УДК 81’322 DOI: 10.14529/ling160404

ПРОЕКТ СОЗДАНИЯ КИТАЙСКО-РУССКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА ОФИЦИАЛЬНО-ДЕЛОВЫХ ТЕКСТОВ С ДИСКУРСИВНО-СТРУКТУРНОЙ РАЗМЕТКОЙ

М.Ю. Мухин, Ян И
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург

Статья посвящена проекту создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой. Данная разметка заключается в описании структуры каждого абзаца в виде сети дискурсивных единиц, соединенных дискурсивными отношениями. Основу первичного наполнения корпуса составляют доклады о работе правительства КНР на китайском языке и их официальные переводы на русский. Выравнивание китайских и русских текстов в корпусе, т. е. их синтаксическое соотнесение, проводится по структуре каждого абзаца. В статье представлены история разработки проблемы создания синтаксических корпусов, общие задачи проекта, его теоретические основания и прикладные перспективы, критерии отбора текстов для корпуса, принципы разметки и выравнивания текстов, а также программное обеспечение для разметки и хранения данных (общая схема данных и интерфейс). Создаваемый корпус может быть в дальнейшем использован для решения задач машинного перевода и других алгоритмов автоматической обработки текста, обучения иностранным языкам, сопоставительной лингвистики, теории перевода и т. д.

Ключевые слова: корпусная лингвистика, параллельный корпус, дискурсивно-структурная разметка, трибанк, китайско-русский корпус, дискурсивное выравнивание, автоматическая обработка текста, машинный перевод.

Введение

Дискурсивно-структурная разметка в лингвистическом корпусе включает информацию о структуре дискурса, в том числе информацию об идентификации элементарных дискурсивных единиц, дискурсивных отношений, организации текста и т. д. Согласно теории риторической структуры, структура дискурса, или «структура связей», отвечает за организацию собственно текста и превращает его из простой последовательности предложений в некое единое целое [3, с. 159; 10]. Перед дискурсивной разметкой, или дискурсивным парсингом, стоит, таким образом, задача идентификации между различными дискурсивными единицами в тексте. На данном этапе теория анализа структуры дискурса, обращенная к обработке естественного языка, главным образом существует в английской исследовательской традиции.

В начале XXI века был создан первый корпус с дискурсивной разметкой – *RST Discourse Treebank* [5], основанный на теории риторической структуры [11] – далее ТРС. Разработанная в 1980-е годы американскими лингвистами У. Манном и С. Томпсон ТРС предлагает описание структуры дискурса в виде сети дискурсивных единиц, соединенных семантическими отношениями, причем, как правило, дискурс имеет древовидную структуру [3, с. 159], что определило появление и дальнейшее использование интернационального термина «дискурсивный трибанк». Иными словами, в дискурсивном трибанке на основе ТРС структура целого текста представлена в виде дерева.

Другой известный проект, *Penn Discourse Treebank* (PDTB), был создан в 2004 г. [13]. В 2008 году был представлен PDTB 2.0, состоящий из 2 304 текстов, в том числе 40 600 маркеров [15]. В этом корпусе были размечены дискурсивные связки, или коннекторы (discourse connectives), и их аргументы (arguments), что обеспечило более детальное описание структуры дискурса. Наряду с термином «дискурсивная связка» употребляются различные синонимы: «дискурсивный маркер», «метка» «связка клауз» и т. д. Под всеми этими терминами подразумеваются слова и фразы, которые необходимы для организации дискурса (состоящего по меньшей мере из двух связанных клауз), а их семантика является частью дискурса [7, 8, 16].

По модели *Penn Discourse Treebank* были созданы корпуса разных языков: *Arabic Discourse Treebank* [4], *Prague Discourse Treebank* [14], *Chinese Discourse Treebank* [17]. Модифицированная Теория риторической структуры была использована для описания структуры русского дискурса [3], устного дискурса в проекте Корпуса устной русской монологической речи [2], а также в проекте Корпуса текстов на русском языке [1].

Для разметки китайского дискурса предложена схема «*Connective-driven Dependency Tree*» – далее CDT, на которую повлияла Теория риторической структуры (идея древовидной структуры и центрального положения дискурсивных единиц), PDTB (участие дискурсивных связок) и китайские традиционные синтаксические теории [9]. Схема описывает структуру каждого абзаца китайского

текста в виде сети клауз, соединенных дискурсивными связками, и применяется в процессе создания Китайского дискурсивного трибанка (Chinese Discourse Corpus with Connective-driven Dependency Tree Structure).

Появление дискурсивных трибанков повлияло и на развитие параллельных корпусов. Их создание является важным направлением, связанным с изучением современного переводоведения и совершенствованием машинного перевода. Сегодня параллельные корпуса используются для решения самых разных теоретических и прикладных задач. В соответствии с требованиями машинного перевода в параллельных корпусах должна проводиться более глубокая синтаксическая разметка.

Первый параллельный корпус с дискурсивно-структурной разметкой был создан в 2000 г. на основе TPC [12]. Разметка в этом корпусе производится отдельно для исходных и переводных текстов. Следует отметить, что выравнивание текстов в этом корпусе, как и в других параллельных корпусах, проводится исключительно по единицам синтаксического уровня (предложениям и абзацам). Для создания более эффективной дискурсивной структуры в *Chinese-English Discourse Structure Parallel Corpus (Китайско-английский параллельный корпус с дискурсивно-структурной разметкой)* Feng предлагает новые принципы выравнивания текстов [6]. Общая идея заключается в том, что «выравнивание структуры двух текстов проводится по дискурсивным единицам и отношениям, а также по их иерархическим структурам» и осуществляется одновременно с процессом разметки [6, с. 159].

Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой, о котором идет речь в этой статье, основан на опыте Китайского дискурсивного трибанка [9] и Китайско-английского параллельного трибанка [6]. Мы используем платформу для разметки и параллельных текстов, созданную Feng Wenhe в 2013 г. [6]. Однако для нового проекта необходимо уточнить принципы разметки и выравнивания китайско-русских текстов, а также выработать особые принципы, определяемые особенностями китайского и русского дискурса.

1. Выбор текстов для корпуса

Различные параллельные корпуса создаются на материале официально-деловых текстов, особенно государственного и межгосударственного уровня – например, корпус слушаний Европарламента (http://www.statmt.org/europarl/) и др. Требования к точности перевода, к максимально полной соотнесенности таких текстов формализуют задачу и облегчают выравнивание корпуса.

На экспериментальном этапе в корпусе размещены четыре «Доклада о работе правительства КНР (с 2012 г. по 2015 г.)» на китайском языке и их переводы на русский. В будущем мы планируем расширить корпус и включить в него еще шесть докладов, а также законы и официально-деловые тексты других жанров. На сегодняшний день объем корпуса составляет 931 абзац текста, 116 668 текстоформ, в том числе 46 190 текстоформ в русской части и 70 478 – в китайской.

При отборе источников учитываются следующие факторы:

1. Чтобы обеспечить качество перевода, исходные документы должны быть переведены известными специалистами или официальными учреждениями. Основной источник нашего материала – официальный сайт правительства КНР (http://cn.theorychina.org/), который обслуживает *Central Compilation & Translation Bureau*, что в значительной степени гарантирует качество переводных текстов.

2. Выбранные документы должны характеризоваться относительной устойчивостью в структурно-семантическом плане. В частности, в «Докладах», с которыми ежегодно выступает премьер-министр, содержится много повторяющихся элементов (от слов до текстовых структур), что имеет большое значение для анализа языков оригинала и перевода и дальнейшего осуществления автоматической разметки в параллельном корпусе.

3. В отличие от перевода художественных произведений, при переводе правительственных документов большое внимание уделяется сохранению исходного смысла и структуры текста. Поэтому в исходном и переводном текстах совпадает порядок следования предложений, а структурные отношения в большинстве случаев являются взаимно-однозначными. Таким образом, облегчается задача выравнивания текстов оригинала и перевода.

2. Принципы разметки в корпусе

Вначале приведем в качестве примера размеченные параллельные тексты (1).

(1)

Исходный Текст (а):

а1[在财政收支矛盾较大的情况下，我们竭诚尽力，] @||@ а2[始终把改善民生作为工作的出发点和落脚点，] @| а3[注重制度建设，] @||@ а4[兜住民生底线，] @||@ а5[推动社会事业发展。]

Переводной Текст (б):

б1[При наличии довольно крупных противоречий между финансовыми доходами и расходами мы со всей искренностью] ||@ б2[неизменно брали за исходную точку и конечную цель всей своей работы улучшение народной жизни,] @| б3[уделяя особое внимание институциональному строительству,] @||@ б4[не допуская выхода за нижний предел обеспечения народной жизни] @||@ б5[и стимулируя развитие социальных сфер.]

(«Доклад о работе правительства КНР», 2014 г.)

Квадратными скобками в этом примере выделены элементарные дискурсивные единицы (ЭДЕ); буквы и цифры между ними обозначают китайские

клаузы, соотносимые с ними русские синтаксические единицы и их порядок. Количество вертикальных черт (знак «|») перед клаузой указывает на уровень иерархии в структурном дереве, к которому она относится. Дискурсивные связки подчеркнуты, а знак «@» обозначает центральное положение ЭДЕ в отношении между клаузами.

По этому примеру видно, что разметка в Китайско-русском параллельном корпусе включает такие параметры, как элементарные дискурсивные единицы и типы отношений между ними, дискурсивные связки и их семантические характеристики, центральное положение ЭДЕ, а также другую информацию об иерархической структуре. Теперь рассмотрим параметры разметки более детально.

Деление на элементарные дискурсивные единицы

В РРС дискурсивная единица определяется рекурсивно: это либо любой отрывок текста, имеющий РРС-структуру, либо элементарная дискурсивная единица [3, с. 163]. В дальнейшем, как правило, мы будем употреблять термины «элементарная дискурсивная единица» (unit), далее – ЭДЕ, говоря о единицах самого низкого уровня, и «дискурсивная единица» (text span), говоря о единицах любого объема.

В нашем корпусе ЭДЕ представляют собой конечные узлы в структурном дереве. Согласно англо-американской традиции, ЭДЕ обычно равна клаузе [3, с. 163; 11, с. 245]. В русском традиционном синтаксисе термин «клауза» почти не используется; и в русской, и в китайской синтаксической науке единого определения клаузы не существует. На сегодняшний день единых критериев универсальных принципов выделения ЭДЕ для текстов, написанных на различных языках (в данном случае на китайском и русском), нет.

