Федеральное государственное автономное образовательное учреждение
высшего образования «Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»
Уральский гуманитарный институт
Кафедра фундаментальной и прикладной лингвистики и текстоведения

На правах рукописи

ЯН И

КИТАЙСКО-РУССКИЙ ПАРАЛЛЕЛЬНЫЙ ДИСКУРСИВНЫЙ КОРПУС ОФИЦИАЛЬНО-ДЕЛОВЫХ ТЕКСТОВ: ТЕОРЕТИЧЕСКИЕ ОСНОВЫ И ОПЫТ СОЗДАНИЯ

5.9.8. Теоретическая, прикладная и сравнительно-сопоставительная лингвистика

Диссертация
на соискание ученой степени кандидата
филологических наук

Научный руководитель:
доктор филологических наук, доцент
Мухин Михаил Юрьевич

Екатеринбург – 2026 г.

Оглавление

ВВЕДЕНИЕ

Диссертационное исследование посвящено созданию китайско-русского параллельного дискурсивного корпуса официально-деловых текстов и сопоставительному исследованию на его основе. В работе особое внимание уделяется теории и практике дискурсивной разметки и ее результатам, системному анализу лингвистических аспектов дискурса, обсуждению ключевых понятий, необходимых для моделирования языка на дискурсивном уровне (в рамках теории зависимостей), а также сопоставлению синтаксической и дискурсивной структур китайских и русских официально-деловых текстов.

Корпусная лингвистика возникла в 1960-е годы и с тех пор произвела в различных областях лингвистических исследований так называемую «корпусную революцию», создав принципиально новую методологическую основу для современной науки. Ориентированная на реальное словоупотребление, корпусная лингвистика подчеркивает важность речевых фактов и использует методы количественного анализа. Для подобных исследований принципиально важны объем и репрезентативность корпуса, что определяет достоверность получаемых результатов. Как отмечает А. Н. Баранов, «чем больше материал, тем выше достоверность выводов, тем шире сфера действия наблюдаемых закономерностей» [Баранов, 2021, с. 121].

Одновременно исследователи осознают значительный потенциал корпусов в компьютерном моделировании языка. Корпусная разметка позволяет извлекать и интерпретировать лингвистическую информацию. По замечанию А. А. Барковича, благодаря метаязыковым данным (то есть разметке, аннотации) стали осязаемыми новые возможности изучения универсально-статистических и специально-предметных «выборок» языковых фактов [Баркович, 2016].

В то же время проблема создания параллельного дискурсивного корпуса обусловлена недостаточной разработкой методологических подходов к разметке дискурса в межъязыковом контексте. Несмотря на активное развитие теорий дискурса и корпусной лингвистики, большинство существующих параллельных

корпусов размечены преимущественно на лексико-грамматическом уровне, а дискурсивные структуры не получают должного внимания. Эти ограничения препятствуют проведению глубокого межъязыкового анализа логико-семантических отношений, дискурсивных связей и структурных особенностей текстов, что существенно сдерживает развитие прикладных исследований, связанных с переводом, автоматической обработкой текстов, а также сравнительным анализом культурных и языковых особенностей.

Несмотря на существование концепции унифицированных схем аннотирования и методов автоматической разметки, вопросы совместимости дискурсивных схем для разных языков, таких как китайский и русский, остаются актуальными. Одной из ключевых задач является разработка универсальных и гибких моделей дискурсивной разметки, позволяющих адекватное фиксирование логико-семантических отношений в тексте. Это особенно важно при создании двуязычных параллельных корпусов.

Объект исследования – параллельный дискурсивный китайско-русский корпус официально-деловых текстов как инструмент сопоставительного лингвистического анализа.

Предмет исследования – разработка и разметка параллельного корпуса с последующим сопоставительным анализом синтаксической и дискурсивной структур китайских и русских официально-деловых текстов.

В рамках предлагаемой концепции дискурсивный анализ используется как метод анализа на уровне крупнейших речевых единиц. Структура дискурса является ключевой проблемой дискурсивного анализа и обработки естественного языка. Интеграция теории дискурсивного анализа в создание дискурсивного корпуса требует интеграции и уточнения формальных теорий моделирования дискурса.

На основе базовых понятий дискурсивного моделирования в диссертации формируется схема разметки дискурсивной структуры и создается дискурсив-

ный корпус. Кроме того, проводится статистический и сопоставительный анализ структурных особенностей китайских и русских деловых текстов.

Актуальность исследования. Моделирование дискурса имеет ключевое значение для решения различных задач семантического анализа при обработке естественного языка. По мере развития концепций машинной обработки объем анализируемых единиц постепенно увеличивается: от фонем и морфем, затем – слов и предложений, и, наконец, до дискурса как более крупной единицы. В результате дискурс становится самой актуальной единицей машинной обработки естественного языка, так как многие значения проявляются только на дискурсивном уровне.

Актуальной научной задачей является разработка методологических, теоретических и технологических основ создания параллельного дискурсивного корпуса, учитывающего логико-семантические и структурные особенности текстов разных языков. Решение этой задачи будет способствовать расширению возможностей межъязыкового исследования дискурса и развитию автоматизированных систем анализа текста.

За последние годы появилось множество современных экспериментальных дискурсивных корпусов (также известных как «дискурсивные банки деревьев», «трибанки», discourse treebanks): Дискурсивный банк деревьев, построенный в рамках теории риторических структур (RST Discourse Treebank) [Carlson, Marcu, Okurowski, 2002], Пенсильванский дискурсивный банк деревьев (Penn Discourse Treebank) [Prasad, Rashmi et al., 2019], Корпуса устной речи и проект «Рассказы о сновидениях» [Кибрик, Подлесская, 2009], русскоязычный дискурсивный корпус Ru-RSTreebank [Pisarevskaya et al., 2017], Пенсильванский дискурсивный банк деревьев китайского языка [Zhou, Xue, 2015], Дискурсивный корпус китайского языка [Li et al., 2014], Дискурсивный корпус связанных клауз китайского языка (Chinese clause relevance structure corpus) [冯文贺 et al., 2020; Lyu, Feng, 2023] и др. Эти дискурсивные корпуса успешно применяются для решения разнообразных практических задач, связанных с интер-

претацией дискурса, однако они еще не получили должного отражения в лингвистических теоретических исследованиях. Тем не менее возможности дискурсивной разметки наглядно демонстрируют потенциал систематизации дискурса с помощью компьютерных технологий и требуют дальнейшего исследования с точки зрения теории языка.

Корпусные исследования структуры языка дали многообещающие результаты во многих междисциплинарных областях исследований. В то же время в большинстве корпусов лингвистическое аннотирование ограничивается уровнями слова и предложения. Для моделирования формальных и семантических структур, возникающих на уровне текста, требуется разметка принципиально иного – дискурсивного уровня. Опыт развития системно-структурного дискурсивного анализа показывает, что дискурсивные элементы и структуры подчиняются универсальным языковым закономерностям, что позволяет рассчитывать на успешность лингвистического аннотирования на уровне дискурса.

На сегодняшний день отсутствует систематизированный подход к интеграции дискурсивных структур в параллельные корпуса. Это ограничивает возможности проведения сравнительного дискурсивного анализа и автоматизированной обработки дискурса на межъязыковом уровне. Возникает необходимость разработки новых методик, которые бы позволили создавать более точные и репрезентативные дискурсивные модели, учитывающие языковые различия и особенности структур дискурса в различных языках.

Степень разработанности проблемы. Типичной единицей формально-структурного анализа является предложение. Значительный вклад в разработку вопросов, связанных с его структурной организацией, внесли представители разных научных традиций. Среди американских дескриптивистов следует отметить Л. Блумфилда, Ф. Боаса, Э. Сепира, Б. Уорфа, Ч. Фриза, З. Харриса, Н. Хомского и др. В российской лингвистике этой проблематикой занимались В. Г. Адмони, Н. Д. Арутюнова, В. А. Белошапкова, Л. А. Булаховский, В. Г. Гак, А. В. Гладкий, Г. А. Золотова, Т. П. Ломтев, Е. А. Лютикова,

И. А. Мельчук, Б. Ю. Норман, Е. В. Падучева, З. Д. Попова, И. П. Распопов, С. Я. Фитиалов, Н. Ю. Шведова, Т. В. Шмелева и др. В китайской лингвистической традиции важный вклад внесли Дин Шушэн (丁树声), Фэн Шэнли (冯胜利), Фэн Чживэй (冯志伟), Ли На (Li Charles N.), Ли Цзиньси (黎锦熙), Лю Хайтао (刘海涛), Лу Цзяньмин (陆俭明), Люй Шусян (吕叔湘), Нин Чуньян (宁春岩), Шэнь Цзясюань (沈家煊), Ши Юйжчи (石毓智), Син Фуи (邢福义), Сюй Лицзюн (徐烈炯), Чжан Боцзян (张伯江), Чжао Шикай (赵世开), Чжао Юаньжэнь (赵元任), Чжу Дэси (朱德熙) и др.

С 1990-х гг. корпусные технологии внесли значительный вклад в развитие всех областей лингвистики, в частности компьютерной и структурной. Формальные лингвистические теории стали основой для создания лингвистически аннотированных корпусов. Особое место среди них занимают теории синтаксической разметки и соответствующие корпуса, которые относятся к наиболее сложным объектам моделирования. Выделяются два представительных типа способов синтаксического анализа предложения. Первый тип – дерево составляющих, отражающее иерархию группирующихся элементов; оно используется в грамматике непосредственных составляющих Л. Блумфилда [Bloomfield, 1933], трансформационной грамматике Н. Хомского [Chomsky, 1957], грамматике китайских клауз Син Фуи (邢福义) [邢福义, 1995; Син Фуи, 2020] и др. Второй тип – дерево зависимостей (или дерево подчинений), фиксирующее связи между структурными элементами; оно разрабатывается в рамках теории валентности глаголов и синтаксической структуры Л. Теньера [Tesnière, 1959], лексического подхода Р. Хадсона [Hudson, 1982], модели «смысл – текст» и синтаксического анализа И. А. Мельчука [Мельчук, 1974; Mel’čuk, 1988], а также в работах Х. Гейфмана [Gaifman, 1965], А. В. Гладкого [Гладкий, 1973], Л. Данлос [Danlos, 2004], Н. Фрейзера [Fraser, 1993], Дж. Хейса [Hays, 1964]; Лю Хайтао (刘海涛) [刘海涛, 1991], Фэн Чживэя (冯志伟) [冯志伟, 2001] и др.

Дискурс как единица формально-структурного анализа оформился в 1950-е гг., когда З. Харрис ввел в научный оборот термин «дискурсивный анализ» (discourse analysis) [Harris, 1952]. В дальнейшем проблемы дискурсивного анализа как особого уровня лингвистики нашли отражение в трудах целого ряда исследователей. Среди зарубежных ученых следует отметить Б. Гроса, К. Сайднера, Дж. Хиршберг, Дж. Хоббса, Э. Хови, Д. Румелхарта, К. Маккьюина и др. В отечественной традиции к данной проблематике обращались Н. С. Валгина, В. Г. Гак, И. Р. Гальперин, С. И. Гиндин, Н. Д. Зарубина, Г. А. Золотова, А. А. Кибрик, Л. М. Лосева, Б. С. Лунев, Т. М. Николаева, А. И. Новиков, В. В. Одинцов, М. И. Откупщикова, Е. В. Падучева, Е. А. Реферовская, Г. Я. Солганик, Л. В. Сухова, З. Я. Тураева, И. А. Фигуровский и др. Существенный вклад внесли и китайские исследователи, среди которых Бай Чуньжэнь (白春仁), Ван Фусянь (王福祥), Ли Сикуй (李锡奎), Люй Шусян (吕叔湘), Ляо Цючжун (廖秋忠), Сон Жоу (宋柔), Сюй Цзиннин (徐晶凝), Сюй Цзюцзю (徐赳赳), Тянь Сяолинь (田小琳), У Вэйчжан (吴为章), Фу Хуэйминь (付慧敏), Хуан Гоувэнь (黄国文), Чжан Цзюй (战菊), Чу Чонси (Chauncey C. Chu), Чэнь Цзе (陈洁), Ши Тицян (史铁强) и др.

Основу для дискурсивного анализа и моделирования заложил синтаксический анализ предложения. По аналогии с синтаксической разметкой дискурсивная разметка реализуется в двух основных формах: в виде деревьев составляющих и деревьев зависимостей. Первая традиция опирается на теорию риторической структуры (Rhetorical Structure Theory [Mann, Thompson, 1988]), которая легла в основу создания дискурсивных банков на разных языках. Разработкой и применением ТРС занимались У. Манн, С. Томпсон, М. Табоада, Л. Карлсон, Д. Марку, М. Э. Окуровский, А. Нойманн, М. Штеде, А. Зельдес, а также российские и китайские исследователи – М. И. Ананьева, И. М. Кобозева, Д. Б. Писаревская, А. А. Кибрик, А. О. Литвиненко, В. И. Подлесская, Лэ Мин (乐明), Лю Ян Джанет (Liu Yang Janet), Пэн Сияо

(Peng Siyao), Чжоу Годун (周国栋), Ли Яньцуй (李艳翠). Вторая традиция связана с переносом идей грамматики зависимостей на уровнь дискурса, что привело к разработке теорий дискурсивной структуры зависимостей и корпусов дискурсивных деревьев зависимостей. В данном подходе дискурс трактуется как бинарная структура, состоящая из языковых единиц определенного уровня. Значительный вклад в ее развитие внесли Ф. Вольф, Э. Гибсон, Л. Данлос, Е. Мильтсакаки, Р. Прасад, Б. Веббер, А. Джоши, Дзюн Судзуки, Ясухиса Есида, а также китайские исследователи – Лю Тин(刘挺), Ли Суцзянь(李素建), У Юнпэн(吴永芃), Конг Фан(孔芳), Фэн Вэньхэ (冯文贺).

Создание и использование параллельных корпусов стало одним из ключевых направлений развития корпусной лингвистики, что проявилось в активном расширении коллекций многоязычных ресурсов, в том числе русско-китайских и китайско-русских (С. П. Дурнева, О. В. Кукушкина, Ю. Н. Кузнецова, М. Ю. Мухин, Б. В. Орехов, К. И. Семенов, В. П. Захаров; Ли Фэн (李峰), Лю Мяо (刘淼), Цуй Вэй (崔卫), Чжан Лань (张岚), Чэнь Сяохуэй, Шао Цин (邵青) и др.). Современное состояние изучения выравнивания в параллельных корпусах характеризуется развитием методов как автоматической, так и полуавтоматической обработки текстов на различных уровнях языка. Значительный вклад в разработку теоретических основ и практических методов внесли такие ученые, как Д. О. Добровольский, А. В. Зубов, И. И. Зубова, Г. Е. Кедрова, Д. В. Сичинава, М. А. Шведова, Бо Сяоцзин (柏晓静), Фэн Миньсюань (冯敏萱), Фэн Вэньхэ (冯文贺), Ван Кэфэй (王克非).

Параллельный корпус активно используется в сопоставительном анализе языка, который направлен на выявление языковых различий и сходств, а также на углубленное изучение семантических, синтаксических и дискурсивных структур в разных языках. Значительный вклад в эту область внесли такие ученые, как Д. О. Добровольский, А. М. Зуров, С. А. Маник; Ф. Бопп, Дж. Гримм, К. Джеймс, Т. Р. Кшешовски, А. МакЭнери, Р. Раск, Р. Сяо, М. А. К. Хэллидей;

Ван Вэньбинь (王文斌), Вэй Найсин (卫乃兴), Конг Лей (孔蕾), Пань Уэньгуо (潘文国), Сюй Юйлин (许余龙), Цинь Хунву (秦洪武) и др. Они предложили основные методы сопоставительного анализа, включая количественные подходы на основе параллельных корпусов, что способствовало развитию сопоставительного языкознания.

Лингвистическое исследование официально-делового стиля имеет глубокие традиции, и их основы заложили Л. Г. Барлас, Т. В. Губаева, Л. Р. Дускаева, Н. В. Егорова, М. Н. Кожина, Н. А. Купина, Т. В. Матвеева, О. В. Протопопова, Д. Э. Розенталь, Ху Юйшю (胡裕树), Ни Баоюань (倪宝元), Чжан Хуэйсэн (张会森), Чжао Цзе (赵洁) и др. Одним из важных фрагментов официально-делового стиля является дипломатический язык, который изучали А. Ф. Абдулсалам, Г. И. Исина, Ю. М. Кукарина, А. С. Мустафина, В. И. Попов, Вань Яньсин, Ма Лимин, У Айхуа (武瑗华) и др. Исследования этих ученых особенно значимы для осуществления анализа структурных особенностей дипломатических текстов.

Таким образом, к наиболее важным сферам, определившим предмет настоящей диссертации, относятся корпусная лингвистика (в частности, создание дискурсивных банков деревьев), структурный дискурсивный анализ, формально-синтаксические теории, а также сопоставительное языкознание.

Цель исследования заключается в создании параллельного дискурсивного китайско-русского корпуса официально-деловых текстов и проведении на его основе сопоставительного синтаксического и дискурсивного анализа.

Достижение поставленной цели предполагает решение следующих **задач**:

1) создать теоретическую основу описания дискурсивной структуры китайских официально-деловых текстов и их переводов на русский язык (проанализировать и сопоставить различные способы анализа дискурсивной структуры, обосновать использование концепции структур зависимостей для дискурсивного анализа и т. д.);

2) определить элементы дискурсивной структуры и их лингвистические характеристики, проанализировать основные понятия описания дискурсивной структуры в рамках теории зависимостей;

3) сформулировать основные этапы и приемы создания параллельного дискурсивного корпуса официально-деловых текстов;

4) выбрать текстовый материал для создания дискурсивного корпуса, выработать конкретные принципы дискурсивной разметки в рамках структуры зависимостей, включая принципы деления на элементарные дискурсивные единицы (ЭДЕ) и их выравнивания, принципы структурированной организации этих единиц, а также принципы разметки особенностей структурных пар;

5) провести дискурсивную разметку и сформировать электронный корпус, в процессе работы адаптировать к материалу специфические принципы дискурсивной разметки;

6) провести сопоставительный статистический анализ и на основе размеченных данных сопоставление синтаксической и дискурсивной структур китайских и русских деловых текстов.

Материал исследования составляют китайские и русские официально-деловые тексты – межправительственные двусторонние документы РФ и КНР, а именно совместные заявления и совместные декларации, извлеченные с сайта правительства Китая (URL: https://www.gov.cn/) и сайта президента России (URL: http://kremlin.ru/). На настоящий момент собрано и размечено 24 текста: по 12 на китайском и русском языках, общий объем которых превышает 64 000 текстоформ[1] (41 784 китайских и 22 602 русских) и включает 429 параллельных абзацев для каждого языка.

Методология и методы исследования. Теоретической и методологической основой диссертации стали научные представления о методологии струк-

[1] Текстоформа в корпусной лингвистике понимается как уникальная словоформа и единица разметки на лексическом уровне.

турной лингвистики, положения корпусной лингвистики и дискурсивного анализа, а также подходы к семантической обработке естественного языка. В работе использовалась группа методов: формально-структурный, корпусный, графико-статистический, когнитивный, межъязыковой сопоставительный, квантитативно-лингвистический, а также общенаучные методы наблюдения, описания и моделирования.

Научная новизна работы заключается в том, что в ней применен оригинальный подход к планированию и разметке параллельного дискурсивного корпуса официально-деловых текстов; созданный корпус послужил основой проведенных сопоставительных исследований особенностей китайского и русского делового дискурса. При осуществлении дискурсивного анализа и корпусной разметки обоснована и использована модель зависимостей. Новизна обнаруживается в ранее не аннотированном материале: для корпуса привлечены документы РФ и КНР, которые создаются совместно правительствами и согласуются в ходе совместной работы, что обеспечивает их семантическую эквивалентность.

Теоретическая значимость диссертации обусловлена концепцией модели зависимостей, которая служит базовой для формирования дискурсивного корпуса. В работе уточняются ключевые понятия дискурсивной разметки – элементарная дискурсивная единица, клауза, дискурсивное отношение, дискурсивный коннектор, дискурсивная вершина и др. Теоретически значимы принципы разметки и механизмы выравнивания параллельного корпуса китайских и русских текстов, отличающихся как синтаксическими структурами, так и результатами сегментации. Обосновано понятие русского синтаксического аналога (РСА) китайской клаузы, проведено сопоставление синтаксических типов сегментов в китайских и русских параллельных текстах. Получены и концептуально обобщены количественные данные дискурсивного сопоставления китайской и русской частей корпуса.

Практическая значимость исследования в первую очередь проявляется в создании впервые китайско-русского параллельного дискурсивного корпуса официально-деловых текстов, который открыт для дальнейших синтаксических, дискурсивных, семантических исследований и в целом для решения различных задач контрастивной лингвистики. Разработанная и адаптированная для корпусной разметки модель зависимостей может быть использована при создании других дискурсивных корпусов, в том числе параллельных. В диссертации предложен ряд механизмов, позволяющих решать проблемы дискурсивной асимметрии двух разных языков в процессе выравнивания, что представляет ценность для последующих корпусных проектов.

Результаты работы позволяют усовершенствовать лингвистические теории формально-структурного анализа на самом сложном уровне языка – дискурса, вносят вклад в развитие корпусной лингвистики, в частности в области синтаксической и дискурсивной разметки, способствуют углубленному изучению языковой структуры и объяснению языковых явлений с точки зрения структуры, а также помогают совершенствовать технологии обработки естественного языка.

Положения и материал диссертации могут быть использованы в преподавании лингвистических дисциплин, связанных с дискурсивным анализом, корпусной и сопоставительной лингвистикой, а также в смежных областях.

Достоверность полученных результатов обеспечивается большим объемом эмпирического материала, который лег в основу создания корпуса и последующего корпусного исследования, комплексным подходом к анализу текстовых данных, а также использованием целого ряда классических и современных методов анализа, соответствующих целям и задачам работы. Научные положения и выводы, сформулированные в диссертации, подкреплены экспериментальными данными и прошли тщательную апробацию. Для наглядной визуализации, верификации полученных результатов и обеспечения открытого до-

ступа к материалам исследования создан интерактивный веб-ресурс корпуса, доступный по адресу: https://www.crpardt.cn/index.html?lang=ru.

Положения, выносимые на защиту:

1. Концепция созданного китайско-русского параллельного дискурсивного корпуса официально-деловых текстов основана на структуре зависимостей как оптимальном способе представления дискурсивной структуры текста, а также на единой схеме дискурсивной разметки, учитывающей специфику китайского и русского языков.

2. Создание параллельного дискурсивного корпуса требует прохождения следующих исследовательских этапов:

- разработка теоретической основы корпуса – обобщение схемы дискурсивной разметки, установление структуры зависимостей в качестве основы анализа дискурсивной структуры, обоснование ключевых понятий анализа дискурсивной структуры (клауза и элементарная дискурсивная единица, дискурсивное отношение, дискурсивный коннектор, дискурсивная вершина);
- разработка принципов отбора текстов, применения технологий структурирования материала и разметки;
- собственно разметка и выравнивание, которое предполагает преодоление асимметрии дискурсивных структур китайских и русских официально-деловых текстов.

3. Сопоставление дискурсивных структур в параллельном корпусе целесообразно на основании двусторонних официально-деловых документов правительств РФ и КНР (совместных деклараций и заявлений). Эти тексты, признанные в международной коммуникации аутентичными, выступают в качестве равноправных оригиналов, что обеспечивает их семантическую эквивалентность.

4. Дискурсивная структура представлена в корпусе в виде бинарного направленного ациклического графа, в котором узлы соответствуют элементар-

ным дискурсивным единицам, а дуги фиксируют структурные пары, определяемые асимметричными логико-семантическими отношениями. Структура включает также такие параметры, как тип дискурсивных отношений, коннекторы и вершины.

5. Механизмы разметки и выравнивания дискурсивных структур китайских и русских текстов включают определение элементарных дискурсивных единиц (китайских клауз и их русских синтаксических аналогов), установление структурных пар внутри абзацев и оптимизацию дискурсивной структуры в параллельных текстах.

6. В корпусе размечаются следующие дискурсивные отношения: двенадцать логико-семантических отношений (пояснение, соединение, причина – следствие, цель, дополнение, оценка, время, противопоставление, сопоставление, условие, градация и уступка) и отдельный синтаксический тип для русских синтаксических аналогов, выделяемых внутри простого предложения.

7. Сопоставительный статистический анализ на базе корпуса выявляет следующие основные параметры дискурсивных структур:

- 89 % абзацев в размеченных текстах содержат от одной до шести ЭДЕ;
- большинство китайских клауз и русских синтаксических аналогов структурно эквивалентны, при этом существует синтаксическая асимметрия в передаче информации на двух языках;
- достигнуто выравнивание 1248 структурных пар, при этом 4 пары остались невыровненными из-за синтаксических и семантических различий;
- наиболее частотными дискурсивными отношениями в официально-деловых текстах являются пояснение, соединение, причина – следствие, цель;
- дискурсивные вершины чаще всего соотносятся с первыми ЭДЕ структурной пары.

Апробация работы. Выводы и основные положения диссертационного исследования обсуждались на заседании кафедры фундаментальной и приклад-

ной лингвистики и текстоведения Уральского федерального университета, а также представлены и апробированы на ряде международных и российских научных конференций, включая: Международная научная конференция «Новая Россия: традиции и инновации в языке и науке о языке» (Екатеринбург, 28–30 сентября 2016 г.); The 5th Conference on Natural Language Processing and Chinese Computing & The 24th International Conference on Computer Processing of Oriental Languages (Куньмин, Китай, 2–6 декабря 2016 г.); Международная научная конференция «Корпусная лингвистика – 2017» (Санкт-Петербург, 27–30 июня 2017 г.); The 16th China National Conference on Computational Linguistics (Нанькин, Китай, 13–15 октября 2017 г.); The 18th Chinese Lexical Semantics Workshop (Лэшань, Китай, 18–20 мая 2017 г.); Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов – 2019» (Москва, 8–12 апреля 2019 г.); Теоретическая семантика и идеографическая лексикография: Словарь. Дискурс. Корпус» (Екатеринбург, 29–30 октября 2019 г.).

По теме диссертации опубликовано 9 работ, включая 3 статьи в рецензируемых научных журналах и изданиях, определенных ВАК РФ и Аттестационным советом УрФУ, в том числе 1 статья, индексированная в международных базах цитирования Scopus, WoS.

Структура работы. Диссертация состоит из введения, четырех глав, заключения, списка сокращений и терминов, списка литературы и двух приложений. Глава 1 («Теоретические основы создания параллельного дискурсивного корпуса») посвящена теоретическим основам предлагаемого исследования: специфике корпуса как лингвистического исследовательского ресурса, определению дискурса как объекта лингвистического моделирования и роли параллельного корпуса в сопоставительной лингвистике. Глава 2 («Параллельный дискурсивный корпус: модель построения, терминология, единицы описания») содержит описание способов моделирования дискурсивных структур, разработку концептуальных основ дискурсивной структуры зависимостей, уточнение терминологии и определение единиц описания. Глава 3 («Опыт создания китай-

ско-русского параллельного корпуса официально-деловых текстов») описывает практические аспекты создания корпуса: этапы отбора материала для корпуса, принципы дискурсивной разметки и выравнивания, используемое программное обеспечение и способы хранения размеченных данных. Глава 4 («Дискурсивные структуры в китайско-русском параллельном корпусе: статистический анализ и интерпретация данных») посвящена количественному и качественному анализу: изучаются параметры абзацев, результаты сегментации и выравнивания элементарных дискурсивных единиц, выравнивания дискурсивных структур зависимостей двух языков и разметки дискурсивных параметров структурных пар. Заключение подводит итоги исследования, формулирует основные выводы и очерчивает перспективы последующей работы. Приложение 1 содержит перечень документов, размеченных в параллельном корпусе, а Приложение 2 наглядно демонстрирует результаты дискурсивной разметки в виде скриншотов веб-ресурса.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ ПАРАЛЛЕЛЬНОГО ДИСКУРСИВНОГО КОРПУСА

Цель данной главы – обосновать теоретическую базу исследования, показать актуальность создания и использования лингвистических корпусов, рассмотреть общие вопросы изучения дискурсивных структур, а также проиллюстрировать связь между параллельным корпусом и сопоставительным анализом языка.

1.1. Корпус как лингвистический исследовательский ресурс

1.1.1. Современное состояние корпусных исследований языка

Интерес к корпусным исследованиям возник в начале 1960-х гг., когда в США был опубликован первый электронный Брауновский корпус [Francis, Kučera, 1964]. Брауновский корпус включает 1 млн словоупотреблений из 500 текстов различных жанров письменной речи американского варианта английского языка: газетных статей, научных трудов, объявлений, книг о хобби, религиозной литературы и др.

Брауновский корпус стал прорывом в американской лингвистике того времени. Однако работа над языковыми корпусами подверглась широкой критике. Полное отрицание значения корпусных исследований в то время можно найти у Н. Хомского. По его словам, «Корпусная лингвистика ничего не значит. Это как сказать, что <…>, предположим, физики и химики решат, что <…> будут записывать на видео, как что-то происходит в мире; они соберут огромное количество видеозаписей обо всем, что происходит, и на основе этого будто бы смогут сделать какие-нибудь обобщения или что-то понять» – цит по: [Andor, 2004, p. 97]. Сторонники Н. Хомского считали, что «создавать Брауновский корпус – значит впустую тратить время» [Leech, Johansson, 2009, p. 6]. Причина критики заключалась в том, что корпуса первого поколения не имели грамматической и иной разметки, поэтому не поддерживали быстрый и многофункциональный поиск, а также системно-структурный анализ языка. По сути, корпус-

ные исследования того времени ограничивались наблюдением за большим объемом языкового материала.

В то же время корпусные исследования вызвали значительный интерес в Европе, во многом соответствуя британской традиции эмпирического анализа языка [Firth, 1957][1]. В 1980-х гг. в Англии был создан ряд корпусов, таких как корпусный проект Исследования употребления английского языка (начиная с 1959 г.)[2] [Quirk, 1966], Ланкастер-Осло-Берген Корпус (1961–1978) [Johansson, Leech, Goodluck, 1978], проект Международный компьютерный архив современного английского языка [Leech, Johansson, 2009], Британский национальный корпус[3] [Leech, 1992] и др.

За более чем полувековое развитие корпусной лингвистики количество корпусов в мире, в том числе в России и Китае, значительно увеличилось. Первый современный корпус русскоязычных текстов – Уппсальский корпус (ныне Уппсальский и Тюбингенский корпус) – был создан в 1980-е гг. в Уппсальском университете (Швеция). Потом появились многие другие проекты: Машинный фонд русского языка [Машинный фонд русского языка: идеи и суждения, 1986], Корпус текстов русских газет конца XX века (более 200 тыс. слов) (2000 г.)[4], Хельсинкский аннотированный корпус [Копотев, Мустайоки, 2003], наиболее известный Национальный корпус русского языка (далее – НКРЯ) [Плунгян, Сичинава, 2004] с открытым доступом с 2004 г. (сегодня уже НКРЯ 2.0 [Савчук и др., 2024]) и др.

В Китае корпусные исследования сначала были преимущественно связаны с лингводидактическими задачами, в частности с обучением английскому

---

[1] Суть этого подхода заключается в том, что значение слова (равно как и другие лингвистические концепты) существует только в контексте (в тексте) [цит. по: Захаров, Богданова, 2020, с. 14].

[2] URL: https://www.ucl.ac.uk/english-usage/index.htm (дата обращения: 31.05.2024).

[3] URL: http://www.natcorp.ox.ac.uk/ (дата обращения: 31.05.2024).

[4] URL: https://www.philol.msu.ru/~lex/corpus/corp_descr.html (дата обращения: 31.05.2024).

языку. Первый корпус – Корпус научно-технического английского языка Шанхайского университета транспорта (JDEST) – был создан в 1980-е гг. [杨惠中, 黄人杰, 1982]. Что касается собственно корпусов китайского языка, то их разработка началась гораздо позже из-за сложности компьютеризации иероглифической письменности. В начале 1990-х гг. был опубликован первый электронный корпус китайского языка – письменный архив газеты «Жэньминьжибао» (кит. 《人民日报》语料库). На основе его базы данных в 1998 г. был сначала построен первый морфологически аннотированный Корпус китайского языка (кит. 《人民日报》标注语料库).

В настоящее время в Китае созданы десятки крупных корпусов; наиболее представительными из них являются следующие: Сбалансированный корпус китайского языка (кит. 语料库在线) [靳光瑾 et al., 2005], Корпус BCC Пекинского университета языка и культуры [荀恩东 et al., 2016], Корпус современного китайского языка Центра китайской лингвистики при Пекинском университете (кит. CCL语料库) [詹卫东 et al., 2019]. В Китае развитие корпусных исследований также сопровождалось разработкой многих специфических корпусов, ориентированных на решение отдельных научных проблем – см.: [谢家成, 2004; 王龙吟, 何安平, 2005; 梁茂成, 2003; 肖维青;, 2005] и др.

Наряду с развитием национальных корпусов формируются многоязычные корпуса, в том числе русско-китайские и китайско-русские. В России с 2016 г. доступен параллельный русско-китайский подкорпус НКРЯ [Сичинава, Шведова, 2010]. В Китае – Политический русско-китайский и китайско-русский параллельный корпус [崔卫, 李峰, 2014]; Русско-китайский параллельный корпус научных текстов гуманитарной области [Тао, Захаров, 2015; 陶源, 2014]; Русско-китайский переводческий корпус [刘淼, 邵青, 2016] и др. Они открыли новую перспективу для теории и практики перевода и сопоставления разных языков.

К концу XX в. корпусная лингвистика сформировалась как отдельное направление. Как отмечает В. А. Плунгян, корпусные исследования стали больше «призывать лингвистику изучать факты, а не конструкты; рассуждать о свойствах наблюдаемых явлений, а не о свойствах моделей» [Плунгян, 2008, с. 8]. При этом постоянно подчеркивается важность количества языковых фактов (текстов): «чем больше материал, тем выше достоверность выводов, тем шире сфера действия наблюденных закономерностей» [Баранов, 2021, с. 121]. Корпус используется как языковой ресурс, обеспечивающий естественные речевые данные для изучения языка. Размер корпуса и его репрезентативность имеют ключевое значение для достоверности исследований. Кроме того, корпус позволяет проводить статистический анализ, который становится обязательным инструментом верификации и воспроизводимости результатов [Копотев, 2021, с. 93].

На сегодняшний день «корпусная революция» произошла во многих областях лингвистических исследований. Корпусные данные и технологии используются как удобный инструмент для решения практических задач, в частности, лексикографических исследований [Беляева, 2013; Добровольский, 2012; Земичева, Иванцова, 2019; Чумарина, 2017], изучения грамматики [Плунгян, Стойнова, Добрушина, 2016; Quirk et al., 1985], и обучения иностранному языку и разработке учебников [李文中, 濮建忠, 2001; Сысоев, 2010; Чернякова, 2011] и т. д. Корпусные технологии дают возможность быстрого поиска, построения конкорданса, получения статистической информации о языковых единицах и о лингвистических категориях и метаданных и т. п. – см.: [Захаров, Богданова, 2020].

По мере развития корпусные подходы становятся одними из важнейших в языкознании, а корпусные исследования подразделяются как минимум на два направления: основанные на корпусных данных (corpus-based study) и вызванные корпусными данными (corpus-driven study) [Tognini-Bonelli, 2001, p. 99]. Однако, некоторые ученые отмечают, что различия между двумя основными

подходами к изучению языковых явлений стираются – см. об этом: [McEnery, Hardie, 2012]. Главная причина состоит в том, что эти два подхода посвящены общему объекту лингвистического исследования – речевым фактам: «в мировой науке не стоит вопрос об обращении к корпусам, вопрос заключается скорее в подходе к этому лингвистическому ресурсу» [Борискина, 2015, с. 24].

Корпуса также оказали значительное влияние на обработку естественного языка (Natural Language Processing, NLP). В «эпоху искусственного интеллекта» (далее – ИИ) компьютерная обработка информации практически всегда опирается на большие аннотированные корпуса. Однако следует отметить, что методы, лежащие в основе современных интеллектуальных технологий, отличаются от традиционных корпусных подходов с появлением больших языковых моделей (далее – БЯМ, Large language models – например, BERT, GPT и др.). Если ранние системы ИИ, такие как ELIZA (диалоговые программы) и SHRDLU (понимание естественного языка), были относительно понятны лингвистам, то современные модели на основе больших данных и глубокого обучения представляют собой своего рода «черный ящик», трудно объяснимый даже для разработчиков.

Тем не менее, если языковые модели демонстрируют высокие способности генерации и понимания текста, это указывает на существование глубоких закономерностей, которые могут быть объяснены с позиций лингвистики [许余龙, 刘海涛, 刘正光, 2020]. Новые перспективы лингвистики, вероятно, заключаются в том, что изучение механизмов работы больших языковых моделей может приблизить ученых к пониманию сущности языка.

1.1.2. Разметка и корпусное моделирование языка

Одной из ключевых особенностей корпуса является наличие разметки, или аннотации [Захаров, Богданова, 2020; Копотев, Мустайоки, 2008; Плунгян, 2003; Савчук, 2011; Савчук и др., 2024]. Разметка определяется как «интерпретирующая (в первую очередь лингвистическая) информация, внесенная в суще-

ствующие электронные языковые данные (устную и/или письменную речь) или приписанная к ним с помощью специальных тегов» [Leech, 1993, p. 275]. В учебнике «Корпусная лингвистика» В. Ю. Захаров и С. Ю. Богданова определяют разметку как исследовательскую работу – «приписывание текстам и их компонентам специальных тегов» и подразделяют разметку на две основные категории – собственно лингвистическую, описывающую лексические, грамматические и прочие характеристики элементов текста, и внешнюю, экстралингвистическую (сведения об авторе и о тексте: автор, название, год и место издания, жанр, тематика) [Захаров, Богданова, 2020, с. 34–35].

За время развития корпусной лингвистики были разработаны различные типы разметки (лингвистической и экстралингвистической), которые предоставляют многоплановую информацию для обработки естественного языка. Разметка понимается как внесение интерпретирующих тегов в текст (письменный или устный) в целях машинного обучения – см. об этом: [Goldberg, 2022; Joshi, 1991; Nadkarni, Ohno-Machado, Chapman, 2011]. Помимо традиционной грамматической разметки, активно развивается описание пауз и интонации, семантических элементов текстов.

