КОРПУС УСТНЫХ ПЕРЕВОДОВ КАК НОВЫЙ ТИП КОРПУСА ТЕКСТОВ

THE INTERPRETING CORPUS AS A NEW TYPE OF TEXT CORPUS

Михайлов М.Н. *(mihail.mihailov@uta.fi),* ***Исолахти Н.Б.*** *(nina.isolahti@uta.fi)*
Тамперский университет, Тампере, Финляндия

Я теперь пишу не роман, а роман в стихах – дьявольская разница.

(из письма А.С. Пушкина П.А. Вяземскому)

Обсуждается разработка корпуса устных переводов на примере судебного дискурса. Корпус сочетает в себе корпус устной речи и параллельный корпус. Разметка должна содержать коммуникативную, просодическую и экстралингвистическую информацию. Данные корпуса интересны для междисциплинарных исследований.

1. На пути к корпусу дискурса

За последние десятилетия корпусная лингвистика сделала гигантский скачок. Корпуса текстов из «роскоши« стали «средством передвижения»: на сегодняшний день редкое исследование по лингвистике обходится без привлечения корпуса текстов. Тем не менее нельзя не отметить, что развитие это идет неравномерно: в первую очередь растут одноязычные общеязыковые корпуса текстов (такие, как British National Corpus, Национальный корпус русского языка и т.п.) при явном отставании параллельных, диахронических корпусов и т.п.

Развитие корпусов устной речи также, по понятным причинам, идет более медленными темпами. Устная составляющая больших корпусов текстов несопоставимо мала по сравнению с письменной, например, устная речь составляет лишь 3,9% от общего объема Русского национального корпуса (http://www.ruscorpora.ru/corpora-stat.html, 08.04.2008). Главной проблемой при создании корпусов устной речи являются крайне медленные темпы ввода данных, поскольку какая-либо автоматизация на данный момент отсутствует. Не менее сложной оказывается и разметка корпуса. Из коллекций транскрибированной[1] устной речи лишь монологи теоретически можно организовать по тем же принципам, что и письменные тексты. Диалоги нельзя хранить в формате «текстовых» корпусов текстов. В противном случае неизбежны очень существенные информационные потери и заметные неудобства.

В корпусе текстов, о котором пойдет речь в нашей статье, представлен еще более неудобный для хранения и аннотирования материал: дискурс с устным переводом. В ходе такого дискурса имеет место обмен репликами на двух языках, часть которых (но не обязательно все!) переводится на другой язык, причем перевод может происходить в обе стороны и с наложением во времени (то есть в некоторых случаях последовательный перевод переходит в синхронный).

Корпус текстов судебных переводов становится в некоторой степени междисциплинарным продуктом. Судебный дискурс интересен для исследователей как особый тип речевого общения. Исследования проводятся в разных направлениях: изучается, например, влияние институционального контекста на деятельность участников дискурса, роли участников дискурса, их интерактивная деятельность, взаимодействие и взаимовлияние, речевое выражение иерархии и конфликта взаимоотношений (напр., Красовская 2006; Välikoski 2004). Устный перевод вносит в коммуникацию дополнительный компонент: роли участников коммуникативной ситуации осложняются межкультурными и межъязыковыми различиями.

Судебный перевод остается малоизученным элементом институционального дискурса. Работы последних лет посвящены изучению роли переводчика в судебном процессе, профессионализма, техники судебного перевода (González, Vásquez & Mikkelson 1991; Mikkelson 2000), однако появились и работы, рассматривающие судебный перевод как часть судебного дискурса (Hale 2004; Moeketsi 1999).

Аутентичный материал, показывающий все так, «как это было», с паузами, исправлениями, повторами и

[1] В данной статье транскрипция понимается в широком смысле как способ передачи устной речи с помощью письменных знаков.

