министерством иностранных дел. Идея, положенная в основу этого проекта, заключалась в том, чтобы сосредоточить внимание на корпусной лингвистике как области, в которой мы могли бы сократить досадный разрыв между научными исследованиями и высшим образованием, нередко возникающий на практике. В нашем случае, в университете Осло, мы стремились создать новую учебную среду для студентов, изучающих русский язык на продвинутом уровне, – среду, которая бы имела прямое отношение к новейшим исследованиям в области лингвистики. Чтобы добиться этой цели, мы пришли к идее создания русско-норвежского параллельного корпуса – корпуса RuN.

Однако нельзя сказать, чтобы корпус RuN возник из ниоткуда. В университете Осло, в частности на Кафедре литературы, страноведения и европейских языков, существуют сильные традиции в области разработки параллельных корпусов, берущие свое начало в англо-норвежском параллельном корпусе, инициированном Стигом Юханссоном в 1994 году. Англо-норвежский корпус был впоследствии расширен и дополнен другими языками, в результате чего появился Ословский многоязычный корпус (Oslo Multilingual Corpus, или OMC). В проекте RuN мы использовали программу для выравнивания параллельных текстов, разработанную командой многоязычного корпуса OMC (Hofland and Johansson 1998). Использование этой программы предполагает полуавтоматическое выравнивание. В отличие от полностью автоматических программ для выравнивания, таких как Hunalign, наша программа обладает интерфейсом, который позволяет человеку вмешаться, если обнаруживаются ошибки в предложенном программой выравнивании. Поэтому мы рассчитываем, что наш корпус содержит меньше ошибок по сравнению с аналогичными корпусами, использующими исключительно автоматические программы. С другой стороны, выбранный нами метод выравнивания (наполовину автоматический, наполовину осуществляемый вручную), разумеется, отнимает много времени, и мы постоянно вынуждены задаваться вопросом, стоит ли он таких усилий. Ответ на этот вопрос зависит от того, имеется ли у нас достаточно финансовых средств, чтобы оплачивать работу ассистентов. Вполне возможно, нам придется отказаться от полуавтоматического метода выравнивания.

В 2010 и 2011 годах мы расширили свой проект, включив немецкий, шведский, французский, итальянский, болгарский, сербохорватский и польский языки и таким образом сделав наш корпус по-настоящему многоязычным. Поскольку нашим новым приоритетом стали европейские языки в целом, нам пришлось изменить название корпуса: прежний корпус RuN теперь называется RuN-Euro.

Абсолютное большинство текстов в корпусе RuN-Euro – это художественная литература. Русский, норвежский и английский остаются главными языками корпуса, который в совокупности насчитывает 9 миллионов слов. При выборе текстов мы стремимся не дублировать те, что уже включены в другие параллельные корпуса, такие как Intercorp, Parasol и Национальный корпус русского языка (НКРЯ), однако некоторые совпадения все же неизбежны.

У корпусов RuN-Euro и ОМС имеется общий веб-интерфейс Glossa, разработанный Текстовой лабораторией при Кафедре лингвистики и скандинавских языков в Осло. Glossa – это графический интерфейс, созданный на основе корпусного менеджера IMS Corpus Workbench. Морфологическая разметка текстов посредством тэгов осуществляется Текстовой Лабораторией с использованием и адаптацией существующих маркировочных моделей (Nygaard et al. 2008).

Поисковый интерфейс позволяет ограничить поиск по дате публикации, автору, жанру и другим параметрам. Однако изначально Glossa не разрабатывалась для работы с параллельными корпусами, поэтому у интерфейса есть некоторые аспекты, которые требуют улучшений, что, вероятнее всего, и будет сделано в ближайшее время. Так, к примеру, в Glossa встроены весьма продвинутые статистические инструменты для обработки результатов, полученных для *исходного языка*. Однако статистическая обработка данных невозможна для языка перевода. Пользователь может просматривать тексты, увеличивать размер контекста и получать доступ к метаданным на языке перевода. Если же требуется дальнейшая обработка результатов, полученных для языка перевода (например, посредством статистического анализа), это придется делать «вручную», вводя данные на языке перевода в соответствующую программу.