Перед нами стоит задача сегментации и выравнивания параллельных текстов. Если в исходном тексте членение на ЭДЕ совпадает с членением на клаузы, то в переводном тексте такое совпадение необязательно. Рассмотрим пример (2):

(2)

(А1)这将鼓舞我们砥砺前行，(А2)不断创造新的辉煌。

(Б1) А это не может не вдохновить нас на неуклонное движение вперед (Б2) к новым блестящим успехам.

В исходном тексте компоненты А1 и А2 являются клаузами, а в переводном соотносимый с А2 компонент Б2 (*к новым блестящим успехам*) клаузой не является. Если наша задача заключается в обработке параллельных текстов, то их членение должно быть взаимообусловлено. С учетом теоретической базы и практического опыта при разметке корпуса мы выделяем ЭДЕ в исходном тексте, исходя из трех параметров:

1) грамматическая структура (ЭДЕ обязательно состоит из одной глагольной группы и одной или нескольких именных групп);

2) семантическая структура (в ЭДЕ содержится как минимум одно суждение);

3) формально-пунктуационная структура (между ЭДЕ обычно ставится знак препинания (запятая, точка с запятой, точка и т. д.), но не любое предложение со знаком препинания членится на ЭДЕ – например, 过去一年，(*в прошлом году*).

Выделение дискурсивных связок (коннекторов)

Дискурсивные связки в структурном дереве представляют собой узловые точки, которые соединяют дискурсивные единицы. В отличие от традиционного понимания функций союзов и союзных слов, дискурсивные связки соединяют разные типы конструкций – не только клаузы и предложения, но и сверхфразовые единства (например, вводное слово *в частности*). Функцию дискурсивной связки могут также выполнять не только традиционные союзы и союзные слова, но и предлоги, наречия (в том числе неоднословные), вводные конструкции. Например, в примере (1), предлог *при* считается дискурсивной связкой, потому что функционально соотнесен с китайским аналогом «在...情况下» в исходном тексте и так же указывает на отношение условия между клаузами б1 и б2.

Итак, при разметке исходного текста дискурсивная связка считается языковой единицей, которая соединяет клаузы или сверхфразовые единства и указывает на дискурсивное отношение между ними.

Функция дискурсивной связки может быть выражена и имплицитно, поэтому дискурсивные связки делят на эксплицитные и имплицитные. Эксплицитные связки подразделяются далее на одиночные (например, *и, но, при, в частности, 并且, 但是, 尤其是* и т. д.) и двойные (*не только... но и..., хотя...но..., 不...不...而是, 既...又..., 虽然...但是...* и т. д.).

Логико-семантические отношения в дискурсивном корпусе

В отличие от синтаксических отношений, дискурсивные отношения обладают прежде всего логико-семантическим характером. Li в докторской диссертации предложила классификацию, включающую 4 группы отношений (в их числе 17 разновидностей), с учетом асимметричности отношений, выделяемых в РРС, и роли дискурсивных связок. Эта классификация основана на традициях китайских синтаксических теорий. Перечислим виды дискурсивных отношений, выделяемых Li.

1. Параллельные отношения (5): соединительные, последовательные, прогрессивные, альтернативные и сравнительные.

2. Противительные отношения (2): противопоставительные и уступительные.

3. Каузальные отношения (6): собственно каузальные, целевые, обстоятельственные, условные, гипотетические, а также отношения умозаключения.

4. Расширительные отношения (4): изъяснительные, заключительные, иллюстрационные и оценочные [9].

Эта классификация отношений была использована для выполнения разметки не только корпуса исключительно китайских текстов, но китайско-английского параллельного корпуса [6]. В существующую платформу для выполнения разметки, которая была создана Feng [6], интегрирован именно такой список отношений. Создавая подобный, но уже китайско-русский корпус, мы решили принять эту классификацию дискурсивных отношений за основу. Однако на практике различные виды отношений приходится адаптировать к материалу (официально-деловым текстам) и понимать расширительно (в особенности это касается русских переводных текстов).

Структурный анализ дискурса при разметке

Размеченная дискурсивная структура представляется в виде дерева зависимостей – иерархического графа, например, тексты (1) можно представить в следующей форме (рис. 1). В сущности, структурный анализ дискурса показывает степень близости соседних дискурсивных единиц в семантическом и грамматическом отношении. Такой подход к анализу параллельных текстов совпадает с общим пониманием структуры текста и процесса перевода, а также обеспечивает выравнивание исходного и переводного текстов по дискурсивной структуре.

Как показано на рис. 1, после членения текстов (1) на ЭДЕ, в данном случае на клаузы, приведенные ранее тексты можно представить иерархически. В узловых точках содержатся дискурсивные связки, которые репрезентируют дискурсивные отношения в текстах.

4. Принципы выравнивания текстов в корпусе

Поскольку оригинальный и переводной тексты проявляют структурную асимметричность, для построения корпуса необходимо обеспечить выравнивание текстов и их структуры. Мы берем за основу принципы выравнивания структуры дискурса для английского и китайского языков [6] и, модифицируя их, предлагаем конкретные принципы выравнивания для китайских и русских текстов, которые касаются параллельных дискурсивных структур, особенностей сегментации текста (т. е. выделения в нем элементарных дискурсивных единиц) и разметки логико-семантических отношений между ЭДЕ.