оговорками, представляет для таких исследований огромную ценность. На кафедре перевода русского языка Института современных языков и переводоведения Тамперского университета создается корпус **CoInCoUT** (Court Interpreting Corpus at the University of Tampere). Материал представляет собой собрание аудиозаписей допросов подсудимых и свидетелей с участием переводчиков языковой пары финский-русский. Записи транскрибируются и аннотируются. На настоящий момент в корпусе имеются аудиозаписи 8 процессов, в которых работало 3 разных переводчика, общая продолжительность записей около 14 часов. Корпус постоянно пополняется.[2]

2. Архитектура корпуса устного перевода

Корпус текстов организован в формате базы данных Postgresql с доступом к ней через веб-интерфейс, написанный на языке PHP 5, и размещен на Linux-сервере Института современных языков и переводоведения mustikka.uta.fi (доступ на сервер ограничен). Тексты аудиозаписей хранятся в виде таблицы, каждая запись которой представляет собой текстоформу, просодический элемент или элемент разметки («тэг»). В связанной с ней таблице хранится информация о каждой реплике: говорящий, момент начала и конца говорения, язык, реплика-«хозяин» (например, если реплика является переводом другой реплики). Леммы текстоформ помещены в отдельные две таблицы, связанные с таблицей текстов — таблицу русских лемм и таблицу финских лемм.[3] Еще одна таблица служит для хранения информации о судебных заседаниях, с которых были получены записи[4].

Таким образом, путем организации корпуса в виде набора связанных таблиц, удается придать некую форму этому нагромождению реплик на разных языках, произносимых разными людьми. Вообще говоря, представляется, что данные такого типа можно на сегодняшний день хранить без существенных потерь только в виде SQL базы данных либо в формате XML.

Для выполнения различных запросов к корпусу текстов написано большое количество различных утилит, объединенных в веб-интерфейс. С помощью него пользователь может получать конкордансы, словники, наборы коллокаций и т.п. Поскольку корпус создан в первую очередь для исследования дискурса и процесса перевода, то и конкорданс в таком корпусе будет несколько отличаться от конкорданса в его традиционном понимании. В зависимости от того, выполнялся ли в реконструируемом фрагменте речи перевод, конкорданс может быть как обычным, так и параллельным. В некоторых случаях возможна и комбинация обоих типов. Для дискурсивно-переводческого корпуса принципиально важно также то, что в запрос, кроме собственно поискового слова, может включаться различная информация по просодике, коммуникативной функции реплик и т.п. Более того, возможны запросы на получение информации типа «Реплики, в которых председательствующий говорит тихо» или «Реплики, в которых переводчик перебивает подсудимого» и т.п.

Пользователь корпуса получает довольно богатые возможности по сбору самой разнообразной статистики, вплоть до темпа речи говорящих. Так, можно сравнивать темп речи и количество пауз разных переводчиков, а также сравнивать темп речи говорящего и переводчика. Программная оболочка постоянно развивается и пополняется новыми функциями и возможностями.

3. Общие принципы транскрибирования и аннотирования

Для транскрибирования аудиозаписей корпуса была выбрана транскрипция на базе стандартной орфографии на обоих языках (финском и русском) без передачи фонетических особенностей речи говорящих. Причина состоит в том, что фонетическая транскрипция заметно усложнила бы и без того трудоемкую работу, не только замедлив собственно процесс ввода данных, но и сделав невозможной автоматическую лемматизацию. Поэтому от фонетической транскрипции было решено отказаться, поскольку на данный момент использовать материалы для исследований в области фонетики и фонологии не предполагается.

С другой стороны, в транскрипции важно выделить те особенности звучащей речи и ситуации общения, которые потребуются при исследовании дискурса и перевода. Так, для дискурсивного анализа необходимо представление о социальной иерархии участников коммуникативной ситуации. Доминирует в зале суда председательствующий, тогда как между сторонами, т.е. обвинителем и защитником ярко выражено

[2] Авторы выражают признательность лицам, оказавшим помощь в сборе материала и разрешившим получить аудиозаписи из архивов судов, а также их использование в исследовательских целях.

[3] Для обработки текстов используются лемматизаторы RusMorph для русского языка (Гельбух А. Ф., Сидоров Г. О. 2005) и Fintwol для финского.

[4] Рассматривается возможность размещения на сервере аудиозаписей, однако на данный момент это технически неосуществимо в связи с трудо- и времяемкостью работы: необходимо не только «порезать» звуковой файл на реплики, но и убрать из него имена собственные и прочие данные, представляющие конфиденциальный характер.

соперничество, а подсудимый находится на низшей ступени этой иерархии. Речь свидетеля, в свою очередь, изначально должна характеризоваться нейтральностью. Переводчик выступает в роли независимого профессионала, передающего высказывания этих участников коммуникации на другом языке. Мы можем предположить, что хороший перевод отразит языковые личности и роли участников, например, авторитетный характер высказываний судьи. Поэтому указание говорящего оказывается чрезвычайно важным. Кроме того, текстовая передача дискурса без указания на автора реплики вообще вряд ли возможна.