Разработка разметки напрямую связана с теоретическими исследованиями языка. Язык разметки представляет собой искусственный метаязык, основанный на лингвистической теории и составляющий ее неотъемлемую часть. Сегодня многие виды разметки одновременно конкретизируют и обобщают различные лингвистические теории. Например, при осуществлении семантической разметки в НКРЯ использовались сведения о значении слов и структуре семантических классов – в частности, Толкового словаря русских глаголов под ред. Л. Г. Бабенко [Толковый словарь русских глаголов, 1999], Системного семантического словаря русского языка Л. М. Васильева [Васильев, 2000] и др. Морфологическая разметка русских корпусов в основном опирается на морфологическую классификацию Грамматического словаря русского языка А. А. Зализняка [Зализняк, 2003]. Функционально-семантическая разметка мо-

жет соответствовать идеологии модели «Смысл – Текст» [Мельчук, 1974][1]. Помимо увеличения объема корпусов, возникают и новые виды разметки – например, введение микросинтаксической разметки [Иомдин, 2008].

Развитие теории разметки тесно связано с формальными теориями языка, предложенными в структурной лингвистике [Гладкий, Мельчук, 1969; Chomsky, 1957; 刘海涛, 2017]. На ранних этапах, из-за технических ограничений, формальные языковые модели отходили от естественных языковых фактов, что вызвало критику со стороны антропоцентрических лингвистических школ. С развитием корпусных технологий, особенно корпусной разметки, стало возможным соотнести результаты абстрактного формального описания языка с естественными языковыми данными. Если формальная лингвистика занимается «математизацией» языка и созданием языковых моделей, то задача корпусной разметки состоит не только в моделировании языка, но и в объяснении этих формальных языковых явлений и выявлении общих закономерностей.

Однако в большинстве исследований разметка рассматривается преимущественно как технологическая операция на предварительном этапе подготовки корпуса. Языковеды очень осторожно относятся к самостоятельной разработке разметки: она считается трудоемкой и требует автоматизации, поскольку качество корпусных исследований сильно зависит от репрезентативности корпуса [Захаров, Богданова, 2020, с. 34]. Разработчики НКРЯ отмечают, что «добавление каждого нового текста сопровождается трудоемкой работой по его разметке»[2], поиск по корпусу возможен только по уже заданным параметрам, что в итоге сужает его возможности.

В последние годы исследования, основанные на лингвистическом материале, показывают, что стремление к бесконечному увеличению объема корпуса утрачивает прежнюю оправданность: после достижения определенного раз-

[1] Подробное см.: URL: https://ruscorpora.ru/page/annotation/ (дата обращения: 22.05.2024).

[2] Об этом см.: URL: https://ruscorpora.ru/page/faq/ (дата обращения: 22.05.2024).

мера в корпусе находится все меньше «новых» языковых фактов. В связи с этим внимание ученых смещается на детализацию и совершенствование разметки.

Типичным примером, сочетающим разметку корпуса и описание человеческих знаний, является электронная база знаний HowNet, созданная китайскими лингвистами Дун Чжэньдуном и Дун Цянем в 1990-х гг. [董振东, 1998; Dong, Dong, 2003]. Главная идея создания HowNet заключается в описании базовых понятий мировых знаний (в виде словарной статьи) через атрибуты или свойства языковых единиц (слов) и отношений между ними. При подготовке проекта HowNet была разработана целая схема метаязыка для аннотирования человеческих знаний через описания значений слов – база знаний семемы (sememe knowledge base) и метаязык KDML (Knowledge Database Markup Language): база знаний семемы, представленная в виде тегов-семем, используется для описания основных атрибутов и свойств понятий (в форме лексемы); а метаязык KDML – для описания отношений между семемами [董振东, 1998, p. 79–81].

Работа Дун Чжэньдуна по созданию HowNet демонстрирует важность разметки, особенно в плане описания семантических отношений между единицами языка. Хотя HowNet формально хранится в словарных статьях, его создание существенно отличается от традиционного тезауруса: «задача составления тезауруса заключается в классификации языковых фактов, а задача проекта HowNet – установление отношений между понятиями и между семемами… и можно сказать, что наша разметка формирует языковым способом описание концепций, лежащих в основе человеческого знания о мире» [董振东, 董强, 2001, p. 5].

Успешное применение схемы разметки в HowNet показывает, что разметка может выступать как относительно самостоятельная и стабильная система, пригодная для описания различных языковых явлений и применения к разным языкам. На основе этой системы можно практически работать со многими языками. Как отмечает Дун Чжэньдун, «именно с целью проверки эффективности

и универсальности данной системы разметки в базе HowNet изначально поддерживаются два языка – китайский и английский» [董振东, 1998, p. 76]. В рамках такого понимания многоязычного корпуса разработка универсальной для различных языков системы разметки рассматривается как подход к лингвистическим исследованиям – в частности, в области типологических и сопоставительных исследований. Аналогично проекты по Универсальным зависимостям (Universal Dependencie[1]) [Nivre et al., 2020; Nivre et al., 2016] и ворднет-подобные тезаурусы WordNet [Fellbaum, 1998] (русский RussNet [Азарова и др., 2002; Azarova и др., 2002], китайский WordNet [Wang, Bond, 2013]) демонстрируют возможность построения лингвистических ресурсов на основе единых универсальных моделей разметки.

Исходя из этого в данной работе при разработке схемы дискурсивной разметки одновременно проводятся теоретическая обработка схемы аннотирования и практика разметки, поэтапно аннотируются тексты корпуса (подробнее см. в п. 3.3.), причем работа с китайским и русским текстами проводится одновременно.

В данной работе мы предполагаем создать параллельный корпус, в котором единая схема аннотирования будет применяться как к китайским, так и к русским текстам. Термин «параллельный» используется здесь, поскольку двуязычные тексты с семантическими соответствиями служат материалом, однако наша цель не в сопоставлении текстов с точки зрения переводоведения, а в объединении сходных текстов в рамках единой дискурсивной разметки и проведении сопоставления на основе ее результатов.

Исходим из следующих ключевых положений:

1) разметка является метаязыком для описания языка и использования языковых единиц в речи;

---

[1] URL: https://universaldependencies.org/ (дата обращения: 22.05.2024).

2) разработка разметки для корпуса фактически предполагает создание теоретических рамок для формального описания и моделирования языка;

3) данные, полученные в ходе разметки и с ее помощью, открывают новые возможности для моделирования языка и изучения его структуры.

1.2. Понятие дискурса как объекта лингвистического моделирования

В последние годы дискурсивный анализ активно развивается. Как отмечают Д. Шиффрин и соавторы, дискурсивному исследованию уделяют внимание различные дисциплины: 1) лингвистика, антропология и этнология, философия, которые изучают способы построения и понимания дискурса, дискурсивное моделирование, подходы к дискурсивному анализу; а также 2) когнитивная психология, социология, искусственный интеллект, которые применяют методы дискурсивного анализа в своих исследованиях [Schiffrin, Tannen, Hamilton, 2001, p. 1–10]. В связи с междисциплинарным характером дискурсивного анализа термины «дискурс» и «дискурсивный анализ» определяются и трактуются по-разному.

Термин «дискурс» (фр. discours, англ. discourse) происходит от латинского слова “discursus” и был впервые использован во введении уровневого анализа языка французским лингвистом Э. Бенвенистом в 1950-е гг. [Бенвенист, 1962]. По его мнению: «С предложением мы покидаем область языка как системы знаков и вступаем в другой мир, в мир языка как средства общения, выражением которого является речь (le discours)» [Там же, с. 138]. Хоть Бенвенист не рассматривал дискус как объект уровневого анализа языка, его рассуждения о дискурсе сами по себе поднимают вопрос о структурном анализе единиц языка выше уровня предложения.

В современной лингвистике дискурс понимают по-разному. Так, П. Серио выделяет восемь значений этого термина:

1) эквивалент понятия «речь»;

2) единица, превышающая по размерам фразу, высказывание в глобальном смысле, то есть то, что является предметом исследования «грамматики текста», которая изучает последовательность отдельных высказываний;

3) в рамках теорий высказывания или прагматики воздействие высказывания на его получателя с учетом ситуации;

4) «беседа» как основной тип высказывания;

5) речь с позиции говорящего в противоположность «повествованию», которое разворачивается без эксплицитного вмешательства субъекта высказывания;

6) употребление языковых единиц, их актуализация в речи;

7) система ограничений, которые в силу определенной социальной или идеологической позиции накладываются на неограниченное число высказываний, свойственных для определенного вида социума;

8) теоретический конструкт, предназначенный для исследований производства текста [Серио, 1999, с. 26–27].

Исходя из различных трактовок дискурсивный анализ в современной лингвистике можно условно разделить на три направления: 1) критический дискурсивный анализ, происходящий из французской философии постмодернизма [Фуко, 1996; Dijk van, 1993; Dijk van, 2004], рассматривающий «дискурс» как отражение идеологии и ментальности, присущих тексту, обладающему целостностью, связностью и погруженностью в контексты (социокультурный, социально-психологический и др.); 2) дискурсивный анализ, ориентированный на языковое употребление, фокусирующийся на влиянии внеязыковых факторов (коммуникативных, социальных и т. д.) на речевую деятельность и ее продукты (письменные и устные); 3) дискурсивный анализ в задачах обработки естественного языка и компьютерной лингвистике, который стремится ввести концептуальные элементы дискурса в рамки формально-структурного анализа и дать машине необходимые знания для моделирования процессов построения

(порождения, синтеза) и понимания (анализа) дискурса, анализируя дискурсивные элементы и их отношения.

Дискурсивный анализ первого типа сильно отличается от понимания, которое релевантно в контексте лингвистических исследований – см. например: [Серио, 1999; Йоргенсен, Филлипс, 2008]. По мнению Ю. С. Степанова, во «французском» понимании дискурс – это не столько языковое произведение, сколько выражение некоторой мифологии, бытующей в культурно-языковой среде [Степанов, 1995, с. 40]. Т. А. ван Дейк ввел это понимание дискурса в более широкий обиход, который постепенно развивается как новое направление научного исследования – критический анализ дискуса [Dijk van, 1993; Dijk van, 2004]. Однако наша работа не связана с пониманием дискурса в рамках этого направления.

Дискурсивный анализ второго типа формировался в первую очередь в британско-американской школе (термин “the British-American school” введен Гай Куком [Cook, 1989]) и включает основные современные теории и подходы к дискурсивному анализу [Brown, Yule, 1983; Chafe, 1992; Cook, 1989; De Beaugrande, Dressler, 1981; Dijk van, 1972; Gee, 1999; Halliday, Hasan, 1976; Harris, 1952; Hobbs, 1993; Hoey, 2001; Hoey et al., 2007; Johnstone, 2001; Renkema, 2004; Schiffrin, 1994; Schiffrin, Tannen, Hamilton, 2001; Sinclair, Coulthard, 2013; Stubbs, 1983].

Несмотря на единую ориентацию на «внеязыковые факторы», ученые продолжают разрабатывать различные подходы к дискурсу. Согласно Р. Стаббсу, анализ дискурса – это лингвистическое изучение «естественно возникающего связного устного или письменного дискурса», объектом которого является языковая единица, превосходящая предложение или фразу, например, устный разговор или письменный текст, и в котором акцент делается на анализе использования языка в социальных контекстах [Stubbs, 1983].

Дж. Синклер и М. Култхард считают, что дискурсивный анализ – это подраздел грамматического анализа языка и что на дискурсивном уровне сле-

дует изучать когезию и когерентность [Sinclair, Coulthard, 2013]. По мнению М. А. Халлидея, цель дискурсивного анализа состоит в том, чтобы выявить, как люди понимают речи друг друга [Halliday, 1974].

В центре внимания британо-американской школы находятся в основном когезия (англ. cohesion) и когерентность (англ. coherence), структура текста/дискурса (англ. textual/discourse structure), тип текста/дискурса (англ. textual/discourse type), грамматика текста/дискурса (англ. text/discourse grammar), теория схем (англ. Schema theory), теория жанров (англ. Genre theory), анализ бытового диалога (англ. Conversation analysis), теория речевых актов (англ. Speech act theory), интерактивная социолингвистика (англ. interactional sociolinguistics), этнография коммуникации (англ. the ethnography of communication), прагматика (англ. pragmatics), вариативный анализ (англ. variation analysis) и др.

Несмотря на доминацию функционализма в большинстве исследований, это не мешает лингвистам проводить структурный анализ самого текста. Фактически, дискурсивный анализ британо-американской школы изначально опирался на структурное описание дискурса, а затем постепенно расширился до функциональных, когнитивных и других интерпретаций.

Со 2-й пол. 1970-х гг. интенсивно развивается структурный дискурсивный анализ, к числу которого относятся следующие теории: модель процессуального анализа Р. де Богранда и В. Дресслера [De Beaugrande, Dressler, 1981], теория когезии М. Халлидея [Halliday, Hasan, 1976], теории Хоббса [Hobbs, 1979; Hobbs, 1985], Теория риторических структур (Rhetorical Structure Theory) [Mann, Thompson, 1988], модель Пенсильванского трибанка (Penn Discourse Treebank) [Miltsakaki и др., 2004; Webber, Joshi, 1998], Теория интенциональной структуры (Intentional Structure Theory) [Grosz, Sidner, 1986; Grosz, Sidner, 1990], Теория тематической прогрессии Ф. Данеша (Thematic Progression Theory) [Daneš, 1974], Теория группы предложений китайского языка [吴为章, 田小琳, 2000], Теория сложных предложений китайского языка [张仕仁, 1994; 邢福义,

2001], Теория обобщенной тематической структуры [宋柔, 2013; 蒋玉茹, 宋柔, 2012; 宋柔, 2022], Теория структуры связанных клауз [冯文贺 et al., 2020; Lyu, Feng, 2023] и др.

Перечисленные теории предоставляют разные подходы к дискурсивному анализу, описывая структурную информацию дискурсивного уровня с различных точек зрения: риторических отношений, тематической прогрессии, референции, функции предложения, развития события. Некоторые формальные теории дискурсивного анализа абстрактно изображают дискурс как совокупность линейных последовательностей или иерархических деревьев, которые могут служить основой компьютерного анализа дискурса.

Дискурсивный анализ третьего типа применяется в компьютерной лингвистике и обработке естественного языка. Здесь под дискурсивным анализом понимается процесс преобразования языка из поверхностной линейной последовательности единиц дискурсивного уровня в глубокое структурированное иерархическое представление, которое отражает взимосвязи между дискурсивными частями через их структуру, тем самым интегрируя внутренние и внешние знания для раскрытия процесса понимания и генерации дискурса. Основная задача такого анализа заключается в выявлении дискурсивной структуры, определении отношений между составляющими ее единицами и объяснении формирования и понимания дискурса как целой единицы с точки зрения анализа структуры [孔芳, 王红玲, 周国栋, 2019; Ананьева, Кобозева, 2016]. Для отличия от второго понимания, распространенного в современной лингвистике, в данной работе этот подход предлагается назвать «анализом дискурсивной структуры».

Анализ дискурсивной структуры можно рассматривать как расширение структурных подходов на дискурсивном уровне анализа языка. По словам А. А. Кибрика, дискурсивный анализ уже стал «уровневым разделом лингвистики, занимающимся языковыми единицами максимального объема» [Кибрик, 2019, с. 127]. В целом, анализ дискурсивной структуры отвечает потребностям

современной компьютерной лингвистики и обработки текстов на естественном языке, охватывая языковые единицы практически всех уровней – от фонетики до дискурса. Здесь под дискурсом понимается языковая единица, состоящая из различных синтаксических сущностей, ключом к анализу которой являются изучение связей между этими сущностями и их иерархии, включая поверхностные и глубинные семантические отношения.

Дискурсивный анализ в области компьютерной лингвистики в значительной степени опирается на семь текстовых признаков, предложенных Р. де Богрaндом и В. Дресслером [De Beaugrande, Dressler, 1981], а именно когезию, когерентность, интенциональность, адресованность, информативность, ситуативность, (типологическую) интертекстуальность. Из них исследования когезии (cohesion) и когерентности (coherence) внесли наибольший вклад в развитие компьютерной обработки дискурсивной структуры – см.: [孔芳, 王红玲, 周国栋, 2019, p. 2055].

Когезия обеспечивает внутреннюю лексико-грамматическую связность текста, то есть связь его элементов, при которой интерпретация одних элементов зависит от других и позволяет адресанту реализовать свою коммуникативную цель с наибольшей точностью и ясностью; а когерентность организует части дискурса таким образом, что авторский замысел становится понятным читателю, то есть реализуется то, что в прагматике называется уместностью [цит. по: Величко, 2016, с. 40]. Когезия включает лингвистические средства (грамматические, лексические, фонетические), благодаря которым предложения в тексте соединены в более крупные единицы на структурном уровне.

На практике дискурсивный анализ прежде всего связан с созданием дискурсивных корпусов, которое началось с конца 1980-х гг. На основе вышеизложенных формальных теорий дискурсивной структуры было завершено много репрезентативных корпусных проектов, включая корпус теории риторических структур (RST corpus) [Carlson, Marcu, Okurovsky, 2001; Marcu, 1996; Marcu, 2000], PDTB [Miltsakaki et al., 2004; Prasad et al., 2008; Prasad et al., 2019], дис-

курсивный банк деревьев зависимостей (discourse dependency treebanks) [Li et al., 2014; Lyu, Feng, 2023; Yang, Li, 2018; Yoshida et al., 2014] и др.

Современные корпусные технологии предоставляют графические средства и решения для сбора и хранения данных дискурсивного анализа, что способствует развитию дискурсивных корпусов. В настоящее время такие корпуса стали ключевым техническим обеспечением и языковым ресурсом в областях компьютерной лингвистики и обработки естественного языка.

Стоит также отметить, что многие исследования дискурсивной структуры наследуют методы синтаксического анализа. Например, создатели PDTB прямо указывают, что схема «коннектор – аргументы» основана на схеме «предикат – аргументы» [Miltsakaki et al., 2004; Zhou, Xue, 2012]. На данный момент формальные подходы к описанию синтаксической структуры предложения остаются одними из наиболее разработанных методов для комплексного анализа языковой структуры [Падучева, 1964, с. 99]. Таким образом, результаты синтаксического анализа структуры предложения также являются важной основой и ориентиром для дискурсивного анализа, проводимого в данной работе.

1.3. Параллельный корпус в контексте сопоставительной лингвистики

Параллельный корпус, или корпус параллельных текстов, представляет собой собрание текстов на одном языке вместе с их переводами на другой язык или другие языки. В настоящее время параллельные корпуса широко применяются в различных областях лингвистики, включая контрастивную лингвистику, типологию, теорию перевода, сравнительное литературоведение, культурологию, автоматическую обработку текста и др. [Добровольский, 2015; Сичинава, 2015; 卫乃兴, 2011; 许余龙, 2009; 陶源, 2015; 柏晓静 et al., 2002; 王克非, 2012; 甄凤超, 2004]. Таким образом, создание и использование параллельных корпусов представляется целесообразным и актуальным направлением прикладных лингвистических исследований.

Российские и китайские исследователи внесли значительный вклад в развитие проектов параллельных корпусов. За последние годы было собрано большое количество параллельных текстов и создано несколько высококачественных корпусов, среди которых: параллельный русско-китайский корпус в составе Национального корпуса русского языка [Семенов, Дурнева, Кузнецова, 2020]; полистилевой русско-китайский и китайско-русский параллельный корпус, создаваемый под руководством Цуй Вэя [崔卫, 李峰, 2014a; 崔卫, 张岚, 2014]; русско-китайский параллельный корпус научных текстов гуманитарной области, создателем которого является китайский ученый Тао Юань [Тао, Захаров, 2015; Тао, 2015]; русско-китайский переводческий корпус, разработанный китайским ученым Лю Мяо [刘淼, 邵青, 2016] и разделенный на три блока: подкорпус рассказов А. П. Чехова, китайско-русский подкорпус художественной литературы, подкорпус обучения русскому языку как иностранному; китайско-русский параллельный корпус официально-деловых текстов с дискурсивно-структурной разметкой, разработчиками которого являются М. Ю. Мухин и Ян И [Мухин, Ян, 2016] и др. – см. подробное об этом: [Чэнь, Кукушкина, 2018].

Ключевым этапом при создании параллельных корпусов является выравнивание (alignment) текстов, то есть установление соответствий между фрагментами текста оригинала и перевода. Благодаря выравниванию тексты разбиваются на содержательно эквивалентные фрагменты, что обеспечивает их хорошую сопоставимость.

Выравнивание параллельных текстов теоретически можно проводить по различным единицам: абзацам, предложениям, клаузам, фразам и даже отдельным словам. Как правило, чем меньше единица выравнивания, тем выше качество параллельного корпуса. В идеале применяется пословное выравнивание, однако, как отмечает М. Копотев, оно «часто оказывается почти невозможным по естественным причинам: наборы лексем, словоформ и устойчивых выражений в разных языках не совпадают» [Копотев, 2014, с. 97]. Наоборот,

выравнивание по более крупным единицам (например, абзацам) не дает достаточно полезной информации для научных лингвистических исследований.

На прктике тексты чаще всего выравнивают на уровне предложения. Однако и этот подход сталкивается с трудностями: членение текста на предложения и абзацы не всегда совпадает в оригинале и переводе. А. В. Зубов и И. И. Зубова выделяют шесть возможных типов соответствий между предложениями двух текстов: 1) одно исходное предложение переводится одним предложением; 2) два исходных предложения переводятся одним предложением; 3) одно исходное предложение переводится двумя предложениями; 4) два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и в тексте перевода не совпадают; 5) предложение исходного текста не переводится; 6) предложение в тексте перевода не имеет эквивалента в тексте оригинала [Зубов, Зубова, 2004]. Таким образом, если два языка существенно различаются, идеально согласованных единиц выравнивания на любом уровне может и не существовать. Выбор единиц выравнивания обычно определяется типом параллельного корпуса и целью его создания.

В параллельных дискурсивных корпусах (в данном случае параллельном корпусе с дискурсивной структурной разметкой) тексты двух языков часто выравниваются по клаузам [Мухин, Ян, 2016; Feng et al., 2018; Li et al., 2020]. В этих корпусах клаузы выполняют двойную функцию: они служат единицами сегментации текста и одновременно единицами для выравнивания. Клаузы рассматриваются как единицы дискурса, обладающие относительно независимой и стабильной синтаксической структурой; они реализуют самостоятельные суждения, а в письменных текстах обычно выделяются или разделяются специальными знаками препинания (подробнее см. п. 2.2.2).

Помимо выравнивания по клаузам, Фэн Вэньхэ и соавторы предлагают осуществлять выравнивание по иерархиям дискурсивной структуры и по дискурсивным отношениям [Feng et al., 2018; 冯文贺, 2013]. В созданном ими ки-

тайско-английском параллельном дискурсивном корпусе оригинальный китайский текст и английский перевод выравнивались в соответствии с иерархическими структурами и отношениями текста перевода. По словам Фэн Вэньхэ, основное предположение состоит в том, что «внутренняя иерархическая структура дискурса и структурные отношения в текстах оригинала и перевода должны точно совпадать. Это объясняется тем, что дискурсивная структура по сути является логико-семантической структурой. Переводной текст должен передавать не только смысл и синтаксический строй текстов оригинала, но и, по возможности, логико-семантические отношения – такие как причина – следствие, цель, сопоставление и др., существующие в исходном тексте, а также их структурную иерархию этих отношений» [冯文贺, 2013, p. 59].

Таким образом, дискурсивная структурная разметка фактически отражает понимание текста читателем (аннотатором), основанное на переводе, и по сути является интерпретацией переводчиком исходного текста. Мы следуем аналогичной схеме при разметке китайско-русского параллельного корпуса с дискурсивно-структурной разметкой.

Изучение параллельного корпуса естественным образом подходит для сопоставительного анализа языка. Во-первых, параллельные корпуса предоставляют языковой материал с переводческими соответствиями, что важно для контрастивных исследований [McEnery, Xiao, 2007]. В сопоставительной лингвистике основой анализа часто служит понятие «эквивалентность перевода» [Krzeszowski, 1984; James, 1980; Halliday, McIntosh, Strevens, 1964]. Как отмечают М. А. К. Хэллидей, А. Макинтош и П. Стревенс: «в результате переводческой деятельности получаются тексты, обладающие смысловой эквивалентностью, а самый простой способ понять эквивалентность – это наблюдать за переводческими отношениями между текстами» [Halliday, McIntosh, Strevens, 1964, p. 115].

На практике, как подчеркивают Цинь Хунву и Чжой Ся, в реальных переводах встречаются различные ошибки под влиянием таких факторов, как раз-

ница в способах языкового кодирования и индивидуальный стиль письма переводчика, поэтому не многие варианты перевода строго семантически эквивалентны [秦洪武, 周霞, 2024]. Следовательно, в параллельный корпус важно включить тексты оригинала и перевода, обладающие как можно более полной семантической эквивалентностью.

Процедура выравнивания, в которой исследователь участвует лично, открывает широкие возможности для самого сопоставительного лингвистического анализа. В ходе выравнивания исследователи наблюдают различия между двумя языками, а понимание этих различий постепенно углубляется при повторном анализе аналогичных языковых явлений. Интерпретация результатов выравнивания параллельного корпуса, особенно различий в свойствах единиц, которые поддаются или, наоборот, не поддаются выравниванию, и составляет основу сопоставительного лингвистического анализа на базе параллельных корпусов.

Параллельный корпус также обеспечивает возможности количественного анализа и служит базой данных для сопоставительных исследований. В последние годы благодаря совершенствованию статистических методов и инструментов корпусов они широко используются в различных видах количественных сопоставительных исследований [Добровольский, 2012; Добровольский, 2015; Добровольский, 2020; Маник, 2019; 卫乃兴, 2011; 许余龙, 2001; 秦洪武, 周霞, 2024].

В статье «Методологические вопросы количественных сопоставительных исследований» китайский лингвист Сюй Юйлун отмечает, что «количественные методы исследований сосредоточены на контроле несоответствий, чтобы сопоставительный анализ имел общую основу» [许余龙, 2001, с. 3]. Он опирается на четыре базовых метода контроля смешивающих переменных, предложенных У. Вирсмой в фундаментальном труде «Методы исследования в образовании» [Wiersma, 1999], а именно: 1) рандомизация выборки, 2) стабилизация

экспериментальных условий, 3) управление назависимыми переменными и 4) статистический контроль ковариант.

Сюй Юйлун убедительно демонстрирует, что эти методы принципиально применимы в количественных контрастивных лингвистических исследованиях [许余龙, 2001]. Использование параллельного корпуса, особенно с фильтрацией по типам разметки и выравниванию, позволяет эффективно контролировать несоответствия различных параметров при сопоставлении двух языков.

Таким образом, в настоящей работе предлагается провести сопоставительный лингвистический анализ на основе созданного параллельного дискурсивного корпуса, в частности, – сравнение сходств и различий в распределении количества и использовании дискурсивных средств в китайском и русском языках.

1.4. Официально-деловые тексты в политической коммуникации как источник данных для параллельного корпуса

Официально-деловой стиль понимается в функциональной стилистике как особая разновидность литературного языка, обслуживающая сферу права, власти, администрации, коммерции внутри- и межгосударственных отношений [Кожина, 2003]. Этот стиль привлекает внимание как российских, так и китайских лингвистов, которые подчеркивают его роль в политической практике, переводе и межгосударственном общении (А. Н. Кожин, Л. Г. Барлас, М. Н. Кожина, Г. Я. Солганик и Т. С. Дроняева, В. Д. Черняк, Д. Э. Розенталь, Н. А. Купина, Т. В. Матвеева, Ху Юйшю (胡裕树), Ни Баоюань (倪宝元), Чжан Хуэйсэн (张会森), Чжао Цзе (赵洁) и др.).

Современная лингвистика отмечает значительное разнообразие направлений и жанров официально-делового стиля. Одна из наиболее исчерпывающих лингвистических классификаций представлена в «Стилистическом энциклопедическом словаре русского языка» под ред. М. Н. Кожиной, где официально-деловой стиль подразделяется на дипломатический, законодательный, юриди-

ческий и административный подстили [Кожина, 2003]. К общим стилистическим чертам официально-делового стиля относятся «предписующе-долженствующий характер, точность, не допускающая инотолкования, стандартизированность, неличностность, официальность и безэмоциональность изложения» [Там же, с.164].

Особое внимание в лингвистике уделяется дипломатическим текстам, которые являются важным средством межгосударственного общения. Благодаря тому, что они часто имеют эквиваленты на двух и более языках, такие тексты становятся удачным материалом для включения в параллельные корпуса. Среди известных проектов – *Europarl parallel corpus*[1], основанный на официальных текстах Европарламента, включающий 21 язык [Koehn, 2005]. Также стоит упомянуть *United Nations Parallel Corpus*[2], который охватывает тексты ООН на множестве языков, включая шесть рабочих языков [Ziemski, Junczys-Dowmunt, Pouliquen, 2016]. Кроме того, они обладают значительным политическим, правовым и институциональным значением, формируют международный имидж государства и способствуют поддержанию стабильности межгосударственных отношений, что обусловливает их особое место в системе международных отношений. В условиях глобализации и растущей потребности в унификации международных документов, изучение языковых особенностей политических текстов и создание на их основе лингвистически аннотированного корпуса становятся особенно актуальными.

Сами дипломатические документы могут быть разделены на несколько подсистем в зависимости от выполняемых функций и особенностей оформления. Китайский русист У Айхуа различает три типа дипломатических документов в межгосударственных отношениях РФ и КНР: во-первых, это дипломатическая переписка, используемая для решения повседневных международных

[1] Об этом см.: URL: https://www.statmt.org/europarl/ (дата обращения: 10.10.2025).

[2] Об этом см.: URL: https://www.un.org/dgacm/en/content/uncorpus(дата обращения: 10.10.2025).

дел, обычно представляемая в виде официальных писем (таких как ноты, меморандумы, письма и верительные грамоты); во-вторых, дипломатические документы, опубликованные для международного сообщества и официально отражающие позицию обоих правительств по важнейшим международным вопросам, такие как совместные декларации, совместные заявления и коммюнике; в-третьих, дипломатические беседы, включая официальные и неофициальные беседы, которые обычно проводятся в рамках двусторонних или многосторонних встреч и направлены на развитие общения и сотрудничества между государствами [武援华, 1998, p. 80]. Ю. М. Кукарина также выделяет четыре жанровых подсистемы: документы, подтверждающие полномочия отдельных должностных лиц дипломатического корпуса (верительные грамоты, отзывные грамоты, консульский патент и др.); международные договоры (договор, конвенция, соглашение, протокол, акт, пакт, статут, хартия, устав, коммюнике и др.); дипломатическая переписка (личная нота, вербальная нота и др.); внутриведомственные дипломатические документы (годовой отчет посольства, шифротелеграмма и др.) [Кукарина, 2020].

Разные виды дипломатических документов значительно различаются по формату. Как отмечает У Айхуа, дипломатическая переписка подчиняется строго установленным требованиям к оформлению, тогда как для дипломатических документов таких жестких стандартов нет [武援华, 1998, p. 79]. Многие из этих документов также отличаются закрытостью и конфиденциальностью, что обусловлено характером дипломатической информации [Кукарина, 2020, с. 265].

Дипломатические документы, особенно такие, как совместные заявления и декларации, обычно публикуются на официальных сайтах правительств обеих стран для общего доступа. Несмотря на отсутствие ярко выраженной формальности и частичное включение в язык дипломатических документов черт публицистического стиля (что проявляется в стремлении к созданию преднамеренной неопределенности и политкорректности) (см. об этом: [武援华, 1998; Абдулсалам,

2023]), такие документы активно используются в дипломатической деятельности и в основном сохраняют черты официально-делового стиля. Они также имеют (хоть и ограниченно) предписывающе-долженствующий характер, характеризуются точностью и официальностью.

В настоящее время лингвистические исследования дипломатического языка в основном фокусируются на лексике, грамматике и стилистике, акцентируя внимание на анализе специфических языковых единиц и структурных особенностей, при этом исследования часто носят сопоставительный характер. Г. И. Исина и А. С. Мустафина рассматривают лексические и грамматические особенности языка дипломатического общения, включая использование официальных протокольных формул, комплиментарной лексики и экспрессивно-эмоциональных слов, а также особенностей употребления категории долженствования и инфинитивов в дипломатической коммуникации [Исина, Мустафина, 2016].

Вань Яньсинь провела глубокий сравнительный анализ лексических особенностей русского и китайского официально-делового стиля, отметив сходства русского и китайского официально-делового стиля в использовании нормативных и терминологических единиц, отсутствии разговорных выражений, стремлении к точности и однозначности, а также различия в употреблении и национально-специфических терминах [Вань, 2023].

Ма Лимин проанализировала использование предлогов в дипломатических подстилях двух языков, подчеркнув их роль в точности передачи информации и предотвращении двусмысленности [Ма, 2024а]. В другой работе она же сравнивает структуру простых предложений в китайских и русских дипломатических документах. Исследование Ма Лимин показало, что обе языковые системы используют сложные структуры для компактности текста, особенно в предложениях с предложными конструкциями и однородными членами. В русском языке часто встречаются причастные и деепричастные обороты, а в китайских документах преобладают сложные детерминативы, модификато-

ры и объемлющие предложения, что отражает различия в языковых системах [Ma, 2024b].

На данный момент системный анализ дискурсивной структуры дипломатических текстов еще не проводился. Также отсутствуют сравнительные исследования, использующие эти тексты в качестве параллельного корпуса для разметки дискурсивной структуры или семантического анализа. В то же время дипломатические тексты являются высоко структурированным и стандартизированным двуязычным материалом, что придает им особую ценность как источнику корпусных данных. Формализованная лексика, устойчивые синтаксические модели и предсказуемые дискурсивные структуры позволяют более точно сопоставлять языковые средства китайского и русского языков. Следовательно, привлекаемые тексты являются не только важным объектом сопоставительного исследования в контексте официально-делового стиля, но и ценным ресурсом для создания параллельного дискурсивного корпуса, обеспечивающего воспроизводимость и сопоставимость результатов анализа.

Нужно отметить, что в отношении материала настоящего исследования используются два термина – «тексты» и «документы». С точки зрения общеязыковых значений отношения между ними можно считать родо-видовыми, т. к. документ – это речевое произведение (то есть текст) официально-деловой коммуникации. Используя здесь эти термины фактически как смысловые эквиваленты, мы акцентируем внимание или на исходном официально вербализованном акте (документ), или на лингвистической единице, представляющей собой последовательность предложений и сегментированной на элементарные дискурсивные единицы и текстоформы (текст).

Выводы по первой главе

В главе были рассмотрены проблемы и задачи, связанные с созданием параллельного дискурсивного корпуса. Актуальность создания и использования лингвистических корпусов обусловлена тем, что корпус не только предоставля-

ет объективные, репрезентативные и масштабные данные о языке, но и формирует систему разметки, которая может использоваться для формального моделирования языка. Разметка выступает в роли метаязыка для описания языка и употребления языковых единиц в речи, ее разработка предполагает создание теоретических оснований для формального описания и моделирования языка и открывает новые возможности для изучения его структуры. В данной работе дискурсивный анализ соотнесен с созданием дискурсивного корпуса, разработкой схемы разметки дискурса и применением ее к китайским и русским текстам.

Анализ дискурсивной структуры понимается как подход, ориентированный на структурное исследование организации дискурса и связанный с задачами компьютерной лингвистики и обработки естественного языка. Этот тип дискурсивного анализа фокусируется на построении и моделировании дискурса как структурированного объекта и на преобразовании его поверхностной, линейной формы дискурса в иерархическое и структурированное представление, отражающее отношения между элементами дискурса. Многие дискурсивные структурные исследования в плане аналитических методов опираются на синтаксические структурные исследования. Создание дискурсивного корпуса связано именно с дискурсивным анализом именно этого типа.

Параллельный корпус служит базой данных и инструментом для проведения сопоставительных лингвистических исследований. Он содержит исходные тексты и их переводы на другой язык и характеризуется выравниванием двух текстов по соответствующим элементам. Параллельный корпус позволяет проводить количественный и качественный анализ: с помощью сопоставления соответствий и несоответствий можно выявлять особенности синтаксиса, морфологии, лексики и в целом дискурса в разных языковых системах. Процедура разметки при выравнивании предполагает сопоставление фрагментов параллельных тестов, что позволяет исследователю наблюдать сходства и различия между двумя языками и постепенно уточнять понимание этих различий при повторении аналогичных языковых явлений. Сопоставительный лингвистический

анализ на основе параллельного корпуса можно понимать как интерпретацию результатов анализа единиц, которые поддаются или не поддаются корпусному выравниванию.

Настоящее исследование построено на материале дипломатических текстов, таких как совместные декларации и заявления, которые играют ключевую роль в международных отношениях и часто имеют эквиваленты на нескольких языках. Двуязычные тексты, создаваемые в рамках межгосударственных соглашений КНР и РФ, служат важным инструментом дипломатии и межгосударственного общения. Формализованные черты официально-делового стиля делают такие тексты ценным ресурсом для создания параллельных корпусных данных, а также обеспечивают воспроизводимость и сопоставимость корпусного анализа.

ГЛАВА 2. ПАРАЛЛЕЛЬНЫЙ ДИСКУРСИВНЫЙ КОРПУС: МОДЕЛЬ ПОСТРОЕНИЯ, ТЕРМИНОЛОГИЯ, ЕДИНИЦЫ ОПИСАНИЯ

Цель данной главы – выработать структурные и концептуальные основы создания дискурсивного корпуса, обосновать эффективность представления дискурсивной структуры в виде деревьев зависимостей, сформировать необходимый терминологический аппарат и обсудить единицы корпусной разметки. В параграфе 2.1 подробно рассматриваются два способа моделирования синтаксических и дискурсивных структур – представление их в виде составляющих и в виде зависимостей. Параграф 2.2 посвящен формированию терминологического аппарата для представления дискурсивных структур.

2.1. Способы моделирования дискурсивных структур

2.1.1. Способы структурно-синтаксического моделирования: зависимости и вложения (набор составляющих)

Для того чтобы выработать принципы моделирования дискурсивной структуры, нужно сначала рассмотреть исходные способы структурно-синтаксического моделирования. В лингвистических исследованиях в качестве оформленной единицы структурного анализа принято рассматривать предложение. Считают, что способы моделирования предложения являются ориентиром для описания единиц других уровней, в том числе и дискурса [Белошапкова, 1977; Син Фуи, 2020; Русская грамматика: синтаксис, 1980], что обеспечивает значимую поддержку и настоящего исследования.

В современной лингвистике используются разные способы представления структуры предложения, из которых наиболее употребительны два – описание структуры предложения через иерархию непосредственных составляющих и через отношения подчинения (зависимости) слов [Гладкий, 1973, с. 282; Падучева, 1964]. Для демонстрации двух способов анализа структуры приведено наше предложение *«Твои книги стоят на полке».* На рис. 2.1 показаны два ва-

рианта представления структуры данного предложения: (А) дерево составляющих и (Б) дерево синтаксических зависимостей.

