А.Б. Кутузов
akutuzov72@gmail.com
в печати

Переводы мужские и женские: есть ли разница?

(на материале Корпуса несовершенных переводов)

В российском переводоведении до сих пор сравнительно редки количественные исследования процессов перевода. Эту досадную лакуну могут исправить разнообразные параллельные корпусы и их изучение объективными статистическими методами.

Один из таких корпусов — Russian Learner Parallel Corpus (RLPC), разрабатываемый группой тюменских лингвистов. Он и послужил материалом для настоящей статьи, которая посвящена исследованию гендерной асимметрии в переводных текстах и сравнению её с аналогичной асимметрией в текстах не-переводных.

Мы уже рассказывали о проекте корпуса несовершенных переводов год назад в [1]. Не повторяя той публикации, напомним основные особенности проекта и прогресс, достигнутый за прошедшее время.

Доступные в настоящее время параллельные англо-русские и русско-английские корпусы не содержат переводов, выполненных непрофессионалами. Между тем, необходимость в таком корпусе несомненна, поскольку ошибки перевода представляют интерес для переводоведов, психолингвистов, методистов и преподавателей. В западном переводоведении примером такого корпуса является Comparable Learner Translation Corpus [2], но в нём нет русских текстов, да и объём его невелик — чуть больше 400 переводов.

Поэтому мы поставили цель создать корпус несовершенных переводов, выровненных с оригиналами, который был бы достаточно велик для того, чтобы давать репрезентативные данные об ошибках непрофессиональных переводчиков. Кроме того, к нему должен быть открыт свободный доступ через Интернет. Конечный планируемый объём корпуса — 10 миллионов словоупотреблений в английских и русских текстах. Каждый текст существует в виде оригинала и перевода (или нескольких переводов).

Где можно отыскать несовершенные переводы? Конечно, ошибки время от времени совершает каждый переводчик, но их число нестабильно и они могут случаться редко. Поэтому мы пришли к выводу, что единственным адекватным источником материала для нас могут быть только переводы студентов. Они постоянно содержат ошибки, они уже существуют в цифровой форме и их сравнительно легко получить на соответствующих кафедрах российских университетов.

Анализ несовершенных переводов может показать, как личность переводчика или выбранные переводческие стратегии влияют на совершенные ошибки, изучить психолингвистические аспекты переводческих решений, выявить переводческие универсалии. Наглядное сравнение нескольких переводов одного оригинала полезно для создания инструментов оценки качества перевода. Поиск реальных переводческих ошибок в больших массивах переводов даёт основу для создания объективных систем классификации этих ошибок.

В настоящий момент корпус доведён до объёма в 1146 текстов (это число включает и оригиналы и переводы). Корпус доступен в форме «сырых» данных по адресу http://tc.utmn.ru/files/trc.zip. Численные данные приведены в таблице 1.

[Табл. 1]

	Оригиналы	**Переводы**	**Общее количество**

Английский	100	513	613
Русский	25	508	533
Общее количество	125	1021	1146

Таким образом, 51% текстов являются английскими и 49% - русскими. Всего эти тексты содержат 467513 словоупотреблений. Мы продолжаем добавлять материал в корпус, в качестве среднесрочной цели поставлен порог в 1 миллион словоупотреблений. После этого мы планируем проанализировать комментарии и предложения от наших пользователей и переработать структуру корпуса в соответствии с ними. После этого начнётся следующий этап проекта, целью которого является набор 10 миллионов словоупотреблений. Будет организован веб-интерфейс для добавления оригиналов и переводов силами интернет-сообщества. При этом мы собираемся сохранить выше указанное соотношение английских и русских переводов.

Сейчас собраны и обработаны тексты из Тюменского государственного университета (около 900 текстов), Московского государственного университета (около 100 текстов) и Удмуртского государственного университета (около 100 текстов). Тексты из Челябинского государственного университета и Нижегородского государственного лингвистического университета находятся в процессе обработки.

Технически корпус организован в виде простых текстовых файлов и заголовочных файлов по стандарту Translational English Corpus (http://www.llc.manchester.ac.uk/ctis/research/english-corpus). Заголовочные файлы содержат мета-информацию к соответствующим текстовым файлам. Поля в заголовочных файлах таковы:

1. Пол переводчика?
2. Курс?
3. Оценка за перевод?
4. Черновик или окончательная версия перевода?
5. Жанр текста?
6. Тип перевода (экзамен или текущая работа)?
7. Ситуация перевода (домашний или в классе)?
8. Год перевода?
9. Оригинал или перевод?
10. Университет?