Для дискурсивного анализа большое значение имеет определение структуры дискурса, выявление дискурсивных пар. Судебный допрос строится, как правило, в форме диалога, когда за вопросом следует ответ, а при судебном переводе, соответственно, ВОПРОС → ПЕРЕВОД → ОТВЕТ → ПЕРЕВОД. Однако, на практике такая идеальная конструкция часто нарушается. Причины нарушения дискурсивных пар могут быть различные, невольные или преднамеренные, как например, нежелание подсудимого отвечать на вопрос. Нарушения дискурсивной пары вопрос-ответ могут также происходить из-за неудачного перевода. В корпусе размечены следующие функции высказываний:

Q	вопрос
A	ответ
Qb	подготовка и обоснование вопроса
Foc	уточнение вопроса
Re	повторение вопроса
St	продолжение вопроса
C	приказ/команда
P	предложение
Com	комментирование предыдущего высказывания
Com/Tr	комментирование и/или исправление перевода

и т.д.

Анализ дискурса требует фиксирования многих интонационных и просодических особенностей устной речи, однако, разработка системы транскрипции подчинена задачам и целям исследования (см., напр., Tainio 1997). Цель нашего исследования требует отражения тех просодических сторон устной речи, которые могут быть каким-либо образом связаны с процессом коммуникации. Так, например, темп речи говорящего и хезитация могут быть маркерами уверенности или неуверенности говорящего в достоверности сведений, а интенсивность речи (громкость), фразовая интонация и акцентное выделение могут быть показателем норм коммуникативной ситуации, иерархии участников, разрешения или обязательности выполнения устного волеизъявления.

Нередко то, **как** сказано, может быть важнее того, **что** сказано. Поэтому судебный переводчик в идеале должен отражать все особенности речи говорящего (Driesen 1992; González, Vásquez & Mikkelson 1991: 16, 272; Hale 2004: 8–9; Moeketsi 1999: 100. С другой стороны, у речи переводчика тоже могут быть особенности. Так, например, хезитация или замедленный темп речи могут быть показателем неуверенности переводчика в выборе правильного варианта перевода. При разметке корпуса применяются следующие элементы:

<up>/<down>	движение тона вверх/вниз
<quiet>/<loud>	тихий /громкий
<quick>/<slow>	быстрый / медленный
(.)	пауза 0,2 секунды или короче
(2.9)	пауза длиннее чем 0,2 секунды и ее длина
(ee ~1.0)	пауза, заполненная гласным звуком, и ее длина
(mm ~1.0)	пауза, заполненная согласным звуком, и ее длина

Элементы разметки при выводе на экран заменяются на шрифтовое оформление (например, разрядка для замедленного темпа речи, мелкий шрифт для тихой речи и т.п.) или графические обозначения (например, стрелки для указания движения тона). Приведем небольшой пример (1), чтобы показать, как выглядит транскрибированный диалог при выводе на экран веб-браузера результатов поиска.

(1)

Обвинитель:	**Переводчик**
(11.3) oliko tuota teillä puhetta siellä kun tapasitte Pp:ssa niin (.) tätä enemmänkin kokaiinista?	(0.8) (ее ~0.5) у вас, когда вы встретились в Рр, (0.2) были какие-то (ее) более подробные разговоры о кокаине?
(1.1) {@@@} täällä esitutkinnassa (0.4) kertoneet että, (0.7) kaveri puhui kokaiinista, (.) ehkä sokeri oli kokaiinia.	(0.7) (ее) вы на предварительном следствии показали что, (.) ваш приятель говорил *о* кокаине, и вот этот вот (ее ~0.4) сахар вероятно (ее) был кокаином.
(4.1) ↑ no puhuko ↓ hän jostakin sokerista siinä sitten?	= но ↑ он говорил ↓ вам тогда о чем-то о сахаре?