Дискурсивные структуры и элементарные дискурсивные единицы

Выравнивание структуры параллельных текстов начинается с членения параллельных текстов на ЭДЕ. По естественным причинам набор словоформ и устойчивых выражений в результате членения оригинальных и переводных тексты на ЭДЕ не совпадают. Например, в (1) структура дискурса оригинального текста (а) была представлена в следующей форме:

(а1), || (а2), | (а3), || (а4), || (а5).

Если не учитывать китайский оригинальный текст, деление русского переводного текста следовало бы представить в другом виде:

(б1) (б2), | (б3), || (б4), || (б5).

По этой причине наш анализ дискурсивных единиц начинается с исходного текста, что обеспечивает общую сегментацию параллельных текстов. Это означает, что сначала необходимо разбить оригинальный (китайский) текст на клаузы, а потом выделять в переводном (русском) тексте соотносимые фрагменты. Поскольку в нашем при-

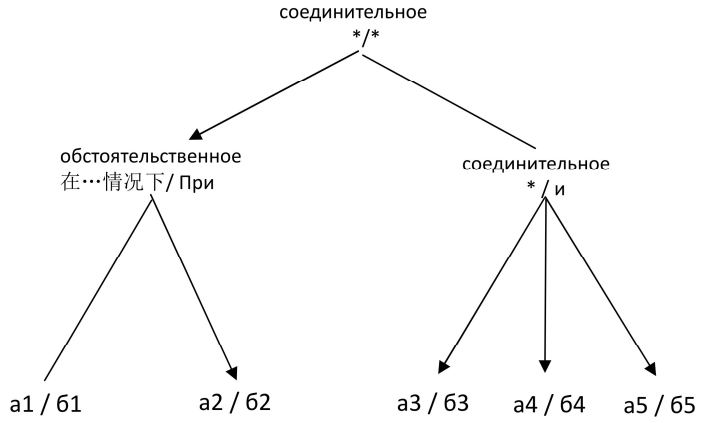

Рис. 1. Результат разметки и выравнивания текстов (1) в корпусе
(Примечание: буквы и цифры обозначают китайские (а) и русские клаузы (б) и их порядок; в каждой узловой точке указан тип отношений и дискурсивная связка; звездочка (*) указывает на наличие имплицитной связки)

мере в оригинальном тексте (а1) и (а2) являются двумя клаузам, иерархическая организация в корпусе будет представлена следующим образом:

(а1/б1), || (а2/б2), | (а3/б3), || (а4/б4), || (а5/б5).

Сущность выравнивания по ЭДЕ заключается в общей проблеме сегментации текста, а вопрос определения границ ЭДЕ сводится к тому, насколько крупными или мелкими могут быть единицы разметки.

Выравнивание по дискурсивным отношениям между ЭДЕ

Исходный текст является для переводчика объективной данностью, поэтому средства выражения дискурсивных отношений в нем также можно считать объективными. Поскольку перевод осуществляет специалист, обладающий индивидуальной языковой способностью, переводной текст в определенной степени отражает понимание переводчиком исходного текста. Иными словами, в переводном тексте выбор средств выражения логико-семантических отношений субъективен. Переводчик может изменить не только сам набор клауз и порядок их следования, но и отношения между ними. Соответственно, набор дискурсивных отношений мы определяем по переводному тексту. Такая методика соотносит разметку корпуса со стратегией перевода. Ср. примеры (3) и (4):

(3) *Исходный текст (а)*:
[在财政收支矛盾较大的情况下，我们竭诚尽力，]
[始终把改善民生作为工作的出发点和落 脚点，]...

Переводной Текст (б): [При наличии довольно крупных противоречий между финансовыми доходами и расходами мы со всей искренностью] [брали за исходную точку и конечную цель всей своей работы улучшение народной жизни,]...

«Доклад о работе правительства КНР», 2014 г.

(4) *Исходный Текст (в)*:
[在结构性矛盾突出的情况下，我们积极作为，
有扶有控，] [多办当前急需又利长远的事，] ...

Переводной Текст (г): [В связи с острыми структурными противоречиями] [мы действовали активно и принимали поощрительные либо ограничительные меры,] ...

«Доклад о работе правительства КНР», 2015 г.

Китайская дискурсивная связка «在…情况下» употребляется в одном и том же значении при соединении клауз в исходных текстах (а) и (в). В переводных текстах (б) и (г) она переведена разными способами – дискурсивными связками (в данном случае предлогами) «*При...*» и «*В связи с...*», которые актуализируют обстоятельственное и каузальное отношения соответственно. Это означает, что и переводчик воспринял дискурсивную связку «在…情况下» в тексте (3) как выразитель обстоятельственного отношения, а в тексте (4) – каузального.

5. Программное обеспечение корпуса

Разметка в данном корпусе проводится вручную с использованием специального программного обеспечения. Интерфейс платформы для структурной разметки параллельного корпуса, разработанный китайским ученым Feng [6], представлен на рис. 2. С помощью этого интерфейса можно пополнять корпус параллельными текстами и выполнять необходимую разметку и выравнивание конкретной пары языков (многоязыковая платформа поддерживает различные шрифты без ограничений в плане кодировки).