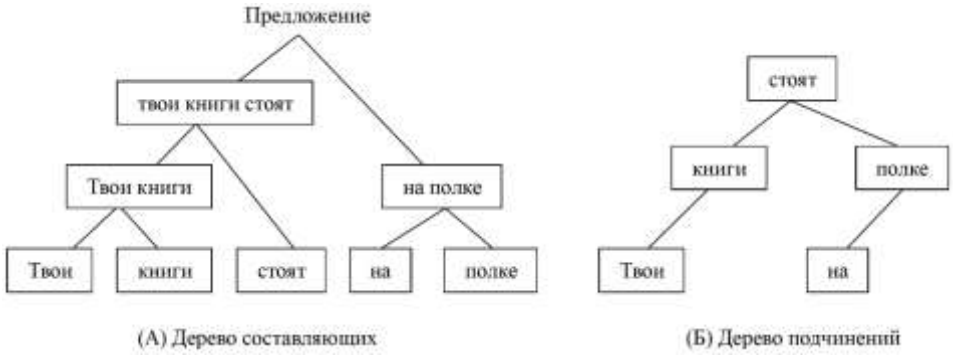


Рисунок 2.1. Дерево составляющих и дерево подчинений.

Примечание: древовидный граф составляется из набора ребер и связанных узлов, причем на разных уровнях узлы получают разные названия: верхний узел дерева называется корнем; узел, не имеющий дочерних элементов, – терминальным или листовым узлом; узел, имеющий дочерние элементы, – внутренним узлом.

Несмотря на то, что два способа представления синтаксической структуры имеют ряд сходств, включая иерархичность, они в сущности отражают принципиально разные способы организации структуры. Внутренние узлы дерева составляющих образованы группировками слов – «*твои книги*», «*на полке*» и «*Твои книги стоят*»; тогда как в дереве подчинений все терминальные единицы (слова) участвуют в организации структуры предложения напрямую. Дерево составляющих предполагает объединение языковых единиц в группировки, а дерево подчинений – построение непосредственных отношений между двумя отдельными словами. Различие между этими подходами становится еще более очевидным на примере, представленном на рис. 2.2.

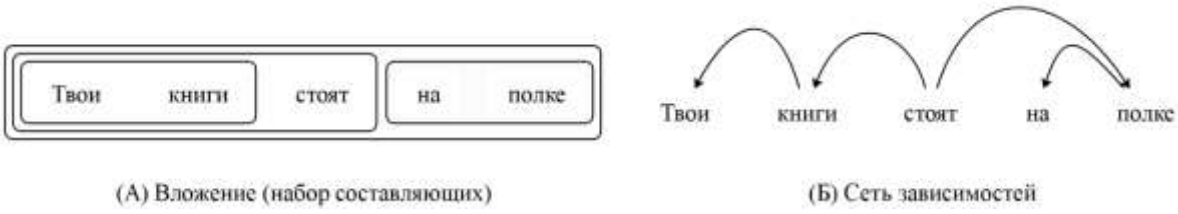


Рисунок 2.2. Комбинация составляющих и сеть зависимостей.[1]

На рис. 2.2 (А) представлен вариант структуры составляющих, где слова объединяются в группировки, которые, в свою очередь – в более крупные группировки вплоть до целого предложения. Рис. 2.2 (Б) отображает структуру подчинений, в которой зависимости (связи) между словами представлены ребрами, а направление ребра (стрелка) указывает организацию отношений, придавая линейной структуре иерархичность. Эти две модели демонстрируют разные подход: один – к анализу вложенных элементов, другой – к анализу связей между элементами.

Эти структуры можно интерпретировать с точки зрения психологического анализа человеческого мышления, а именно через понятия «группирования» и «сети» соответственно. Концепция «группирования» (англ. chunking) была впервые предложена американским психолингвистом Дж. Миллером [Miller, 1956]. С помощью стратегии группирования отдельные дискретные фрагменты вводимой информации объединяются в более крупные группы, тем самым уменьшая число смысловых агрегатов [Лук, 1978]. Многие современные лингвистические исследования развивались под влиянием идеи группирования, включая метод непосредственных составляющих [Bloomfield, 1933; Bloomfield, 1983], порождающую грамматику Н. Хомского [Chomsky, 1957] и грамматику с фразовой структурой, движимую управляющим словом [Müller et al., 2021].

[1] На всех последующих рисунках приняты следующие соглашения: структура зависимостей отображается ориентированными ребрами (форма которых – дуга или прямая – зависит от компоновки узлов), а структура вложения – иерархическими деревьями.

Однако, подобно как стратегия группирования формально допускает объединения только соседних элементов, анализ в рамках вложения по умолчанию предполагает, что вложения формируются только соседними языковыми единицами[1].

Дерево зависимостей, строящееся на основе бинарных отношений, представляет собой более гибкий и динамический способ структурирования языкового объекта. Идея зависимостей оказала влияние на формирование многих известных грамматических теорий. Если понимать зависимости как основу всех методов, основанных на бинарных асимметричных связях, то можно вспомнить такие концепции, как теория «смысл⇔текст» [Мельчук, 1974; Мельчук, 1995], лексический подход к грамматике [Hudson, 1984], теория функционального генеративного описания [Sgall, Hajicová, Panevová, 1986] для построения Пражского синтаксического банка (Prague Dependency Treebank) [Böhmová et al., 2003] и др. Сегодня, когда мы говорим об идее зависимостей, речь идет не столько о конкретной теории, сколько о целом подходе в лингвистике.

Структура зависимостей с прагматической точки зрения является более эффективным способом описания языкового объекта, что обусловлено тем, что она состоит лишь из узлов (элементы языка) [Мельчук, 1964, с. 13]. Простые бинарные связи способствуют построению сетевых структур независимо от линейного порядка языковых единиц и, следовательно, обеспечивают создание единых теоретических рамок для анализа различных языков и языковых явлений разного уровня. Например, в настоящее время многие языковые модели создаются именно на основе структуры зависимостей. К ним относятся WordNet и ворднет-подобные тезаурусы, то есть лексические сети [Fellbaum, 1998; Азарова, Митрофанова, Синопальникова, 2003; Azarova et al., 2002; 张俐 et al., 2003;

[1] Комбинация между двумя соседними единицами X-Y, а также между двумя соседними группами (XY)-Z。

董振东, 董强, 郝长伶, 2007], синтаксические сети [Lai, Huang, 1999; Hajič et al., 2004; Kromann, Mikkelsen, Lynge, 2003; Hays, 1977; McDonald et al., 2013; Lu, Liu, 2020; Mel'čuk, 1988; 刘海涛, 1991; 刘挺, 马金山, 2009; 陈芯莹, 刘海涛, 2011; 李正华, 2014], фонетические сети [Chan, Vitevitch, 2009; Eguíluz et al., 2005; Mukherjee et al., 2008; Samuel, Strogatz, Vitevitch, 2010], семантические сети [Borge-Holthoefer, Arenas, 2010; Hills et al., 2009; Liu, 2009; Steyvers, Tenenbaum, 2005], а также дискурсивные сети [Danlos, 2004; Li et al., 2014a; Lyu, Feng, 2023; Wolf, Gibson, 2005; Yang, Li, 2018a; Yoshida et al., 2014]. Эти исследования демонстрируют эффективность использования структуры зависимостей при анализе и решении теоретических и прикладных задач лингвистики. Таким образом, структура зависимостей представляет собой универсальный способ моделирования разных языковых явлений.

2.1.2. Способы структурного моделирования дискурса

Исходя из концепции структур составляющих и зависимостей, существующие исследования дискурсивных структур можно разделить на три типа. Первый из них – дискурсивная структура составляющих, представленная теорией риторических структур (далее – ТРС) [Taboada, Mann, 2006; Mann, Thompson, 1988a] и корпусами, построенными в рамках ТРС [Marcu, 1996; Carlson, Marcu, Okurowski, 2002; Carlson, Marcu, Okurowski, 2003; 陈莉萍, 2008; 李艳翠, 周国栋, 2015; Литвиненко, 2001; Кибрик, Подлесская, 2009; Pisarevskaya et al., 2017]. Второй тип – дискурсивные структуры зависимостей, в том числе схема «дискурсивный коннектор и аргументы» [Forbes-Riley, Webber, Joshi, 2006; Miltsakaki et al., 2004a; Miltsakaki et al., 2008a; Webber, 2004] для создания Пенсильванского дискурсивного древовидного банка (Penn Discourse Treebank, далее – PDTB) [PDTB-Group, 2008a; Prasad et al., 2004; Zhou, Xue, 2012 и др.], структура связанных клауз [冯文贺 et al., 2020; Lyu, Feng, 2023] и др. Третий тип – это смешанные типы структурного моделирования дискурса, представ-

ленные моделью направленного ациклического графа (Directed acyclic graph) [Danlos, 2004; Danlos, 2005; Danlos, 2008] и цепочечного графа [Wolf, Gibson, 2005; Wolf, Gibson, 2006]. Вышеупомянутые исследования предложили различные способы представления дискурсивных структур, что отразилось на создании дискурсивных корпусов (дискурсивных трибанков, англ. – discourse treebanks).

Рассмотрим далее различные модели дискурсивных структур, чтобы обосновать основы создания нашего корпуса. Для того чтобы пояснить разницу между ними, используем для примера один и тот же дискурсивный фрагмент (2.1).

(2.1)

C1. Max experienced a lovely evening last night.

C2. He had a fantastic meal.

C3. He ate salmon.

C4. He devoured lots of cheese.

C5. He won a dancing competition.

Примечание: следует отметить, что в разных теориях по-разному определяются единица дискурса, дискурсивное отношение и дискурсивный коннектор. В данном параграфе диссертации нам нужно представить разные способы организации дискурсивной структуры, поэтому используется один и тот же метод, предложенный Л. Данлосом [Danlos, 2005] для деления единиц дискурса и определения дискурсивных отношений. Во всех примерах данной работы элементарные единицы расположены в ряду с последовательной нумерацией.

2.1.2.1. Дискурсивная структура вложения: теория риторических структур

ТРС была разработана в 1980-е гг. американскими лингвистами У. Манном и С. Томпсоном [Mann, Thompson, 1987; Mann, Thompson, 1988b;

Mann, Matthiessen, Thompson, 1989a]. Первоочередная цель теории заключается в «выявлении основания автоматического порождения и функционирования текста, достижения связного характера текстового произведения» [Mann, Thompson, 1988a; Ковальчук, Володина, 2016, с. 107]. Таким образом, с момента возникновения ТРС авторы стремились разработать более формальную теоретическую схему модели дискурса.

ТРС предоставляет единую четкую аналитическую схему для анализа структуры дискурса. По модели ТРС дискурсивная структура сводится, с одной стороны, к множеству образующих его элементов – дискурсивных единиц, а с другой, к множеству связывающих эти элементы «риторических отношений». Основное положение теории заключается в том, что любая дискурсивная единица связана хотя бы с одной другой единицей данного дискурса посредством некоторых риторических отношений [Кибрик, Плунгян, 2002, с. 309].

Деление на дискурсивные единицы считается необходимым первым шагом моделирования. Дискурсивные единицы могут быть разными по размеру: от элементарных дискурсивных единиц (далее – ЭДЕ) до дискурсивных фрагментов. По У. Манну и С. Томпсону, типичная ЭДЕ определяется как клауза. С помощью риторических отношений ЭДЕ объединяются в дискурсивные фрагменты, которые в свою очередь могут объединяться в более крупные, и так до уровня целого текста. Риторические отношения имеют логико-семантический характер и выполняют текстообразующую функцию (подробнее см. в подпараграфе 2.2.2). Дискурсивная структура по ТРС представлена в виде иерархического дерева (см. рис. 2.3).

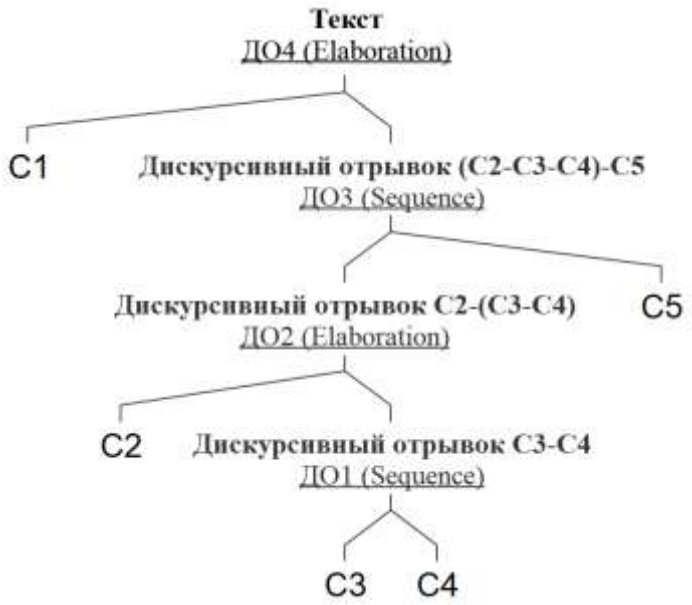


Рисунок 2.3. Иерархическая структура дискурсивного фрагмента (2.1) по TPC.

Как видно из рис. 2.3, дискурсивная риторическая структура представляет собой иерархическую модель вложения, основанную на группировании ЭДЕ. В древовидной модели TPC внешними узлами являются ЭДЕ, а внутренними – дискурсивные фрагменты; отношения устанавливаются между соседними единицами.

В настоящее время TPC развивается как описательный подход к анализу дискурсивной структуры, который получает теоретическую и прикладную детализацию. С основой на TPC был создан ряд дискурсивных корпусов разных языков [Peng, Liu, Zeldes, 2022; Кибрик, Подлеская, 2009; Carlson, Marcu, Okurowski, 2003; 陈莉萍, 2007; Ананьева, Кобозева, 2016a; Cao, Cunha, Iruskieta, 2018; Chistova et al., 2021; Marcu, 1996; Pisarevskaya et al., 2017; Stede, 2004; Stede, Neumann, 2014; Toldova et al., 2017; 乐明, 冯志伟, 2006; 张培佳, 冯德正, 2018]. Кроме того, широкое применение эти модели получили в таких сферах, как моделирование и генерация текстов [Hovy, 1993a; Nicholas, 1994 и др.], автоматическое реферирование [Батура, Бакиева, 2018; Marcu, 1997b; Marcu, 1997c], автоматическая обработка текстов [Бакиева, 2017; Бакиева, Батура, 2017a; Бакиева, Батура, 2017b], моделирование дискурса [Сусов, 2006] и др.

2.1.2.2. Локальная дискурсивная структура зависимостей

В Пенсильванском дискурсивном трибанке (Penn Discourse Treebank, PDTB) [Miltsakaki et al., 2004a; PDTB-Group, 2008a] предложен другой вариант решения для описания дискурсивных единиц и их отношений – модель «дискурсивный коннектор (далее – ДК) – аргументы» [Forbes-Riley, Webber, Joshi, 2006; Miltsakaki et al., 2004a; Miltsakaki et al., 2008a]. По своей концепции модель «ДК – аргументы» аналогична схеме «предиката – аргументов» для анализа предложения [Kingsbury, Palmer, 2002; Дудчук, Подобряев, 2004, с. 13–26]. На уровне дискурса «предикат» считается как ДК, который рассматривается как «предикативное» слово дискурсивного уровня и главный показатель семантических отношений дискурса [Miltsakaki et al., 2004a; Miltsakaki et al., 2004b]; а «аргументы» – элементарные единицы, распределенные по обе стороны от коннектора (см. рис. 2.4).

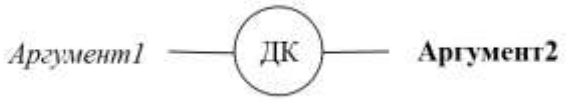


Рисунок 2.4. Локальная структура по модели PDTB.[1]

В целом создатели PDTB не строили древовидную структуру, а лишь определяли локальные бинарные отношения между двумя аргументами. По их замечанию, модель «предикат – аргументы» относится к «независимому» подходу к структурному анализу дискурса [Prasad et al., 2008b, p. 2961], поскольку они изучают локальную дискурсивную структуру в терминах семантики ДК, а не модели целого дискурса [Scheffler, Stede, 2016, p. 242]. Так как центр внимания модели «ДК – аргументы» сосредоточен на семантических исследованиях ДК, общий анализ дискурсивной структуры начинается с распознавания

[1]В модели «предикат – аргументы» дискурсивная структура не представлена графически. Приведенный здесь бинарный дискретный граф с узлом коннектора построен на основе результатов нашего анализа.

дискурсивных коннекторов и их аргументов, при этом нет необходимости предварительного деления на ЭДЕ. Анализ дискурсивного фрагмента (2.1) в рамках концепции «ДК – аргументы» приведен в (2.2).

(2.2)

C1. *Max experienced a lovely evening last night*. [Implicit = AltLex] **C2. He had a fantastic meal.** [Relation = Elaboration]

C2. *He had a fantastic meal*. [Implicit = AltLex] **C3. He ate salmon.** [Relation = Elaboration]

C3. *He ate salmon*. [Implicit = AltLex] **C4. He devoured lots of cheese.** [Relation = Sequence]

C4. *He devoured lots of cheese*. [NoRel] **C5. He won a dancing competition.**

Примечание: в примере мы соблюдаем принцип разметки ДК и аргументов в корпусе PDTB [PDTB-Group, 2008a]. ДК подчеркнуты, аргумент1 выделен курсивом, а аргумент2 – жирным шрифтом.

Как видно из (2.2), PDTB описывает локальную бинарную структуру дискурса, состоящую из двух соседних аргументов. ДК – это и основа для формирования локальной структуры дискуса, и объект описания дискурсивной структуры.

Модель PDTB получила широкое распространение в компьютерной лингвистике, был создан ряд дискурсивных корпусов: Надкорпусные базы данных [Зализняк и др., 2015], Кросслингвистическая база данных для аннотирования логико-семантических отношений в тексте [Дурново, Зацман, Лощилова, 2016], HIT-CDTB [Zhang, Qin, Liu, 2014], Chinese Discourse Corpus with Connective-driven Dependency Tree Structure [Li et al., 2014c], Chinese Discourse Treebank [Zhou, Xue, 2015a] и мн. др.

2.1.2.3. Глобальная дискурсивная структура зависимостей

В противоположность локальной структуре, глобальная дискурсивная структура представляет собой членение на крупные составляющие (в первую очередь абзацы в письменном тексте) – см. например: [Дейк ван, 1989; Кибрик, 2003; Падучева, 1965]. Репрезентативной моделью является структура связанных клауз, предложенная китайскими лингвистами Фэн Вэньхэ и др. [冯文贺 et al., 2020; Lyu, Feng, 2023]. Структура связанных клауз выявляет структурно-семантические отношения между двумя клаузами в одном абзаце. Формально структура связанных клауз соотносится с представлением структуры зависимостей предложения (см. рис. 2.5). В структуре связанных клауз терминальный узел указывает на клаузу, а ребро – на зависимость между двумя клаузами; стрелка ребра отражает направления зависимости. Итак, структура связанных клауз по типу относится к структуре зависимостей.

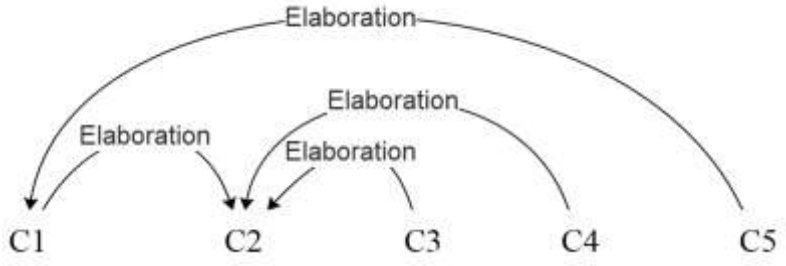


Рисунок 2.5. Структура связанных клауз.

Как показано на рис. 2.5, в данной структуре семантические отношения непосредственно связывают две клаузы. Вследствие отсутствия понятия «дискурсивный фрагмент» в рамках модели зависимости применяются особые решения. Например, отношение между ЭДЕ C2 со вложенной частью дискурсивного фрагмента {C3, C4}, занимающей внутренние узлы в TPC, представлено здесь как два непосредственных отношения между клаузами C2-C3 и между клаузами C2-C4.

2.1.2.4. Смешанные типы структурного моделирования дискурса

Наряду со структурами вложения и зависимости выделяются смешанные структуры, включающие в себя вложение и зависимость одновременно. Они представлены моделью направленного ациклического графа [Danlos, 2004; Danlos, 2005; Danlos, 2008] и моделью цепочечного графа [Wolf, Gibson, 2005; Wolf, Gibson, 2006]. На рис. 2.6 показан структурный анализ дискурсивного фрагмента (2.1) в рамках направленного ациклического графа (А) и цепочечного графа (Б).

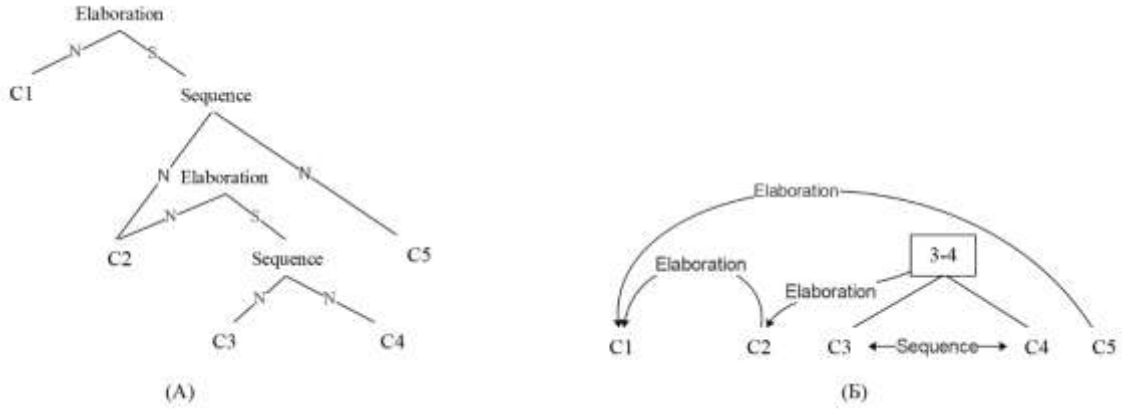


Рисунок 2.6. Смежные структуры дискурса: структура DAGs и цепочечная структура.

На рис. 2.6 (А) видно, что структурный анализ дискурса по Л. Данлосу [Danlos, 2005] сохранил иерархическую форму изображения дискурсивной структуры. В этой иерархии аналогичным образом также выделены асимметричные отношения между ядерными (N) и сателлитными (S) составляющими, описанными в TPC. Однако модель направленного ациклического графа представляет собой более гибкую структуру, поскольку она позволяет отображать отношения между прерывистыми ЭДЕ. Например, на *(A)* видно отношение между прерывистыми клаузами *C2* и *C5*. Множественные родительские узлы также входят в структуры составляющих. Например, клауза *C2* одновременно

участвует в последовательности клауз *C2* и *C5* (*Sequence*) и отношении детализации (*Elaboration*).

В цепочечном графе Вольфа и Гибсона *(Б)* [Wolf, Gibson, 2005] соединены разные способы представления дискурсивной структуры. В этой схеме ЭДЕ расположены не иерархически, а в линейном порядке. ЭДЕ соединены дугами (и прямыми линиями), соответствующими дискурсивным отношениям. Следует отметить, что в цепочечном графе Вольфа и Гибсона допускается объединение соседних ЭДЕ в группу [Wolf, Gibson, 2005], которую можно объединить с другой ЭДЕ или группой и строить бинарные отношения. Например, на *(Б)* ЭДЕ С2 и группа {С3-С4} формируют отношение детализации (Elaboration).

Обсужденные выше модели структуры дискурса формируют структурные основы создания нашего Корпуса. В частности, TPC предоставляет «стандартизированную» концептуальную схему для анализа дискурсивных структур – ЭДЕ и дискурсивных отношений. Модель «ДК – аргументы» фокусируется на дискурсивных служебных словах и их значениях, формируя потенциальную дискретную структуру дискурса. Структура связанных клауз задает формальную схему для модели структуры дискурса в рамках теории зависимостей. Смешанные структурные модели помогли нам заново оценить преимущества и недостатки традиционных концепций.

Исходя из различных способов организации структуры дискурса можно определить важные составляющие моделирования дискурсивной структуры:

1) выделение дискурсивных единиц, которые являются непересекающимися частями текста (ЭДЕ обычно приравнивается к клаузе);

2) определение структурных отношений, которые устанавливаются между дискурсивными единицами;

3) выявление способов определения дискурсивных отношений (в частности, дискурсивных коннекторов и их значений, центров отношений и т. п.).

2.1.3. Обоснование представления структуры дискурса в рамках модели зависимостей

Создатели рассмотренных выше моделей объясняют их целесообразность с разных точек зрения. Сторонники модели структуры вложения придерживаются мнения, что вложение элементарных единиц отражает специфику порождения дискурса, объединяющего более мелкие единицы в крупные, что не только увеличивает длину текста в линейной последовательности, но и усложняет иерархическую структуру текста [彭宣维, 2011]. Вложение является самым простым способом для представления структуры дискурса [Carlson, Marcu, Okurowski, 2003; Egg, Redeker, 2008; Egg, Redeker, 2010; Moser, Moore, 1996; Polanyi, 1997]. При организации модели вложения тексты сначала разделяют на дискурсивные сегменты (ЭДЕ), которые объединяются дискурсивными отношениями в большие сегменты (дискурсивные фрагменты), образуя, таким образом, N-арные или двоичные деревья, расположенные в иерархическом порядке [Dijk, Kintsch, 1983; Duchier, Gardent, 2001; Egg, Redeker, 2008; Egg, Redeker, 2010; Grosz, Sidner, 1986; Longacre, 1983; Mann, Thompson, 1988a; Moser, Moore, 1996; Polanyi, 1988a; Polanyi, 1997].

Однако оппоненты концепции составляющих утверждают, что естественный дискурс настолько сложен, что структуры вложения не способны четко описать прямые связи между дискурсивными единицами и не позволяют в более свободной форме представить эти отношения [冯文贺 et al., 2020; Asher, 2008; Danlos, 2008; Lee et al., 2008; Li et al., 2014a; Lyu, Feng, 2023; Wolf, Gibson, 2005; Wolf, Gibson, 2006]. Такие естественные ограничения обусловлены тем, что в структуре вложения допускаются только отношения между соседними единицами, но не множественные родительские узлы[1] и перекрестные отноше-

[1] Множественные родительские узлы показывают, что одна и та же единица может участвовать в двух или более двух дискурсивных отношениях. См., например, рис. 2.7 (B).

ния[1] [冯文贺 et al., 2020; Dijk, Kintsch, 1983; Grosz, Sidner, 1986; Longacre, 1983; Mann, Thompson, 1988a; Marcu, 2000; Polanyi, 1997]. Например, на рис. 2.7 только первые две структуры *(А)* и *(Б)* могут быть организованы в рамках вложения, а последние две структуры *(В) и (Г)* не могут. Однако ученые утверждают, что последние две структуры действительно характерны для текстов на естественном языке. Так, структура *(В)* описывает дискурсивную структуру, аналогичную фрагменту *а* в (2.3), а структура *(Г)* описывает дискурсивную структуру, аналогичную фрагменту *б* в (2.3).

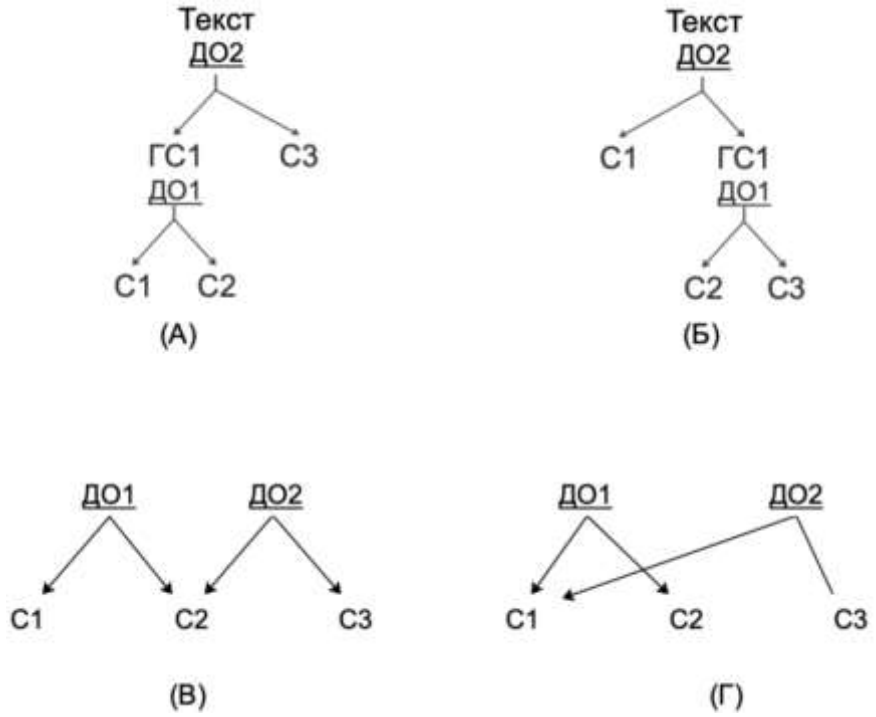


Рисунок 2.7. Дискурсивные структуры вложения и структура зависимостей.

[1] Наличие перекрестных отношений в дереве обусловлено тем, что в одном отношении, которое имеет множественные родительские узлы, объединяются не только соседние дискурсивные сегменты. Например, на рис. 2.7 (Г) отношения ДО1 и ДО2 пересекаются.

(2.3)

а.

1. Mary is in a bad mood

2. because her son is ill.

3. Specifically, he has an attack of bronchitis.

б.

1. Fred prepared a pizza

2. to please Mary.

3. Next, he took a nap.

Кроме того, вложение дискурсивных единиц, организованное «снизу вверх», может вызывать нечеткое изображение структур дискурса. Это также отмечают сторонники структуры вложения: для одного и того же текста может быть построен более чем один граф риторической структуры [Кибрик, Плунгян, 2002, с. 313; Mann, Matthiessen, Thompson, 1989b, p. 4]. В TPC такая нечеткость обычно объясняется принципиальной возможностью различных трактовок дискурсивной структуры. Однако основная проблема такой неопределенности структуры, на наш взгляд, состоит в недостатках самого способа вложения, а не в возможности различных трактовок. Например, та же проблема возникает при построении структуры вложения предложения «*Твои книги стоят на полке*». Вложенные структуры предложения могут быть представлены как на рис. 2.8 (А), так и на рис. 2.8 (А1). На самом деле семантическая структура этого предложения однозначна и не предполагает разницы в трактовках.

Рисунок 2.8. Сравнение реализации структур вложения.

Недостаток структуры вложения состоит еще в том, что иногда возникают проблемы с определением отношений между каждой единицей внутри вложения – *(X, Y)*, и единицей за пределами этого вложения – *Z*, поскольку отношение между единицами *X* и *Z* может отличаться от отношения между единицами *Y* и *Z*. Здесь проблема заключается в том, что вложенные структуры образуются комбинациями соседних элементов и формируются отношениями между вложениями и узлами, а не только двумя узлами. Разумеется, что такая проблема может не проявляться на синтаксическом уровне в связи с тем, что элементы предложения связаны конкретными явными и определенными синтаксическими правилами. Напротив, семантическая интерпретация отношений между единицами на уровне дискурса более разнообразна и сложна. Поэтому нам нужны уточненные структурные пары, а также отношения между единицами. Фактически Д. Марку столкнулся с подобной проблемой при создании дискурсивного корпуса теории риторических отношений, причем он предложил дополнительные правила к теории риторических структур для адаптации к выявлению дискурсивных отношений – принцип ядра [Marcu, 2000].

На наш взгляд, для разметки дискурсивных структур следует выработать более гибкие, простые и четкие универсальные принципы, которые имеют операциональный характер, и с этой точки зрения концепция бинарных зависимостей более удачна, чем концепция составляющих.

В первую очередь описание дискурсивной структуры в рамках идеи зависимостей отвечает потребностям описания сложных структурных явлений дискурса. В отличие от структуры вложения, которая в значительной степени опирается на линейный порядок единиц, концепция зависимостей отделяет порядок слов от структуры, при этом не предъявляя особого требования к линейному расположению языковых единиц. Структура зависимости может описывать связи как между соседними, так и между отдаленно расположенными языковыми единицами [Richard-Zappella, 1995, p. 87–88; цит. по: 刘海涛, 1991]. Следовательно, структура зависимостей приспособлена к описанию дискурсивной структуры, которая имеет более свободный порядок единиц, чем синтаксическая структура предложения.

Далее, структура зависимостей предполагает достаточно простые решения при осуществлении анализа дискурса. Несмотря на то, что TPC предложила универсальный набор отношений для описания дискурсивных структур разного уровня [Mann, Thompson, 1988a], на практике всегда приходится предварительно определять близость отношений между ЭДЕ, а затем упорядочивать эти отношения иерархически: чем ближе отношения между клаузами, тем ниже они располагаются в общей иерархии. Как правило, чем сложнее аналитическая процедура, тем чаще возникают неоднозначности. К тому же иерархическое упорядочивание не обязательно для всех структур, а иногда оно вызывает излишние споры при интерпретации. Как показывает рис. 2.8, нет необходимости спорить о порядке объединения элементов предложения «*Твои книги стоят*» и «*стоят на полке*». В отличие от концепции составляющих, бинарная структура зависимостей упрощает процедуру анализа, поскольку в этой структуре есть только элементы и ребра.

Важно и то, что структура зависимостей способна представить кросс-языковые и кросс-уровневые языковые явления в единой схеме. Теория структуры зависимостей была впервые применена для анализа синтаксических структур [Тестелец, 2001; Кубрякова, 1970; Robinson, 1970; Gaifman, 1965; Hays,

1964; Mel’čuk, 1988; Osborne, 2019; Tesnière, 1959; 刘海涛, 1991] и до сих пор имеет широкое распространение в лингвистике. В частности, в 2016 г. проект Универсальных зависимостей применяет рамки структуры зависимостей для описания синтаксических структур множества языков (296 подкорпусов по 168 языкам включены в недавно выпущенной версии *Universal Dependencies 2.15* (15 ноября 2024 г.), в значительной степени демонстрируя ее межъязыковую адаптивность. Кроме того, в последние годы также появился ряд проектов корпусов с разметкой дискурсивной структуры зависимостей [Li et al., 2014a; Li et al., 2014b; Lyu, Feng, 2023; Yang, Li, 2018a; Yoshida et al., 2014; 张牧宇 et al., 2013]. По мнению Вана Юэлуна, аспиранта национального университета Сингапура, «структура зависимостей постепенно стала адаптируемой системой для анализа различных языковых явлений на разных уровнях языка» [王跃龙, 2012, p. 11].

Кроме того, в настоящее время алгоритмы искусственных нейронных сетей (нейросетей) [McCulloch, Pitts, 1943] широко применимы для машинного обучения [Ptukhin, Khrushkov, Bozhko, 2019]. Как математическая модель, включающая элементы (нейроны) и ребра (связи нейронов), нейросеть имеет много общего с сетевой структурой зависимостей. Благодаря наличию сетевого внешнего представления, структура зависимостей хорошо приспособлена для машинного обучения [Bontempi, Flauder, 2019; Sharma, Sharma, Biswas, 2015]. Поэтому у нас есть основания полагать, что дискурсивный анализ может быть эффективным при компьютерной обработке естественного языка.

Наконец, в отличие от иерархического анализа дискурса, который сильно зависит от программного обеспечения [Marcu, 2000; Кибрик, Подлесская, 2009; 李艳翠, 周国栋, 2015], анализ в рамках зависимостей не предъявляет высоких требований к программному обеспечению для разметки. Например, в данной работе для анализа дискурсивной структуры зависимостей используются только возможности *Microsoft Excel* (подробнее см. п. 3.3. Программное обеспечение и хранение размеченных данных). В этом случае бинарная зависимость

представлена парой цифр, которые легко порождать и модифицировать, что значительно снижает сложность описания. Более того, размеченные данные, записанные в *Microsoft Excel*, могут быть трансформированы в формат *XML* [Захаров, Богданова, 2020, с. 37–40] для адаптации к любому стандарту разметки.

Опыт создания одноязычного дискурсивного корпуса показывает, что идея структуры зависимостей демонстрирует более высокую степень согласия между аннотаторами (см. в Таблице 2.1). Так, при разметке корпуса PDTB этот уровень достигает 90,2 %. В смешанной структуре он также выше, чем в структурах вложения.

Таблица 2.1. Согласие между аннотаторами при разметке дискурсивных корпусов разных типов.

	Типы структуры	**Согласие между аннотаторами**		**Язык**
		по структурам	**по дискурсивным отношениям**	
PDTB [Miltsakaki et al., 2004b]	Структура зависимостей	90,2 %		Англ.
Структура связанных клауз [冯文贺 et al., 2020]	Структура зависимостей	90,1 %		Кит.
Chain graph [Wolf, Gibson, 2005]	Смешанная структура	84,2 %	83,5 %	Англ.
Структура иерархии клауз [李艳翠, 周国栋, 2015]	Структура вложения	77,4 %	82,3 %–84,5 %	Кит.
RST DT [Carlson, Marcu, Okurowski, 2003]	Структура вложения	77,9 %–92,9 %	69,5 %–88,2 %	Англ.

Итак, по указанным причинам считаем, что в нашем исследовании структура дискурса должна получить описание в рамках концепции зависимостей.

2.2. Концептуальные основы дискурсивной структуры зависимостей при создании параллельного дискурсивного корпуса

Приведенные выше модели структуры дискурса системно представляют на уровне дискурса конструктивные единицы и их отношения. Несмотря на различия в способах организации, общими для них являются следующие понятия: 1) основная единица анализа дискурсивной структуры; 2) зависимость, или отношение между дискурсивными единицами; 3) показатели дискурсивных отношений – дискурсивные коннекторы; 4) асимметричные и симметричные дискурсивные отношения. Исходя из этого выстроен данный параграф. В подпараграфе 2.2.1 обсуждаются базовые ограничения формирования дискурсивной структуры зависимостей. В п. 2.2.2 рассматриваются подходы к делению текста на элементарные дискурсивные единицы. В п. 2.2.3 анализируются свойства дискурсивных отношений (зависимостей) и способы их классификации. П. 2.2.4 посвящен свойствам и типологии дискурсивных коннекторов. В п. 2.2.5 обсуждается понятие «дискурсивная вершина» в структуре зависимостей дискурса.