Обвинитель:	**Переводчик**
(1.7) ↑ kun olette tässä sanonut ↓, että (1.1) että kaveri käski laittaa ↑ repussa olleen sokerin jääkaappiin ↓, tein kuten käski ↑ hän käski ↓, (0.6) kaverini puhui kokaiinista, ehkä sokeri oli kokaiinia.	(0.3) (ее ~0.4) вы (ее) сказали на предварительном следствии, что ваш (.) приятель сказал (0.3) положить (0.3) (ее ~0.3) (.) сахар (0.6) в (.) морозилку, (0.4) и (ее ~0.5) вероятно (0.3) сахар и был кокаином.
(1.0) oletteko kuitenkin (.) siinä vaiheessa ymmärtäneet suomenkielen sanat (.) sokeri (.) jääkaappi (.) kokaiini?	(0.6) вы тогда ↑ все-таки понимали ↓ слова (.) сахар (.) холодильник и (.) кокаин?

4. Вопрос о единице устного перевода

Поскольку корпус предназначен для изучения в первую очередь перевода, основным структурным элементом корпуса должны быть сопоставимые между собой отрезки исходного сообщения и его перевода. Такими **базовыми единицами** в CoInCoUT являются **ИРО** (исходный речевой отрезок) и **ПРО** (перевод речевого отрезка). ИРО – это переводимый сегмент дискурса, т.е. отрезок исходного высказывания/реплики, ограниченный с двух сторон переводом или репликами других участников дискурса. Данный сегмент следует отличать от применяемой в переводоведении «единицы перевода«, которая не имеет четкого формального выражения и может быть как отдельным словом, высказыванием, так и текстом в целом (см., напр., Витренко 2006).

Исходные речевые отрезки (ИРО) и их переводы объединяются в более крупные структуры – **реплики участников** (РУ). Минимальная РУ = ИРО + ПРО, т.е. один исходный речевой отрезок и его перевод. Однако так происходит только в идеале. При реальном судебном допросе используются вопросы различных типов (см. Hale 2004), которые подразумевают ответы разной длины и структуры. Так, например, вопрос типа *Вы были там в 12 часов?* предполагает односложный ответ *Да/Нет.* В свою очередь, на вопрос типа *Расскажите своими словами, как Вы провели этот день?* ожидается достаточно длинный ответ в форме монолога. В РУ такого типа может быть несколько десятков ИРО и ПРО.

Кроме того, в зале суда нередко кипят страсти, и участники коммуникативной ситуации могут перебивать друг друга или переводчика, уточнять или пояснять свое же предыдущее высказывание. В таком случае может происходить нарушение пар ИРО → ПРО. Для разметки наложения реплик участников коммуникации применяется следующая нотация:

[ ] наложение реплик
[/ наложение реплик происходит в середине слова
/] наложение реплик заканчивается в середине слова
= реплика начинается непосредственно после предшествующей без какой-либо паузы.

Далеко не всегда дела обстоят так благополучно, как в примере (2), где обвинитель говорит достаточно медленно и переводчику удается передать подсудимому содержание сообщения. Случается, что русскоязычный участник дискурса, владеющий финским языком, начинает исправлять переводчика, не дослушав перевода до конца. Так, в примере (2) обвинитель спрашивает у свидетеля о событиях, связанных со спором о воспитании ребенка (*lapsen huoltajuuskiista),* переводчик начинает переводить и использует русский термин *спор о воспитании ребенка*, соответствующий употребленному обвинителем финскому термину. Свидетель, не дослушав перевода, перебивает переводчика, говоря о том, что вопрос шел не о воспитании ребенка, а об алиментах, и рассказывает, что тогда произошло. Таким образом, перевод прерван, вопрос обвинителя переведен лишь частично, и на вопрос обвинителя *Стал ли Хх выяснять, где находятся деньги тогда, когда зашел спор* ***о воспитании ребенка****?* был получен ответ *Хх узнал о деньгах, когда встал вопрос* ***об алиментах****.* Ситуацию можно истолковать двояко: (а) свидетель, возможно, исправляет неверное представление обвинителя о положении дел, или (б) свидетель считает неверным вариант перевода, выбранный переводчиком, и исправляет переводчика.