3. Области применения

3.1 Курсы для магистерских и аспирантских программ

По итогам проекта RuN были разработаны новые курсы для магистерских программ, на которых студенты активно работают с корпусом RuN-Euro, занимаясь исследованиями в области контрастивной лингвистики. Среди прочего наши студенты пишут зачетные работы – в том числе магистерские и кандидатские диссертации – занимаясь сопоста-

вительным анализом русского, норвежского (и английского) языков и используя данные корпуса.

Разницу между квантитативным и квалитативным подходами можно продемонстрировать на примерах из диссертации Марии Филюшкной Краве (PhD, 2011) о русских деепричастиях и их эквивалентах в переводах на английский и норвежский языки. Для начала обратите внимание на тот факт, что русские деепричастия делимитативного способа действия переводятся с эксплицитными маркерами предшествования в 70% случаев (см. 1P/1E), тогда как семельфактивные деепричастия переводятся с эксплицитными маркерами предшествования лишь в 1% случаев. Вместо этого мы видим в переводе комитативные абсолютные конструкции и т.п. (см. 2P/2E):

(1P) – Нет, – *подумав*, отвечал Левин.
(Толстой, «Анна Каренина»)
(1E) No, answered Levin, *after* an instant's thought.

(2P) Сергей сидел в углу, *закинув* ногу на ногу, и курил.
(Пелевин, «Поколение П»)
(2E) Sergei sat in the corner *with* his legs crossed, smoking.

Теоретически (временные) свойства делимитативов vs. семельфактивов должны позволить нам объяснить этот контраст, однако точные цифры (70% против 1%) здесь не так важны. Следующие два примера более интересны и, кстати, оказались единственными примерами такого рода во всем корпусе. Вопрос заключается в том, можно ли свести значение деепричастия совершенного вида, семантику суффикса *-в*, к значению предшествования.

(3P) В конце девятого класса Ника завела увлекательный роман с молодежным поэтом, […] укатила с ним в Коктебель, *сообщив* об этом телеграммой *постфактум*, уже из Симферополя.
(Улицкая, «Медея и ее дети»)
(3E) At the end of the ninth grade Nike embarked on a headlong romance with a youth poet […] flounced off with him to Koktebel, *announcing* this ex *post facto* by telegram when she was already in Simferopol.

(4P) По ночам […] они вели долгие содержательные разговоры, *сохранив с тех пор* на всю жизнь глубокое чувство душевной близости.
(Улицкая, «Медея и ее дети»)

(4E) At night […] they engaged in long, deeply meaningful conversations *and retained from that time* for the whole of the rest of their lives a deep emotional bond.

При квалитативном подходе одного-единственного исключения достаточно для того, чтобы создать трудности для анализа. Однако это те самые моменты, когда возникают наиболее интересные, нетривиальные вопросы. Более подробный анализ и возможные решения вопросов, поставленных в примерах (3P) и (4P), см. в работе Krave 2011.

Среди прочих тем, изучаемых магистрантами и аспирантами в университете Осло: вид в славянских императивах (Alvestad 2013), реалии в переводах с русского на норвежский (Kharina, PhD 2013-15), дискурсивные частицы в сопоставлении, интерпретация и перевод местоимений с *-то/-нибудь*, перевод на норвежский глаголов начинательного способа действия с приставками *по-* и *за-* и т.д. Корпус RuN-Euro также активно используется при написании зачетных работ на бакалаврском уровне. В общей сложности за последние пять лет под моим научным руководством студенты написали около ста работ по русскому языку с привлечением лингвистического анализа контрастивных данных из нашего корпуса. Большинство из этих работ содержит таблицы и статистические данные, однако наиболее интересные проблемы почти всегда обнаруживаются именно в мелочах. Лучшие студенты предъявляют новые данные, которые ставят под сомнение сложившиеся представления. Как в случае с диссертацией Краве – одного или двух исключений бывает достаточно, чтобы поднять новые научные вопросы.