Конечно, эти поля далеко не исчерпывают всех возможных «вопросов», которые исследователь может иметь к переводу. Но в реальности чрезвычайно редко мы знаем о конкретном студенческом переводе нечто больше, чем выше приведённый список. Зачастую даже и этой экстралингвистической информацией мы обладаем не полностью: например, знаем оценку за перевод, но не знаем пола переводчика, или наоборот. Тем не менее, во всех случаях мы стараемся максимально заполнить заголовочный файл, используя всю имеющуюся информацию.

Мы будем использовать веб-интерфейс, который позволит пользователям фильтровать результаты поиска по любому из вышеуказанных параметров. Например, пользователь может искать переводы определённой фразы, которые сделали студенты ТюмГУ мужского пола в 2009 году и сравнить их с переводами, сделанными студентами женского пола.

Для поиска переводов по оригиналам, собранные нами тексты необходимо будет «выровнять» (to align) по предложениям. Тем не менее, даже пока это ещё не сделано, наш корпус уже можно использовать как сравнительный, содержащий набор текстов на одном

языке и набор их переводов на другой. Ниже приводится пример такого использования.

Как известно, женская и мужская речь (как устная, так и письменная) отличается друг от друга по количественным показателям. Так, исследователи сходятся в том, что женский словарь беднее [3. С.5, 4. С.19]. Это значит, что женщины с более высокой вероятностью, чем мужчины, используют уже ранее использованные в тексте слова. Параметрически это выражается в более близком к единице соотношении количества словоупотреблений и словоформ в «мужских» текстах. В англоязычной литературе этот параметр обычно носит название TTR (type to token ratio).

Ещё одно отличие, связанное с гендером, состоит в том, что средняя длина предложения в устной речи у мужчин также выше, чем у женщин [5. С.144]. В письменной речи значимых различий в длине предложения не выявлено, но зато женщины употребляют больше местоимений и частиц. [6. С.9]

Однако, все вышеупомянутые наблюдения были сделаны на материале оригинальных первичных текстов. С точки зрения переводоведения, важен вопрос, сохраняются ли эти различия во вторичных текстах, а именно — в переводах. Можно предположить, вслед за Моной Бейкер, что переводческая универсалия нейтрализации сглаживает эту асимметрию. [7]. Большой корпус переводных текстов, размеченный по признаку пола переводчика, даёт возможность проверить это предположение.

Итак, мы выделили из набранного нами на сегодняшний день корпуса те тексты, для которых присутствуют как мужские, так и женские переводы. Это дало нам 27 текстов и 240 переводов, из которых 39 — мужские и 201 — женский. Количественное преобладание женских переводов вполне объяснимо, если вспомнить тотальное преобладание девушек на переводческих факультетах российских вузов. Тем не менее, мужских переводов достаточно много для того, чтобы применить к ним методы статистики.

Для каждого перевода мы подсчитали количество словоформ и словоупотреблений (и, следовательно, параметр TTR), а также количество предложений (и, следовательно, среднюю длину предложения в словах). В большинстве случаев мы имели дело с несколькими «однополыми» переводами одного текста, поэтому мы также вычисляли средний TTR и среднюю длину предложения для всех «однополых» переводов данного текста. При этом мы использовали не среднее арифметическое, а медиану, для того, чтобы снизить влияние случайных отклонений (например, студент перевёл не весь текст, а лишь половину, поэтому количество словоупотреблений у него значительно ниже, чем в других переводах того же текста).

Для вычисления параметров текстов использовалась утилита анализа корпусов Corsis (http://corsis.sourceforge.net). Затем данные заносились в электронные таблицы LibreOffice Calc, где и производились последующие подсчёты. Лемматизация не проводилась.

Пример таблицы для одного из текстов (14 переводов, выполненных женщинами и 2 перевода, выполненных мужчинами):

[Табл. 2]