2.2.1. Базовые ограничения структуры зависимостей на уровне дискуса

Как уже было отмечено выше, структура зависимостей уже была введена в анализ дискурса. Однако большинство таких работ сосредоточено либо на алгоритмах преобразования деревьев составляющих в деревья зависимостей [Li et al., 2014a; Yang, Li, 2018a; Yoshida et al., 2014], либо на технических аспектах построения дискурсивных деревьев зависимостей [冯文贺 et al., 2020; Lyu, Feng, 2023]. При этом модель зависимостей используется преимущественно как внешняя форма для описания дискурсивных явлений, тогда как ее фундаментальные ограничения на дискурсивном уровне и лингвистическая специфика этих формальных элементов остаются нерассмотренными.

Мы планируем построить и проанализировать структуру зависимостей для русского и китайского языков в параллельном корпусе. Так как дискурсив-

ные исследования пока еще редко осуществляются в рамках структуры зависимостей, основные ограничения на формирование синтаксических зависимостей введены для объяснения общих особенностей структуры зависимостей.

Общепринятые ограничения структуры зависимостей определены в статье «Структура зависимости и правила перехода» [Robinson, 1970]:

Правило I. В предложении только один элемент является независимым.

Правило II. Все остальные элементы непосредственно зависят от какого-либо элемента.

Правило III. Ни один элемент не может непосредственно зависеть более чем от одного другого.

Правило IV. Если элемент А непосредственно зависит от элемента Б, и между ними в линейном порядке строки находится элемент В, то элемент В должен непосредственно зависеть от А, или от Б, либо от какого-либо другого элемента, который также находится между А и Б[1].

На приведенные правила ориентируются многие исследования по синтаксической структуре зависимости [Теньер, 1988; Тестелец, 2001; Gaifman, 1965; Hays, 1964; Hudson, 1984; Matthews, 1981; Mel'čuk, 1988; Tesnière, 1959; 刘海涛, 1991], потому что именно эти ограничения определяют минимальные требования к формированию ориентированного ациклического дерева, обладающего тремя ключевыми свойствами: древовидность, ацикличность и связность. Правило I, или ограничение единственного корня, говорит о том, что в структуре может быть лишь один элемент, который не зависит ни от какого другого – он выступает в качестве корня глобальной структуры. Правило II относится

[1] В оригинальном тексте эти четыре правила выражены следующим образом: 1) One and only one element is independent; 2) All others depend directly on some element; 3) No element depends directly on more than one other; 4) If A depends directly on B and some element C intervenes between them (in linear order of string), then C depends directly on A or on B or on some other intervening element [Robinson, 1970, p. 260].

к связности всех узлов структуры, то есть каждый элемент должен иметь прямую зависимость от другого элемента – он должен быть соединен с каким-либо другим элементом направленным ребром. Правило III, или ограничение единственной вершины (родительского узла), означает, что каждый элемент синтаксической структуры (кроме корня) может иметь только одну вершину – единственный элемент, от которого он зависит напрямую. Правило IV формулирует ограничение проективности (или непрерывности) в структуре зависимостей, которое иногда описывается как требование отсутствия пересечений зависимостей [Hays, 1964; Gaifman, 1965].

За этими формальными ограничениями, как мы утверждаем, на самом деле стоит фундаментальное лингвистическое понимание языковой структуры (на уровне предложения). Для построения достоверной и интерпретируемой дискурсивной структуры необходимо вернуться к трактовкам этих формальных структурных ограничений и дополнить их положениями, специфичными для дискурсивных структур.

При применении вышеуказанных четырех ограничений к анализу дискурсивной структуры зависимостей практически бесспорным является правило II, которое говорит о связанности языковой единицы. Сам дискурс представляет собой смысловое и структурное единство, которое характеризуется целостностью и связностью [Дейк, 1989; Арутюнова, 2002; Гальперин, 1981; Кибрик, 1994; Лосева, 1980; Макаров, 2003; Храпченко, 1986; Хурматуллин, 2009]. В соответствии с ограничением этого структурного правила связи устанавливаются между всеми сегментами дискурса, что обеспечивает целостность структурного анализа.

Однако для оформления структур зависимости на дискурсивном уровне остаются открытыми остальные вопросы:

1. Какое лингвистическое значение или практическое значение имеет корень? При каких условиях образуется единственный корень структуры зависи-

мостей? Как определить единственный корень для дискурсивной структуры зависимостей?

2. Почему возникает ограничение на наличие единственного корня? Что именно ограничивает это требование? На каком основании определяются дискурсивные вершины в локальной структурной паре?

3. Как следует понимать принцип проективности? Какова роль проективности в формировании дискурсивной структуры? Каким образом можно оптимизировать структуру дискурса, чтобы она соответствовала требованиям проективности?

Интерпретация единственного независимого корня. Ограничение единственного корня является ключевым для сохранения формальной корректности и связности синтаксических структур зависимостей. Корень, или корневой узел, – это единственный узел в дереве, который не зависит от других и служит исходной точкой для всех зависимостей в структуре: в иерархической структуре он является самым верхним узлом, как показано на рис. 2.1 (Б), или – в сглаженной структуре – начальным узлом, связанным ребрами без стрелки (см. рис. 2.2 (Б)).

Корень выполняет роль глобального центра управления всеми синтаксическими зависимостями в структуре предложения. Это означает, что именно через корень реализуется интеграция всех элементов предложения в единое целое, как на синтаксическом, так и на семантическом уровнях. Л. Теньер называл корень «центральным узлом, образованным словом, которое подчиняет себе – прямо или косвенно – все слова предложения» [Теньер, 1988, с. 26]. В терминологии зависимой грамматики корень часто (хотя и не всегда) идентифицируется с финитным глаголом-предикатом, не зависящим ни от одного другого слова в предложении. Он определяет валентностную структуру предложения, – список ожидаемых семантических ролей (агент, пациент, бенефициар и т. п.) – и организует вокруг себя остальные компоненты (подлежащее, дополнения, обстоятельства). По мнению И. А. Мельчука, «корень выступает в роли семанти-

ческого предиката, который формирует ядро высказывания», в рамках которого каждый компонент предложения получает свое место и функцию – см.: [Mel'čuk, 1988, с. 78–80]. Именно корень обеспечивает единство как формальной, так и семантической интерпретации одного предложения.

Ограничение единственного корня вытекает из определения структуры зависимостей как связного ациклического графа [Hays, 1964; Mel'čuk, 1988; Nivre, 2005]. Наличие единственного независимого узла обеспечило связность и древовидность дерева: без него граф распадается на несвязанные компоненты (лес) и уже не может рассматриваться как корректное дерево зависимостей. Л. Теньер даже утверждает, что корень в некотором смысле «отождествляется со всем предложением»: именно «корень обеспечивает структурное единство предложения тем, что связывает все его элементы в единый пучок» [Теньер, 1988, с. 26]. В задачах автоматической обработки текста корень играет роль опорной точки: через корень осуществляется «раскрытие» структуры – самые далекие ветви от корня указывают на границы предложения. Как отмечает Й. Нивре в своей «универсальной грамматике», каждому предложению ставится в соответствие набор базовых зависимостей, образующих корневое дерево, которое маркирует границы автономной синтаксической единицы как целостного предложения и отправной точкой для работы парсера [Nivre, 2005; Nivre, 2015; Nivre и др., 2020b]. Таким образом, ограничение единственного корня обеспечивает не только формальную корректность структуры зависимостей, но и практическую применимость при создании и применении синтаксических парсеров.

Таким образом, при описании структуры зависимостей для дискурса следует также соблюдать ограничение на единственный корень. В первую очередь необходимо определить границы автономной единицы для дискурсивного анализа, называемой «сложное целое» дискурса. Известно, что в лингвистических трактовках сложное целое может изучаться как «сложное синтаксическое целое» (далее – ССЦ) или «абзац» [Солганик, 2013; Водясова, 2013; Гальперин,

1981; Лосева, 1980; Пешковский, 2001; Поспелов, 1948; Реферовская, Десницкая, 1983]. Обсудим возможность использования ССЦ и абзаца как сложного целого при создании дискурсивной структуры.

Одно из наиболее распространенных определений ССЦ – «группа предложений, объединенных по смыслу и грамматически и выражающих более или менее законченную мысль» [Солганик, 2013, с. 38]. Исходя из этого определения, ССЦ понимается как смысловая, структурная и функциональная единица дискурса. Рассматривая ССЦ как сложное целое для структурного анализа, в принципе необходимо найти дискурсивную единицу, которая рассматривается как смысловой, структурный и функциональный центр, то есть корень этого сложного целого. Проблема, однако, состоит в том, что на практике сложно найти границы дискурсивных единиц и определить корневой узел ССЦ. В дискурсе, характеризующемся целостностью и связностью, не только предложения могут быть объединены в ССЦ, но и ССЦ могут быть объединены в более крупные единицы, пока объединение не достигнет самого текста. В этом случае реальная единица выражения корневого узла может измениться.

В лингвистике текста абзац представляет собой типографский термин [Пешковский, 2001, с. 459], под которым понимается «отрезок письменного или печатанного текста от одной красной строки до другой красной строки» [Розенталь, Теленкова, 1985]. Лингвистическая значимость абзаца остается спорной, однако, как формально-структурный элемент, абзац широко изучается во многих прикладных исследованиях, связанных с обработкой естественного языка [Gelbukh, Sidorov, 2006a; Church, 1993; Coquenet, Chatelain, Paquet, 2021; Gelbukh, Sidorov, 2006b; Gelbukh, Sidorov, Vera-Félix, 2006; Gupta, Pala, 2012; Le et al., 2014; Tiedemann, 2011; 李维刚 et al., 2003]. При создании дискурсивного корпуса абзац также рассматривается как структурная единица дискурса [吴永芃 et al., 2018; Danlos, 2004; Lyu, Feng, 2023; Wolf, Gibson, 2005; 李艳翠, 周国栋, 2015].

Рассматривая знак абзаца (или красную строку) как естественное ограничение сложного синтаксического целого, будем исходить в данной работе из приоритета абзацной структуры. Как уже говорилось выше, в ССЦ всегда допускается один корень, а когда ССЦ образуют еще более крупное синтаксическое целое, положение корня может измениться. Но, согласно существующим исследованиям, границы ССЦ и абзацев не всегда совпадают [Солганик, 2013; Лосева, 1980; Пешковский, 2001; Реферовская, Десницкая, 1983]. Если абзац является единицей более крупной, чем ССЦ, то обязательно в его структуре появляется несколько отдельных «корневых» узлов. Тогда как же определить «единственный корень» в качестве корневого узла среди этих отдельных «корневых» узлов в одном абзаце? Если абзац является единицей меньше, чем ССЦ, то какую позицию занимает единственный корень этого абзаца? На поставленные вопросы должны ответить исследования структуры зависимостей, построенной путем рассмотрения абзацев как автономной единицы дискурсивного анализа.

Интерпретация единственной вершины. Вершина (англ. head), или по-другому «родительский узел», «управляющий» или «локальный центр», является доминирующим узлом в одной структурной паре, состоящей из двух связанных языковых единиц[1]. Вершина является ключевым компонентом формирования древовидной структуры [Tesnière, 1959; Hudson, 1990; Hudson, 2007; Mel’čuk, 1988; Мельчук, 1974], поскольку она определяет зависимость, то есть асимметричное отношение между двумя языковыми единицами, входящими в структурную пару, что проявляется в доминировании вершины над периферией.

[1] Понятие «вершина» связано с понятием «периферия», что получило разные названия в теориях зависимостей: regissant и subordonne [Tesnière, 1959]; head и modifier [Hudson, 1984], governor и dependent [Hays, 1964; Mel’čuk, 1988].

С точки зрения формальной теории графов, ограничение единственной вершины является дополнительной гарантией строгого деревообразного представления зависимостей и его ацикличности. В этом случае только после определения единственной вершины в каждой локальной структурной паре можно определить единственный корень глобальной структуры дерева: вершины определяют направление каждого ребра (со стрелкой, указывающей на периферию) и, таким образом, определяют корень – конечный вершинный узел.

В синтаксических теориях грамматики зависимостей разработаны три лингвистические интерпретации понятия «вершина»: морфологическая, синтаксическая и семантическая [Tesnière, 1959; Zwicky, 1985; Hudson, 1987; De Marneffe, Nivre, 2019]. Морфологическая интерпретация определяет вершину как слово, которое характеризуется самым большим количеством морфологических признаков. Например, в предложении *Она прочитала эту книгу* глагол *прочитала* реализует морфологические категории времени, рода и числа, а значит, занимает доминирующую позицию. С точки зрения синтаксической функции, в структурной паре «существительное – глагол» глагол выступает в качестве вершины; в структурной паре «прилагательное – существительное» в качестве вершины выступает существительное. В соответствии с семантической интерпретацией вершина может быть соотнесена с семантической категорией структурной пары. Например, в паре *эту – книгу* слово *книга* рассматривается как вершина, потому что оно определяет семантическую категорию этой структурной пары как предмет; а в структурной паре *прочитала – книгу* слово *прочитала* рассматривается как вершина, потому что оно определяет семантическую категорию этой структурной пары как действие. Таким образом, вершина может быть морфологическим управляющим центром, синтаксическим узлом, обеспечивающим объединение всех элементов в единое целое, а также семантическим ядром этой структуры.

Опыт синтаксического анализа показывает, что при определении вершин необходимо четко различать эти три уровня. Как отмечает И. А. Мельчук, каж-

дый из этих уровней обладает собственной структурной организацией, отражающей природу отношений между элементами в рамках соответствующего уровня. Он предлагает разграничение трех уровней языковой зависимости – семантического, синтаксического и морфологического (см. рис. 2.9) [Mel'čuk, 1988, p. 16–25]. На семантическом уровне зависимости представляют собой «неупорядоченную сеть», где узлы обозначают семемы, а ребра отражают семантические отношения. Здесь возможны циклы, множественные связи и перекрестные отношения, обусловливающие семантическую гибкость. На синтаксическом уровне структура принимает форму ориентированного дерева, которая соответствует канонической модели зависимостей и служит основой для синтаксического анализа и генерации. На морфологическом уровне зависимости между морфемами организуются в цепи, отражающие линейную последовательность морфологических элементов внутри слова – см.: [Mel'čuk, 1988; Mel'čuk, 2014; Dependency in linguistic description, 2009].

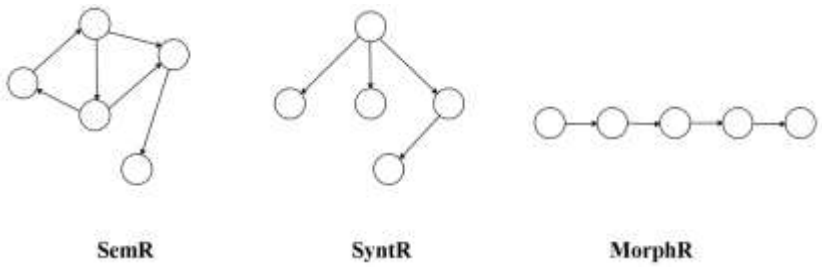


Рисунок 2.9. Три уровня зависимости: семантический, синтаксический и морфологический.

Некоторые другие работы по анализу структуры зависимостей, не различающие разные уровни вершин, допускают наличие множественных вершинных узлов. Например, в работе Р. Хадсон [Hudson, 1984] при построении дерева зависимостей предложения «*John seems to like Mary*» одновременно связывается с «*seems*» «*to*» и «*like*», и в результате анализа возникло несколько родительских узлов у слова «*John*» (см. рис. 2.9), а скорее, зависимость между «*John*» и «*seems*» является морфологической, между «*John*» и «*like*» – семантической.

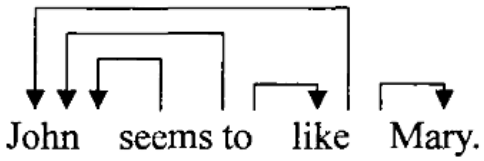


Рисунок 2.10. Множественные вершинные узлы.

Следует отметить, что нарушение ограничения единственной вершины приводит в структуре к появлению не только циклов с несколькими вершинами, но и множественных корней (*seems* и *to*). Такая множественность корней принципиально нарушает иерархическую организацию предложения, что влечет за собой ряд проблем как для лингвистической интерпретации, так и для решения прикладных задач. Во-первых, наличие более чем одного корня порождает синтаксическую неоднозначность: становится неясно, какой из элементов выполняет роль ядра синтаксической структуры. Это затрудняет установление семантической направленности и управления в предложении. Во-вторых, в контексте автоматической обработки текста многокорневая структура приводит к техническим затруднениям. Алгоритмы, основанные на древовидной модели, теряют возможность однозначно выделить независимые единицы анализа, поскольку границы между деревьями становятся неочевидными. В результате структурная разметка утрачивает как интерпретативную четкость, так и прикладную полезность. Иными словами, наличие множественных вершин и, как следствие, множественных корней разрушает логическую целостность древовидного графа, превращая его в произвольный направленный граф, утративший лингвистическую интерпретируемость и структурную строгость.

Итак, необходимо определить уровень анализа дискурсивных вершин для корректного установления единственной дискурсивной вершины для локальной структурной пары. Анализ дискурсивной структуры во многом представляет собой разновидность структурного анализа на семантическом уровне, в котором дискурсивные единицы (например, клаузы) рассматриваются как уз-

лы, а семантические отношения между этими единицами – как ребра. В то время как дискурсивные вершины отражают семантическую важность одной дискурсивной единицы относительно другой в локальной структурной паре. Как пишет в вышеуказанном исследовании И. А. Мельчук, эти семантические отношения могут изначально формировать неупорядоченные сети: параллельные, перекрестные, взаимозависимые. Однако естественные тексты организованы не беспорядочно – они следуют логической иерархии, структурируя информацию согласно причинно-следственным, временным, контрастивным или другим отношениям. Чтобы отразить эту иерархичность, необходимо преобразовать исходную семантическую сеть в ориентированное дерево. Иными словами, на основе ненаправленной сети зависимостей необходимо для каждой структурной пары установить единственную вершину и построить иерархию зависимостей.

Как уже говорилось при обсуждении корневого узла, в нашем исследовании необходимо ввести ограничения, чтобы сохранить единственный корень и тем самым обеспечить древовидность и связность глобальной структуры дискурса. Однако, из-за отсутствия четких формальных показателей на дискурсивном уровне, подобных тем, что существуют на синтаксическом уровне, необходимо выработать более объективные и четкие основания для определения и разметки дискурсивных вершин. Эту проблему должно решить наше корпусное исследование.

Интерпретация ограничения проективности. Ограничение проективности говорит о том, что все слова, находящиеся между вершиной и ее периферией, обязаны прямо или опосредованно входить в это же поддерево и не должны образовывать зависимости с единицами за его пределами. Практически такое требование гарантирует, что при графическом изображении зависимости не пересекаются – дерево зависимостей можно нарисовать на плоскости без перекрещивающихся ребер. В терминах теории графов все ребра в дере-

ве зависимостей укладываются над линией предложения без пересечений (см. рис. 2.11).

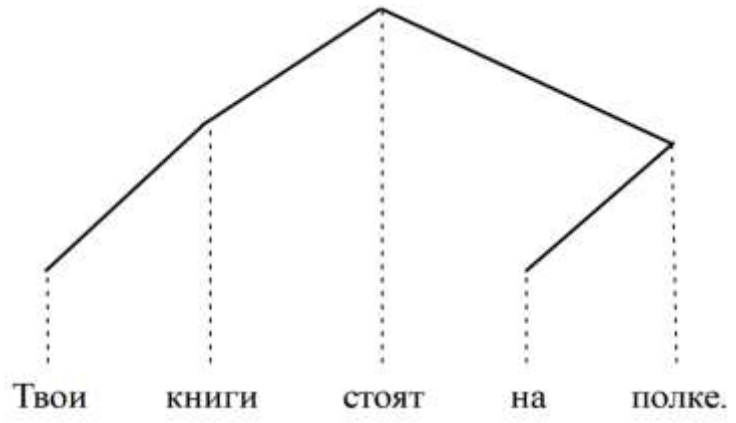


Рисунок 2.11. Проективная структура зависимостей предложения «*Твои книги стоят на полке*».

Из четырех базовых правил именно ограничение проективности вызывает наибольшее число дискуссий и разногласий в лингвистических исследованиях. С одной стороны, много ученых указывают на то, что нарушения правила встречаются в реальных языках, особенно в языках с относительно свободным порядком слов (русский, немецкий, чешский, латинский) [Hajičová, 1991; Hudson, 1982; Mel'čuk, 1988]. Некоторые эмпирические исследования [Kuhlmann, Nivre, 2006; Nivre, 2005] показывают, что около 5–15 % зависимостей в реальных корпусах оказываются непроективными, особенно в письменной речи или с усложненной синтаксической структурой.

С другой стороны, при соблюдении условия проективности дерево зависимостей оказывается формально эквивалентным дереву составляющих [Hays, 1964; Gaifman, 1965; Hudson, 1984]. Многие исследования подтверждают, что проективная структура зависимостей и формально иерархическая вложенная структура имеют прямое преобразовательное отношение. Это наблюдение восходит к работам Д. Хейса и Х. Гейфмана, которые первыми показали, что каждое проективное дерево зависимостей может быть изоморфно дереву конституи-

ент (дереву составляющих), построенному по правилам контекстно-свободной грамматики [Gaifman, 1965; Hays, 1964].

Иными словами, проективность обеспечивает такое упорядочение синтаксических связей, при котором зависимые элементы формируют непрерывные подцепочки, и таким образом позволяет организовать синтаксическую структуру по вложенности. Отсюда Р. Хадсон отмечает, что требование проективности можно понимать как требование «последовательности» (sequentiality), при котором линейный порядок слов должен соответствовать иерархическому порядку синтаксических зависимостей [Hudson, 1984, p. 98]. Кроме того, количество отношений в проективной структуре соответствует количеству отношений в структуре составляющих: для дерева зависимостей, состоящего из n единиц, количество отношений (s) должно быть равно $n-1$; если $s < n-1$, то появляются клаузы, которые не включены в структуру зависимости. А если $s > n-1$, то структура не проективна [冯文贺 и др., 2020, p. 25].

В современных исследованиях правило проективности чаще рассматривают как модельное допущение, а не как безусловную грамматическую универсалию. Как отмечается в ряде работ, дерево может быть связным, ацикличным и с единственной вершиной, но при этом содержать непроектные (перекрещивающиеся) связи [Gaifman, 1965; Hudson, 1990; Kuhlmann, Nivre, 2006; Mel’čuk, 1988]. Например, И. А. Мельчук подчеркивает, что проективность как «непересечение дуг» – это следствие линейного упорядочивания, а не свойство самой грамматической зависимости [Mel’čuk, 1988]. В этом смысле одна и та же структура может быть проективной при одном порядке слов и непроективной – при другом. Таким образом, проективность становится не свойством грамматической структуры, а функцией линейной реализации. В этом смысле правило проективности ближе к вычислительной эвристике или к когнитивному упрощению, чем к универсальному грамматическому правилу [Debusmann, 2006; Kuhlmann, Nivre, 2006; Nivre, Nilsson, 2005].

В компьютерной лингвистике и обработке естественного языка проективная структура считается «оптимальной» для вычислительной обработки языка. Оптимальная структура зависимостей, как правило, сохраняет наиболее базовую информацию, необходимую для анализа предложения, и при этом потребляет минимум вычислительных ресурсов. Многие алгоритмы синтаксических парсеров [Eisner, 1996; Nivre, 2003] показывают, что проективное дерево дает значительный выигрыш в скорости и предсказуемости анализа.

В отличие от синтаксических структур, дискурсивные структуры характеризуются большей гибкостью. Дискурсивные элементы, не ограниченные жесткими формальными правилами, такими как морфологические или грамматические правила, могут быть расположены не так строго, как в синтаксисе, что обеспечивает более свободный порядок их организации – см.: [Danlos, 2004; Wolf, Gibson, 2005]. Это означает, что дискурсивная структура зависимостей может в значительной степени не соблюдать проективные ограничения. Однако на практике степень гибкости такой структуры трудно контролировать, поскольку в дискурсивном фрагменте любые два предложения могут быть связанными. Таким образом, важно, чтобы зависимости в дискурсивном дереве не пересекались, чтобы структура оставалась понятной и логичной.

Для поддержания оптимальной дискурсивной структуры необходимо исключать из дискурсивного дерева, включающего все возможные связи между дискурсивными элементами, те ребра, которые выполняют одну и ту же функцию, чтобы избежать избыточности и дублирования информации. Важно определить, какие структурные пары имеют явную, эксплицитную функцию и должны быть включены в дерево, а какие могут быть имплицитными или опущенными без потери смысловой нагрузки. Вопрос оптимизации дискурсивных структур в контексте приоритетов выделения связей требует практического опыта в процессе дискурсивного структурирования.

2.2.2. Элементарная дискурсивная единица

Деление текста на элементарные дискурсивные единицы является первым этапом структурного анализа дискурса. Однако существуют разногласия по поводу того, как определять ЭДЕ. Одни утверждают, что сегментами дискурса должны быть клаузы [Chu, 1998; Givón, 1983; Grimes, 1975; Haiman, Thompson, 1988; Hovy, 1993b; Mann, Thompson, 1988a], другие говорят об интонационных единицах [Кибрик, Подлесская, 2006; Grosz, Sidner, 1986; Hirschberg, Litman, 1987], фразовых единицах [Lascarides, Asher, 1993; Longacre, 1983; Webber et al., 1999] или предложениях [廖秋忠, 1991; Hobbs, 1985; Polanyi, 1988b]. В большинстве случаев, чтобы найти баланс между детализацией структурного анализа дискурса и возможностью согласовать терминологию, ЭДЕ определяют как «клаузы» [乐明, 2008; 吴云芳, 徐艺峰, 王恺然, 2015; 陈莉萍, 2008; Ананьева, Кобозева, 2016a; Литвиненко, 2001; Carlson, Marcu, Okurovsky, 2001; Mann, Thompson, 1988a; Marcu, 1996].

Термин «клауза» заимствован из английского языка (“clause”) и обычно обозначает простое предложение или конструктивную часть в составе сложного предложения. Этот термин часто встречается в китайских и русских грамматических исследованиях, однако в конкретных теориях и языках существует разногласие относительно выделения синтаксических единиц, составляющих «клаузу».

В ранних грамматических исследованиях китайского языка уже отмечена синтаксическая единица с объемом меньшим, чем предложение. Китайский лингвист Лю Фу в 1920 г. впервые предложил термин «клауза» (кит.小句, дословный перевод «маленькое предложение») для обозначения составных частей сложного предложения [刘复, 1920]. Первым, кто рассматривал клаузу как грамматическую единицу, был Люй Шусян [吕叔湘, 1979]. Он выделял статические единицы (слово, конструкция) и динамические единицы (клауза, предложение), отмечая, что клауза является элементарной динамической единицей,

при этом простое предложение состоит из одной клаузы, а сложное – из двух или более клауз [Там же, p. 29].

В последние годы клауза – это центральное понятие в грамматических исследованиях китайского языка. Предложенная Син Фуи мысль о «клаузе как центральном узле предложения» [Син Фуи, 2020; Xing, 2017; 邢福义, 2016; 邢福义, 1995] утвердила особый статус клаузы в языковой системе китайского языка. По мнению Син Фуи, среди языковых единиц различных типов и уровней только клауза обладает непосредственной связью с прочими языковыми единицами и поэтому выполняет роль связующего центра в системе. Точнее говоря, с точки зрения языковой реализации клауза связана с модальностью; с точки зрения внутреннего устройства клауза связана со словами и синтагмами; с точки зрения внешней дистрибуции клауза связана со сложным предложением и группой предложений [Син Фуи, 2020, с. 31–33].

В отличие от понимания клаузы в теоретическом контексте, фокусирующемся на ее функциях в системе языка, клауза как ЭДЕ чисто практически выделяется по формально-пунктуационным признакам. При анализе китайского дискурса знаки препинания широко используются в качестве эксплицитного формального маркера [李艳翠 et al., 2013; Jin et al., 2004; Li, Feng, Feng, 2015; Xue, Yang, 2011]. Из всех знаков пунктуации знаки конца предложения (точка, вопросительный знак, восклицательный знак и точка с запятой) всегда рассматриваются как разделители ЭДЕ. Каплевидная запятая (отражающая паузу между словами) никогда не считается разделителем ЭДЕ. Вопрос о том, может ли использоваться в качестве разделителя ЭДЕ при делении текста китайского языка запятая, остается спорным.

Ученые придерживаются разных мнений о том, следует ли использовать запятые в качестве знаков, выделяющих ЭДЕ. Лэ Мин считает, что запятые не должны использоваться в качестве таких маркеров [乐明, 2008]. Согласно статистическим данным, из всех знаков препинания в тексте запятые в китай-

ском тексте составляют почти 40 %, причем в 27,5 % они стоят между подлежащим и сказуемым.

У Йонпэн и др., наоборот, утверждают, что запятая должна использоваться как разделитель ЭДЕ [吴永芃 et al., 2018]. В большинстве случаев единицы, выделенные запятыми, являются предложениями или их функциональными аналогами [Там же, p. 77].

Чтобы устранить неоднозначность использования запятой при сегментации ЭДЕ, Люй Гуоин и др. устанавливают два этапа анализа: 1) деление текста на первичные дискурсивные единицы с использованием запятых, точек и других знаков препинания; 2) распознавание фундаментальных дискурсивных единиц на основе результатов первичной сегментации, в процессе которого некоторые синтаксические единицы объединяются [吕国英 et al., 2015].

В некоторых работах предлагается выделять ЭДЕ без учета знаков препинания. Например, обсуждая тематическую структуру дискурса (discourse topic structure analysis), Си Сюэфэн и др. в качестве ЭДЕ рассматривают микротему [奚雪峰 et al., 2017]. Это единица, содержащая необходимые элементы для формирования простого предложения (в первую очередь подлежащее и сказуемое), причем все такие конструкции рассматриваются как отдельные ЭДЕ, например:

1) 两名俄罗斯航天员**进入**航天飞机。

(дословный перевод[1]: Два российских космонавта вошли в космический корабль.) – одна микротема (ЭДЕ)

2) 两名俄罗斯航天员**进入**航天飞机**开始**准备升空。

(дословный перевод: Два российских космонавта вошли в космический корабль, начали подготовку к взлету.) – две микротемы (ЭДЕ)

---

[1] Здесь и далее дословный перевод на русский язык выполнен для иллюстративности с сохранением порядка слов и грамматических особенностей китайской фразы.

Примеры взяты из [奚雪峰 et al., 2017]

В процессе деления письменного текста на ЭДЕ с помощью знаков препинания как полное исключение запятой, так и использование только запятой в качестве маркера деления текста неизбежно приводит к большому количеству неоднозначных элементов [Xue, Yang, 2011]. С одной стороны, сегментация ЭДЕ без учета запятых приводит к слишком большим единицам, которые являются сложными предложениями или группами предложений, что не позволяет проанализировать всю структуру дискурса. С другой стороны, деление на ЭДЕ с постоянным влиянием запятой приводит в результате к появлению мнимых синтаксических единиц, таких как подлежащее, сказуемое, дополнение, обстоятельство и др. Как несоразмерно большие, так и неоправданно маленькие ЭДЕ приводят к ошибкам в структурном анализе дискурса. Последнее статистическое исследование по работе Ли Яньцуя и др. на основе корпуса показало, что совпадение выделения ЭДЕ по знакам препинания (в том числе запятой) с клаузами может достигать 89,2 % [Li, Feng, Feng, 2015], что дает основание для решения о предварительном делении текстов на китайские клаузы с помощью пунктуации.

В традиции русской грамматики термин «клауза» используется относительно недавно. В отдельных случаях он понимается как «элементарное предложение в составе сложных предложений» [Тестелец, 2001]. С развитием технологий обработки естественного языка и связанных с ними лингвистических теорий термин «клауза» постепенно стал использоваться широко. Например, А. О. Литвиненко при описании дискурсивной структуры русского языка в рамках TPC определяет «клаузу» как «элементарное предложение, состоящее из одной глагольной группы и одной или нескольких именных групп» [Литвиненко, 2001, с. 163].

При анализе дискурсивной структуры русского языка в прототипическом случае ЭДЕ совпадает с клаузой, то есть каждая предикация выделяется в отдельную ЭДЕ [Там же, с. 161]. Но при делении на русские ЭДЕ для создания

дискурсивного корпуса отмечаются дополнительные случаи. Например, А. Л. Литвиненко предусматривает следующие правила для распознавания ЭДЕ для русского дискурса:

а) однородные сказуемые представляют собой разные единицы;

б) сентенциальные актанты, кроме придаточных при глаголах мысли, речи и ощущений, принадлежат к той же ЭДЕ, что и матричный предикат;

в) придаточные дополнительные при глаголах мысли, речи и ощущений выделяются в отдельную ЭДЕ;

г) прямая речь выделяется в отдельную ЭДЕ (разбивается на несколько отдельных ЭДЕ, если содержит более 1 клаузы);

д) определительные придаточные и причастные обороты, являющиеся нерестриктивными определениями, выделяются в отдельную ЭДЕ (или несколько ЭДЕ);

е) рестриктивные определения, независимо от их формы, принадлежат к той же ЭДЕ, что и определяемое слово [Литвиненко, 2001, с. 161–162].

При создании корпуса текстов на русском языке в рамках TPC М. И. Ананьева и М. В. Кобозева в качестве ЭДЕ выделяют:

а) финитные клаузы (кроме клауз, являющихся сентенциальными актантами и при этом не входящих в косвенную речь);

б) деепричастные обороты с причинно-следственным и уточняющим значением;

в) описательные (нерестриктивные) причастные обороты;

г) предложные группы со значением причины, следствия, уступки и контраста [Ананьева, Кобозева, 2016, с. 2].

Дискурсивные единицы, выделенные в соответствии с вышеуказанными принципами, можно назвать «русскими клаузами» и их аналогами.

Следует отметить еще раз, что понимание клауз как ЭДЕ для анализа дискурсивной структуры может отличаться от ее трактовки в грамматических теориях. Как подчеркивают У. Манн и С. Томпсон, все разделенные единицы

независимо от их размера должны непосредственно или опосредованно участвовать в генерации текста [Mann, Thompson, 1988].

В целом, несмотря на споры, в большинстве моделей анализа структуры дискурса как в китайской, так и в русской практике в качестве основной единицы рассматривается клауза, а не предложение. Такое понимание принято и в данной работе, а в п. 3.3.1. приведены основные принципы деления на ЭДЕ при создании китайско-русского параллельного корпуса. Следует также отметить, что при создании параллельного дискурсивного корпуса важной целью является межъязыковое структурное выравнивание текстов, которое изначально производится на уровне ЭДЕ. При этом надо понимать, что определить во всех случаях четкие соответствия ЭДЕ (или клауз) в двух языках практически невозможно. Сегментация параллельных текстов в нашем случае будет основана на выделении китайских клауз и соотношении с ними русских синтаксических аналогов.

2.2.3. Дискурсивные отношения и их типы

Модели структуры зависимостей, рассматривающие дискурсивные отношения, ориентированы в большей степени на семантические, а не грамматические признаки. Исследования по дискурсивным отношениям (далее – ДО) в рамках TPC и модели PDTB закладывают основу для обсуждения этих семантических зависимостей. В TPC дискурсивное семантическое отношение называется «риторическим отношением», которое устанавливается между ЭДЕ в целях обеспечения последовательности и целостности структуры дискурса [Mann, 1984; Mann, Thompson, 1988a]. В теории перечислен список из 24 видов риторических отношений[1] (см. Таблицу 2.2) и дано подробное определение каждого отношения с трех точек зрения: 1) ядро отношения, 2) специфические ограничения присвоения этих статусов, а также 3) эффект отношения, производимый

[1] В TPC включаются 23 риторических отношения и одна схема – *Joint (Конъюнкция)*.

на читателя (слушателя). По количеству ядер можно разделить риторические отношения на симметричные и асимметричные [Mann, Thompson, 1988a].

Таблица 2.2. Исходный набор риторических отношений[1]

Circumstance (Обстоятельство)	Antithesis (Антитезис)
Solutionhood (Решение)	Concession (Уступка)
Elaboration (Развитие, или Детализация)	Condition (Условие)
Background (Фон)	Otherwise (Альтернатива)
Enablement (Обеспечение возможности)	Interpretation (Интерпретация)
Motivation (Мотивация)	Evaluation (Оценка)
Evidence (Свидетельство)	Restatement (Переформулировка)
Justify (Обоснование)	Summary (Резюме)
Purpose (Цель)	Sequence (Последовательность)
Volitional Cause (Волитивная Причина)	Contrast (Противопоставление)
Volitional Result (Волитивный Результат)	Joint (Конъюнкция)
Non-Volitional Cause (Неволитивная Причина)	
Non-Volitional Result (Неволитивный Результат)	

Авторы РРС отмечают, что одни и те же риторические отношения можно проследить на всех уровнях дискурсивной иерархии [Mann, Thompson, 1988a]. Следовательно, стандартные повторяющиеся отношения из одного определенного набора устанавливаются как между клаузами в сложных предложениях, так и между группами предложений. Это неудивительно: фактически РРС распространяет типологию семантико-синтаксических отношений между клаузами на отношения в дискурсе. Для РРС несущественно, каким именно образом выражено данное отношение и соединяет ли оно независимые предложения или группы предложений [Олешков, 2006, с. 110–111]. Здесь отметим различие между дискурсивным отношением в тексте и синтаксическим отношением в сложных предложениях; оно заключается в том, что дискурсивное риторче-

[1] В скобках указаны русские соответствующие переводные термины разных отношений [Кибрик, Плунгян, 2002; Литвиненко, 2002].

ское отношение функционально по отношению к клаузам. Риторическое отношение, как пишут А. А. Кибрик и В. А. Плунгян, «указывает на то, что каждая единица существует не сама по себе, а добавляется говорящим к некоторой другой для достижения определенной цели» [Кибрик, Плунгян, 2002, с. 309]. Например, в TPC риторические отношения определяются со специфическими «ограничениями» для дискурсивных единиц и «эффектом» данного отношения, производимым на читателя (слушателя).

Кроме того, создатели TPC не задают набор риторических отношений жестко, то есть в случае конкретного анализа текста разных жанров можно добавлять в перечень дополнительные отношения или какие-то, наоборот, удалять из него. Это позволяет использовать данную схему для анализа структуры текста разных типов.

В модели PDTB дискурсивные отношения представлены трехуровневой системой [Miltsakaki et al., 2008b; Prasad et al., 2008b; PDTB-Group, 2008b], в которой категоризация семантики коннекторов расположена от общей до детальной. Первый уровень состоит из четырех общих типов отношений: временные, сравнительные, расширительные и условные. На втором и третьем уровнях выделяются некоторые типы отношений (см. Таблицу 2.3).

Таблица 2.3. Трехуровневая система в PDTB для описания дискурсивных отношений.