(2)

Обвинитель: Вопрос	(1.5) ymmärsinkö (0.2) oikein että, (0.6) kun (0.6) ↑tuli tämän **lapsen huoltajuuskiista** ↓(3.0) niin (0.8) silloin Xx (0.6) ryhtyi (0.3) selvittämään (1.1) näiden (0.9) rahojen kohtaloa, (0.3) missä ne ovat, (0.4) oliko tilillä, (0.4) oli {@@@} ?	**Переводчик:**	(0.6) я правильно понял, что когда (.) возникли (0.2) **спор (0.2)** ***о*** **воспитании ребенка** , тогда [Хх стал] …
Свидетель: Ответ	[об **алиментах**]! ↑не о **воспитании** ↓об **алиментах** ! тогда (.) ↑встал вопрос ↓, ↑она сама ↓сказала, ↑что у нее на счету восемьдесят ↓, (0.3) ↑у ребенка на счету ↓восемьдесят тысяч.	**Переводчик:**	(0.7) *se* ei ollut (.) **lapsen huoltajuuskiista** , (.) se oli **lapsen elatuskiista** , (0.4) ja Yy itse (ee) on maininnut että on lapsen ↑tilillä on ↓(.) tämmöisiä varoja .

ИРО и ПРО нередко бывают достаточно длинные, состоят из множества просодически и интонационно обособленных единиц (это видно уже из приведенного выше примера (2)). Для упрощения чтения текста мы вынуждены делить ИРО и ПРО на **элементарные дискурсивные единицы (ЭДЕ)**. Таким образом, базовые единица ИРО и ПРО могут разбиваться на элементарные дискурсивные единицы ИРО = n ЭДЕ или ПРО = n ЭДЕ.

4. Выводы

Из вышесказанного ясно, что корпус устного перевода технически нельзя организовать по той же схеме, по которой создается корпус письменных текстов: в этом случае материалом будет неудобно пользоваться, и значительная часть информации окажется недоступной. При работе над корпусом устной речи, и в особенности – над корпусом устного перевода, становится очевидной многоуровневость и многослойность данных, которые необходимо или желательно представить в таком корпусе текстов. Корпус оказывается необычайно разнообразным и в плане разметки, причем лишь часть ее связана с языковой стороной коммуникации.

Мы назвали корпус устного перевода новым типом корпуса текстов потому, что такой корпус является как бы гибридом одноязычного и параллельного корпуса текстов и требует отражения в разметке информации, связанной с ролями участников общения и ситуациями, в которых это общение происходит. При этом хочется отметить, что часть информации, которую мы пытаемся представить в корпусе **CoInCoUT**, может оказаться весьма полезной и в корпусах письменных текстов. Например, аннотирование диалогической части художественных произведений позволило бы искать примеры употреблений каких-либо выражений в прямой речи. Что касается аннотирования драматических произведений, то здесь применяемая нами разметка оказывается еще более полезной.

Список литературы

1. Витренко А.Г. Что же все-таки такое «единица перевода»? // Вопросы филологии. 2006: 2 (23), с. 53 - 61.

2. Гельбух А. Ф., Сидоров Г. О. К вопросу об автоматическом морфологическом анализе флективных языков // Труды международной конференции «Диалог-2005». М., 2005. (http://www.dialog-21.ru/Archive/2005/Gelbukh%20Sidorov/ GelbukhA.htm, 4.4.2008).

3. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования. // НТИ. сер. 2. 2003, № 10.

4. Красовская О.В. 2006. Судебный диалог как конвенциональная коммуникативная форма. // Вопросы языкознания. 2006, № 5.

5. Driesen Ch. Status und Funktion des Gerichtsdolmetschers/-übersetzers in Deutschland. // Mitteilungsblatt: Österreicher Ubersetzer- und Dolmetscherverband «UNIVERSITAS«. 1992. S. 7 - 13.

6. Hale S. The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter. Amsterdam/Philadelphia: John Benjamins, 2004.

7. FINTWOL: http://www.csc.fi/english/research/software/fintwol (4.04.2008)

8. González R., Vásquez Victoria F., Mikkelson H. Fundamentals of Court Interpretation. Theory, Policy, and Practice. Durham, North Carolina: Carolina Academic Press, 1991.

9. Mikkelson H.. Introduction to Court Interpreting. Manchester, UK: St. Jerome, 2000.

10. Moeketsi R.H. Discourse in a Multilingual and Multicultural Courtroom: A Court Interpreter's Guide. Pretoria: Van Schaik Publishers, 1999.

11. Tainio L. (ред.) Keskustelunanalyysin perusteet. Tampere: Vastapaino, 1997

12. Välikoski T.-R. The Criminal Trial as a Speech Communication Situation. Tampere: Tampere University Press, 2004.