Таким образом, параллельные корпуса не только предоставляют материал для иллюстрации лингвистического анализа, но и способны заставить исследователя задаться новыми вопросами при столкновении с неоднозначными реальными данными. Хотя языки вообще и переводы в параллельных корпусах в частности отличаются порой самым неожиданным образом, нас не должно это пугать и обескураживать. В следующем разделе мы предлагаем метод для изучения проблемы «переводизмов» (“translationese”) и параллельных корпусов.

3.2 Интерактивный перевод RuN

Стандартное возражение против использования параллельных корпусов – проблема «переводизмов» – по сути не относится к тем из нас, кто не занимается квантитативной корпусной лингвистикой. Мы можем попросту игнорировать неудобные или неверные данные. Тем не менее мне бы хотелось внести свой вклад в поддержку квантитативных исследований, продемонстрировав один из способов применения нашего

параллельного корпуса – разработку интерактивного переводческого теста, RuN-IT.

Эта программа для перевода создавалась, главным образом, в целях ее использования в преподавании иностранных языков и самообучении русскому языку на продвинутом уровне, и в этом качестве она активно применяется как студентами-бакалаврами, так и магистрантами. Студенты могут ввести свои варианты перевода в режиме онлайн, а затем сверить их по базе данных с аутентичными переводами носителей языка, а также ознакомиться с комментариями, сделанными исследователями проекта RuN, и списками типичных ошибок, допущенных другими студентами.

Помимо этого, полученная в результате база данных может оказаться ценной для лингвистических исследований (в контрастивной лингвистике, переводоведении), а также предоставить нам убедительные доказательства в дискуссии о том, насколько можно доверять параллельным корпусам в лингвистических исследованиях.

Идея, лежащая в основе Интерактивного Перевода RuN, проста (хотя, разумеется, довольно сложна на практике из-за недостатка финансовых/человеческих ресурсов): для начала мы отобрали 12 текстовых фрагментов норвежско-русских переводов из корпуса RuN-Euro. Далее мы попросили 8–10 носителей русского языка перевести те же самые фрагменты на русский. В итоге мы получаем первичные данные такого содержания:

1. Sofie Amundsen var på vei hjem fra skolen.
   (Jostein Gaarder, *Sofies verden*/*Мир Софии*)
1a. София Амуннсен возвращалась домой из школы.
1b. София Амундсен шла по дороге домой из школы.
1c. София Амундсен возвращалась из школы домой.
1d. София Амундсен возвращалась домой из школы.
1e. Софи Амуннсен шла со школы домой.
1f. София Амундсен шла из школы домой.
1g. София Амуннсен шла домой из школы.
1h. София Амундсен шла домой после школы.
1i. София Амундсен шла домой из школы.

Предложение 1a является подлинным профессиональным переводом из корпуса RuN-Euro. Непрофессиональных переводчиков попросили перевести оригинальные тексты на хороший и естественный по стилю русский язык, сохраняя содержание оригинала. Проводя систематические сравнения корпусных переводов с независимыми переводами носителей языка, можно получить более полное представление о том, какой *лингвистической* ценностью (для контрастивных лингвистиче-

ских исследований) обладают переводы, сделанные профессиональными русскими переводчиками художественной литературы. Насколько близко к тексту оригинала они переводят? В целом, считается, что русские переводчики склонны делать более свободные, литературные переводы по сравнению, скажем, с норвежскими переводчиками. Материал, собранный с помощью нашей интерактивной программы для перевода, может помочь нам прояснить подобные вопросы. Используя соответствующие статистические программы, можно измерить разрыв между приведенными выше переводами в соответствии с различными параметрами.