Первый уровень	Второй уровень	Третий уровень	Первый уровень	Второй уровень	Третий уровень
Temporal	Synchronous	——	**Comparison**	Pragmatic contrast	——
				Pragmatic concession	——
	Asynchronous	Precedence		Contrast	Juxtaposition
					Opposition
		Succession		Concession	Expectation
					Contra-Expectation
Expansion	Exception	——	**Contingency**	Pragmatiy cause	——
	List	——		Pragmatic condition	Relevance
	Conjunction	——			Implicit assertion
	Instantiation	——		Cause	Reason
	Restatement	Specification			Result
		Equivalence			Hypothetical
		Generalization			General
	Alternative	Conjunctive			Unreal Present
		Disjunctive			Unreal Past
		Chosen Alternative			Factual Present
					Factual Past

Стоит отметить, что анализ ДО по модели PDTB осуществляется с помощью дискурсивного семантического показателя – дискурсивного коннектора (см. п. 2.2.3.). На основании наличия или отсутствия явного дискурсивного коннектора ДО делятся на эксплицитные и имплицитные. Независимо от наличия или отсутствия коннекторов анализ семантических связей между аргументами может быть осуществлен только на основе контекста или соответствующих общих знаний.

Список ДО как в TPC, так и в модели PDTB напоминает традиционный список типов обстоятельственных придаточных. Это неудивительно, поскольку «фактически TPC распространяет типологию семантико-синтаксических отношений между клаузами на отношения в дискурсе» [Олешков, 2006, с. 111]. На самом деле в ранних исследованиях сложных предложений как в китайском языке, так и в русском уже было показано, что типология логико-семантических отношений в группах предложений[1] китайского языка (или ССЦ в русской традиции) в целом аналогична типологии отношений в сложном предложении [陈洁, 1997; 徐赳赳, 1997; 廖秋忠, 1991; Валгина, 2003; Лосева, 1980; Син Фуи, 2020; 刘辰诞, 赵秀凤, 2011; 吴为章, 田小琳, 2000; 张凤珍, 陈洁, 2005; 陈洁, 2007]. Например, в «Грамматике китайского языка» Сина Фуи к основным отношениям относятся причинные, сочинительные и противительные [Син Фуи, 2020, с. 551–557]. У Вэйчжан и Тянь Сяолинь в работе «Китайская группа предложений» выделяют двенадцать типов: сочинительный (并列), временной (时间) и распространенный последовательный (时空顺序), градационный (递进), выбора (选择), обобщения (总分), пояснительный (解证), причины и следствия (因果), целевой (目的), условия (条件), противительный (转折), условный (假设) и уступительный (让步) [吴为章, 田小琳, 2000]. Проектируе-

[1] Группа предложений – это грамматическая единица китайского языка, состоящая из двух и более предложений, сосредоточенных вокруг одного значения [Син Фуи, 2020, с. 545–557].

мая кросслингвистическая база данных для аннотирования логико-семантических отношений предлагает следующие логико-семантические отношения: временные, причинные, противительные, соединительные, сопоставительные, условные, мереологические и другие виды отношений – см.: [Дурново, Зацман, Лощилова, 2016, с. 124–125].

Помимо дискурсивных корпусов, некоторые традиционные теории по дискурсивным отношениям также способствуют нашему пониманию дискурсивных отношений при анализе дискурсивной структуры. Исходя из них типы ДО в определенных текстах или в текстах определенных жанров, прежде всего, должны быть исчерпывающим набором. Как известно, классификация ДО в разных исследованиях сильно различаются: от 2 [Grosz, Sidner, 1986] до более 400 [Hovy, Maier, 1992] различных дискурсивных отношений [цит. по: Wolf, Gibson, 2005, p. 249]. Иногда говорят о том, что типология ДО настолько сложна, что трудно полностью описать их все. Впрочем, в последние годы некоторые исследования показывают, что в классификации ДО наметилась тенденция к упрощению. Швейцарские лингвисты О. Инькова и Э. Манзотти, исходя из рационального мышления и мышления в целом, предлагают оригинальную классификацию логико-семантических отношений: генерализация, спецификация, исключение из множества и аддитивность [Инькова, Манзотти, 2019].

ДО также многозначны. С одной стороны, полисемия ДО состоит в том, что одно отношение могут содержать одновременно несколько специфические особенности. Например, по мнению китайского лингвиста Шэна Сяолуна, отношения «причина» и «следствие» одновременно содержат временной последовательный характер [申小龙, 2005, p. 225]. С другой стороны, ДО многозначны, потому что связи между дискурсивными единицами могут понимать поразному разные аннотаторы. Субъективные факторы при оценке ДО являются основными факторами, влияющими на достоверность дискурсивных корпусов. В целях снятия влияния субъективных факторов ученые предлагают различные

способы их распознавания при создании дискурсивных корпусов. Например, TPC дает каждому отношению трехстороннюю характеристику: ядро/сателлит, ограничения и эффект, – чтобы уменьшить неоднозначность в понимании дискурсивных отношений [Mann, Thompson, 1988a]. Фэн Вэньхэ и др. предлагают методы разметки признаков ДО с целью унификации восприятия разных аннотаторов [冯文贺, 徐钰仪, 李青春, 2020]. Кроме того, для уменьшения расхождений в оценках ДО дискурсивное аннотирование обычно проводится многократно или одновременно с участием разных лиц, что максимально обеспечивает качество и достоверность разметки ДО дискурсивного корпуса.

Таким образом, вышеизложенные работы служат теоретической и практической основой для нашего анализа ДО в корпусе китайских и русских официально-деловых текстов.

2.2.4. Дискурсивные коннекторы и их типы

Дискурсивный коннектор (англ. discourse connective, ДК), или по-другому «дискурсивный маркер», «дискурсивная связка», «коммуникативная единица», «средство межфразовых связей», «текстовое связующее», является одной из важных составляющих дискурсивного анализа наряду с дискурсивным отношением. Несмотря на терминологическое разнообразие, все исследователи подчеркивают, что ключевым свойством таких единиц является способность структурировать дискурс, сигнализируя о логико-семантических отношениях между его компонентами. При выборе термина «дискурсивный коннектор» в данной работе основное внимание уделяется его связующей роли (англ. connectivity) – то есть концептуальной основе для выделения ДК как особого класса единиц в дискурсивной грамматике.

Традиционная грамматика, изучающая союз как соединительное средство, фокусируется на грамматической связности того, что мы называем ДК. Согласно Русской грамматике 1980 г., союз определяется как «служебная часть речи, при помощи которой оформляется связь между частями сложного предложения,

между отдельными предложениями в тексте, а также связь между словоформами в составе простого предложения» [Шведова, 1980, с. 712]. Исходя из синтаксической функции союзы в грамматических исследованиях классифицируются на сочинительные и подчинительные. Данное определение и классификация подчеркивают синтаксические связующие функции ДК, но их дискурсивной функции здесь не уделяется достаточного внимания. В то же время выражение грамматических связей не универсально для разных языков, в том числе и в отношении союзных средств [黄伯荣、廖序东, 2017].

ДК стали объектом пристального внимания в исследованиях после становления функциональной грамматики в 1970-х гг. Согласно М. А. К. Халлидею и Р. Хасану, ДК представляют собой один из ключевых механизмов связности текста: их основная функция заключается в установлении логико-семантической связи между предложениями или фрагментами текста с помощью эксплицитных маркеров [Halliday, Hasan, 1976]. Е. А. Золотова выделяет особую группу слов (союзы, союзные слова, частицы, наречия, вводные слова и конструкции), обеспечивающих связность текста и сигнализирующих о смене перспективы, модальности и коммуникативной установки [Золотова, 1973]. Она утверждает, что эти единицы не только объединяют высказывания в связное целое и обеспечивают переходы между фразами и сегментами речи, но и сигнализируют о смене ракурса, модальности или коммуникативной установки. Аналогичные функции дискурсивных единиц отмечаются и в китайской грамматике. По Ляо Цючжуну, дискурсивные единицы функционируют одновременно как организаторы текста и как средства управления интерпретацией, сигнализируя логическое продвижение дискурса [廖秋忠, 1991]. Таким образом, роль ДК по сути заключается в обеспечении структурной организованности и семантической связности дискурса посредством четко выраженных маркеров.

Следует отметить, что ограничение анализа ДК исключительно лексико-синтаксическим уровнем представляется недостаточным для полного понимания их связующей функции. ДК функционируют как маркеры дискурсивной

структуры, обеспечивающие когезию на уровне дискурса, но иногда являются синтаксически факультативными, то есть необязательными с точки зрения структуры предложения – см.: [Fraser, 1999]. Это свидетельствует о том, что их основная функция лежит за пределами синтаксического уровня и связана с обеспечением когерентности текста в целом. Таким образом, для адекватного описания ДК необходимо учитывать их связующую функцию и выход за рамки предложения. Это в свою очередь требует обращения к анализу дискурсивной структуры.

Исследования, посвященные дискурсивному анализу и выполненные с целью обеспечения потребностей автоматической обработки текста, создают структурную основу для изучения ДК. В теории создания PDTB, как было выше сказано, предполагается структурная связь между соседними дискурсивными единицами, а ДК рассматривается как служебное слово дискурсивного уровня и главный показатель выражения ДО [Miltsakaki et al., 2004a; Miltsakaki et al., 2004b]. ДК также были аннотированы при создании структуры связанных клауз [Lyu, Feng, 2023]. В отличие от PDTB, в теории структуры связанных клауз ДК рассматриваются как дополнения к дискурсивной структуре и используются для анализа ДО между дискурсивными единицами.

В данной работе мы будем исходить из того, что структура, обладающая отношениями, которые оформляются с помощью ДК, является основой для разметки и изучения дискурсивной единицы. Иными словами, чтобы изучить ДК, при разметке корпуса необходимо сначала построить дискурсивную структуру, выделить две дискурсивные единицы, а затем найти языковые средства (то есть ДК), которые соединяют их и выражают семантические отношения.

Ряд работ по созданию дискурсивного корпуса с разметкой ДК китайского и русского языков служит непосредственной эмпирической основой для данной работы. В этих работах на основе корпусных данных обобщаются конкретные типы ДК китайского и русского языков. В качестве ДК китайского языка рассматриваются следующие лексические средства: союзы (например, *和*

(и), 而且 (и), 并且 (и), 不但…而且 (не только… но и…), 因为 (потому что), 如果 (если) и т. д.), предлоги (например, *为了 (в целях), 由于 (в связи с), 除了 (кроме), 对于 (по отношению к)* и т. д.), наречия (например, *同时 (одновременно), 还有 (еще), 也 (и), 仅仅 (только),尤其 (особенно)* и т. д.), а также фразы, представляющие собой комплекс упомянутых ДК (например, *同时也 (одновременно и), 正因为 (именно поскольку), 这表明 (это и значит, что)* и т. д.) [李艳翠, 孙静, 周国栋, 2015, p. 310].[1]

В исследовании И. М. Кобозевой и Н. В. Сердобольской, основанном на материале базы Рускон, рассматриваются источники появления новых коннекторов в современном русском языке [Кобозева, Сердобольская, 2024]. В их работе анализируются новые коннекторы, не отмеченные в источниках как союзы, но имеющие метки «модальный уточнитель», «наречие в функции союза», «аналог союза» и др. Выявлены основные источники пополнения коннекторов: абстрактные существительные и образованные от них предлоги (например, модель «в + окатив»: *в отношении что, в той связи что, в результате* и др.; «в + аккузатив»: *в первую очередь, в то время как, в то же время*; «к + датив»: *к примеру, к слову*; и мн. др.); наречия, выражающие временные, частотные и иерархические значения последовательности событий (*далее, иногда, потом, затем*; *реже, чаще; особенно, преимущественно, тем более что*); комплексы союзов и союзов с частицами (*а еще, но ведь*); сочетания союзов с демонстративом «тот» (*из-за того что, в силу того что, по той причине что*) – см. [Там же, с. 68–72].

В данной работе корпусная разметка ДК будет учитывать вышеперечисленные лексические средства. Однако следует подчеркнуть, что в нашей раз-

[1]Русские переводы в скобках иногда имеют только условное семантическое соответствие с китайским ДК; в некоторых случаях найти полноценное грамматическое соответствие между оригиналом и переводом невозможно.

метке определяющим принципом отнесения единицы к разряду дискурсивных коннекторов является ее способность участвовать в организации дискурсивной структуры и выражении семантических отношений. К основным критериям относятся: 1) соединение двух относительно независимых дискурсивных единиц без утраты их синтаксической автономности при удалении коннектора; 2) эксплицитное выражение логико-семантических отношений между соединяемыми единицами.

2.2.5. Дискурсивные вершины и их определения

Как говорилось выше, в п. 2.2.1, особую роль в структуре зависимостей играет вершина, под которой подразумевается та структурая единица, которая в бинарной зависимости выступает в качестве главенствующего компонента [Tesnière, 1959; Hudson, 1990; Hudson, 2007; Mel'čuk, 1988; Мельчук, 1974]. На синтаксическом уровне вершина проявляется в виде главного слова, управляющего зависимым словом в структурной паре и определяющего его грамматическое поведение, а на дискурсивном уровне – в главной ЭДЕ, которая играет определяющую, синтаксически или семантически главенствующую роль в дискурсивной струкутре.

В анализе дискурсивной структуры вершина часто интерпертируется с учетом семантических и прагматических отношений между дискурисвными единицами. Одним из первых формализованных подходов, в котором отчетливо выделяются управляющие и подчиненные дискурсивные единицы, является TPC. Авторы этой концепции подчеркивают необходимость различения риторической важности дискурсивных единиц, участвующих в риторических отношениях. Согласно их утверждению, в таких отношениях семантически доминирующая единица рассматривается как «ядро» (англ. nucleus), то есть «вершина», а другая, периферийная, – «сателлит» (англ. satellite) [Mann, Thompson, 1988a]. Ядро представляет собой носитель основной идеи внутри дискурсивной пары и организует интерпретацию подчиненной единицы (сателлита) и часто выра-

жает риторическую цель автора. В этом случае речь идет не о грамматических, а о риторических и коммуникативных отношениях между дискурсивными единицами. Кроме того, TPC делит риторические отношения на три типа в зависимости от расположения и количества вершин (ядер): «левовершинные» (N-S), правовершинные (S-N) и двухвершинные (N-N) [Там же]. Первые два типа являются асимметричными отношениями (например, отношение причины и отношение результата), а третий – симметричными, например, соединительное отношение.

Следует отметить, что в классической TPC вершина фиксируется в каждом риторическом отношении. Например, в отношениях причины вершина определяется как ЭДЕ, которая объясняет причину; в отношениях результата вершина – это ЭДЕ, которая объясняет результат [Mann, Thompson, 1988a]. Это означает, что для определения вершины необходимо разработать полный список определений каждого возможного типа дискурсивных отношений. Такой подход в отношении дискурсивной разметки приводит к очевидным проблемам. С одной стороны, неопределенность вершин увеличивает выделяемые типы отношений, которые сложно различить. Так, в TPC выделяется четыре типа причинно-следственных отношений: волитивная причина, волитивный результат, неволитивная причина, неволитивный результат (см. Таблицу 2.2). Сложный список отношений на практике может вызвать разногласия между аннотаторами, что в свою очередь влияет на согласованность разметки. С другой стороны, смена типа текста также может породить новые типы отношений, и это, несомненно, усложнит анализ дискурсивной структуры и процесс разметки дискурсивного корпуса, а также приведет к другим дискуссионным вопросам.

С развитием автоматической обработки текста и вычислимых моделей дискурса проблема дискурсивной вершины получила новое осмысление. Современные дискурсивные исследования, опирающиеся на модель зависимостей, стремятся предложить новые критерии для ее определения, такие как языковые,

формальные, логические и вычислимые [Marcu, 1996; Wolf, Gibson, 2005; Li et al., 2014; 吴永芃 et al., 2018; Lyu, Feng, 2023 и др.].

В рамках TPC Д. Марку предлагает определять ядро как дискурсивную единицу, которая «передает наиболее существенное содержание с точки зрения авторского замысла» [Marcu, 1996, p. 4]. Исходя из этого, при определении ядра учитывается значимость (salience) ЭДЕ в тексте: удаление дискурсивных единиц с высокой значимостью (ядра) приводит к серьезному нарушению понимания, тогда как удаление сателлита может оставить общую информативную целостность неизменной. Важно подчеркнуть, что в работе не предлагается формализованный алгоритм автоматического выделения ядра. Вместо этого акцент делается на семантической и прагматической функции ядра как смыслового центра, определяемого через анализ авторского замысла и относительной важности информации.

В работе У. Вольфа и Э. Гибсона предлагается подход к выделению тематически центрированных сегментов (topically centered) [Wolf, Gibson, 2005]. В их работе дискурсивные сегменты группируются по тематическому признаку, и соответственно, центральной единицей может быть признан тот сегмент, вокруг которого строится серия тематически связанных высказываний. Для установления тематической сплоченности учитываются атрибутивные связи: сегменты, ссылающиеся на один и тот же источник информации или выражающие один и тот же подтекст, как правило, подчинены центральному элементу, обобщающему основное содержание. Для выделения центральных единиц также используются структурные и лексические сигналы, такие как сочинительные союзы и другие маркеры организации текста, которые указывают на структурную значимость сегмента, особенно в случаях, когда речь идет о выводах или обобщениях. Структурная организация сегментов в виде иерархии – с выделением главных и подчиненных предложений – также служит ориентиром при интерпретации того, какой сегмент выполняет роль смыслового центра в дискурсе. Хотя авторы не вводят термин «дискурсивная вершина» или «ядро»,

их подход подразумевает индуктивное выявление центральной единицы дискурса путем анализа связей и иерархии содержания.

В исследовании Ли Яньцуй и др. [Li и др., 2014b] дискурсивная ядерная единица способна представлять смысловые связи всего дискурса, устанавливать отношения с другими сегментами и передавать ключевое содержание. На практике это означает, что при создании дискурсивной структуры ядерная единица обладает наибольшей связностью и способна служить опорой для других единиц, то есть функционально и семантически занимает центральную позицию. Авторы подчеркивают, что выбор ядра не исходит из автономной значимости отдельного предложения, а, напротив, опирается на его роль в общей дискурсивной структуре. Важную роль играют коннекторы, которые помогают идентифицировать основные типы дискурсивных отношений (например, следствие, противопоставление, пояснение и т. д.), и тем самым выделить вершинную единицу, которая структурно и логически доминирует над другими. Важно отметить, что их модель представляет собой синтез структурной и семантической оценки при выборе дискурсивной вершины с акцентом на глобальное значение сегмента в пределах всего текста, а не на его локальную релевантность.

Китайские исследователи У Инопэна и др. [吴永芃 и др., 2018] в рамках дискурсивного дерева зависимостей определяют центральное предложение (дискурсивную вершину) с опорой на структурные и семантические характеристики текста. По их утверждению, центральное предложение занимает важную позицию в тексте (например, начальное или обобщающее предложение), и в дереве зависимостей оно часто выступает в роли корневого узла или узла с наибольшим числом зависимостей. Кроме того, авторы учитывают содержательную нагрузку предложения: насколько этот сегмент выражает ключевую мысль, инициирует последующие сегменты или подводит итоги. Предложения с высокой информативностью и функцией смыслового фокуса рассматриваются как кандидаты на роль дискурсивного ядра. Таким образом, исследователи вычисляют показатели важности предложения – например, количество зависимых

элементов, семантическую насыщенность, степень участия в дискурсивных отношениях, которые позволяет автоматически или полуавтоматически идентифицировать вершины в дискурсе.

При создании китайской дискурсивной структуры связанных клауз Фэн Вэнхэ и др. указывают, что «в китайском языке отсутствуют четкие грамматические маркеры для определения главного или подчиненного предложения» [Lyu, Feng, 2023, p. 83], в результате чего традиционные методы становятся затруднительными и могут привести к ошибкам при идентификации дискурсивной головы. В связи с этим авторы предлагают сосредоточиться на семантической взаимосвязи между клаузами, не прибегая к предварительному определению «головы». Как отмечают авторы, «только после определения всей релевантности клаузы мы можем определить дискурсивную голову на основе значимости ЭДЕ в глобальной дискурсивной структуре» [Там же, p. 86]. Исходя из этого они предложили вычислительный подход к идентификации вершин – меру семантической связанности на основе глобальной структуры, то есть количества связей зависимостей двух ЭДЕ, входящих в одну структурную пару (при этом ЭДЕ с бóльшим количеством связей является вершиной данной структурной пары). В итоге вершины отражают семантическую важность одной ЭДЕ по отношению к другой, что определяется количеством связей одной ЭДЕ с другими в глобальной дискурсивной структуре.

Таким образом, динамические дискурсивные вершины, определяющие в дискурсивной структуре, включая корневой узел, выступают ключевыми элементами, позволяющими выявить глобальную семантическую связанность текста. Ведь именно через перемещение вершины между предыдущей и следующей ЭДЕ становятся видны относительные семантические связи, которые иначе могли бы остаться скрытыми при фиксированной локальной структуре. В дискурисвном фрагменте (2.1) такое распределение подтверждается: некоторые отношения детализации (Elaboration) ведут к вершинам, находящимся в предыдущей ЭДЕ, а некоторые – в следующей ЭДЕ. Итак, дискурсивная вершина

в конкретном отношении не имеет фиксированную позицию, а определяется взаимной релевантностью ЭДЕ в глобальной структуре текста. Семантическая значимость той или иной единицы может варьироваться в зависимости от ее участия в более широких дискурсивных связях, а не только от локального порядка или формальных маркеров. Именно поэтому, на наш взгляд, вычислительный подход к идентификации вершин, предложенный Фэн Вэнхэ и др., позволяет учитывать вариативность когезионных связей и адаптировать метод к различным языкам. В данной работе мы используем вычислительный подход для определения дискурсивной вершины.

Итак, сформулируем главные признаки дискурсивных вершин, а также основные ориентиры для разметки главных ЭДЕ.

1. Дискурсивная вершина определяет семантически главную ЭДЕ в бинарном дискурсивном отношении и играет важную роль в понимании структурного дерева дискурса.

2. Семантическая важность вершины определяется посредством измерения и сравнения количества семантических связей в глобальной дискурсивной структуре участвующих ЭДЕ, входящих в одну структурную пару.

3. Вершина выделяется в каждой структурной паре; даже в случае симметричных отношений ЭДЕ, таких как соединение, мы сможем выявить вершину с помощью измерения их глобальной семантической связанности.

4. В корпусе при формировании дискурсивных деревьев ребра со стрелками зависимостей должны быть направлены от вершины к периферии.

Выводы по второй главе

В главе рассмотрена теоретическая база создания параллельного дискурсивного корпуса. Представлены две основные модели в виде составляющих и зависимостей; исходя из этого существующие теории дискурсивного анализа подразделены на четыре категории: дискурсивная структура составляющих, представленная теорией риторической структуры [Mann, Thompson, 1988a]; ло-

кальная дискурсивная структура зависимостей, представленная моделью PDTB [Miltsakaki et al., 2004a; Prasad et al., 2004; Prasad et al., 2005; Prasad et al., 2008a; Webber et al., 2019]; глобальная дискурсивная структура зависимостей, представленная теорией структуры связанных клауз [冯文贺 et al., 2020; Lyu, Feng, 2023] и смешанные типы структурного моделирования дискурса.

Способ представления зависимостей является более простым, гибким и способен эффективно описывать сложные дискурсивные явления, что подтверждается следующими аргументами: 1) способ зависимостей не ограничен линейным следованием языковых единиц и может быть использован для описания дискурсивных единиц с более свободным порядком; 2) структура зависимостей содержит только элементы и ребра, что упрощает процесс анализа сложных языковых структур по сравнению со способом составляющих; 3) структура зависимостей приспособлена к изучению кросс-языковых и кросс-уровневых языковых явлений; 4) описание дискурсивной структуры зависимостей способствует обработке естественного языка на более высоком уровне, т. к. способ зависимостей как разновидность сетевой структуры имеет много общего с нейросетевыми технологиями; 5) разметка структуры зависимостей может быть выполнена в общедоступном программном обеспечении, что значительно облегчает лингвистическое исследование; 6) проекты создания дискурсивного корпуса в рамках концепции зависимостей показали высокий уровень совпадения экспертных оценок.

В главе подробно рассмотрены основные требования к формированию дискурсивной структуры зависимостей и ключевые понятия, необходимые для разметки дискурсивной структуры. Клаузы при создании дискурсивной структуры зависимостей рассматриваются как элементарные дискурсивные единицы, которые могут быть выделены в письменном тексте с учетом знаков препинания. Суть зависимости заключается в семантических отношениях, которые возникают между двумя элементарными дискурсивными единицами.

Дискурсивные коннекторы являются функциональными словами или сочетаниями слов на дискурсивном уровне: они фактически играют роль структурных связок и семантических индикаторов в формировании дискурса. Анализ дискурсивных коннекторов должен основываться на анализе всей дискурсивной структуры.

Наконец, в каждой структурной паре ЭДЕ существует своя дискурсивная вершина, которая определяет семантически наиболее важную ЭДЕ в бинарном отношении. Статус наиболее важной ЭДЕ подтверждается количеством связей, участвующих в глобальной дискурсивной структуре зависимостей, и может быть объективирован вычислительным путем.

ГЛАВА 3. ОПЫТ СОЗДАНИЯ КИТАЙСКО-РУССКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА ОФИЦИАЛЬНО-ДЕЛОВЫХ ТЕКСТОВ

С учетом общих принципов теории зависимостей [Tesnière, 1959; Hudson, 1984; 刘海涛, 1991; Robinson, 1970; Gaifman, 1965; Hays, 1964] и опыта моделирования дискурсивных структур [Marcu, 1996; Webber, Joshi, 1998; Carlson, Marcu, Okurowski, 2003; Prasad et al., 2008a; Zhou, Xue, 2012; Li et al., 2014b; Poláková et al., 2013; Мухин, Ян, 2016; Zhou, Xue, 2015b; Chistova et al., 2021], при осуществлении данного диссертационного исследования был создан параллельный дискурсивный корпус китайских и русских официально-деловых текстов, в котором размечена структурная информация о дискурсивных единицах и отношениях между этими единицами. Цель главы – конкретизировать принципы дискурсивной разметки и сопоставительного дискурсивного анализа, обеспечить прозрачность и верифицируемость разметки и полученных данных, а также в целом обобщить практический опыт создания данного параллельного дискурсивного корпуса.

В параграфе 3.1. приведены критерии отбора материала для создания параллельного корпуса. Параграф 3.2. посвящен основным принципам разметки, включая деление текстов на ЭДЕ и их выравнивание, модели структурной организации и разметки ДО и ДК; здесь же обобщены основные проблемы разметки и их решения. В параграфе 3.3. описано программное обеспечение, а в параграфе 3.4. представлены формат хранения данных и информация о качестве разметки данного корпуса.

3.1. Этап отбора материала для корпуса

В качестве материала нашего исследования были выбраны двусторонние документы правительств РФ и КНР, выпускаемые регулярно с 1994 г., включая совместные декларации и заявления глав государств, такие как совместная российско-китайская декларация (3 сентября 1994 г.), пекинская декларация Российской Федерации и Китайской Народной Республики (18 июля 2000 г.), мос-

ковское совместное заявление глав государств России и Китая (16 июля 2001 г.) и другие аналогичные документы. На данный момент размечены 24 документа, 12 документов на русском и 12 на китайском языке (далее – документы, перечень см. в Приложении 1). Двенадцать русских документов состоят из 429 абзацев (общий объем – 22 602 текстоформы); китайские состоят также из 429 абзацев (общий объем – 41 784 слова). Корпус не является большим по объему, но при этом он достаточно репрезентативен для собственного анализа дискурсивной структуры привлеченных официально-деловых текстов.

При отборе документов как исследовательского материала учитывались следующие факторы. Во-первых, материал соответствует стандартам письменного изложения и адекватно передает межъязыковые семантические соответствия, что позволяет избежать проблем, связанных с возможными «ошибками» перевода при обсуждении результатов исследования. Во-вторых, создаваемые документы носят периодический характер. С момента объявления об основах взаимоотношений между РФ и КНР в 1992 г. главы двух государств регулярно встречаются и почти ежегодно подписывают аналогичные документы с 1994 г. Однотипные тексты демонстрируют стабильность в словоупотреблении, синтаксических конструкциях и структурной организации, что впоследствии способствует ручной разметке и будущей автоматической обработке структуры дискурса. В-третьих, в отличие от других дипломатических текстов, совместные декларации и заявления публикуются для изучения широкой международной аудиторией, что делает их более доступными и открытыми для анализа.

Особенно важно, что «параллельные» тексты таких жанров, как совместное заявление и совместная декларация, обладают семантической эквивалентностью и не рассматриваются в качестве оригинала и перевода. Если мы признаем какой-либо текст переводным, то это говорит о неполном равноправии оригинального и переводного документов. Однако совместные заявления и совместные декларации оформляются и согласовываются параллельно. Как отмечает К. А. Бекяшев, «текст двустороннего договора чаще всего составляют

на языках обеих договаривающихся сторон. Оба языковых варианта признаются при этом аутентичными, то есть имеющими одинаковую силу, равно подлинными» [Бекяшев, 2020, с. 194]. Таким образом, привлеченные тексты принципиально отличаются от переводных, которые представляют собой трансформацию с одного языка на другой с четким направлением. Конечно, мы должны признать, что такие тексты могут обладать определенными переводческими свойствами (с точки зрения цели эквивалентной передачи смысла), но такая выборка позволяет минимизировать или вообще исключить обсуждение языковых проблем, связанных переводом, таких как качество перевода или ошибки перевода.

3.2. Принципы дискурсивной разметки и выравнивания

Дискурсивная структура в данной работе представляется в форме направленного ациклического графа в рамках концепции зависимостей; в этой структуре абзацы рассматриваются как единое целое. Аналитическая процедура включает следующие этапы:

1) деление китайских и русских параллельных текстов на ЭДЕ;

2) установление структурных пар (зависимостей между двумя ЭДЕ) и построение модели структуры зависимостей;

3) разметка характеристик структурных пар, в том числе синтаксических вариантов соотношения двух единиц в составе структурных пар, семантических типов дискурсивных отношений, дискурсивных коннекторов и дискурсивных вершин.

3.2.1. Принципы деления на ЭДЕ и выравнивания текстов

Деление на ЭДЕ и выравнивание текстов в будущем параллельном корпусе является первым шагом построения дискурсивной структуры. В уже сложившейся традиции разметки параллельных дискурсивных корпусов выравнивание обычно начинается с деления на клаузы в тексте оригинала, после чего

в переводном тексте выделяются соответствующие синтаксические аналоги клауз исходного текста [冯文贺, 2019; Мухин, Ян, 2016]. Однако в применении к нашему сегодняшнему материалу такая идея в чистом виде, без дополнения непригодна: в создаваемом параллельном корпусе нет как такового направления перевода. В связи с этим для осуществления выравнивания аннотатор должен сам определить, в какой части корпуса (китайской или русской) он проводит деление на клаузы, а в какой выделяет синтаксические аналоги.

В данной работе принято решение осуществлять деление на клаузы в китайских текстах, а затем обозначать соответствующие китайским клаузам русские синтаксические аналоги (далее – РСА). Такое решение обусловлено следующими соображениями. Во-первых, работа с китайскими текстами облегчает аннотаторам – носителям китайского языка определить структурные и пропозициональные свойства и критерии деления текста на клаузы. Во-вторых, изучение синтаксических структурных особенностей китайских клауз и РСА более приемлемо для пользователей – носителей русского языка, поскольку, независимо от своей формы и объема, РСА всегда соответствуют китайским клаузам, которые имеют устойчивые синтаксические признаки.

3.2.1.1. Деление на ЭДЕ в китайском тексте

Понимание ЭДЕ в китайском тексте в данной работе во многом совпадает с определением термина «клауза» в работах Син Фуи, Ли Яньцуй и Фэн Уэньхэ и др.: клауза содержит как минимум одну предикативную группу и является простым предложением или конструктивной частью сложного предложения; каждая клауза соответствует одному событию или ситуации и является самостоятельной единицей, которая не служит грамматическим компонентом других клауз; клауза заканчивается знаком препинания [Син Фуи, 2020; Li et al., 2014b; Lyu, Feng, 2023]. Однако в данной работе не предъявляются строгие требования к пунктуации. С одной стороны, наше исследование посвящено лингвистическому изучению функций дискурсивных единиц, а не разработке прин-

ципов их автоматической разметки; с другой стороны, в целях сопоставления ЭДЕ китайские клаузы должны обладать относительно фиксированной синтаксической структурой.

Таким образом, клауза в нашей работе понимается как минимальная синтаксическая, семантическая и функциональная единица дискурса. В ходе создания нашего параллельного корпуса были сформированы следующие конкретные принципы сегментации китайских клауз:

1. Основными синтаксическими типами ЭДЕ являются простые предложения и простые предложения в составе сложных предложений, которые признаются основными синтаксическими единицами в традиционной грамматике китайского языка.

1.1. Все простые предложения выделяются как китайские ЭДЕ. Каждое простое предложение имеет один и только один предикат и заканчивается знаком препинания. Например, в дискурсивном фрагменте (3.1) представлено типичное китайское предложение, в котором слово «*趋势 (тенденция)*» – подлежащее, а слово «*发展 (развиваться)*» – сказуемое.

(3.1)

C96-49. 世界多极化趋势在发展。/[1]

R96-49. Развивается тенденция к многополярности мира. /

1.2. Все простые предложения в составе сложных предложений выделяются как китайские ЭДЕ. В большинстве случаев между простыми предложениями в составе китайских сложных предложений ставится запятая. Они как

[1] Буква и цифры обозначают одну выделенную ЭДЕ: буква «C» – китайскую клаузу, «R» – соответствующую РСА, а цифры указывают год и порядковый номер ЭДЕ в документе (например, «96-49» – сорок девятую ЭДЕ в китайско-российской совместной декларации 1996 г.). Для облегчения восприятия текста читателями, чей родной язык не является китайским, в примерах китайского языка предикат выделен точками снизу.

единые целые не участвуют в составе других синтаксических компонентов, но иногда имеют общий компонент предложения (обычно подлежащее). Например, в (3.2) показано китайское сложносочиненное предложение, состоящее из конструктивных частей 96-111 и 96-112. Две части имеют общее подлежащее «*双方(Стороны)*», но по синтаксической структуре они являются независимыми компонентами, поэтому рассматриваются как две отдельные ЭДЕ.

(3.2)

С96-111. 双方认为，环境保护已成为一个全球性的重要问题，/

С96-112. 并且决心为此加强双边和多边的合作。/

R96-111. Стороны считают, что охрана окружающей среды стала важной глобальной проблемой, /

R96-112. и полны решимости укреплять двустороннее и многостороннее сотрудничество в этих целях. /

1.3. Длинное обстоятельство причины, цели и условия, расположенное в начале предложения, рассматривается как ЭДЕ, поскольку оно обычно выделяется запятыми и в китайской традиции считается самостоятельной единицей, обладающей предикативной конструкцией: если опустить вводный союз (например, 为, 由于, и др.), такие конструкции, как правило, сохраняют форму завершенного высказывания, что позволяет рассматривать их как ЭДЕ. Например, в (3.3) ЭДЕ 94-13 является обстоятельством цели, а ЭДЕ 94-13 – обстоятельством условия.

(3.3)

С94-13. 为进一步确立新型的相互关系，/

С94-14. 从两国关系的远景出发，/

С94-15. 双方决心采取积极和全面的步骤：/

R94-13. В целях дальнейшего утверждения нового качества своих взаимоотношений /

R94-14. и исходя из долгосрочных перспектив отношений между двумя странами, /

R94-15. Стороны преисполнены решимости предпринимать активные и разносторонние шаги: /

2. При выделении ЭДЕ мы имеем в виду следующие знаки препинания: точку (。), точку с запятой (；), вопросительный знак (？), восклицательный знак (！) и запятую (，). Первые четыре можно однозначно использовать в качестве разделителей ЭДЕ, а запятую вероятностно, потому что ее основная функция в китайском языке – соотнесение с паузами внутри предложения, а не просто выделение синтаксической единицы. Итак, структурные единицы, выделяемые запятыми, не всегда можно квалифицировать как ЭДЕ. На основе конкретных примеров, встретившихся в корпусе, приведем случаи, когда запятая не признается нами разделителем ЭДЕ.

2.1. В длинных предложениях запятая иногда ставится как знак паузы для упрощения чтения – см.: [Горелов, 1982, с. 253–254]. При этом удаление такой запятой не нарушает смысловой и синтаксической целостности конструкции. Как показано в дискурсивном фрагменте (3.4), в простом предложении предикативная группа «*中国和俄罗斯 (Китай и Россия) …发展(развитие) …合作(сотрудничества)*» перед запятой играет роль подлежащего, глагол-связка «*是(являться)*» – сказуемого, а «*扩大方向之一 (одним из направлений)*» – именной

части сказуемого[1]. Запятая здесь ставится для разделения длинного подлежащего. Запятая в конце подлежащего используется исключительно как средство графической разгрузки длинной конструкции.

(3.4)

C001-54 中国和俄罗斯进一步全面综合发展经贸、科技和军技领域的合作，是扩大和深化中俄平等信任的战略伙伴关系的重要方向之一。/

R001-54 Дальнейшее развитие всестороннего комплексного сотрудничества между Россией и Китаем в торгово-экономической, научно-технической и военно-технической областях является одним из главных направлений углубления и расширения российско-китайских отношений равноправного доверительного партнерства и стратегического взаимодействия. /

2.2. Запятая не рассматривается как маркер ЭДЕ, когда она отделяет находящееся в начале предложения обстоятельство времени или места. В дискурсивном фрагменте (3.5) запятая ставится после обстоятельства времени «*当前 (сегодня)*», а в дискурсивном фрагменте (3.6) – после обстоятельства места «*在……领域 (в энергетической сфере)*».

(3.5)

当前，世界正在经历大变局，/……

Сегодня мир переживает масштабные перемены, /

---

[1] Следует отметить, что подобные конструкции в китайском и русском языках анализируются по-разному. В примере (3.4) в русском переводе глагол-связка «является» трактуется как вспомогательный глагол в составе составного именного сказуемого, а в китайском предикат «是» соединяет подлежащее и номинативный компонент без необходимости дополнительной структурной перестройки. В китайском языке возможно использование сложных именных групп в позиции подлежащего без нарушения синтаксической целостности. Такие структуры типичны для письменного китайского языка и считаются завершенными, несмотря на их протяженность. См. также: [Горелов, 1982, с. 253–254].

(3.6)

在能源领域，中俄双方均采取多元化战略。/

Стороны реализуют стратегию диверсификации в энергетической сфере. /

2.3. Если запятая стоит после союза в начале предложения, она также не является маркером клаузы, как в дискурсивном фрагменте (3.7).