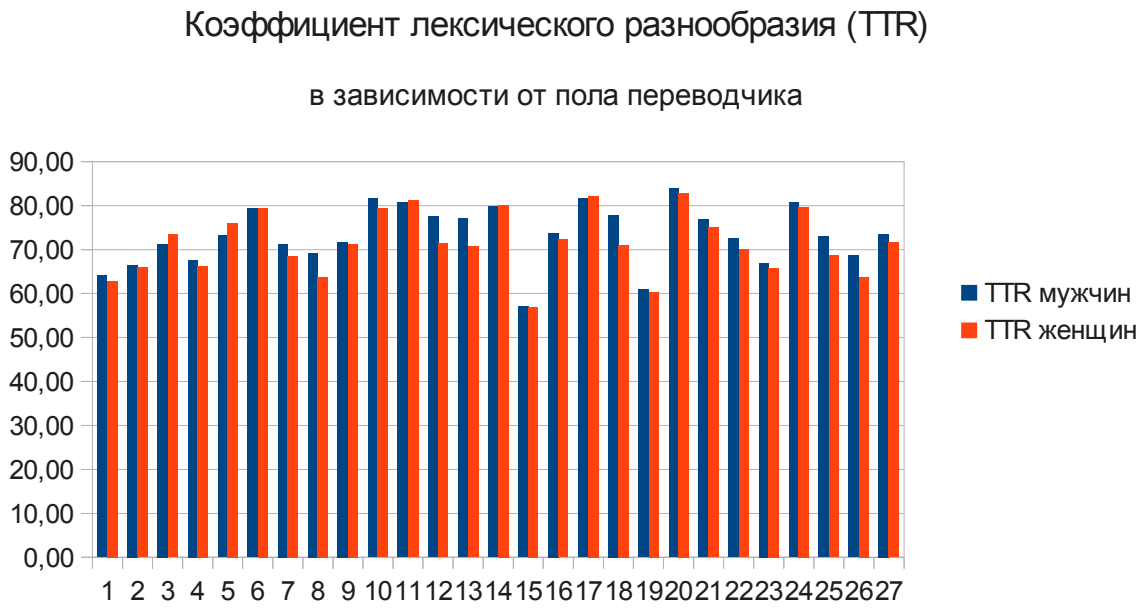
	Tokens	**Types**	**TTR, %**	**Sentences**	**Average sentence length, words**
Females					
1	325	211	64.92	14	23.21
2	443	266	60.05	22	20.14
3	320	207	64.69	13	24.62
4	308	194	62.99	14	22.00
5	325	209	64.31	30	10.83
6	326	214	65.64	14	23.29
7	554	347	62.64	20	27.70

8	629	379	60.25	42	14.98
9	498	296	59.44	31	16.06
10	429	269	62.70	42	10.21
11	458	292	63.76	25	18.32
12	330	205	62.12	26	12.69
13	522	310	59.39	19	27.47
14	327	209	63.91	13	25.15
Average	**379.50**	**240.00**	**62.85**	**21.00**	**21.07**
Males					
1	459	287	62.53	20	22.95
2	326	214	65.64	14	23.29
Average	**392.50**	**250.50**	**64.09**	**17.00**	**23.12**

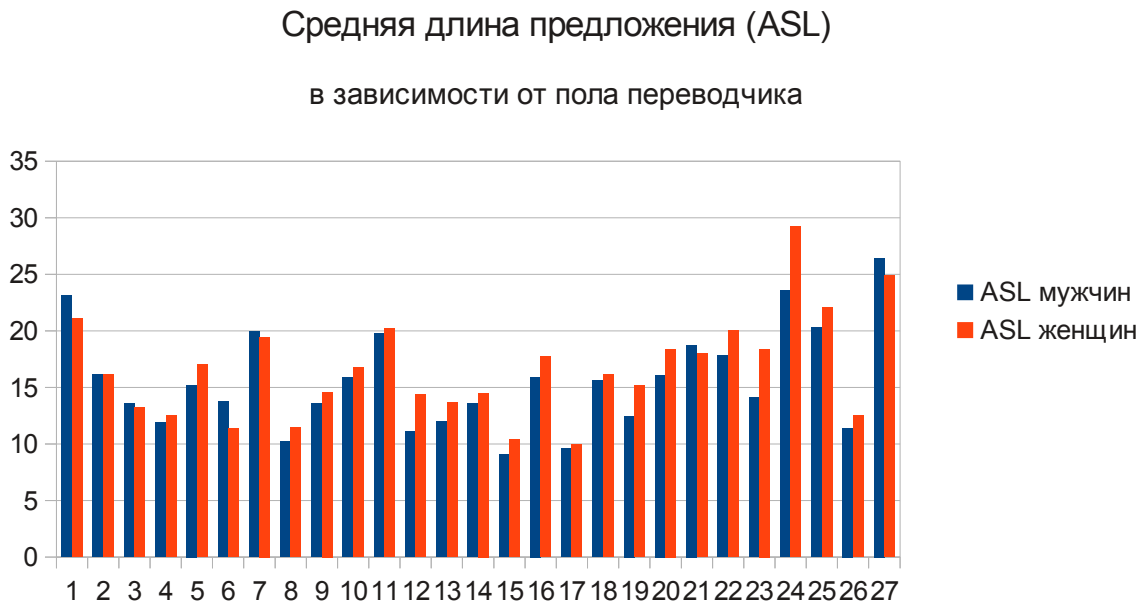
Можно видеть, что при переводах этого текста у женщин показатель лексического разнообразия TTR несколько ниже, чем у мужчин (62,85% против 64,09%). Средняя длина предложения также несколько ниже — 21,07 слова против 23,12.