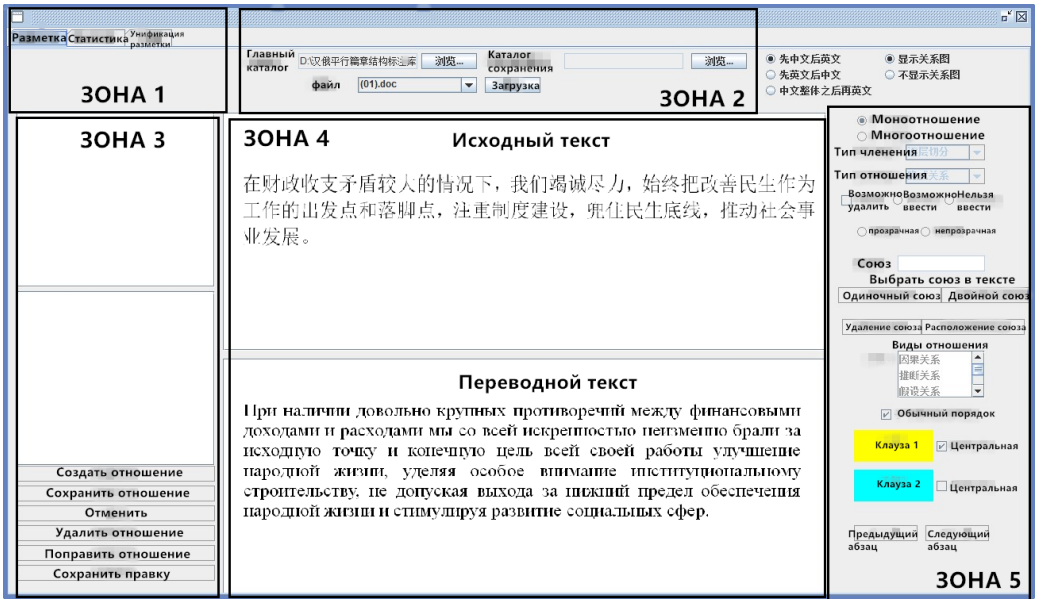

Рис. 2. Платформа китайско-русского параллельного корпуса с дискурсивно-структурной разметкой

В зоне 1 сосредоточены три функции платформы: разметка, статистика и унификация разметки (сравнение разметки, которая проведена разными пользователями). Зона 2 позволяет загрузить текст и выбрать каталог сохранения размеченных данных. Зона 3 представляет собой редактор отношений, которые можно создавать, сохранять, удалять и редактировать. В верхней части зоны 3 можно видеть размеченные отношения. В зоне 4 представлены исходный и переводной тексты. Зона 5 предназначена для выбора размечаемых свойств дискурсивной структуры: дискурсивная связка, тип отношения и другие параметры. Размеченные данные на основе русского и китайского текстов хранятся в XML-файлах. На рис. 3 представлены фрагменты таких файлов.

Формат хранения данных начинается символом <P ID="1"> и заканчивается символом </P>. Разметка отношения начинается символом <R и заканчивается символом />, кроме того, она включает следующие основные типы данных:

- *StructureType* – тип членения (последовательное и параллельное);
- *ConnectiveType* – тип дискурсивной связки (эксплицитная или имплицитная);
- *Connective* – дискурсивная связка;
- *RelationType* – вид отношения (вышеуказанные 17 видов отношений – см. ранее);
- *ConnectivePosition* – место дискурсивной связки (если во фразе можно усмотреть имлицитную связку, то надо также указать ее место).

Такая платформа дает возможность одновременно выполнять разметку и выравнивание структур дискурсов для китайского и русского параллельных дискурсов. Она обеспечивает ввод двух параллельных текстов, деление на ЭДЕ, разметку и выравнивание структуры дискурса, разметку дискурсивных связок и дискурсивных отношений. Накопленная информация, которая хранится в формате XML, обеспечивает дальнейший поиск в корпусе.