(3.7)

因此，中国和俄罗斯坚决反对这一计划。

Поэтому Россия и Китай решительно выступают против такого плана.

Как было отмечено выше, наличие знаков препинания не является строгим требованием к выделению ЭДЕ – в процессе работы некоторые предложения без запятой также были разделены на ЭДЕ. Например, предложение в дискурсивном фрагменте (3.8) не содержит запятой, но в нем есть две независимые предикативные группы, которые рассматриваются как ЭДЕ.

(3.8)

C96-44 重申愿意保持军事技术合作应有的透明度 /

C96-45 并向联合国常规武器转让登记制度提供有关信息 。

R96-44 подтверждают готовность поддерживать надлежащую транспарентность ВТС

R96-45 и предоставлять соответствующую информацию в Регистр обычных вооружений ООН.

В нашем корпусе китайская клауза определяется исходя из трех аспектов: синтаксического, семантического и функционального. Деление текстов на ЭДЕ

требует, чтобы сегменты были единицами с цельной синтаксической структурой, завершенным смыслом и выраженной функцией.

3.2.1.2. Деление на ЭДЕ и их характеристика в русском тексте

Деление русскоязычного текста на ЭДЕ осуществляется в соответствии с китайскими клаузами. Соответственно здесь в качестве ЭДЕ выделяются русские синтаксические аналоги (РСА).

По естественным причинам набор словоформ и устойчивых выражений в результате деления китайских и русских текстов на ЭДЕ не совпадает, и это значит, что РСА могут быть не только клаузами в традиционном понимании русской грамматики, но и единицами различного объема: фрагментами предложения, обособленными членами, группами предложений и др. Для того чтобы понять, какие синтаксические единицы русского языка становятся эквивалентами китайских клауз, необходимо их описать и каталогизировать.

На практике выделенные РСА настолько разнообразны по своей синтаксической структуре, что иногда трудно определить их в терминах русской грамматики. Однако на самом деле нас больше интересует специфика предикативности РСА, а не их конкретные синтаксические названия, поскольку каждый РСА соответствует китайской клаузе, содержащей одну и только одну предикатную группу. По специфике предикативности можно выделить четыре типа РСА:

1) монопредикативные РСА, к которым относятся: а) простые предложения, б) простые предложения в составе сложных предложений, в) конструкции

с каждым из однородных сказуемых при одном подлежащем[1] и г) выделяемые фрагменты предложения, содержащие грамматическую основу;

2) РСА без предиката, к которым относятся все фрагменты предложения без предиката или предикативной группы;

3) полупредикативные РСА, к которым относятся причастные и деепричастные обороты;

4) полипредикативные РСА, к которым относятся сложные предложения и группы предложений, а также предложения с деепричастным или причастным оборотом и предложения с несколькими сказуемыми при одном подлежащем.

Приведем пример дискурсивного фрагмента (3.9) из документа «Совместная декларация Российской Федерации и Китайской Народной Республики (26 марта 2007 г.)», в котором выделено четыре РСА. К монопредикативной единице относятся *PCA96-81, PCA96-83, PCA96-86, PCA96-87* и *PCA96-88*, к РСА без предиката – *PCA96-84* и *PCA96-85*, к полупредикативной – *PCA96-80* и к полипредикативной – *PCA96-82*.

(3.9)

C96-80. 双方同意，在提高联合国的效率和行动能力方面加强合作。/

C96-81. 双方指出，联合国对维护国际和平与安全做出了贡献。/

C96-82. 双方认为，联合国是为和平、发展、安全进行合作的独特的机制，/

C96-83. 肩负迎接二十一世纪全球性挑战的使命；/

C96-84. 为适应业已变化的国际形势，/

---

[1] В данной работе предложения с несколькими сказуемыми при одном подлежащем рассматриваются как сложные, поэтому отдельные конструкции в этих предложениях рассматриваются как монопредикативные РСА. Такое понимание соответствует «Русской грамматике» 1980 г. [Русская грамматика: синтаксис, 1980, с. 461–462]. Подобная точка зрения встречается также в китайской грамматической традиции [冯文贺, 李青青, 2022; 宋柔, 2012].

C96-85. 提高工作效率，/

C96-86. 联合国及其机构应进行适当的改革，/

C96-87. 以更好地履行联合国宪章所赋予的职责；/

C96-88. 联合国的工作及其决策过程应更好地体现联合国全体成员国的共同愿望和集体意志。

R96-80. Отмечая вклад ООН в дело поддержания международного мира и безопасности, /

R96-81. Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности. /

R96-82. Стороны считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности, /

R96-83. что на ее плечах лежит миссия дать ответ на глобальные вызовы XXI века; /

*R96-84. **в целях** адаптации к изменившейся международной обстановке /*

R96-85. и повышения эффективности работы ООН /

R96-86. необходимо провести соответствующую реформу ООН и ее органов, /

*R96-87. **что** позволило бы им еще лучше исполнять обязанности, предусмотренные Уставом ООН; /*

R96-88. деятельность ООН и процесс принятия ее решений должны еще лучше отражать общие чаяния и коллективную волю всех стран – членов ООН.

Ниже на материале данного текста будет продемонстрирован общий подход к построению дискурсивной зависимости, а также конкретные принципы, лежащие в его основе.

3.2.1.3. Предварительное деление китайских и русских текстов на ЭДЕ с помощью знаков препинания

Задача предварительной обработки текстов заключается в автоматизированном делении текстов на ЭДЕ. Тексты предварительно обрабатываются перед тем, как начать разметку структурной информации, что упрощает дальнейшую корректировку сегментации и структурный анализ текстов.

Предварительное деление текстов на ЭДЕ осуществляется с помощью опций поиска и замены в текстовых редакторах, поддерживающих регулярные выражения – таких как *Microsoft word* и *Editplus*. Сначала производится сплошная обработка ЭДЕ китайского языка, затем сплошная обработка ЭДЕ русского языка и – далее – ручная проверка сегментации.

На первом этапе мы в программе Microsoft Word с помощью инструмента «поиск и замена текста» делим тексты на ЭДЕ, используя знак «/» как разделитель ЭДЕ, заменяем на него знаки препинания в китайском языке, а именно точки (。), восклицательные знаки (！), вопросительные знаки (？), запятые (，), точки с запятой (；) и двоеточия (：); в русских текстах – точка (.), двоеточие (:) и точка с запятой (;)[1].

Так как полученные после автозамен сегменты не всегда являются ЭДЕ в нашем понимании – например, члены предложения, разделенные запятой в китайском языке (см. в подпараграфе 3.2.1.1), или инициалы (*В. В. Путин*), мы импортируем результаты первого предварительного деления в программу *EditPlus* и с помощью правил регулярных выражений выполняем сплошное преобразование. В китайских текстах при этой процедуре сегменты слова и словосочетания объединяются в бóльшие единицы; в русских текстах исправляются точки, которые используются при инициалах личных имен и отчеств.

---

[1] Обращаем внимание на то, что в русских текстах запятая не является точным разделителем ЭДЕ, потому что она может ставиться не только между клаузами, но и между словами или словосочетаниями.

Ручная проверка и выравнивание ЭДЕ осуществляются в программе *Microsoft Excel*, в итоге формируется нумерованный список ЭДЕ в китайских и русских текстах.

3.2.2. Принципы установления структурных пар и создания оптимальной дискурсивной структуры зависимостей

На втором этапе построения дискурсивной структуры устанавливаются зависимости между двумя ЭДЕ, образующими структурные пары. Основным критерием для формирования этих пар служит наличие логико-семантической связи между двумя клаузами в пределах абзаца. На рис. 3.1 представлены структурные пары в дискурсивном фрагменте (3.9) – см. дискурсивный фрагмент выше. ЭДЕ расположены в отдельных строках, структурные пары обозначены ребрами. Видно, что в дискурсивном дереве зависимостей корпусного фрагмента выделено 8 структурных пар.

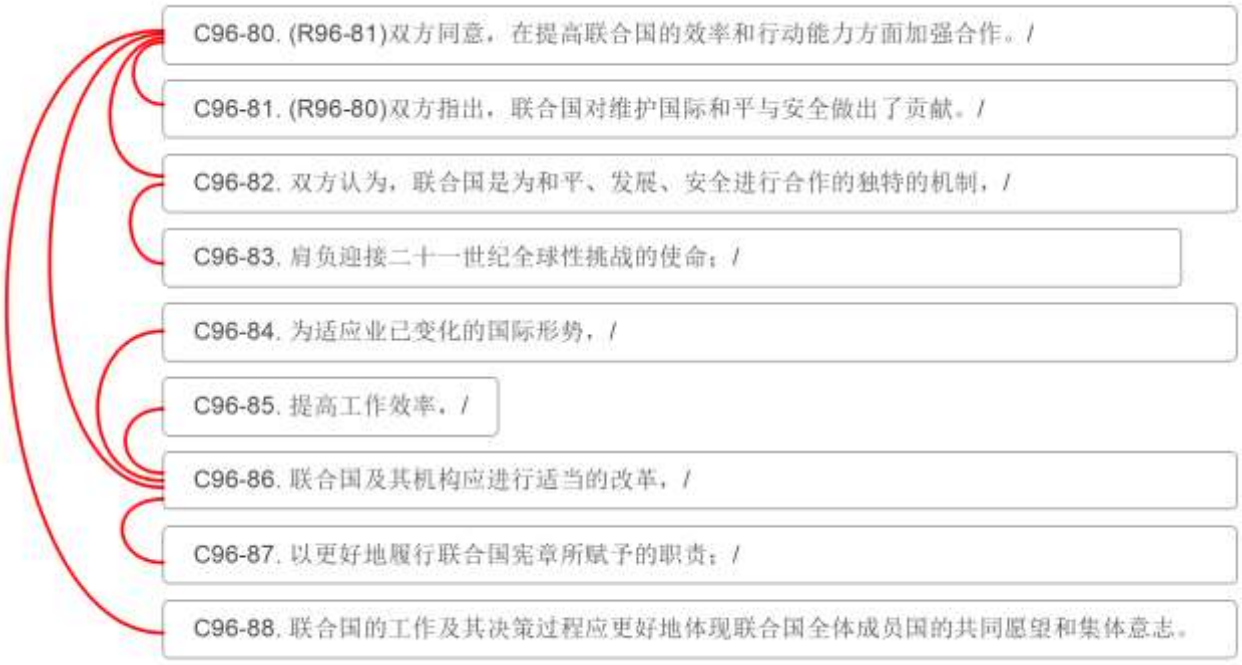


Рисунок 3.1. Структурные пары в текстах (3.9).

Установление дискурсивных зависимостей основывается на специфике когерентности (coherence) и когезии (cohesion) [Halliday, Hasan, 1976]. Два ЭДЕ в одной структурной паре должны быть семантически связаны друг с другом (на этапе выделения структурных пар неважно, какими являются конкретные семантические отношения). Для проверки семантической связанности между двумя ЭДЕ применяется принцип «минимального сложного предложения», предложенный Фэн Вэньхэ [冯文贺 et al., 2020; Lyu, Feng, 2023], согласно которому элементы структурной пары могут быть переформированы в минимальное сложное предложение посредством минимальной модификации, включающей добавление или удаление союзов, подлежащего или знаков препинания. В (3.10) приведены некоторые структурные пары текстов (3.9), которые при определенных условиях образуют минимальные сложные предложения. В скобках приво-

дятся удаленные или добавленные элементы, которые необходимы для удовлетворения формальных требований к сложным предложениям.

(3.10)

R96-81 – R94-82: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности* ***(. Стороны , поскольку они)*** *считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности* ***(, .)***

R96-81 – R94-86: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности* ***(, для этого)*** *необходимо провести соответствующую реформу ООН и ее органов* ***(, .)***

R96-81 – R94-88: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности* ***(, чтобы)*** *деятельность ООН и процесс принятия ее решений должны еще лучше отражать общие чаяния и коллективную волю всех стран – членов ООН.*

…

Следует отметить, что, если применять принцип «минимального сложного предложения» к любым двум ЭДЕ в дискурсе, можно в результате получить циклическую структуру, содержащую множество перекрестных связей (подробнее об интерпретации ограничения проективности см. в п. 2.2.1.). Как правило, многие из этих перекрестных связей не столь важны для идентификации и понимания дискурсивной структуры. Например, на рис. 3.1 видно, что была исключена структурная пара R*96-84* – R*96-85*, поскольку обе ЭДЕ связаны с R*96-85* в дискурсивной структуре этого фрагмента. Это значит, что для создания ациклического цикла, в котором зависимости не должны пересекаться, следует оптимизировать структурные пары – сохранить необходимые для формирования ациклического цикла и убрать лишние.

С этой целью мы разработали принципы подбора структурных пар для оптимизации дискурсивной структуры:

Принцип 1. Приоритетно сохранение структурной пары с эксплицитным коннектором, поскольку мы считаем, что коннекторы устанавливают прямую и прозрачную структурную связь между двумя ЭДЕ.

Принцип 2. Приоритетно сохранение структурной пары с полной ЭДЕ, а именно формально и семантически завершенной единицы; при этом предпочтение в образовании структурных пар с другими ЭДЕ отдается основной ЭДЕ.

Принцип 3. Возможно удаление структурных пар с соединительными отношениями. Если между двумя ЭДЕ, входящими в одну структурную пару, имеется соединительное отношение, и обе они структурно связываются с другой или другими ЭДЕ (например, в структурной паре 96-84 – 96-85 обе ЭДЕ96-84 и ЭДЕ96-85 связываются с ЭДЕ96-86 в текстах (3.9)), то текущая структурная пара может быть удалена.

3.2.3. Принципы разметки дискурсивных параметров структурных пар

Последним шагом в создании фрагмента параллельного дискурсивного корпуса является разметка сложившихся структурных пар, которая отражает их дискурсивную специфику и создает базу для лингвистического дискурсивного анализа и сопоставления двух языков. В корпусной разметке определяются синтаксические варианты соотношения двух единиц в составе структурных пар, семантические типы (то есть типы дискурсивных отношений), средства связи (то есть дискурсивные коннекторы), а также вершины. Рассмотрим далее эти характеристики.

3.2.3.1. Синтаксические варианты соотношения единиц в составе структурных пар

Разметка синтаксических вариантов соотношения единиц структурных пар (далее – синтаксических вариантов) отражает то, что две ЭДЕ, составляющие структурную пару, находятся в одном завершенном предложении или в разных (то есть графически разделены запятой или точкой). Цель разметки

синтаксических вариантов заключается в том, чтобы лучше отразить структурные особенности дискурса и создать основу для дальнейшего этапа анализа ДО.

Исходные соотношения между двумя китайскими клаузами в составе структурных пар проявляются в двух случаях: это или две клаузы в одном сложном предложении (то есть ЭДЕ китайского языка являются простыми предложениями в составе сложного предложения), или две клаузы в разных предложениях. Так как китайским клаузам могут соответствовать в русском тексте синтаксические единицы различного типа – РСА, соотношения между РСА в составе структурных пар могут быть трех типов: два РСА в сложном предложении, например, в тексте (3.9) в структурной паре *R96-82 – R96-83: Стороны считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности, / что на ее плечах лежит миссия дать ответ на глобальные вызовы XXI века;* , два РСА в разных предложениях, например, в тексте (3.9) в структурной паре *96-81 – 96-82*: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности. / Стороны считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности*, а также два РСА внутри простого предложения, например, в дискурсивном фрагменте (3.9) в структурное паре *R96-84 – R96-86*: *в целях адаптации к изменившейся международной обстановке / необходимо провести соответствующую реформу ООН и ее органов…*

Соотношения, установленные между парами ЭДЕ в одном сложном предложении и в разных предложениях, безусловно, значимы и на дискурсивном уровне. Однако связи между РСА внутри простого предложения зачастую представляют собой чисто синтаксические отношения, такие как связь подлежащего и сказуемого, которая не соотносится с самостоятельной дискурсивной функцией. Грамматическая основа, безусловно, является важной частью предложения и текста, но сама по себе она не формирует дискурсивных отношений, поскольку дискурс подразумевает взаимодействие на уровне смысловых

и коммуникативных связей между более крупными единицами текста. Например, в (3.11) между РСА *R05-9* и *R05-10* чисто синтаксическая связь между подлежащим и сказуемым. Эти синтаксические варианты вторичны в параллельном дискурсивном корпусе из-за первичной сегментации ЭДЕ в китайском тексте.

(3.11)

C05-9. 世界多极化和经济全球化作为当前人类发展阶段的重要趋势，/

C05-10. 其发展进程存在不平衡和矛盾的现象。/

R05-9. Процессы становления многополюсного мироустройства и экономической глобализации, являющиеся важными тенденциями современного этапа развития человечества, /

R05-10. протекают неравномерно и противоречиво. /

Мы, конечно, не отрицаем, что два РСА внутри простого предложения также могут быть семантически связаны (например, связь между *R96-84* и *R96-86*), но эти связи должны признаваться как синтаксические отношения внутри простого предложения, а не на уровне дискурса.

Итак, мы утверждаем, что эти синтаксически связанные структурные пары должны фигурировать в нашей модели с целью полного описания и сопоставления дискурсивной структуры текстов на разных языках, но они не должны рассматриваться при последующем обсуждении дискурсивных отношений и дискурсивных коннекторов (поскольку последние являются признаками уровня дискурса). Однако мы учитываем, что структурные пары с двумя РСА внутри простого предложения могут быть важны для сопоставительного анализа и что в лингвистике серьезно обсуждаются механизмы перехода с межсентенциальных связей в одном языке во внутрисентенциальные в другом языке. Поэтому для синтаксического варианта, предполагающего наличие двух РСА внутри простого предложения, дополнительно выделим следующие подтипы:

1. РСА, находящиеся внутри простого предложения без предлога, а именно РСА подлежащего – сказуемого и РСА сказуемого – дополнения.

2. РСА, находящиеся внутри простого предложения с предлогом.

3. РСА, находящиеся внутри простого предложения с деепричастным или причастным оборотом.

3.2.3.2. Семантические типы дискурсивных отношений и дискурсивных коннекторов

В структуре зависимостей дискурсивные отношения (зависимости) понимаются как логико-семантические связи между ЭДЕ. Предлагаем выделять следующие типы дискурсивных отношений для разметки официально-деловых текстов: причина – следствие, цель, условие, уступка, время, оценка, соединение, дополнение, пояснение, противопоставление, сопоставление. Данный перечень дискурсивных отношений выстроен нами на основе логико-семантических классификаций, предложенных в рамках TPC [Mann, Thompson, 1988a] и PDTB [Prasad и др., 2008a], и соотнесенных с ними китайских и русских теорий. Кроме того, при его формировании были учтены результаты исследований ССЦ в русской лингвистической традиции [Бондарко, 1987; Валгина, 2003], а также структурно-семантические классификации отношений в рамках китайских сложных предложений [吴为章, 田小琳, 2000; 朱德熙, 1982; 邢福义, 2001; 黄伯荣, 廖序东, 2002].

Так как в различных теориях отмечается многозначность дискурсивных отношений и субъективный фактор при аннотировании корпусов, в данной работе многозначные ДО маркируются с помощью описания их первичного и вторичных признаков. Однозначному ДО приписывается только первичный тип. Для неоднозначного ДО как первичный отмечен только наиболее определенный вариант, а затем в качестве вторичных размечаются другие предполагаемые типы, максимум два. При разметке знак «++» используется для обозначения первичного типа, а «+» – вторичных.

Объективацию разметки ДО также обеспечивает лексический фактор, в основном дискурсивные коннекторы или лексический контекст. Если две ЭДЕ объединяются эксплицитным ДК, семантический тип ДО преимущественно определяется благодаря связующему их коннектору. В случае отсутствия эксплицитного ДК мы либо пытаемся добавить коннектор в целях проверки связи между двумя ЭДЕ, либо определяем дискурсивное отношение на основе лексического контекста – слов или фраз, содержащихся в самих ЭДЕ. В Таблице 3.1 перечислены семантические типы ДО и способы их определения для каждой структурной пары в дискурсивном фрагменте (3.9). Как показано в таблице, эксплицитные ДК выделены жирным шрифтом; имплицитные ДК, используемые для проверки отношений между ЭДЕ, даются на сером фоне; лексические контексты, используемые для проверки отношений, подчеркнуты (одинарное и двойное подчеркивание используется для разграничения различных отношений, устанавливаемых группами слов).

Таблица 3.1. Разметка дискурсивных отношений.

Структурные пары	**Семантические типы ДО**	**Способы определения ДО**
96-80 – 96-81: *Отмечая вклад ООН в дело поддержания международного мира и безопасности, / Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности.*	Время++ Причина – следствие +	ДО «Время++» устанавливается между двумя РСА с помощью деепричастного оборота: *Стороны сначала отмечали вклад ООН ..., а потом согласились укреплять сотрудничество ...* ДО «Причина – следствие+» может быть названа через добавление коннектора «поэтому»: *вклад ООН в дело поддержания..., поэтому надо укреп-*

		лять сотрудничество для повышения ее эффективности и дееспособности.
96-81 – 96-82: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности. / Стороны считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности,*	Причина – следствие ++	ДО «Причина – следствие++» определяется с помощью лексического контекста: *Стороны согласились укреплять сотрудничество ..., потому что они считают, что ООН представляет собой уникальный механизм для сотрудничества ...*
96-82 – 96-83: *Стороны считают, что ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности, / что на ее плечах лежит миссия дать ответ на глобальные вызовы XXI века; /*	Соединение++ Причина – следствие+	ДО «Соединение++» определяется между двумя РСА с учетом конструкции предложения: *Стороны считают, что ..., что ...* ДО «Причина – следствие+» может быть названа, потому что *ООН представляет собой уникальный механизм для сотрудничества во имя мира, развития и безопасности, поэтому на ее плечах лежит миссия дать ответ на глобальные вызовы XXI века.*

96-81 – 96-86: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности. / необходимо провести соответствующую реформу ООН и ее органов, /*	Цель++	ДО «Цель++» устанавливается между двумя РСА, потому что *для того, чтобы повышать ее эффективность и дееспособность, необходимо провести реформу ООН и ее органов.*
96-84 – 96-86: ***в целях** адаптации к изменившейся международной обстановке / необходимо провести соответствующую реформу ООН и ее органов, /*	Цель++	ДО «Цель++» устанавливается с помощью эксплицитного коннектора «*в целях*».
96-85 – 96-86: *и повышения эффективности работы ООН / необходимо провести соответствующую реформу ООН и ее органов, /*	Цель++	ДО «Цель++» устанавливается с помощью эксплицитного коннектора «*в целях*».
96-86 – 96-87: *необходимо провести соответствующую реформу ООН и ее органов, / **что** позволило бы им еще лучше исполнять обязанности, предусмотренных Уставом*	Цель++	ДО «Цель++» устанавливается с помощью эксплицитного дискурсивного коннектора «*что*».

ООН; /		
96-81 – 96-88: *Стороны согласились укреплять сотрудничество в области повышения ее эффективности и дееспособности. / деятельность ООН и процесс принятия ее решений должны еще лучше отражать общие чаяния и коллективную волю всех стран – членов ООН.*	Цель++	ДО «Цель++» определяется с помощью лексического контекста: *согласиться повышать эффективность и дееспособность ООН, чтобы ее решения еще лучше отражать общие чаяния и коллективную волю всех стран – членов ООН.*

Корпусная разметка не предъявляет особых формальных требований к коннекторам, поэтому ДК в принципе считаются все лексические единицы, которые могут объединять две ЭДЕ и выражать логико-семантические отношения между ними. Например, такие русские ДК, как *и, а, в то же время, в этом контексте, для этого, в этих целях, прежде всего, и на этом фундаменте* и др. Аналогично размечены китайские ДК, такие как: *并 (и), 以 (в целях / чтобы / с целью…), 但 (но), 但是 (но / однако), 为 (в целях / чтобы / с целью…), 同时 (одновременно), 包括 (включая / в том числе ...), 其中包括 (в том числе включая), 无论 (несмотря на то, что), 考虑到 (учитывая), 鉴于 (ввиду того, как)* и др. Следует отметить, что описание союзов, соединяющих два слова или две фразы внутри простого предложения, выходит за рамки данного исследования.

3.2.3.3. Разметка вершин структурных пар

Дискурсивная зависимость есть семантическая ассиметричная связь между ЭДЕ, которая представлена вершиной в структуре зависимостей. Дискурсивная вершина в нашем понимании, с одной стороны, относится к локальному дискурсивному отношению структурной пары, а с другой – к глобальной дискурсивной структуре абзаца. И в том, и в другом аспекте она обладает уникальностью: в каждом дискурсивном отношении допускается единственная главная ЭДЕ, и в каждой дискурсивной структуре допускается единственный корень.

В данной работе при разметке дискурсивных вершин мы отказались от их фиксирования для каждого типа дискурсивных отношений, как в TPC [Mann, Thompson, 1988a]; в целях получения интерпретируемых и относительно объективных дискурсивных вершин (см. об этом в п. 2.2.5) используется вычислительный подход, предложенный в [冯文贺 et al., 2020; Lyu, Feng, 2023]. Следуя их формуле измерения вершин, если удалить связь между элементами структурной пары ЭДЕ-1 и ЭДЕ-2, глобальная структура будет разделена на две части, то есть S-1 и S-2; соответственно, количество связей в S-1 есть глобальная семантическая связанность единицы ЭДЕ-1, а количество связей в S-2 – глобальная семантическая связанность единицы ЭДЕ-2. В соответствии с этим методом мы определяем вершины для каждой из структурных пар в (3.9) (см. Таблицу 3.2). Например, при определении вершины структурной пары *R96-80 – R96-81* устанавливается, что в случае исключения данной пары из глобальной структуры *ЭДЕ R96-80* прямо и косвенно участвует в нуле структурных пар (то есть глобальная семантическая связанность *ЭДЕ R96-80* равна 0), тогда как *ЭДЕ R96-81* – в семи структурных парах (то есть глобальная семантическая связанность *ЭДЕ R96-81* равна 7). Кроме того, основываясь на количестве вхождений ЭДЕ в качестве вершины во всех структурных парах, определяем единственную корневую вершину этого абзаца. Так как *R96-81* в качестве вершины встречается 4 раза в дискурсивной структуре, она является наиболее частой

вершинной ЭДЕ и, следовательно, рассматривается как корневой узел этого абзаца.

Таблица 3.2. Разметка дискурсивных вершин.

Структурные пары (ЭДЕ-1 – ЭДЕ-2)	Измерение глобальной семантической связанности ЭДЕ		Определение дискурсивной вершины
	ЭДЕ-1	ЭДЕ-2	
R96-80 – R96-81	0	7	*R96-81*
R96-81 – R96-82	6	1	*R96-81*
R96-82 – R96-83	7	0	*R96-82*
R96-81 – R96-86	4	3	*R96-81*
R96-84 – R96-86	0	7	*R96-86*
R96-85 – R96-86	0	7	*R96-86*
R96-86 – R96-87	7	0	*R96-86*
R96-81 – R96-88	7	0	*R96-81*

Этот подход по сути определяет семантическую важность вершинной ЭДЕ локальной структурной пары путем измерения количества отношений, в которые прямо и косвенно входит ЭДЕ в дискурсивной структуре. Он позволяет гарантировать, что как корень, так и вершины представляют собой формально и семантически значимые ЭДЕ: вершины семантически выражают основную идею локальной структурной пары, а корень – главную идею абзаца в целом; при этом они должны быть максимально полными с точки зрения синтаксической структуры (т. к. при построении дискурсивной структуры зависимостей приоритетно сохраняется ЭДЕ, имеющая относительно полную синтаксическую структуру).

Таким образом, после добавления к структурным парам системной лингвистической информации (ДО, ДК, дискурсивные вершины и корень) разметка дискурсивной структуры зависимостей завершается. Полученные дискурсивные структуры зависимостей можно представить в виде следующей схемы с помощью программного обеспечения (см. рис. 3.2). На схеме, как и ранее, ЭДЕ расположены в строках, структурные пары соединены ребрами со стрелкой. Далее стрелки ребра направлены от вершины к периферии, а над ребрами

обозначены типы отношений и коннекторы. Корень обозначен оператором ROOT.

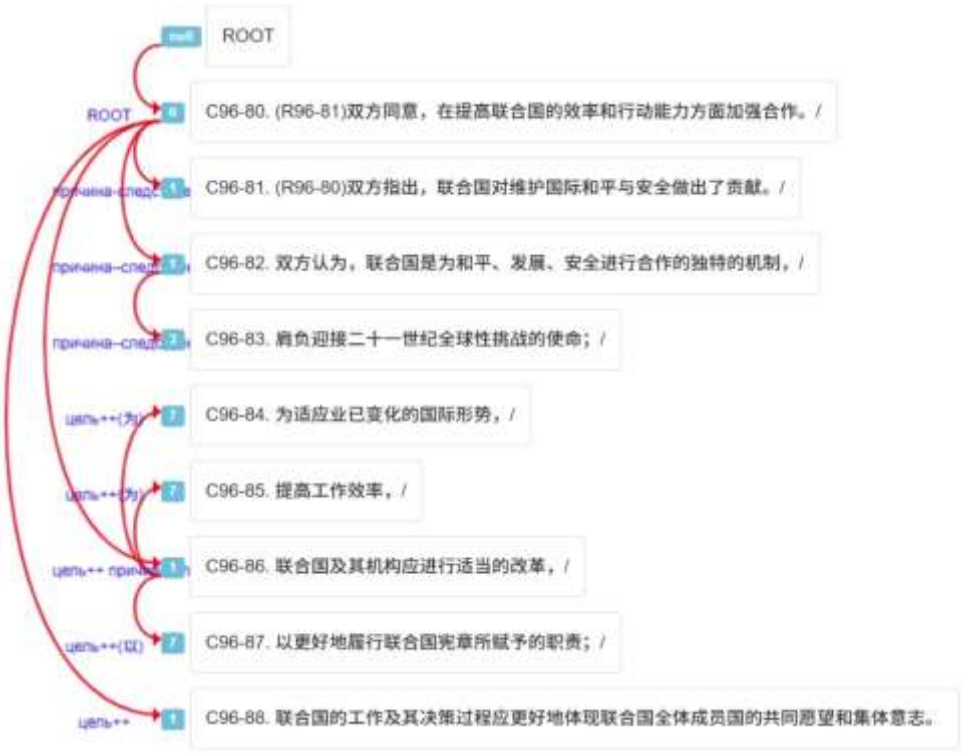


Рисунок 3.2. Схема дискурсивной структуры зависимостей дискурсивного фрагмента (3.9)[1]

Дискурсивные вершины придают наглядность ациклической диаграмме дискурса. По рис. 3.2 видно, что дискурсивный цикл со стрелкой фактически объясняет порядок ЭДЕ при генерации дискурса и последующем его восприятии. По направлению стрелок в тексте (3.9) можно определить четыре основных направления смысловой организации дискурса, начиная с корня *R96-81*:

1. *R96-81 → R96-80;*

2. *R96-81 → R96-82 → R96-83;*

3. *R96-81 → R96-86 → 96-84; R96-81 → R96-86 → 96-85; R96-81 → R96-86 → 96-87;*

4. *R96-81 → R96-88.*

В итоге такая направленная бинарная ациклическая дискурсивная структура с нашей разметкой становится лингвистически читаемой и интерпретируемой.

3.2.4. Выявленные проблемы разметки дискурсивной структуры

В данном подпараграфе обсуждаются проблемы, возникающие на этапе разметки и выравнивания корпуса, а также способы их решения.

3.2.4.1. Проблема несовпадения порядка следования ЭДЕ в китайских и русских текстах

Несовпадение порядка следования ЭДЕ двух языков может вызвать ряд проблем при разметке параллельной структуры дискурса. Во-первых, оно приводит к проблемам деления текстов на ЭДЕ и их выравнивания, без которого

[1] Нулевой корень (null) и отношение «ROOT» – это технические маркеры, встроенные в программное обеспечение, которое мы используем для представления нашего дискурсивного дерева зависимостей.

невозможен параллельный корпус. Исходя из опыта разметки, можно констатировать, что в параллельных текстах подобное несовпадение встречается нередко. Например, в дискурсивном фрагменте (3.9) *R96-80* является РСА, соответствующим китайской клаузе *C96-81*, а *R96-81 – C96-80*. Для того чтобы обеспечить выравнивание, мы добавляем к разметке ЭДЕ одного языка порядковый номер соответствующей ЭДЕ другого языка (см. диаграммы на рис. 3.2).

Кроме того, из-за несовпадения порядка следования ЭДЕ русские синтаксические аналоги не всегда являются полными смысловыми эквивалентами китайских клауз. Например, в дискурсивном фрагменте (3.12) обстоятельство образа действия «在双边及多边基础上 (*на двусторонней и многосторонней основе*)» находится в китайском тексте в клаузе *C97-107*, а в русском тексте – в РСА *R96-108*. Такое несовпадение в большинстве случаев принципиально не влияет на разметку дискурсивной структуры.

(3.12)

C96-106.双方表示，应同任何形式的恐怖活动和有组织的跨国犯罪作坚决的斗争，

C96-107.并将**在双边及多边基础上**经常交流经验，

C96-108.加强合作。

R96-106. Стороны, отмечая необходимость решительной борьбы со всеми видами терроризма и транснациональной организованной преступности,

R96-107. будут осуществлять регулярный обмен опытом

R96-108. и укреплять сотрудничество ***на двусторонней и многосторонней основе.***

Наконец, несовпадение порядка следования ЭДЕ иногда может объясняться естественными различиями в текстовых структурах: общей дискурсивной организации, языковых средств объединения ЭДЕ и текстовой сегментации.

Например, в случае (3.13) в структуре корпусного фрагмента русского текста, представленной на рис. 3.3, семантическая зависимость устанавливается между *R07-132* и *R07-133(C07-136)*, поскольку они тесно связаны друг с другом таким синтаксическим средством, как деепричастный оборот. Однако между соответствующими им китайскими ЭДЕ *C07-132* и *C07-136 (R07-133)* мы не можем установить прямую семантическую связь, потому что клауза *C07-136* «双方对此表示担忧。 (*Стороны выражают озабоченность)*» не связана очевидным образом с *C07-132 «*双方注意到包括互联网在内的信息通信技术、系统以及手段的迅猛发展和广泛使用带来的广阔前景，/ (*Стороны отмечают широкие возможности, предоставляемые стремительным развитием и массовым использованием информационно-коммуникационных технологий, систем и средств, включая Интернет)»*. И, наоборот, первая клауза связана с C07-134: «但也带来现实威胁，/ (*процессы глобальной информатизации приведут реальные угрозы)»*. Итак, в данной работе при разметке параллельных текстов на двух языках создаются отдельные структуры. Наша главная задача – изобразить дискурсивные структуры в том реальном виде, в котором они проявляются в тексте, учитывая при этом специфику межъязыкового выравнивания.

(3.13)

C07-132. 双方注意到包括互联网在内的信息通信技术、系统以及手段的迅猛发展和广泛使用带来的广阔前景，/

C07-133. (R07-134)同时认为，全球信息化进程虽有许多优点，/

C07-134. (R07-135)但也带来现实威胁，/

C07-135. (R07-136)即信息产业成果可能用于不符合保障军民领域国际安全与稳定的目的，

C07-136. (R07-133)双方对此表示担忧。

R07-132. Сознавая широкие возможности, предоставляемые стремительным развитием и массовым использованием информационно-коммуникационных технологий, систем и средств, включая Интернет, /

R07-133. (C07-136) Стороны выражают озабоченность /

R07-134. (C07-133) в связи с тем, что наряду с преимуществами, которые открывают процессы глобальной информатизации, /

R07-135. (C07-134) появляются реальные угрозы /

R07-136. (C07-135) использования достижений в информационной сфере в целях, несовместимых с задачами обеспечения международной стабильности и безопасности как в гражданской, так и в военной сферах.

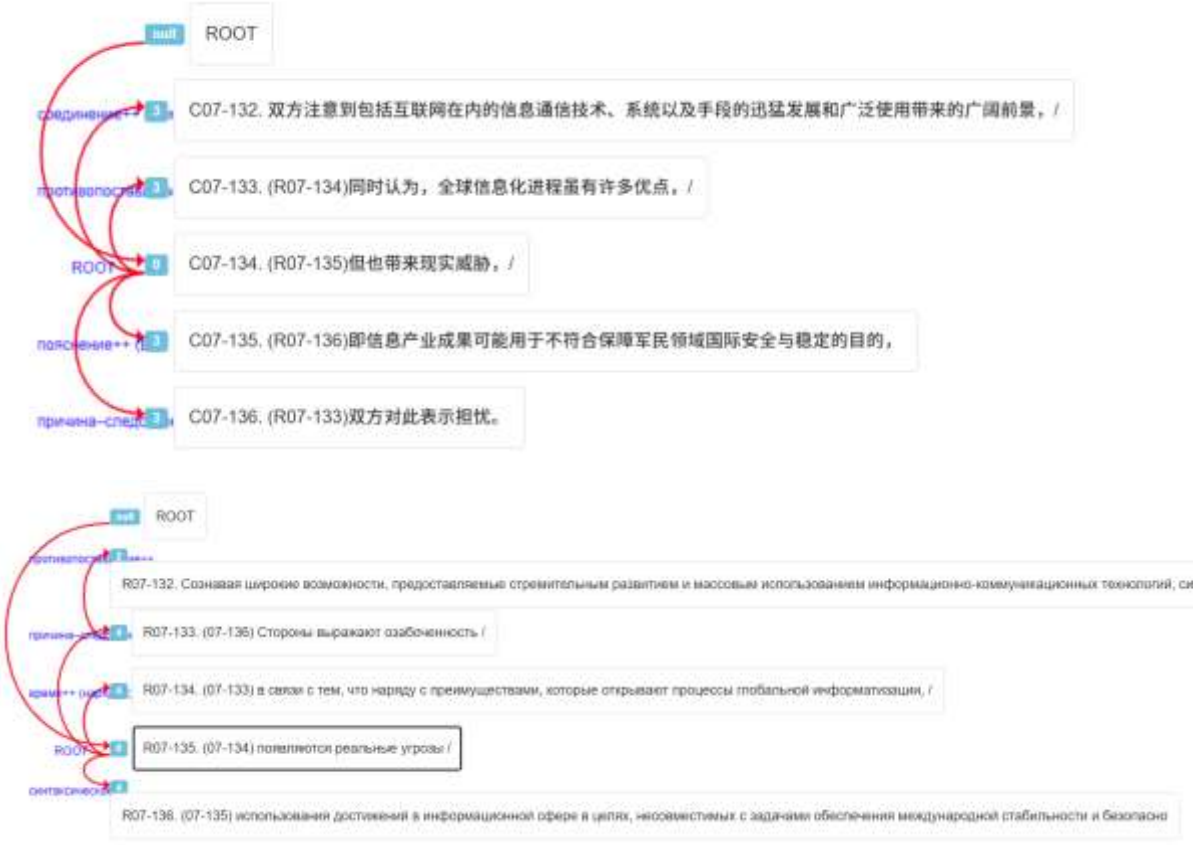


Рисунок 3.3. Дискурсивная структура зависимостей текстов (3.13).