Всего таких таблиц получилось 27, по числу текстов. Данные по их переводам были сведены в две таблицы — по параметру TTR и по средней длине предложения. Визуализировать получившиеся данные можно так:

[Рис.1]



[Рис. 2]



Среднее арифметическое TTR у мужчин составило 73,25%, у женщин — 71,49%. Среднее арифметическое ASL у мужчин — 15,59 слова, у женщин — 16,64 слова. Впрочем, сами по себе значения среднего арифметического в данном случае не являются убедительным доказательством, поскольку они могут быть статистически не значимы — как видим, разница в показателях совсем небольшая.

Поэтому мы провели проверку значимости расхождений между выборками по t-критерию (он же критерий Стьюдента). Эта проверка показала, что расхождения между выборками по TTR значимы с вероятностью ошибки не более 0,1 процента. Расхождения между выборками по средней длине предложения также значимы, но уже с вероятностью ошибки 0,5 процента. Практически, это означает, что можно с уверенностью утверждать: фактор пола переводчика влияет на исследуемые параметры текста перевода.

Что следует из этих результатов? Во-первых, как и ожидалось, в переводных текстах воспроизводится гендерная асимметрия по признаку коэффициента лексического разнообразия. Мужчины продолжают более активно использовать новые слова, а женщины по-прежнему делают это реже. То есть, влияние нейтрализации здесь отсутствует, переводные тексты по этому признаку не отличаются от оригинальных, а тезис о более богатой лексически речи мужчин, получает дополнительное подтверждение.

Более интересен результат по параметру «средняя длина предложения». Согласно нашим данным, в переводных текстах, выполненных женщинами, предложения в среднем длиннее, чем в тех, которые выполнили мужчины. Между тем, по результатам, изложенным в [6], значимых различий по средней длине предложения в оригинальных текстах, порождённых мужчинами и женщинами, нет. Это противоречие можно интерпретировать двояко:

1. Мы имеем дело с неким фундаментальным свойством перевода в направлении «английский-русский», из-за которого женщины-переводчики начинают порождать более длинные предложения, чем в своей обычной речи.
2. В [6] отмечается, что женщины порождают более длинные предложения, когда они пытаются имитировать речь мужчин. Можно предположить, что в ситуации перевода женщина подсознательно пытается подражать мужской речи, поскольку патриархальные стереотипы подсказывают ей, что так текст будет «лучше».

Второе объяснение, на наш взгляд, обладает большей объясняющей силой. Вероятно, чтобы

подтвердить его, необходимо сравнить мужские и женские переводы по другим синтаксическим параметрам, например, по коэффициенту синтаксической сложности предложения. Кроме того, для большей объективности всех изложенных выше результатов их желательно пересчитать после лемматизации текстов (то есть, приведения всех словоупотреблений к начальным формам). Это позволит считать, например, словоупотребления «видит» и «видела» двумя манифестациями одной леммы «видеть». В настоящем исследовании лемматизация не проводилась (по соображениям недостатка времени) и поэтому подобные словоупотребления расценивались анализатором как разные словоформы. Не исключено, что анализ лемматизированных текстов даст иные результаты.

Литература

1. Кутузов А.Б. Корпус несовершенных переводов: необходимость проекта // Сборник научных трудов "Проблемы теории, практики и дидактики перевода", вып.14, т.1. - Нижний Новгород, 2011
2. Kübler N. A Comparable Learner Translator Corpus: creation and use. Proceedings of the Comparable Corpora Workshop of the LREC Conference, May 31 2008, Marrakech, Maroc, pp 73-78
3. Goroshko O. Differentiation in Male and Female Speech Styles. Research Support Scheme Network Chronicle. – 1998. - N6. 2
4. Диасамидзе Л.Р. Способы конструирования гендерной идентичности в интернет-дискурсе (на материале англоязычных и русскоязычных текстов политических сетевых дневников (блогов) // автореферат диссертации на соискание учёной степени кандидата филологических наук. - Тюмень, 2010
5. Ерофеева Т.И. Речевой портрет города в экспериментальном исследовании. // Филологические заметки, т.2. - Пермь, 2007
6. Енгалычев В.Ф., Белянин В.П., Константинова Е.С., Ощепкова Е.С. Психолингвистические особенности «мужского» и «женского» языков // Труды регионального конкурса научных проектов в области гуманитарных наук. Выпуск 2. – Калуга: Эйдос. – 2001. – С. 177-187.
7. Baker M. Corpus-based Translation Studies: The challenges that lie ahead // Terminology, LSP and Translation Studies in Language Engineering. Amsterdam/ Philadelphia: John Benjamins, 1996, cc. 175-186