Заключение

В данной работе мы представили новый современный проект – китайско-русский параллельный корпус официально-деловых текстов с дис-

```
<?xml version="1.0" encoding="GB2312"?>
- <DOC>
  - <P ID="1">
      <R ID="1" UseTime="29" ParentId="-1" ChildList="2|4" Center="1" SentencePosition="1...198|200...357" Sentence="При
        наличии довольно крупных противоречий между финансовыми доходами и расходами мы со всей
        искренностью неизменно брали за исходную точку и конечную цель всей своей работы улучшение
        народной жизни, |уделяя особое внимание институциональному строительству, не допуская выхода за
        нижний предел обеспечения народной жизни и стимулируя развитие социальных сфер. "
        LanguageSense="true" RoleLocation="abnormal" ConnectiveAttribute="不可添加" ConnectivePosition="" RelationType="并
        列关系" Connective="" RelationNumber="单个关系" Layer="1" ConnectiveType="隐式关系" StructureType="并列切分"/>
      <R ID="2" UseTime="76" ParentId="1" ChildList="3" Center="2" SentencePosition="1...104|106...198" Sentence="При
        наличии довольно крупных противоречий между финансовыми доходами и расходами мы со всей
        искренностью |неизменно брали за исходную точку и конечную цель всей своей работы улучшение
        народной жизни, " LanguageSense="true" RoleLocation="abnormal" ConnectiveAttribute="不可添加"
        ConnectivePosition="" RelationType="并列关系" Connective="" RelationNumber="单个关系" Layer="2" ConnectiveType="隐式
        关系" StructureType="逐层切分"/>
      <R ID="3" UseTime="78" ParentId="2" ChildList="" Center="2" SentencePosition="1...80|82...104" Sentence="При
        наличии довольно крупных противоречий между финансовыми доходами и расходами| мы со всей
        искренностью " LanguageSense="true" RoleLocation="normal" ConnectiveAttribute="不可删除" ConnectivePosition="1...
        4" RelationType="背景关系" Connective="При " RelationNumber="单个关系" Layer="3" ConnectiveType="显式关系"
        StructureType="逐层切分"/>
      <R ID="4" UseTime="38" ParentId="1" ChildList="" Center="3" SentencePosition="200...255|257...318|320...357"
        Sentence="уделяя особое внимание институциональному строительству, |не допуская выхода за нижний
        предел обеспечения народной жизни |и стимулируя развитие социальных сфер. " LanguageSense="true"
        RoleLocation="normal" ConnectiveAttribute="不可删除" ConnectivePosition="320...320" RelationType="并列关系"
        Connective="и" RelationNumber="单个关系" Layer="2" ConnectiveType="显式关系" StructureType="并列切分"/>
    </P>
</DOC>

<?xml version="1.0" encoding="GB2312"?>
- <DOC>
  - <P ID="1">
      <R ID="1" UseTime="11" ParentId="-1" ChildList="2|4" Center="1" SentencePosition="1...41|42...64" Sentence="在财政收支
        矛盾较大的情况下，我们竭诚尽力，始终把改善民生作为工作的出发点和落脚点，|注重制度建设，兜住民生底线，推动社会事业发展。 "
        LanguageSense="true" RoleLocation="abnormal" ConnectiveAttribute="不可添加" ConnectivePosition="" RelationType="并
        列关系" Connective="" RelationNumber="单个关系" Layer="1" ConnectiveType="隐式关系" StructureType="逐层切分"/>
      <R ID="2" UseTime="6" ParentId="1" ChildList="3" Center="2" SentencePosition="1...21|22...41" Sentence="在财政收支矛盾
        较大的情况下，我们竭诚尽力，|始终把改善民生作为工作的出发点和落脚点，" LanguageSense="true" RoleLocation="abnormal"
        ConnectiveAttribute="不可添加" ConnectivePosition="" RelationType="并列关系" Connective="" RelationNumber="单个关系"
        Layer="2" ConnectiveType="隐式关系" StructureType="逐层切分"/>
      <R ID="3" UseTime="22" ParentId="2" ChildList="" Center="2" SentencePosition="1...14|15...21" Sentence="在财政收支矛盾
        较大的情况下，|我们竭诚尽力，" LanguageSense="true" RoleLocation="normal" ConnectiveAttribute="不可删除"
        ConnectivePosition="1...1&&11...13" RelationType="背景关系" Connective="在...情况下" RelationNumber="单个关系"
        Layer="3" ConnectiveType="显式关系" StructureType="逐层切分"/>
      <R ID="4" UseTime="24" ParentId="1" ChildList="" Center="3" SentencePosition="42...48|49...55|56...64" Sentence="注重
        制度建设，|兜住民生底线，|推动社会事业发展。 " LanguageSense="true" RoleLocation="normal" ConnectiveAttribute="可添加"
        ConnectivePosition="56-" RelationType="并列关系" Connective="并" RelationNumber="单个关系" Layer="2"
        ConnectiveType="隐式关系" StructureType="并列切分"/>
    </P>
</DOC>
```

Рис. 3. Формат хранения данных в корпусе

курсивно-структурной разметкой. На сегодняшний день определены критерии отбора текстов для корпуса, принципы их разметки и выравнивания, а также особенности программного обеспечения. На этой базе создание такого корпуса представляется вполне осуществимым.

Мы предполагаем, что данный корпус может быть использован для машинного перевода, обучения иностранному языку (в данном случае русскому и китайскому), сопоставительного анализа структуры и перевода двух языков и решения других актуальных задач.

Способы разработки параллельного корпуса с дискурсивно-структурной разметкой еще нельзя назвать совершенными. Подлежит уточнению классификация типов логико-семантических отношений с точки зрения китайской и русской лингвистической традиции. Пока не однозначно формализована сегментация текстов и выделение ЭДЕ, а текстовая вариативность заставляет уточнять и принципы выравнивания исходного и переводного текстов. Кроме того, сама платформа должна быть более свободной, т. е. давать возможность аннотаторам добавлять в нее дополнительные виды дискурсивных отношений.

В ближайшем будущем мы постараемся решить указанные проблемы и пополнить корпус новыми текстами с переводом не только в направлении «китайский → русский», но и «русский → китайский».

Исследование выполнено при поддержке Программы повышения конкурентоспособности Уральского федерального университета (номер соглашения 02.A03.21.0006) и «China Scholarship Council».

Литература

1. Ананьева, М.И. Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур / М.И. Ананьева, М.В. Кобозева // Тр. междунар. конф. «Диалог», 2016. – http://www.dialog-21.ru/media/3460/ananyeva. pdf (дата обращения: 29.07.2016).

2. Кибрик, А.А. Рассказы о сновидениях: Корпусное исследование устного русского дискурса / А.А. Кибрик и В.И. Подлесская. – М.: Litres, 2014. – 736 с.

3. Литвиненко, А.О. Описание структуры дискурса в рамках Теории Риторической Структуры применение на русском материале / А.О. Литвиненко // Труды Международного семинара Диалог, 2001. – С. 159–168.