Приведем несколько примеров. Как и следовало ожидать, переводчики не всегда используют одни и те же лексические единицы: *возвращалась* vs. *шла*. Тем не менее употребление глагола несовершенного вида в данном контексте (1) не вызывает у переводчиков разногласий. Это приводит нас к следующему вопросу: насколько часто в среднем переводчики и носители русского языка будут употреблять формы одного и того же вида (в заданном контексте)? Из приведенных ниже двух предложений мы видим, что видовая вариативность в переводах действительно встречается:

2. Det første stykket hadde hun gått sammen med Jorunn. 3. De hadde snakket om roboter.
2a. Первый отрезок пути она шла вместе с Йорунн, 3a. и девочки говорили про роботов. (нсв+нсв)
2b. Первую часть пути она прошла вместе с Йорунн. 3b. Разговаривали они о роботах. (св+нсв)
2c. Первую часть пути она шла вместе с Йорунн. 3c. Они говорили о роботах. (нсв+нсв)
2d. Первую часть пути они прошли вместе с Йорун, 3d. разговаривая о роботах. (св+нсв)
2e. Первую часть пути она прошла вместе с Ёрун. 3e. Они говорили о роботах. (св+нсв)
2f. Часть пути она шла вместе с Юрунн, 3f. и они разговаривали о роботах. (нсв+нсв)
2g. Первую часть пути она прошла вместе с Йорунн. 3g. Они говорили о роботах. (св+нсв)
2h. Первую половину пути она прошла с Юрунном. 3h. Они говорили о роботах. (св+нсв)
2i. Первую часть пути она прошла вместе с Юрунн. 3i. Они разговаривали о роботах. (нсв+нсв)

В отношении выбора видовых форм в профессиональном переводе (a) и в переводах восьми носителей языка (b-i) можно построить следую-

щую матрицу данных на основе указанных трех предложений (н = несовершенный вид; с = совершенный вид):

a н н н
b н с н
c н н н
d н с н
e н с н
f н н н
g н с н
h н с н
i н н н

Впоследствии полный набор данных (сотни предложений) по употреблению вида можно использовать для вычисления расстояния Хэмминга между парами переводчиков (по методологии, использованной в работе Waldenfels 2012 о категории вида в славянских императивах):

> 1 – (число контекстов с одним и тем же видом / общее число контекстов)

Очевидно, что расстояние между *a* и *c* в упрощенном примере, приведенном выше, равно 0, поскольку (1 – (3/3) = 0), тогда как расстояние Хэмминга между *a* и *b, d* или *e* равняется 0,33, то есть (1 – (2/3) = 1/3).

Наша база данных параллельных переводов на один и тот же язык позволит нам отныне изучать подобные вопросы более систематически, с учетом различных параметров и тем самым получить лучшее представление о том, насколько можно полагаться на данные корпусных переводов в лингвистических исследованиях.

Литература

Hofland, Knut and Stig Johansson. 1998. “The Translation Corpus Aligner: A program for automatic alignment of parallel texts”. In *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, edited by Stig Johansson and Signe Oksefjell. Amsterdam: Rodopi, 87–100.

Krave, Maria F. 2011. *Converbs in Contrast: Russian converb constructions and their English and Norwegian counterparts.* PhD thesis. University of Oslo.

Nygaard, Lars, Joel Priestley, Anders Nøklestad and Janne Bondi Johannessen. 2008. “Glossa: A multilingual, multimodal, configurable user interface”. In *Language Resources and Evaluation Conference*. Marrakech.
http://www.hf.uio.no/tekstlab/LREC-glossa_2008.pdf

Von Waldenfels, Ruprecht. 2012. “Aspect in the imperative across Slavic – a corpus driven pilot study”. In *The Russian Verb* (*Oslo Studies in Language* 4, no. 1), edited by Atle Grønn and Anna Pazelskaya, 141–154.