3.2.4.2. Проблема организации равноправных ЭДЕ в структуре зависимостей

Существуют дискурсивные структуры, в которых рядом друг с другом соседствуют три или более трех равноправных дискурсивных единиц, объединяемых отношениями однородности. Согласно TPC и другим исследованиям, такие структуры называются многоядерными и встречаются нередко. Однако

проблема заключается в том, что структура зависимостей имеет единственную вершину, а отношения в ней всегда являются неравноправными. Это заставляет нас подумать о том, как можно включить многоядерные структуры в рамки зависимостей, то есть как в асимметричной структуре объединить равноправные ЭДЕ.

С целью представления всех видов дискурсивных структур в корпусе проводятся адаптационные процедуры. Для решения поставленного вопроса вводится простая математическая модель для описания дискурсивных отношений, где *A* и *B* – две группы ЭДЕ, $\{A_1, A_2, \ldots, A_n\}$ – многоядерная группа, в которой все ЭДЕ равноправны, и $\{B_1, B_2, \ldots, B_n\}$ – другая многоядерная группа, в которой единицы $B_1, B_2, \ldots, B_n$ также связаны друг с другом равноправными отношениями.

Если фрагмент текста (абзац) как единое целое представляет собой многоядерную группу $\{A_1, A_2, \ldots, A_n\}$, то это означает, что нам надо описать отношения между равноправными единицами в рамках концепции зависимостей. Дискурсивные отношения устанавливаются между первой ЭДЕ и каждой последующей ЭДЕ по принципу 2 (см. в п. 3.2.2), и в результате получаются такие структурные пары, как A_1-A_2, A_1-A_3, ... A_1-A_n. При таком варианте структура многоядерной группы показана на рис. 3.4.

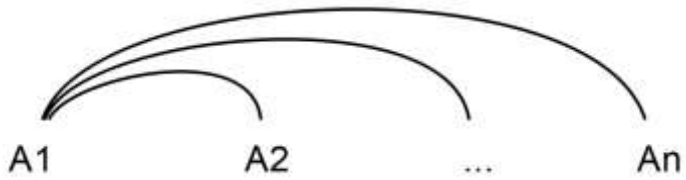


Рисунок 3.4. Многоядерная структура в рамках зависимости: модель (I).

В дискурсивном фрагменте (3.14) и на рис. 3.5 приведен пример анализа текста такого типа из параллельного корпуса.

(3.14)

C01-57. 中俄元首指出，国家独立、主权和领土完整是国际法的核心要素，/

C01-58. 是规范国际关系的根本原则，/

C01-59. **也**是每个国家存在的必要条件。/

…

R01-57. Главы государств России и Китая отмечают, что государственная независимость, суверенитет и территориальная целостность являются важнейшими элементами международного права, /

R01-58. основополагающими принципами, регулирующими международные отношения, /

R01-59. ***а также*** *необходимыми условиями существования каждого государства. /*

…

Рисунок 3.5. Дискурсивная структура зависимостей текстов (3.14) по модели (I).

При необходимости установления отношения между внутренней ЭДЕ многоядерной группы $\{A_1, A_2, \ldots, A_n\}$ и отдельной ЭДЕ B нужно в начале опре-

делить конкретные типы дискурсивных отношений между равноправными ЭДЕ *{A_1, A_2, …, A_n}* и наличие дискурсивного коннектора в этой группе.

1. При наличии отношения соединения между ЭДЕ и существующего ДК в группе *{A_1, A_2, …, connect. A_n}* отношения сначала устанавливаются между равноправными ЭДЕ группы по модели I, а затем из многоядерной группы выбирается синтаксически полная ЭДЕ (обычно – первая) для оформления дискурсивного отношения с ЭДЕ *B*. Таким образом, дискурсивная структура зависимостей учитывает наличие ДК, но в то же время избегает лишнего пересечения отношений. Графическая модель этой структуры приведена на рис. 3.6.

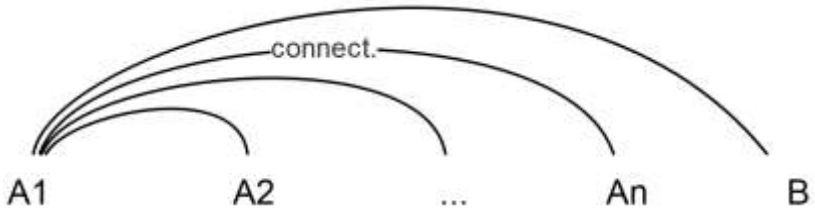


Рисунок 3.6. Многоядерная структура в рамках зависимости: модель (II).

Соответственно, при создании дискурсивной структуры ЭДЕ многоядерной группы корпусного фрагмента (3.14) с внегрупповой ЭДЕ *C01-60 (对任何破坏上述原则的企图和行径进行坚决回击，是每个国家的合法权利。)* и *R01-60 (Решительный отпор любым замыслам и действиям, направленным на подрыв вышеуказанных принципов, - законное право каждого государства*.) получаем в результате схему, показанную на рис. 3.7.

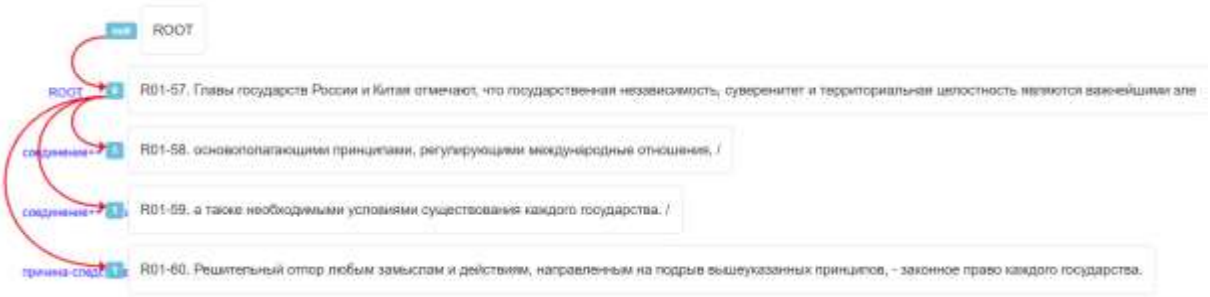


Рисунок 3.7. Дискурсивная структура зависимостей по модели (II).

2. При наличии отношения соединения между ЭДЕ и отсутствии ДК отношения обычно устанавливаются между каждой ЭДЕ из группы *{A_1, A_2, …, A_n}* и ЭДЕ *B* и в результате получаются такие структурные пары, как *A_1-B*，*A_2-B*，… *A_n-B*. В данном случае отношения между ЭДЕ внутри многоядерной группы не устанавливаются с целью избежания пересечения и зацикливания дискурсивных отношений в рамках зависимостей. Графическая модель этой структуры приведена на рис. 3.8.

Рисунок 3.8. Многоядерная структура в рамках зависимости: модель (III).

Пример анализа фрагмента такого типа из параллельного корпуса приведен в (3.15) и на рис. 3.9.

(3.15)

C10-43. 两国将根据各自现代化和经济发展战略需要，完善双边贸易结构，/

C10-45. 规范和转变双边贸易增长方式，/

C10-46. 扩大机电产品和高科技产品贸易，/

C10-47. 建设现代化物流和贸易平台，/

C10-48. 加强在建立经济特区、保护知识产权等领域的交流与合作，/

C10-49. 加快实施双边大型合作项目，/

C10-50. 为中俄经贸合作持续健康稳定发展创造条件。

R10-43. Россия и Китай будут на основе своих национальных стратегий модернизации и экономического развития совершенствовать структуру, /

R10-45. упорядочивать и изменять модель роста двусторонней торговли, /

R10-46. расширять масштабы поставок машинотехнической и высокотехнологичной продукции, /

R10-47. создавать современные логистические и торговые площадки, /

R10-48. расширять обмены и сотрудничество в области создания особых экономических зон и защиты интеллектуальной собственности, /

R10-49. ускорять реализацию крупных двусторонних проектов, /

R10-50. формировать условия для последовательного, здорового и устойчивого развития российско-китайского торгово-экономического сотрудничества.

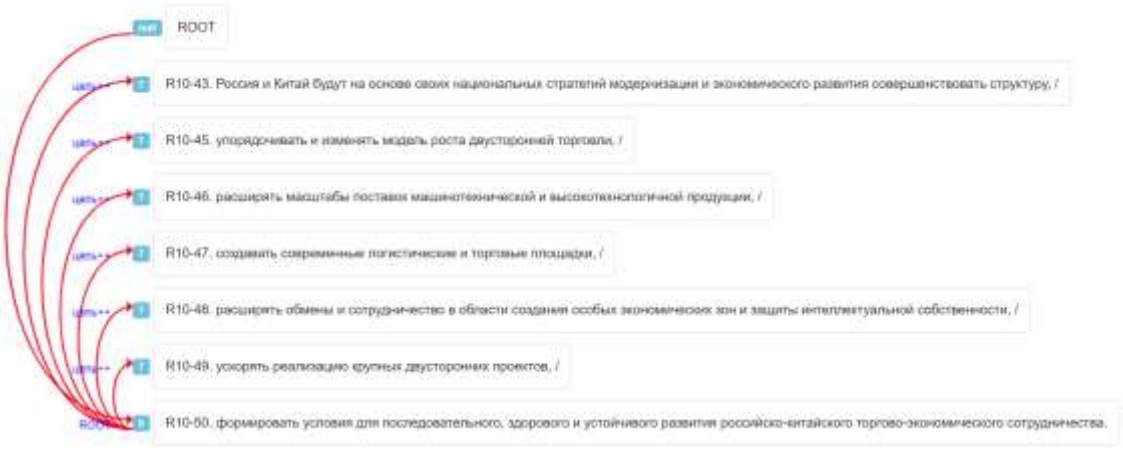


Рисунок 3.9. Дискурсивная структура зависимостей дискурсивных фрагментов (3.15) по модели (III).

3. При наличии соединительных отношений как первичных с другими вторичными отношениями или при наличии других первичных отношений с соединительными как вторичными в многоядерной группе $\{A_1, A_2, \ldots, A_n\}$ структурные пары сначала устанавливаются между каждой ЭДЕ группы *A*, а затем для оформления дискурсивного отношения с ЭДЕ *B* выбирается одна ЭДЕ из этой группы, имеющая самую тесную связь с ЭДЕ *B*. Графическая модель этой структуры приведена на рис. 3.10.

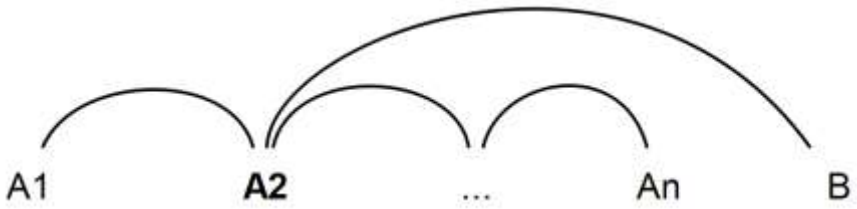


Рисунок 3.10. Многоядерная структура в рамках зависимости: модель (IV).

Фрагмент {*R96-114, R96-115, R96-116*} в дискурсивном фрагменте (3.16) является примером анализа текста такого типа. В этом фрагменте между *R96-114* и *R96-115* устанавливается отношение причины – следствия как первичное, а отношение соединения – как вторичное; обе ЭДЕ имеют семантическую связь с ЭДЕ *R96-116*. При этом мы сначала объединяем *R96-114* и *R96-115*, а потом

R96-115 как ЭДЕ, прямо связанная с ЭДЕ *R96-116*, выбирается для оформления отношения с ней. В результате получается дискурсивная структура зависимостей, показанная на рис. 3.11.

(3.16)

C96-114. 双方认为，冷战后亚太地区政治相对稳定，/

C96-115. 经济快速增长，/

C96-116. 在未来世纪将起重要作用。/

C96-117. 中俄两国愿为亚太地区的和平、稳定与发展继续作出自己的努力。

R96-114. Стороны считают, что после окончания «холодной войны» в азиатско-тихоокеанском регионе сохраняется относительная политическая стабильность, /

R96-115. в экономике наблюдается быстрый рост, /

R96-116. что в следующем веке АТР будет играть важную роль. /

R96-117. Китай и Россия готовы и в дальнейшем прилагать свои усилия в интересах мира, стабильности и развития в АНР.

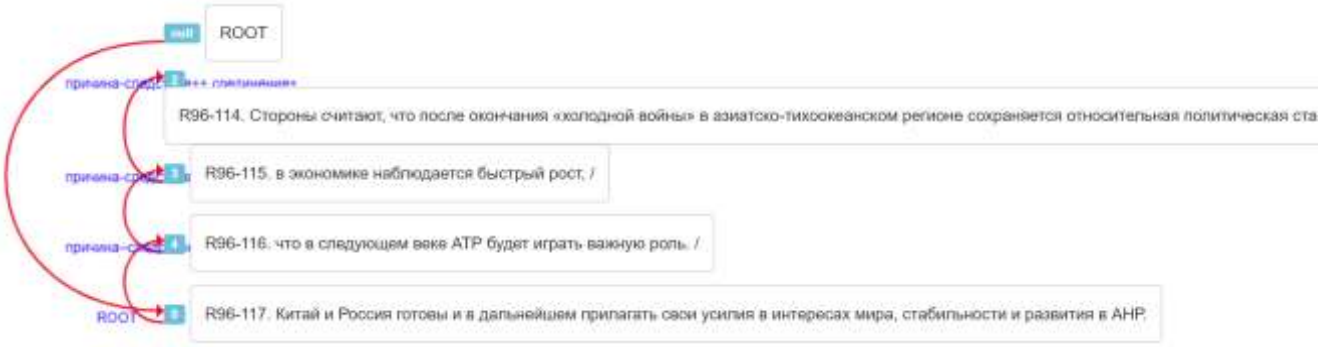


Рисунок 3.11. Дискурсивная структура зависимостей дискурсивных фрагментов (3.16) по модели (IV).

3.2.4.3. Проблема выравнивания структурных пар в китайских и русских текстах

Наряду с выравниванием ЭДЕ в параллельном дискурсивном корпусе тексты также выравниваются по структурным парам, и на их основе выравниваются глобальные дискурсивные структуры параллельных текстов. При этом выравнивание структуры дискурса китайского и русского языков показывает, что параллельные тексты могут различаться по своей дискурсивной организации. В подобных случаях при выравнивании структурных пар предпочтение отдается структурам того языка, которые обладают очевидными дискурсивными структурным признаками: приоритетно сохранение структурных пар с эксплицитным коннектором и синтаксически полной ЭДЕ, а также возможно опущение структурной пары с соединительным отношением (см. об этом в п. 3.2.2.).

Например, в данном корпусе для обеспечения выравнивания дискурсивной структуры двух языков при наличии ДК в одном тексте и его отсутствии в другом дискурсивные структуры двух языков выравниваются по тексту с ДК. Так, в (3.17) в китайском языке при отсутствии ДК структурные пары изначально должны быть установлены между *C07-143 – C07-144* и *C07-143 – C07-145*. Однако для выравнивания с русским текстом при наличии ДК «*и*» в *R07-*

145 для создания дискурсивной структуры китайского языка были выбраны ЭДЕ *C07-143 – C07-144* и *C07-144 – C07-145*, см. рис. 3.12.

(3.17)

C07-143. 中俄强调，伊朗核问题只能通过谈判和平解决。

C07-144. 中俄重申恪守核不扩散体系，/

C07-145. 强调国际原子能机构在解决伊朗核问题过程中的重要作用。

R07-143. Россия и Китай подчеркивают, что проблема ядерной программы Ирана должна решаться исключительно мирным, переговорным путем.

R07-144. Россия и Китай подтверждают приверженность обеспечению незыблемости режима ядерного нераспространения /

R07-145. ***и*** *подчеркивают важную роль, которую играет МАГАТЭ в урегулировании иранской ядерной проблемы. /*

Рисунок 3.12. Выравнивание дискурсивных структур зависимостей текстов (3.17).

3.2.4.4. Проблема разметки вершины структурных пар

Согласно рассуждению в п. 3.2.3.3, дискурсивные вершины устанавливаются с помощью измерения «глобальной семантической связанности» ЭДЕ [冯文贺 et al., 2020]. Для того чтобы вычисленные таким образом вершины действительно отражали семантическую важность ЭДЕ, выработан уже обсужденный выше ряд принципов построения дискурсивной структуры. При организации равноправных ЭДЕ для представления многоядерных структур были разработаны модели (I), (II), (III) и (IV) (см. п. 3.2.4.2).

Однако по причине разнообразия дискурсивных структур не все семантические вершины могут быть вычислены с помощью формального структурного измерения, особенно в случае равных показателей у структурных пар. Одним из типичных примеров является фрагмент дискурса в рамках абзаца, состоящего только из двух ЭДЕ. В дискурсивном фрагменте (3.18) именно такая ситуация: в китайском тексте два простых предложения в составе сложного, а в русском – два отдельных предложения. Из этих двух единиц, безусловно, построена структурная пара, но ее вершина может быть определена вручную только в более широком контексте, дальше этого абзаца. Поскольку последующий фрагмент посвящен содержанию «*официальных переговоров*», ЭДЕ 00-1-2 определена как вершина, а также корень этого корпусного фрагмента.

(3.18)

00-1-1. 应中华人民共和国主席江泽民的邀请，俄罗斯联邦总统普京于2000年7月17日至19日对中华人民共和国进行了国事访问，/

00-1-2. 两国元首在北京举行了正式会谈。

00-1-1. По приглашению Председателя Китайской Народной Республики Цзян Цзэминя Президент Российской Федерации В.В. Путин 17 - 19 июля 2000 года совершил официальный визит в Китайскую Народную Республику. /

00-1-2. В Пекине состоялись официальные переговоры между главами государств России и Китая.

Еще раз отметим, что логика определения вершин на основе структурных вычислений заключается в том, что семантически важные ЭДЕ должны быть структурно связаны с как можно большим количеством других ЭДЕ в абзаце. Формальное измерение структурной связанности ЭДЕ представляет собой относительно объективную меру, которая в большинстве случаев верно определяет верифицируемые вручную вершины, а также значительно повышает эффективность разметки дискурсивного корпуса. Но в случае несовпадения семантического и формального критериев при построении дискурсивной структуры и определения вершин в первую очередь нужно следовать критерию семантическому.

3.3. Программное обеспечение для хранения размеченных данных и их визулиации

Разметка в параллельном корпусе проводится вручную с использованием текстового процессора *Microsoft Word*, табличного процессора *Microsoft Excel* и универсального текстового редактора *EditPlus 4.3*.

Перед дискурсивной разметкой каждый исходный текст предварительно обработан в программах *Microsoft Word* и *EditPlus* (см. в п. 3.2.1.3). После предварительного разделения полученные ЭДЕ располагаются построчно и помечаются порядковым номером документа в начале и разделителем («/») в конце.

Разметка дискурсивных структур осуществляется в *Microsoft Excel*. ЭДЕ сортируются в соответствии с их расположением в абзацах, структурные пары помечаются с помощью нумерации ЭДЕ, а специфика дискурсивных отношений отражается в колонках таблицы (см. рис. 3.13). Благодаря открытому ха-

рактеру структуры зависимостей графическое представление корпуса осуществляется с помощью программного обеспечения[1], которое было разработано группой программистов Ли Суцзяна Пекинского университета [Li et al., 2014]. Данный инструмент, основанный на общей теории зависимостей, обеспечивает простейшую схему разметки и позволяет свободно добавлять новые типы отношений. С помощью этого инструмента можно построить и изобразить дискурсивные структуры зависимостей, определив структурные пары. Однако это только графика, а не интеллектуальное представление корпуса. Следует отметить, что цель использования этого приложения – графическая визуализация размеченных данных о дискурсивной структуре зависимостей, а не выполнение самого дискурсивного структурного анализа.

Формат хранения размеченных данных представлен в табличном виде (см. рис. 3.13). Китайские и русские ЭДЕ выравнены и расположены в колонке *Клаузы* и *РСА*; в столбце *Порядок ЭДЕ* указан порядок следования китайских и русских ЭДЕ в абзаце; в целях облегчения последующего поиска и фиксации перед каждой ЭДЕ указывается порядковый номер документа, который состоит из года и порядка ЭДЕ, например, «94-10» обозначает десятую ЭДЕ в документе китайско-российской совместной декларации 1994 г. В столбце *Структурные пары* обозначены сформированные структурные пары; в столбце *ДК* – дискурсивные коннекторы; типы дискурсивных отношений помечаются в столбцах *Первичный тип ДО* и *Вторичный тип ДО1* и *Вторичный тип ДО2*. Столбец *Синтаксические варианты* содержит данные о синтаксических вариантах соотношения двух ЭДЕ в составе структурных пар, а столбец *Вершины* обозначает вершины структурных пар: 1 – вершиной является первая ЭДЕ, 2 – вторая. В колонке *Типы РСА* значение «1» соответствует монопредикативному РСА, «2» – полипредикативному РСА, «0.5» – полупредикативному РСА, «0» – РСА без предиката.

[1] URL: http://123.56.88.210/demo/depannotate/ (дата обращения: 31.05.2025).

	B	C	D	E	F	G	H	I	J
1	Порядок в абзаце	Клаузы	Структурные пары	ДК	Первичный тип ДО	Вторичный тип ДО1	Вторичный тип ДО2	Синтаксические варианты	Вершины
2	1	94-1. 一中华人民共和国和俄罗斯联邦	1-2		Соединение++			Клаузы в одном с	1
3	2	94-2. 有利于维护和加强亚洲和世界的和平、稳定与发展。							
4	1	94-3. 二双方高度评价１９９２年第一	1-2	并	Соединение-	Время+		Клаузы в одном с	2
5	2	94-4. 并认为两国已具有新型的建设性	2-3	即	Пояснение++			Клаузы в одном с	1
6	3	94-5. 即建立在和平共处各项原则基础	2-4		Пояснение++			Клаузы в одном с	1
7	4	94-6. 既不结盟，/	4-5	既不…	Соединение++			Клаузы в одном с	1
8	5	94-7. 也不针对第三国。							
9	1	94-8. 三双方重申，恪守１９９２年１	1-3		Цель++			Клаузы в одном с	2
10	2	94-9. 决心面向二十一世纪。/	2-3		Время++			Клаузы в одном с	2
11	3	94-10. 把两国关系提高到一个崭新的	3-5	并且	Соединение++			Клаузы в одном с	1
12	4	94-11. 并且最大限度地发挥和利用中	4-5		Цель++			Клаузы в одном с	2
13	5	94-12. 为促进两国国内改革和发展经济的重大任务以及在亚太地区和全世界建立持久和平提供有利的						Клаузы в одном	сложном пр
14	1	94-13. 四为进一步确立新型的相互关	1-3	为	Причина-сле	Цель+		Клаузы в одном с	2
15	2	94-14. 从两国关系的远景出发，/	2-3		Условие++			Клаузы в одном с	2
16	3	94-15. 双方决心采取积极和全面的步骤							
17	1	94-16. 在政治关系方面							
18	1	94-17. ——以和平共处各项原则为基	1-3		Цель++	Условие+		Клаузы в одном с	2
19	2	94-18. 从社会制度和观点的不同不妨	2-3		Цель++	Условие+		Клаузы в одном с	2
20	3	94-19. 始终如一地维护和发展长期睦	3-4		Причина-сле	Цель+		Клаузы в одном с	1
21	4	94-20. 保持经常和多方面的对话，/	3-6		Причина-сле	Цель+		Клаузы в одном с	1
22	5	94-21. 根据公认的国际法准则，/	5-6		Условие++			Клаузы в одном с	2
23	6	94-22. 本着坦诚、信任和考虑相互利益的精神解决出现的问题；							
24	1	94-23. ——严格遵守中俄国界协定，/	1-2		Соединение-	Цель+		Клаузы в одном с	2
25	2	94-24. 公正合理地解决遗留的边界问	2-3		Цель++			Клаузы в одном с	1
26	3	94-25. 按期完成勘界立标工作，/	3-4		Цель++			Клаузы в одном с	1

	K	L	M	N	O	P	Q	R	S	T
1	Порядок РСА в абзаце	РСА	Типы РСА	Структурные пары (р.)	ДК (р.)	Первичный тип ДО (р.)	Вторичный тип ДО1 (р.)	Вторичный тип ДО2 (р.)	Синтаксические варианты (р.)	Вершины (р.)
2	1	94-1. Российская федерация и Китайская Народная Республика,	2	1-2		Соединение++			РСА в сложном п	1
3	2	94-2. благоприятствует сохранению и укреплению мира, стабиль	1							
4	1	94-3. II. Оценивают динамичное и успешное развитие отношений	1	1-2		Соединение+	Время+		РСА в сложном п	2
5	2	94-4. считают, что между ними сложились новые отношения кон	2	2-3		Пояснение++			РСА в сложном п	1
6	3	94-5. — подлинно равноправные отношения добрососедства, др	1.5	2-4		Пояснение++			РСА в разных пр	1
7	4	94-6. Эти отношения не носят союзнического характера /	1	4-5	и	Соединение++			РСА в сложном п	1
8	5	94-7. и не направлены против третьих стран.	1							
9	1	94-8. III. Стороны подтверждают твердую приверженность принц	1	1-3	и	Цель++			РСА в сложном п	2
10	2	94-9. обращаясь в XXI век, и полны решимости, /	1.5	2-3		Время++			РСА внутри прос	2
11	3	94-10. поднять отношения между двумя странами на качественн	1	3-5	тем самым	Причина-сле	Соединение+		РСА в сложном п	1
12	4	94-11. тем самым максимально раскрывая и используя значител	0.5	4-5		Цель++			РСА внутри прос	2
13	5	94-12. создать благоприятные условия для содействия решению	1							
14	1	94-13. IV. В целях дальнейшего утверждения нового качества св	0	1-3	В целях	Причина-сле	Цель+		РСА внутри прос	2
15	2	94-14. и исходя из долгосрочных перспектив отношений между д	0	2-3	и исходя из	Причина-сле	Цель+		РСА внутри прос	2
16	3	94-15. Стороны преисполнены решимости предпринимать актив	1							
17	1	94-16. 1В области политических отношений	0							
18	1	94-17. - основываясь на принципах мирного сосуществования, /	0.5	1-3		Цель++	Условие+		РСА внутри прос	2
19	2	94-18. исходя из общего понимания, что различия в общественн	0	2-3		Цель++	Условие+		РСА внутри прос	2
20	3	94-19. неуклонно отстаивать и развивать взаимоотношения долг	1	3-4		Соединение++			РСА в сложном п	1
21	4	94-20. поддерживать интенсивный и разносторонний диалог, /	1	4-5		Соединение++			РСА в сложном п	1
22	5	94-21. решать возникающие проблемы на основе общепринятых	1	5-6		Соединение+	Пояснение+		РСА внутри прос	1
23	6	94-22. в духе открытости, доверия, учета взаимных интересов;	0							
24	1	94-23. -строго соблюдать соглашения о российско-китайской гос	1	1-2		Соединение+	Цель+		РСА в сложном п	2
25	2	94-24. на справедливой и рациональной основе решать оставша	1	2-3		Цель++			РСА в сложном п	1
26	3	94-25. в намеченные сроки завершить демаркацию границы, /	1	3-4		Цель++			РСА в сложном п	1

Рисунок 3.13. Организация основных корпусных данных.

Для обеспечения качества разметки было сделано следующее. Сначала проведена индивидуальная слепая разметка одной трети документов. На основе этого опыта были сформированы уточненные принципы разметки. Затем была осуществлена разметка остальных текстов. Далее – этап ручной проверки размеченных данных. Разметка и проверка корпусных данных осуществлялись

с определенным временным интервалом, при этом строго соблюдались принципы разметки, обсужденные в данной главе. Таким образом, многоэтапная разметка и ручная проверка данных в некоторой степени восполнили недостатки индивидуальной, а не коллективной разметки и обеспечили качество создаваемого корпуса.

Таким образом, в соответствии с теоретическими основаниями и разработанными принципами разметки создан китайско-русский параллельный дискурсивный корпус официально-деловых текстов. Следующая глава диссертции будет посвящена интерпретации данных, которые были получены по результатам корпусной разметки.

Выводы по третьей главе

В данной главе изложен опыт создания параллельного дискурсивного корпуса.

1. В качестве материала для создания китайско-русского параллельного дискурсивного корпуса были выбраны 24 совместных декларации и совместных заявления правительств РФ и КНР – документов, которые регулярно оформляются и согласовываются параллельно с двух сторон. Данные тексты, подготовленные и согласованные двумя государствами, семантически эквивалентны и не являются оригиналами и переводами в классическом понимании, что исключает многие споры, возникающие в связи с переводом.

2. Дискурсивный анализ и выравнивание параллельных текстов проводится по абзацам. Дискурсивная структура зависимостей строится в виде направленного ациклического графа, в котором узлами являются элементарные дискурсивные единицы (ЭДЕ), а зависимостями – логико-семантические дискурсивные отношения между двумя ЭДЕ.

3. Анализ и выравнивание дискурсивных структур зависимостей состоит из трех этапов: 1) деление на ЭДЕ и выравнивание текстов, 2) формирование структурных пар и создание дискурсивной структуры зависимостей и 3) раз-

метка структурных пар, то есть снабжение их лингвистической информацией, необходимой для дискурсивного анализа.

4. Разработаны подробные принципы при разметке корпуса на каждом этапе: 1) принципы деления на ЭДЕ и выравнивания текстов по ЭДЕ; 2) принципы установления структурных пар и создания дискурсивной структуры зависимостей; 3) принципы разметки структурных пар.

Каждый из этих принципов получил практическое освоение, которое привело к множеству корректирующих решений при разметке корпуса.

5. Проблемы разметки вызваны несовпадением порядка следования ЭДЕ при сегментации китайских и русских текстов, организацией равноправных ЭДЕ при создании структуры зависимостей, а также определением вершины для равнозначных ЭДЕ исходя из вычисления количества дискурсивных связей. Для каждой из проблем предложено решение в процессе аннотирования текстов.

6. Разметка параллельного корпуса выполнена с помощью доступного программного обеспечения, что обеспечивает прямой доступ к размеченным данным, а не к его формальным графическим представлениям.

ГЛАВА 4. ДИСКУРСИВНЫЕ СТРУКТУРЫ В КИТАЙСКО-РУССКОМ ПАРАЛЛЕЛЬНОМ КОРПУСЕ: СТАТИСТИЧЕСКИЙ АНАЛИЗ И ИНТЕРПРЕТАЦИЯ ДАННЫХ

Цель данной главы – описать особенности дискурсивных структур официально-деловых текстов в созданном корпусе на основе данных, полученных благодаря осуществленной разметке. Рассматриваемые особенности включают количественные параметры абзацев, сегментацию и выравнивание китайских и русских ЭДЕ, выравнивание дискурсивных структур, синтаксические варианты соотношения ЭДЕ структурных пар, типы дискурсивных отношений, соотношение между дискурсивными отношениями и дискурсивными коннекторами, характеристики дискурсивных вершин.

4.1. Количественные параметры абзаца

В корпусе аннотированы 12 параллельных текстов, включающих 429 рассматривамых отдельно китайских и русских абзацев. С точки зрения дискурсивной структуры величину абзаца определяет количество входящих в него ЭДЕ. В табл. 4.1 представлены количественные параметры абзацев.

Таблица 4.1. Величина абзацев в корпусе[1]

Количество ЭДЕ абзаце	Количество абзацев	Количество абзацев, %	Итого
1	58	13,5	**89,0**
2	77	17,9	
3	73	17,0	
4	73	17,0	

[1] В параллельном Корпусе ЭДЕ в китайском и русском текстах выравниваются, поэтому количество ЭДЕ в соотнесенных фрагментах совпадает.

5	55	12,8	
6	**46**	**10,7**	
7	16	3,7	**11,0**
8	10	2,3	
9	9	2,1	
10	4	0,9	
11	4	0,9	
12	2	0,5	
13	2	0,5	
Итого	429	100,0	

Из таблицы видно, что в корпусе величина абзаца варьируется от 1 до 13 ЭДЕ, в основном они сосредоточены в диапазоне от 1 до 6 ЭДЕ: количество абзацев длиной менее 7 ЭДЕ – 382, что составляет 89,0 % от общего числа; количество абзацев длиной больше 6 ЭДЕ – 47, что составляет 11,0 % от общего числа абзацев. Это говорит о том, что в официально-деловых текстах китайского и русского языков длина абзацев обычно не превышает 6 клауз, в то же время это показывает, что дискурсивная структура в нашей работе построена преимущественно на основе абзацев с количеством клауз до 6.

4.2. Результаты сегментации и выравнивания китайских и русских ЭДЕ

В Корпусе насчитывается 1681 китайских клауз и соответствующих им 1681 РСА. В процессе выравнивания ЭДЕ двух языков обнаружилось, что синтаксические единицы русского языка, соответствующие китайским клаузам (РСА), не всегда являются собственно клаузами. Иногда это сочетания клауз (сложные предложения и группы предложений) или же компоненты клауз, фрагменты предложения. Для того чтобы составить адекватное представление об элементарных единицах в дискурсивных структурах разных языков, мы провели сопоставительный статистический анализ китайских ЭДЕ и русских РСА.

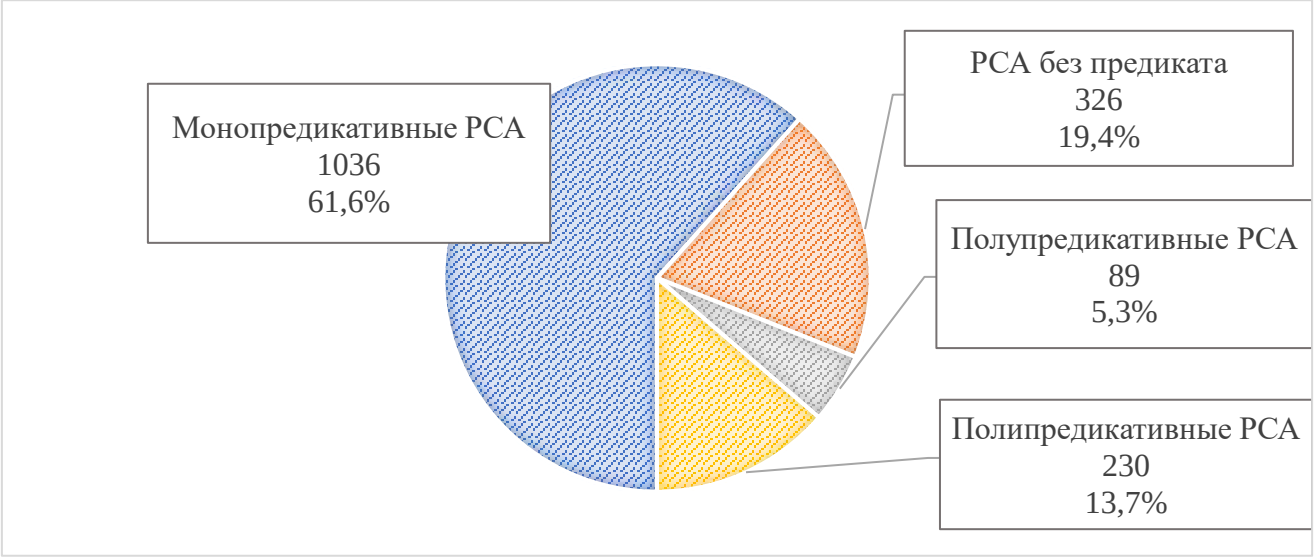


Диаграмма 4.1. Типы РСА в корпусе.

Как следует из диагр. 4.1, монопредикативные РСА составляют более половины от общего числа (61,6 %), что свидетельствует о том, что китайским клаузам в большинстве случаев соответствуют РСА с аналогичными синтаксическими параметрами, то есть с одним предикатом. Также очевидно, что китайские клаузы могут соотноситься с РСА без предиката (19,4 %) и несколькими предикатами (полипредикативные, 13,7 %); определенную долю имеют полупредикативные РСА (5,3 %). Нетрудно понять, почему монопредикативных аналогов больше всего, это и есть залог эквивалентности. Какие же синтаксические и дискурсивные условия определяют распределение нулевых, полу- и полипредикативных РСА русского языка, нам пока неизвестно.

Корпусные данные показывают, что «снижение» уровня синтаксической структуры предполагает соответствие китайских клауз предложно-падежным сочетаниям и именным группам русского языка. Это также относится к полупредикативным структурам – причастным и деепричастным оборотам. «Повышение» уровня синтаксической структуры предполагает соотношение китайских клауз с более крупными единицами – сложными предложениями и группами предложений.

Соотношение ЭДЕ в разных языках можно рассматривать как варианты применения синтаксических средств для эквивалентной передачи одного и того же объема информации. В этом случае параллельные тексты на разных языках могут рассматриваться как эквиваленты по объему передаваемой информации. Синтаксическая асимметрия отражает межъязыковые различия при передаче одинакового объема информации: «снижение» уровня синтаксической структуры китайских клауз в РСА без предиката или полупредикативные РСА предполагает, что русский язык способен передавать ту же информацию более компактными синтаксическими средствами, тогда как в китайском языке для этого требуется использование полных синтаксических единиц. Таким образом, русский язык в этом случае имеет более плотную синтаксическую структуру, чем китайский, и, соответственно, в параллельных текстах количество единиц с полной синтаксической структурой в русском языке меньше, чем в китайском. Например, в дискурсивном фрагменте (4.1) две отдельные клаузы китайского языка (*C00-1-62* и *C00-1-63*) соотносятся с одной клаузой (простым предложением) русского языка.

(4.1)

C00-1-62. 1999年12月9日签署的《中华人民共和国政府和俄罗斯联邦政府关于对界河中个别岛屿及其附近水域进行共同经济利用的协定》是史无前例的，/ *(дословный перевод: Соглашение между Правительством Российской Федерации и Правительством Китайской Народной Республики о совместном хозяйственном использовании отдельных островов и прилегающих к ним акваторий пограничных рек от 9 декабря 1999 года является беспрецедентным,)*

C00-1-63. 它的顺利实施在使中俄边界发展成为一条睦邻友好的纽带方面又迈出了重要的一步。*(дословный перевод: его успешная реализация является еще одним важным шагом, направленным на превращение российско-китайской границы в полосу добрососедства и дружбы.)*

R00-1-62. *Успешная реализация не имеющего прецедента в истории Соглашения между Правительством Российской Федерации и Правительством Китайской Народной Республики о совместном хозяйственном использовании отдельных островов и прилегающих к ним акваторий пограничных рек от 9 декабря 1999 года /*

R00-1-63. *является еще одним важным шагом, направленным на превращение российско-китайской границы в полосу добрососедства и дружбы.*

Повышение уровня ЭДЕ, наоборот, предполагает, что для передачи одной и той же информации в языке Б используются более развернутые синтаксические единицы – например, сложные предложения или группы предложений. В дискурсивных фрагментах (4.2) каждой клаузе китайского языка соответствует одно сложное предложение в русском тексте.