4. AlSaif, A. The leeds arabic discourse treebank: Annotating discourse connectives for arabic / A. AlSaif and K. Markert // In Language Resources and Evaluation Conference, 2010. – http://www.comp.leeds.ac.uk/markert/Papers/LREC2010-LADTB.pdf (дата обращения: 16.08.2016).

5. Carlson, L. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory / L. Carlson, D. Marcu, M.E. Okurowski // In Current Directions in Discourse and Dialogue. – Kluwer Academic Publishers, 2003. – http://www.aclweb.org/anthology/W01-1605 (дата обращения: 28.07.2016).

6. Feng, W. Alignment and Annotation of Chinese-English Discourse Structure Parallel Corpus / W. Feng // Journal of Chinese Information Processing. – 2013. – 27(6). – P. 158–164. – http://jcip.cipsc.org.cn/CN/abstract/abstract1795.shtml (дата обращения: 26.07.2016).

7. Forbes, K. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG / K. Forbes-Riley, B. Webber, A. Joshi // Journal of Semantics. – 2006. – 23. – P. 55–106.

8. Forbes, K. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar / K. Forbes, E. Miltsakaki, R. Prasad et al. // Journal of Logic: Language and Information. – 2003. – 12(3). – P. 261–279.

9. Li, Y. Building a Chinese Discourse Corpus with Connective-driven Dependency Tree Structure / Y. Li, W. Feng, J. Sun et al. // In Proceedings of the 2014 conference on Emporical Methods in Natural Language Processing. – Doha: Qatar, 2014. – P. 2105–2114. – http://emnlp2014.org/papers/pdf/EMNLP2014224.pdf (дата обращения: 26.07.2016).

10. Mann, W. Rhetorical structure theory and text analysis / W. Mann, C. Matthiessen, S.A. Thompson // Amsterdam: Discourse Description, 1992. – P. 39–78.

11. Mann, W.C. Rhetorical Structure Theory: Toward a functional theory of text organization / W.C. Mann, S.A. Thompson // Text. – 1987. – 8(3). – P. 243–281. – http://www.sfu.ca/rst/pdfs/Mann_Thompson_1987.pdf (дата обращения: 02.08.2016).

12. Marcu, D. The Automatic Translation of Discourse Structures / D. Marcu, L. Carlson, M. Watanabe // In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, 2000. – P. 9–17. – http://www.aclweb.org/anthology/A00-2002 (дата обращения: 02.08.2016).

13. Miltsakaki, E. The Penn Discourse Treebank / E. Miltsakaki, R. Prasad, A. Joshi, B. Webber // In Proceedings of the 4th International Conference on Language Resources and Evaluation. – Lisbon: Portugal, 2004. – http://www.cis.upenn.edu/~elenimi/lrec04-lisbon-miltsakaki.pdf (дата обращения: 02.08.2016).

14. Poláková, L. Introducing the Prague Discourse Treebank 1.0 / L. Poláková, J. Mírovský, A. Nedoluzhko et al. // In Proceedings of the 6th International Joint Conference on Natural Language Processing. – Japan, 2013. – P. 91–99.

15. Prasad, R. The Penn Discourse Treebank 2.0 / R. Prasad, N. Dinesh, A. Lee et al. // In Proceedings of the 6th International Conference on Language Resources and Evaluation. – Marrackech: Morocco,

2008. – https://www.seas.upenn.edu/~pdtb/papers/pdtb-lrec08. pdf (дата обращения: 02.08.2016).

16. Webber, B. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse / B. Webber, A. Joshi // Montreal: ACL/COLING Workshop on Discourse Relations and Discourse Markers, 1998. – P. 8–92. – http://arxiv.org/pdf/cmp-lg/ 9806017v1.pdf (дата обращения: 02.08.2016).

17. Zhou, Y. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations / Y. Zhou and N. Xue // In Language Resources and Evaluation. – 2015. – 49(2). – P. 397–431.

Мухин Михаил Юрьевич, доктор филологических наук, директор департамента лингвистики, профессор кафедры современного русского языка и прикладной лингвистики, Уральский федеральный университет им. Б.Н. Ельцина (Екатеринбург), mu-hi@ya.ru

Ян И, аспирант кафедры современного русского языка и прикладной лингвистики, младший научный сотрудник Проблемной лаборатории компьютерной лексикографии, Уральский федеральный университет им. Б.Н. Ельцина (Екатеринбург), xwyang@mail.ru

Поступила в редакцию 17 октября 2016 г.

DOI: 10.14529/ling160404

BUILDING A CHINESE-RUSSIAN PARALLEL DISCOURSE STRUCTURE CORPUS OF OFFICIAL TEXTS

M.Yu. Mukhin, *mu-hi@ya.ru*
Y. Yang, *xwyang@mail.ru*
Ural Federal University named after B.N. Yeltsin, Yekaterinburg, Russian Federation

This paper is devoted to building a Chinese-Russian Parallel Discourse Structure Corpus of Official Texts (CRPDT) that aims at producing a discourse treebank, in which Chinese and Russian parallel texts are manually annotated and aligned at the level of discourse structure. In this corpus, discourse units and their discourse relations are annotated for each paragraph in the parallel texts. Experimental research is based on the material of 4 Chinese source texts "Reports on the work of the Government" and their Russian translations. The paper presents the history and development of building discourse treebanks, the principles of annotation for building parallel discourse treebanks. This paper shows how to work on the discourse segmentation for Chinese-Russian parallel texts. Annotation and alignment tools take from Chinese-English Parallel Discourse Treebank. We postulate that the corpus might be useful for machine translation, language learning, translation studies, discourse analysis of Chinese and Russian texts and future Natural Language Processing.

Keywords: corpus linguistics, parallel corpus, discourse structure annotation, treebank, Chinese-Russian corpus, discourse-level alignment, natural language processing, machine translation.

References

1. Ananyeva M.I. Razrabotka korpusa tekstov na russkom yazyke s razmetkoy na osnove teorii ritoricheskih struktur [Study on the Development of Russian Textual Corpus Based on the Theory of Rhetorical Structures. M.I. Ananyeva, M.V. Kobozeva]. *International Conference «Dialogue»,* 2016. URL: http: //www.dialog-21.ru/media/3460/ananyeva.pdf (accessed: 29.07.2016).

2. Kibrik A.A. *Rasskazy o snovideniyah: Korpusnoe issledovanie ustnogo russkogo diskursa* [Stories about Dreams: Corpus Study on the Russian Spoken Language]. Moscow, Litres, 2014, 736 p.

3. Litvinenko A.O. Opisanie strukturyi diskursa v ramkah Teorii Ritoricheskoy Strukturyi primenenie na russkom materiale [Description of Structure of Discourse in the Theory of Rhetorical Structures Use on Russian Material]. *Proceedings of the International Workshop Dialogue,* 2001, pp. 159–168.

4. AlSaif A. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. Language Resources and Evaluation Conference, 2010. URL: http://www.comp.leeds.ac.uk/markert/Papers/LREC2010-LADTB.pdf (accessed: 16.08.2016).

5. Carlson L. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory Current Directions in Discourse and Dialogue*, Kluwer Academic Publishers, 2003. URL: http://www.aclweb.org/anthology/W01-1605 (accessed: 28.07.2016).

6. Feng W. Alignment and Annotation of Chinese-English Discourse Structure Parallel Corpus. *Journal of Chinese Information Processing,* 27(6), 2013, pp. 158–164. URL: http://jcip.cipsc.org.cn/CN/abstract/abstract1795.shtml (accessed: 26.07.2016).

7. Forbes K. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics,* 23, pp. 55–106.

8. Forbes K. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic: Language and Information, 12(3), 2003*, pp. 261–279.

9. Li Y. Building a Chinese Discourse Corpus with Connective-driven Dependency Tree Structure. *Proceedings of the 2014 conference on Emporical Methods in Natural Language Processing, Doha: Qatar, 2014*, pp. 2105-2114. URL: http://emnlp2014.org/papers/pdf/EMNLP2014224.pdf (accessed: 26.07.2016).

10. Mann W. *Rhetorical Structure Theory and Text Analysis*. Amsterdam, Discourse Description, 1992, pp. 39–78.

11. Mann W.C. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text, 8(3),* 1987, pp. 243–281. URL: http://www.sfu.ca/rst/pdfs/Mann_Thompson_1987.pdf (accessed: 02.08.2016).

12. Marcu D. The Automatic Translation of Discourse Structures. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 2000, pp. 9–17. http://www.aclweb.org/anthology/A00-2002 (accessed: 02.08.2016).

13. Miltsakaki E. The Penn Discourse Treebank. *Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon: Portugal,* 2004. URL: http://www.cis.upenn.edu/~elenimi/lrec04-lisbon-miltsakaki.pdf (accessed: 02.08.2016).

14. Poláková L. Introducing the Prague Discourse Treebank 1.0. *Proceedings of the 6th International Joint Conference on Natural Language Processin*g. Japan, 2013, pp. 91–99.

15. Prasad R. The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrackech: Morocco,* 2008. URL: https://www.seas.upenn.edu/~pdtb/papers/pdtb-lrec08.pdf (accessed: 02.08.2016).

16. Webber B. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. Montreal, ACL/COLING *Workshop on Discourse Relations and Discourse Markers,* 1998, pp. 8–92. URL: http://arxiv.org/pdf/cmp-lg/9806017v1.pdf (accessed: 02.08.2016).

17. Zhou Y. The Chinese Discourse TreeBank: a Chinese Corpus Annotated with Discourse Relations. *Language Resources and Evaluation.* 49(2), 2015, pp. 397–431.

Mikhail Yu. Mukhin, Doctor of Philology, Director of the Department of Linguistics, Professor of Chair of Modern Russian Language and Applied Linguistics, Ural Federal University named after B.N. Yeltsin (Yekaterinburg), mu-hi@ya.ru

Yang Yi, PhD student, Chair of the Modern Russian Language and Applied Linguistics, Junior Research Fellow at the Laboratory for Computational Lexicography, Ural Federal University named after B.N. Yeltsin (Yekaterinburg), xwyang@mail.ru

Received 17 October 2016

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Мухин, М.Ю. Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой / М.Ю. Мухин, Ян И // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – Т. 13, № 4. – С. 23–31. DOI: 10.14529/ling160404

FOR CITATION

Mukhin M.Yu., Yang Y. Building a Chinese-Russian Parallel Discourse Structure Corpus of Official Texts. *Bulletin of the South Ural State University. Ser. Linguistics.* 2016, vol. 13, no. 4, pp. 23–31. (in Russ.). DOI: 10.14529/ling160404