(4.2)

C10-143. 双方表示高度重视全球气候变化问题，/

R10-143. *Стороны заявляют, что придают большое значение проблемам глобального изменения климата, /*

…

C10-146. 为加强国际合作、共同应对全球气候变化挑战作出积极贡献。/

R10-146. *вносить активный вклад в укрепление международного сотрудничества и совместное противодействие вызовам, которые создает глобальное изменение климата. /*

Статистика показывает, что в корпусе соответствие китайских клауз РСА без предиката составляет 19,4 % от общего числа, а, наоборот, полу- и полипредикативным РСА – 19,0 %. Можно сделать вывод, что китайские и русские ЭДЕ демонстрируют разную синтаксическую плотность в определенных случаях, но разница в показателях невелика. Это говорит о том, что синтаксическая плотность дискурсивных единиц китайского и русского языков в разной степе-

ни увеличивается или снижается при передаче одинакового объема информации (с учетом объема информации, который может быть выражен в одной клаузе).

4.3. Результаты выравнивания дискурсивных структур зависимостей

В корпусе насчитывается 1252 структурные пары в китайском языке и 1252 – в русском. Из них 1248 структурных пар выравниваются нормально, а еще четыре структурные пары при структурном анализе не могут быть выравнены (см. табл. 4.2). В корпус были включены и эти четыре структурные пары, однако при статистической обработке данных по выравненным структурам они не принимаются во внимание ввиду невозможности их структурного сопоставления.

Таблица 4.2. Выравнивание структурных пар двух языков.

Выравнивание структурных пар	Количество
Удачное выравнивание	1248
Неудачное выравнивание	4
Итого	1252

Эти четыре невыравненные структурные пары отражают языковую асимметрию между китайским и русским языками и представляют собой значимый материал для межъязыкового сопоставления. Исходя из анализа этих четырех примеров, основная проблема выравнивания структурных пар связана с особенностями синтаксических конструкций РСА, которые были выделены в соответствии с китайскими клаузами и с естественными различиями в способах организации китайского и русского дискурса. В китайском тексте вследствие отсутствия необходимых средств словоизменения (в данном случае глагола) дискурсивные структурные отношения определяются на основе логико-семантических связей между клаузами. Между тем в русском языке структур-

ные отношения между ЭДЕ иногда выражаются благодаря глагольному словоизменению.

Как показано в примере 4.1, в китайском языке в соответствии с семантическими связями структурные пары необходимо образовать между клаузами *C00-2-25* и *C00-2-27* (*необходимо использовать политические, юридические и дипломатические методы для того, чтобы предотвратить распространение оружия массового уничтожения и средств его доставки*) и между клаузами *C00-2-26* и *C00-2-27* (*наращивать международные усилия для того, чтобы предотвратить распространение оружия массового уничтожения и средств его доставки*). В русском языке соответствующими РСА вследствие сегментации китайских клауз являются деепричастный оборот (*R00-2-25*), главные члены простого предложения (*R00-2-26*) и второстепенные члены простого предложения (*R00-2-27*). Структурные отношения между ними, по-видимому, были обусловлены синтаксическими отношениями, то есть деепричастный оборот (*R00-2-25*) должен образовывать структурную пару с предикатом предложения (*R00-2-26*), главные члены предложения (*R00-2-26*) должны образовывать структурную пару со второстепенными членами (*R00-2-27*). В данном случае невозможно ни создать дискурсивную структуру китайского языка по русским структурным парам, ни, наоборот, создать дискурсивную структуру русского языка по китайским.

(4.3)

C00-2-25. 必须通过政治、法律和外交手段，/ (*дословный перевод: необходимо использовать политические, юридические и дипломатические методы,*)

C00-2-26. 加强国际努力，/ (*дословный перевод: наращивать международные усилия,*)

C00-2-27. 防止大规模杀伤性武器及其运载工具的扩散，/ (*дословный перевод: предотвратить распространение оружия массового уничтожения и средств его доставки,*)

C00-2-28. 探讨逐步形成全球防止导弹及其技术扩散监控体系的可能性，/ (*дословный перевод: изучать возможность поэтапного формирования глобальной системы контроля за нераспространением ракет и ракетных технологий,*)

C00-2-29. 并在此领域开展广泛的、非歧视性的对话与合作。(*дословный перевод: а также развивать широкий и недискриминационный диалог и сотрудничество в данной сфере.*)

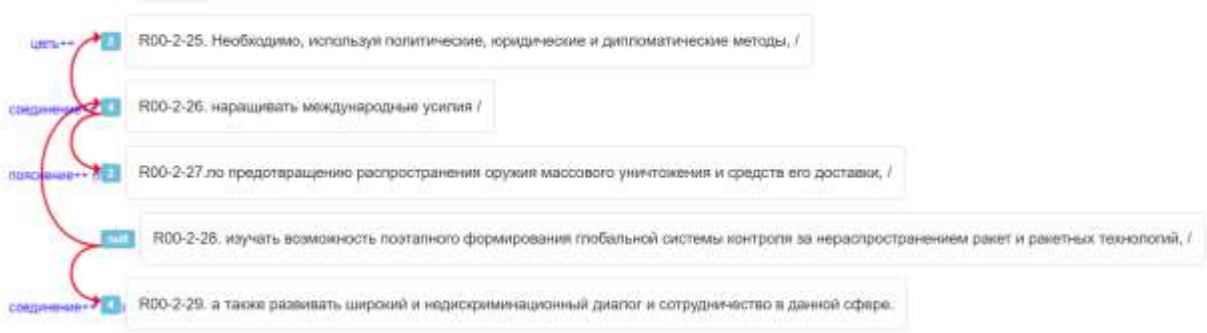
R00-2-25. Необходимо, используя политические, юридические и дипломатические методы, /

R00-2-26. наращивать международные усилия /

R00-2-27. по предотвращению распространения оружия массового уничтожения и средств его доставки, /

R00-2-28. изучать возможность поэтапного формирования глобальной системы контроля за нераспространением ракет и ракетных технологий, /

R00-2-29. а также развивать широкий и недискриминационный диалог и сотрудничество в данной сфере.



Пример 4.1. Структурные пары дискурсивных фрагментов (4.3).

Таким образом, подобные невыравненные структурные пары помечены отдельным образом и, ввиду отсутствия необходимой структурной сопоставимости для сопоставительного анализа, исключены из последующего статистического анализа.

4.4. Результаты разметки дискурсивных параметров структурных пар

4.4.1. Синтаксические варианты соотношения ЭДЕ структурных пар

В нашем корпусе выделяются три синтаксических варианта соотношения ЭДЕ, которые связаны в структурной паре: 1) ЭДЕ в одном сложном предложении, 2) ЭДЕ в разных предложениях и 3) ЭДЕ (только РСА) внутри простого предложения (см. об этом в подпараграфе 3.2.3.1). Как видно из статистических данных в табл. 4.3, в китайском тексте 62,5 % структурных пар размечены внутри предложений, а 37,5 % – между разными предложениями. В русском тексте 61,6 % структурных пар размечены внутри предложений, и среди них 29,2 % – внутри сложных предложений, 32,4 % – внутри простого предложения. 38,4 % составляют РСА, выделенные в разных предложениях. Напомним, что РСА выделяются вторично, в соответствии с сегментированием китайских клауз, и количество ЭДЕ, выделенных внутри предложений, в китайских и русских текстах почти совпадает: 62,5 и 61,6 % соответственно.

Таблица 4.3. Статистика по синтаксическим вариантам структурных пар в корпусе.

Синтаксические варианты ЭДЕ	Китайские тексты		Русские тексты	
	кол-во	%	кол-во	%
ЭДЕ в одном сложном предложении	780	62,5	365	29,2
ЭДЕ в разных предложениях	468	37,5	479	38,4
РСА внутри простого предложения			404	32,4
Итого	1248	100,0	1248	100,0

Таблица 4.3 показывает, что как в русском, так и в китайском языках дискурсивные единицы, входящие в структурную пару, в большинстве случаев располагаются внутри одного предложения, что говорит о том, что свыше 60 % структурных пар в параллельном корпусе образованы между ЭДЕ внутри одного предложения. Подобный результат представляется ожидаемым: с одной стороны, дискурсивные отношения в большинстве случаев устанавливаются между смежными или близко расположенными (в пределах одного предложения) единицами; с другой стороны, несмотря на типологические различия между языками, передаваемый смысл должен оставаться эквивалентным, а значит, и синтаксические средства его выражения в большинстве случаев оказываются близкими.

Для дальнейшего изучения соотношения синтаксических вариантов в китайском и русском языках мы разделили синтаксические варианты двух языков на их совпадающие и несовпадающие реализации в китайских и русских текстах. Совпадающая реализация синтаксических вариантов означает, что как структурная пара китайских клауз, так и соответствующая ей структурная пара РСА находятся внутри одного сложного предложения или, наоборот, в разных предложениях. Таблица 4.4 отражает то, что совпадающие реализации синтаксических вариантов составляют 59,5 % из всех выравненных структурных пар, а несовпадающие – 40,5 %. Это, с одной стороны, говорит о синтаксической

разнице между языками, но, с другой стороны, и о том, что во многих случаях структурные пары китайского и русского языков формируются из ЭДЕ одного и того же синтаксического уровня.

Таблица 4.4. Реализация синтаксических вариантов структурных пар в двух языках.

Совпадающая реализация синтаксических вариантов в двух языках				
			Количество	%
Две ЭДЕ (китайских клаузы и РСА) в одном сложном предложении			330	26,4
Две ЭДЕ (китайских клаузы и РСА) в разных предложениях			413	33,1
Итого			743	59,5
Несовпадающая реализация синтаксических вариантов в двух языках				
	Две китайских клаузы в одном сложном предложении		Две китайских клаузы в разных предложениях	
Два РСА в одном сложном предложении	–		35	2,8
Два РСА в разных предложениях	66	5,3	–	
РСА внутри простого предложения	384	30,8	20	1,6
Итого	450	36,1	55	4,4

В таблице выше отражены четыре типа несовпадающих реализаций синтаксических вариантов:

– тип (I): две китайских клаузы в одном сложном предложении, соответствующие двум РСА в разных предложениях;

– тип (II): две китайских клаузы в одном сложном предложении, соответствующие двум РСА внутри простого предложения;

– тип (III): две китайских клаузы в разных предложениях, соответствующие двум РСА в одном сложном предложении;

– тип (IV): две китайских клаузы в разных предложениях, соответствующие двум РСА внутри простого предложения.

В параллельном дискурсивном корпусе тип I составляет 5,3 % от всех выравненных структурных пар. В дискурсивном фрагменте (4.4) показаны структурные пары такого типа.

(4.4)

C00-1-78. 俄罗斯联邦总统普京邀请中华人民共和国主席江泽民 2001 年方便的时候对俄罗斯联邦进行国事访问，/

C00-1-79. 江泽民主席表示感谢并接受邀请，/

R00-1-78. *Президент Российской Федерации В.В. Путин пригласил Председателя Китайской Народной Республики Цзян Цзэминя посетить Российскую Федерацию с официальным визитом в удобное время в 2001 году. /*

R00-1-79. *Приглашение было с благодарностью принято. /*

Тип II (30,8 %) в корпусе представлен как соответствие между двумя китайскими клаузами в составе одного сложного предложения и двумя РСА, которые связаны между собой различными синтаксическими функциями. Условно данный тип можно разделить на несколько подвидов в зависимости от характера соотношений между русскими сегментами. Например, в дискурсивном фрагменте (4.5) фигурирует соответствие между грамматической основой (*04-188*) и обстоятельством (*04-189*). В дискурсивном фрагменте (4.6) наблюдается соответствие между двумя главными членами предложения – подлежащим (*07-92* и *07-93*) и сказуемым (*07-94*). В дискурсивном фрагменте (4.7) отражено соответствие между обособленным членом и основной частью предложения.

(4.5)

C04-188. 双方将致力于促进亚太地区形成完整的安全合作体系，/ (*дословный перевод: Стороны будут способствовать формированию в Азиатско-Тихоокеанском регионе целостной кооперативной системы безопасности и сотрудничества /*)

C04-189. 所有参加国享有平等权利。/ (*дословный перевод: все государства-участники имеют равные права*)

R04-188. *Стороны будут способствовать формированию в Азиатско-Тихоокеанском регионе целостной кооперативной системы безопасности и сотрудничества /*

R04-189. *с равными правами для всех государств-участников. /*

(4.6)

C07-92. 中俄对国际政治的重大原则问题立场一致，/ (*дословный перевод: Китай и Россия занимают совпадающие позиции по ключевым принципиальным вопросам международной политики,*)

C07-93. 在主要地区和国际问题上的立场相同或相近，/ (*дословный перевод: (Стороны) занимают одинаковые или близкие позиции по основным региональным и международным проблемам,*)

C07-94. 这使两国能够更有效地参与国际合作，/ (*дословный перевод: это позволяет двум странам более эффективно участвовать в международном сотрудничестве.*)

R07-92. *Единство подходов к принципиальным вопросам мировой политики, /*

R07-93. *совпадение или близость позиций России и Китая по основным региональным и международным проблемам /*

R07-94. *позволяют им все более эффективно участвовать в международном сотрудничестве, /*

(4.7)

10-231. 总体积极评价阿富汗和平重建取得的进展，/ (*дословный перевод: (Стороны) в целом позитивно оценивают прогресс в деле мирного восстановления Афганистана,*)

10-232. 表示将继续积极参与这一重建进程。/ (*дословный перевод: (Стороны) выражают намерение продолжить активное участие в этом процессе*)

10-231. *В целом позитивно оценивая прогресс в деле мирного восстановления Афганистана, /*

10-232. *Стороны выражают намерение продолжить активное участие в этом процессе. /*

В корпусе очень редко встречаются тип III (2,8 %), то есть соответствие двух китайских клауз в разных предложениях двум русским простым предложениям в составе сложного предложения, и тип IV (1,6 %) – соответствие двух китайских клауз в разных предложениях членам простого предложения в русском тексте.

Тип III можно найти в корпусе в структурных парах *08-76 – 08-77, 08-76 – 08-78*, и *08-76 – 08-79* – см. дискурсивный фрагмент (4.8). В этих трех структурных парах две китайские клаузы находятся в разных предложениях, а два РСА являются простыми предложениями в составе сложного предложения, соединенными коннектором «*для обеспечения которого*».

(4.8)

08-76. 双方认为，可持续发展是国际合作的重要领域。/

08-77. 各国应加强经验交流，/

08-78. 保护自然资源和生物多样性，/

08-79. 努力建立环境友好型、资源节约型社会。

08-76. *Стороны считают, что устойчивое развитие является важной сферой международного сотрудничества, /*

08-77. ***для обеспечения которого*** *все страны должны наращивать обмен опытом, /*

08-78. *охранять природные ресурсы и биологическое разнообразие, /*

08-79. *предпринимать усилия по формированию общества, дружественного к окружающей среде и бережно использующего ресурсный потенциал.*

Тип IV представлен в корпусе как соответствие между двумя китайскими клаузами, расположенными в разных предложениях, и двумя РСА в пределах одного русского предложения. Подобное соответствие может включать различные синтаксические реализации, включая подлежащее и сказуемое, основную часть предложения и обособленный член и т. п. Примером может служить структурная пара 06-189 – 06-191 в дискурсивном фрагменте (4.9): китайские клаузы находятся в разных предложениях, а соответствующие им два РСА – подлежащее и сказуемое.

(4.9)

C06-189. 上海合作组织五周年峰会将于 2006 年 6 月在上海举行。/ (*дословный перевод: Шанхайская организация сотрудничества пятая годовщина саммит будет в 2006 году в июне в Шанхае проводиться***.**)

C06-190. 峰会将总结该组织的工作经验，/ (*дословный перевод: Саммит подведет итоги этой организации работы опыта,*)

C06-191. 根据各成员国通过的《上海合作组织宪章》规定的任务和原则为组织的发展注入新的活力。(*дословный перевод: согласно всеми государствами-членами принятой «Шанхайской организации сотрудничества Устав» определенных задач и принципов, для организации развития придать новый импульс.*)

R06-189. *Предстоящий в июне 2006 года в Шанхае юбилейный саммит ШОС /*

R06-190. *призван на основе обобщения опыта работы Организации /*

R06-191. *придать новый импульс ее дальнейшему укреплению и развитию в соответствии с задачами и принципами, заложенными государствами-учредителями в Хартии ШОС. /*

Итак, согласно статистическим данным, наибольшую долю среди несовпадающих вариантов имеет тип II, то есть соответствие двух китайских клауз в составе одного сложного предложения двум РСА в одном простом предложении (30,8 %). Это свидетельствует о том, что часть смысловых отношений, выраженных между независимыми единицами в китайском языке, в русском языке может вполне оформляться на уровне структуры простого предложения. Как показано выше в дискурсивных фрагментах (4.5), (4.6), и (4.7), логико-семантические отношения между китайскими клаузами реализуются в русском тексте посредством внутренних синтаксичесикх связей между РСА.

Подобные соответствия, представленные типами II и IV (всего 404 случая, что составляет 32,4 %) отражают соответствие межклаузных семантических связей и внутрипредложенческих синтаксических структур, выявленное при осуществлении дискурсивной разметки. Эти случаи соотношения дискурсивных средств выражения имеют особую исследовательскую ценность с точки зрения сравнения структурных уровней китайских и русских текстов. Поскольку ДО и ДК в подобных случаях выходят за рамки традиционного дискурсивного анализа, будем называть их ДО-аналогами и ДК-аналогами и рассматривать здесь и далее отдельно от прочих ДО и ДК.

4.4.2. Типы дискурсивных отношений

Как уже упоминалось в п. 3.2.3.2, для обеспечения качества разметки корпуса и уменьшения субъективности при определении ДО в их разметке указывался один первичный и не более двух вторичных типов. В табл. 4.5 представлены статистические данные по первичным типам ДО китайского и русского языков. Разметка включала тринадцать типов ДО, в том числе двенадцать

основных логико-семантических отношений: пояснение, соединение, причина – следствие, цель, дополнение, оценка, время, противопоставление, сопоставление, условие, градация и уступка. Для разметки русских ДО в данном корпусе к списку было добавлено тринадцатое, обобщенно названное синтаксическим отношением внутри простого предложения. К этому типу мы относим отношения между подлежащим и сказуемым, сказуемым и дополнением, они встречаются только в разметке структурных пар РСА – членов простого предложения.

Таблица 4.5. Первичные типы ДО в китайских и русских официально-деловых текстах.

Типы ДО (первичные)	**ДО**				**ДО-аналоги**			
	между китайскими клаузами		**между РСА**		**между китайскими клаузами**		**между РСА**	
Пояснение++	231	18,5 %	238	19,1 %	86	6,9 %	100	8,0 %
Соединение++	201	15,9 %	217	14,9 %	82	4,1 %	87	4,7 %
Причина – следствие++	198	16,1 %	186	17,4 %	51	6,6 %	59	7,0 %
Цель++	98	7,9 %	88	7,1 %	133	10,7 %	96	7,7 %
Дополнение++	32	2,6 %	31	2,5 %	4	0,3 %	3	0,2 %
Оценка++	23	1,8 %	22	1,8 %	4	0,3 %	1	0,1 %
Время++	29	2,3 %	30	2,4 %	10	0,8 %	11	0,9 %
Противопоставление++	16	1,3 %	17	1,4 %	4	0,3 %	3	0,2 %
Сопоставление++	8	0,6 %	8	0,6 %	–	–	–	–
Условие++	5	0,4 %	5	0,4 %	29	2,3 %	30	2,4 %
Градация++	3	0,2 %	2	0,2 %	–	–	–	–
Уступка++	–	–	–	–	1	0,1 %	1	0,1 %
Синтаксическое отношение внутри простого предложения++*	–	–	–	–	–	–	13	1,0 %
Итого	844	67,6 %	844	67,6 %	404	32, 4%	404	32,4%

Из таблицы видно, что четырьмя наиболее частотными первичными логико-семантическими отношениями в официально-деловых текстах являются пояснение, соединение, причина – следствие и цель, в то время как остальные

встречаются гораздо реже. Подавляющее большинство ДО в корпусе встречаются с очень похожей частотой в китайских и русских текстах (с разницей ≤ 1,3 %), за исключением ДО «цель».

Вторичные типы ДО выбираются из перечисленных выше двенадцати логико-семантических типов, и статистические данные по ним приведены в табл. 4.6. В итоге сравнительно небольшое количество ДО имеет вторичные типы: 174 русских ДО (из них 5 имеют два вторичных типа) и 160 китайских ДО (из них 4 имеют два вторичных типа); 53 русских ДО-аналога (из них 1 ДО было дополнено двумя вторичными типами) и 72 китайский ДО-аналог (из них 1 ДО было дополнено двумя вторичными типами). Статистика говорит о том, что в корпусе большинство ДО являются однозначно определяемыми и только небольшое количество ДО нуждается в разметке вторичных типов.

Таблица 4.6. Разметка вторичного типа ДО в китайских и русских официально-деловых текстах.

	между РСА		между китайскими клаузами	
ДО без разметки вторичных типов	670		684	
ДО-аналоги без разметки вторичных типов	351		332	
	вторичный 1	вторичный 2	вторичный 1	вторичный 2
ДО с разметкой вторичных типов	169	5	156	4
ДО-аналоги с разметкой вторичных типов	52	1	71	1

В табл. 4.7 представлены все первичные типы ДО, которые были дополнены вторичными типами. Как следует из таблицы, из двенадцати типов девять первичных ДО были дополнены разметкой вторичного типа, а остальные три (противопоставление, сопоставление и уступка) не были. Эти три типа ДО, как правило, имеют более сильные и яркие семантические оттенки, и благодаря

этому они могут быть идентифицированы без добавления других типов отношений.

Наиболее распространенным из дополненных вторичными типами ДО является «причина – следствие++» – всего 151 раз, что составляет 32,7 %; далее следуют ДО «соединение++» (119 раз, 25,9 %) и «пояснение++» (77 раз, 16,7 %). Отметим, что наибольшее разнообразие вторичных типов (7) дополняет ДО «соединение++». Кроме того, ДО «соединение+» как вторичное сочетается с большинством первичных типов, за исключением ДО «дополнение++» и «условие++».

Таблица 4.7. Разметка первичных и вторичных типов ДО и ДО-аналогов в дискурсивном корпусе.

Первичные типы	Вторичные типы	ДО		ДО-аналоги		Итого
		между РСА	между китайскими клаузами	между РСА	между китайскими клаузами	
Причина – следствие++	Цель+	24	26	31	19	151, 32,7 %
	Соединение+	13	8	1	2	
	Время+	5	5	–	–	
	Пояснение+	3	4	–	1	
	Условие+	1	1	–	1	
	Оценка+	1	1	1	1	
	Противопоставление+	1	1	–	–	
Соединение++	Пояснение+	19	17	3	4	119, 25,9 %
	Цель+	18	12	2	4	
	Причина – следствие+	12	10	–	2	
	Время+	4	4	–	2	
	Градация+	1	2	–	–	
	Сопоставление+	1	1	–	–	

	Оценка+	1	–	–	–	
Пояснение++	**Соединение+**	15	16	–	1	77, 16,7 %
	Причина – следствие+	6	5	–	–	
	Цель+	3	3	6	15	
	Оценка+	2	2	–	–	
	Время+	2	1	–	–	
Время++	**Соединение+**	18	14	3	4	44, 9,5 %
	Причина – следствие+	1	1	1	1	
	Градация+	1	–	–	–	
Цель++	Пояснение+	1	5	–	1	32, 6,9 %
	Соединение+	5	3	–	4	
	Условие+	1	1	2	4	
	Оценка+	1	1	–	–	
	Причина – следствие+	1	1	–	1	
Оценка++	Пояснение+	3	3	–	–	15, 3,2 %
	Причина – следствие+	2	2	–	1	
	Соединение+	1	1	–	–	
	Время+	–	–	1	1	
Дополнение++	Пояснение+	5	6	–	–	11, 2,4 %
Градация++	**Соединение+**	2	2	–	–	4, 0,9 %
Условие++	Цель+	–	–	3	3	6 1,3 %
Итого		174	159	54	72	459

Для более точного понимания степени совпадения ДО в двух языках на диагр. 4.2 представлена сводная статистика реализации ДО в китайских

и русских официально-деловых текстах. Диаграмма наглядно показывает, что в подавляющем большинстве случаев (88,4 %) ДО в разметке китайских и русских структурных пар совпадают. Таким образом, можно говорить о высокой степени структурного соответствия между двумя языками на уровне дискурсивных отношений.

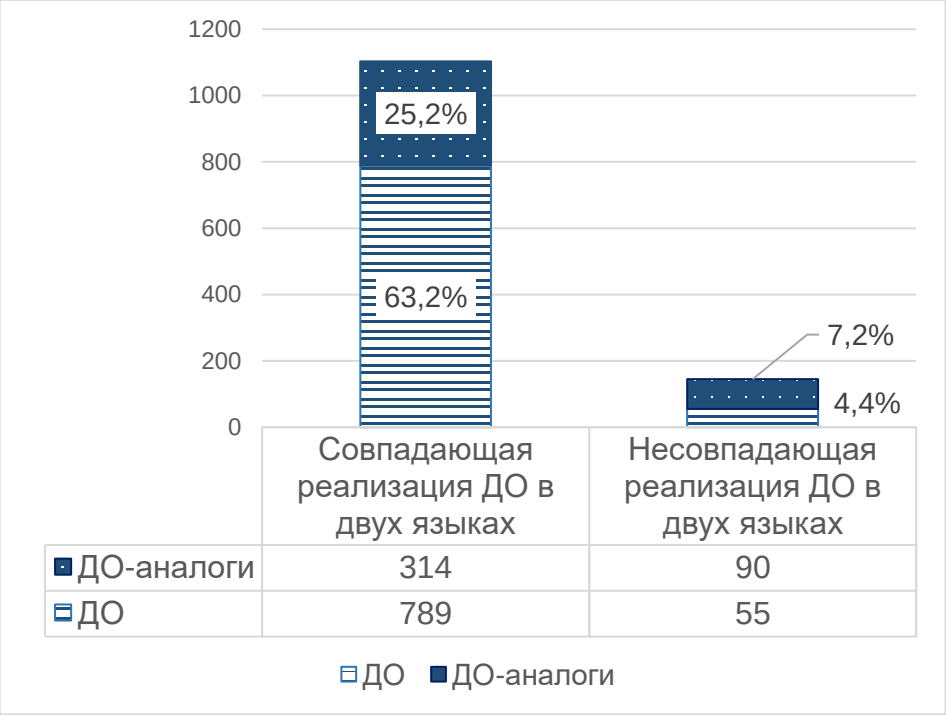


Диаграмма 4.2. Реализация ДО в китайском и русском языках.

4.4.3. Дискурсивные отношения и дискурсивные коннекторы

Как отмечалось ранее, дискурсивные отношения, выражаемые при помощи дискурсивных коннекторов, являются эксплицитными, а без коннекторов – имплицитными. Статистические данные по эксплицитным и имплицитным ДО представлены на диагр. 4.3. Можно говорить о том, что большинство ДО в официально-деловых текстах являются имплицитными, то есть реализуются без выраженного дискурсивного коннектора. Однако в русских текстах ДК используются чаще, чем в китайских. Результаты в целом соответствуют общему пониманию характеристики китайского языка, который в большей степени фокусируется на функции и значении, что приводит к необязательности коннекторов – см. об этом: [连淑能, 1993, p. 48–75].

Диаграмма 4.3. Распределение имплицитных и эксплицитных ДО.

Список русских дискурсивных коннекторов представлен в табл. 4.8. В корпусе встречается 39 русских ДК, в общей сложности 198 вхождений, и 43 ДК-аналогов, в общей сложности – 177.

Таблица 4.8. Русские дискурсивные коннекторы.

№	ДК	кол-во	№	ДК-аналоги	кол-во
1	и	91	1	В целях	27
2	который	16	2	и	22
3	что	11	3	по	19
4	а также	11	4	Для	13
5	в этой связи	7	5	с целью	12
6	В этих целях	6	6	в том числе	8
7	чтобы	6	7	а также	8
8	с тем чтобы	5	8	в интересах	6
9	а	4	9	на основе	5
10	для обеспечения которого	3	10	в целях	5
11	Однако	3	11	в соответствии с	5
12	В этом контексте	3	12	только	3

13	По-прежнему	2	13	прежде всего	3
14	и в этих целях	2	14	в особенности	3
15	тем самым	2	15	можно лишь	3
16	В то же время	2	16	во имя	3
17	и ... в этих целях	2	17	с учетом	2
18	С другой стороны	1	18	на	2
19	Одновременно	1	19	без	2
20	В этих целях	1	20	в духе	2
21	в то время как	1	21	через	2
22	поскольку	1	22	при	1
23	во имя	1	23	благодаря	1
24	с тем чтобы	1	24	тем самым	1
25	и с этой целью	1	25	как	1
26	для того, чтобы	1	26	в строгом соответствии с	1
27	а также	1	27	в связи с	1
28	Для этого	1	28	в полном соответствии с	1
29	как бы ни	1	29	в частности	1
30	Поэтому	1	30	перед лицом	1
31	как ... так и	1	31	вне зависимости от	1
32	в связи с тем, что	1	32	По мере	1
33	какие бы ни	1	33	наряду с	1
34	также	1	34	прежде всего	1
35	в какой	1	35	не … а	1
36	и в этом контексте	1	36	с	1
37	но	1	37	не … но	1
38	и далее	1	38	в том числе и	1
39	и на его основе	1	39	Одновременно	1
Итого		**198**	40	в рамках	1

	41	особенно	1
	42	и исходя из	1
	43	и прежде всего	1
	Итого		**177**

В табл. 4.9 перечислены китайские дискурсивные коннекторы. В корпусе были размечены 34 разновидных китайских ДК, которые встречаются в 206 случаях. В итоге китайские ДК встречаются реже, чем русские – как по виду, так и по частоте.

Таблица 4.9. Китайские дискурсивные коннекторы.

№	ДК	кол-во	*№*	ДК	кол-во	*№*	ДК	кол-во
1	并	53	*13*	因此	2	*25*	并…为此目的	1
2	以	24	*14*	以便	2	*26*	另一方面	1
3	为	21	*15*	考虑到	2	*27*	但是	1
4	为此	18	*16*	不是…而是	2	*28*	以及	1
5	包括	13	*17*	尤其	2	*29*	并且	1
6	即	11	*18*	首先	2	*30*	不管…都	1
7	同时	10	*19*	其中包括	2	*31*	既不…也不	1
8	只有…才	9	*20*	并且…为此	1	*32*	首先是	1
9	特别是	7	*21*	以及为此	1	*33*	鉴于	1
10	也	6	*22*	并在此基础上	1	*34*	使	1
11	但	3	*23*	虽…但	1	**Итого**		**206**
12	无论	2	*24*	并……为此	1			

По граф. 4.1 видно, что лишь некоторые виды ДК используются с высокой частотой (≥ 10 раз); большинство видов ДК используются очень редко (≤ 3

раз) как в русском, так и в китайском языках. Общее распределение частот ДК в основном соответствует закону Ципфа [Zipf, 1932] – одному из базовых представлений о распределении лексических частот.

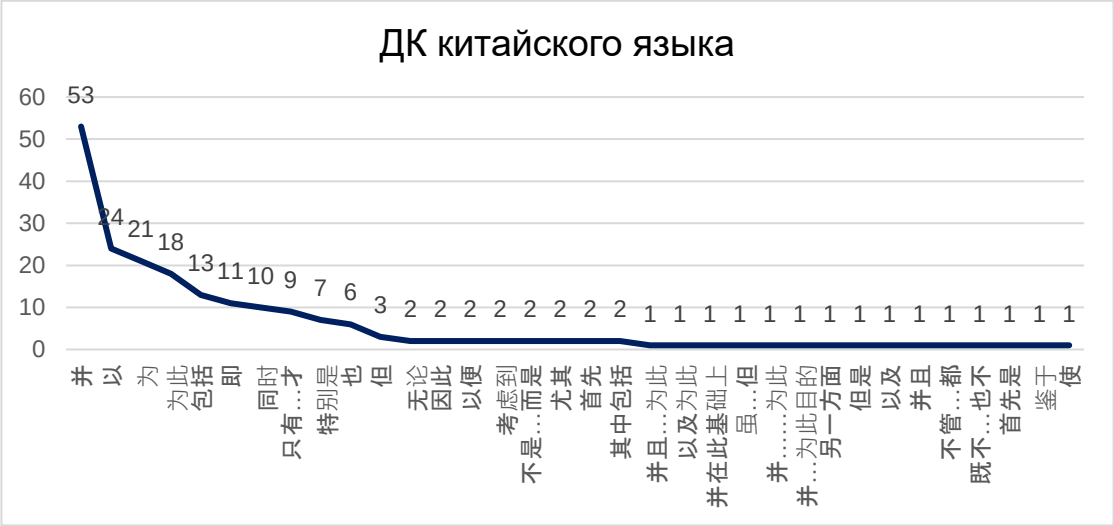


График 4.1. Частотное распределение использования китайских и русских дискурсивных коннекторов.

4.4.4. Дискурсивные вершины структурных пар

Дискурсивные вершины каждой структурной пары были вычислены статистически, исходя из количества семантических связей конкретных ЭДЕ (см. об этом в п. 3.2.3.3.). Полученное распределение дискурсивных вершин показано ниже, на диагр. 4.4. Структурные пары классифицируются в зависимости от положения вершины: в одних вершиной является первая ЭДЕ (ЭДЕ-1), в других – вторая (ЭДЕ-2).

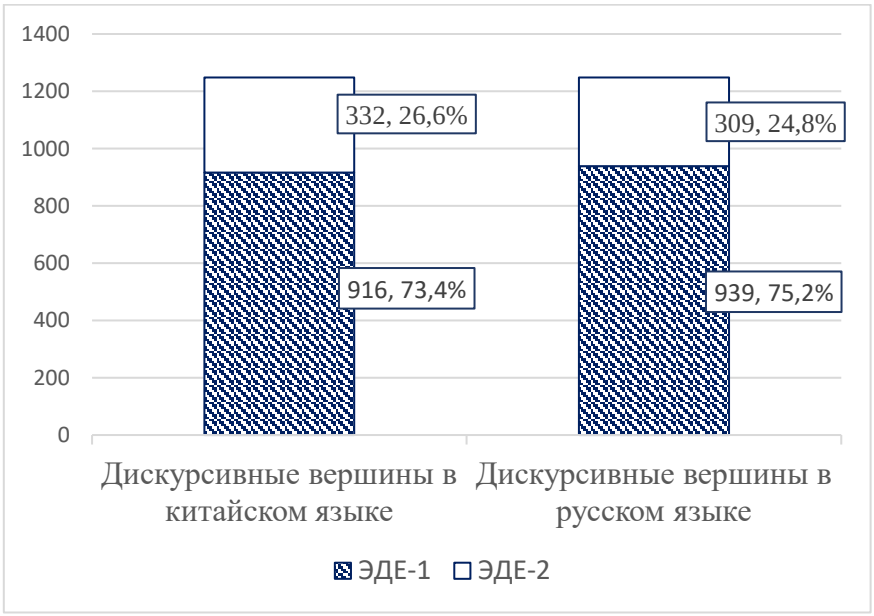


Диаграмма 4.4. Распределение дискурсивных вершин.

Диагр. 4.4 показывает, что большинство дискурсивных вершин в китайских и русских официально-деловых текстах соответствуют ЭДЕ-1, и их частота в китайском и русском языках в целом совпадает. Такое сходство неудивительно в параллельном дискурсивном корпусе, в котором произведено выравнивание текстов. Однако в русских официально-деловых текстах представлено относительно больше структурных пар с вершинной ЭДЕ-1, чем в китайских. Это говорит о том, что при выравнивании структур дискурсивные вершины могут не совпадать.

Чтобы глубже изучить причины несовпадения дискурсивных вершин в параллельных текстах, мы подсчитали их распределение в разных синтаксических вариантах структурных пар (см. табл. 4.10).

Таблица 4.10. Совпадение дискурсивных вершин в китайском и русском языках.

	Совпадающая реализация дискурсивных вершин в двух языках	Несовпадающая реализация дискурсивных вершин в двух языках
ЭДЕ (клаузы и РСА) в одном сложном предложении	359	6
ЭДЕ (клаузы и РСА) в разных предложениях	477	2
РСА внутри простого предложения	373	31
Итого	1209	39

Были выявлены две основные причины несовпадения дискурсивных вершин в структурных парах с РСА внутри простого предложения.

Во-первых, несовпадение дискурсивных вершин в основном вызвано разницей в порядке следования ЭДЕ. Например, в (4.10) первая китайская ЭДЕ (*C96-96*) семантически соответствует второй русской ЭДЕ (*R96-97*), а вторая китайская (*C96-97*) – первой русской (*R96-96*). В результате дискурсивная вер-

шина созданной параллельной структурной пары в китайском и русском текстах (структурная пара *ЭДЕ 96-96 – 96-97*) будет соответствовать ЭДЕ-1 в одном языке (*R96-96*) и ЭДЕ2 – в другом (*C96-97*) (см. пример 4.2).

(4.10)

C96-96.(=R96-97) 双方认为，应在保持联合国安理会对维护世界和平与安全负主要责任的情况下，按照联合国宪章第八章，/ *(дословный перевод: Стороны считают, что необходимо при сохранении главной ответственности за Советом Безопасности за поддержание международного мира и безопасности, в соответствии с главой VIII Устава ООН, /*)

C96-97.(=R96-96) 促进联合国和地区性组织之间在防止与和平调解争端和冲突方面进行合作而做出的努力；/ *(дословный перевод: содействовать усилиям по налаживанию сотрудничества между ООН и региональными организациями в предотвращении и мирном урегулировании споров и конфликтов)*

C96-98. 促进从事经济和社会的发展、人道主义援助等问题的非政府组织同联合国及其在上述领域的专门机构的工作中进行更具建设性和健康的协调。*(дословный перевод: способствовать более конструктивной и здоровой координации деятельности неправительственных организаций, занимающихся проблемами экономического и социального развития и гуманитарной помощи, с работой ООН и ее специализированных учреждений в этих областях.*)

R96-96.(=R96-97) Стороны считают, что необходимо содействовать усилиям по налаживанию сотрудничества между ООН и региональными организациями /

R96-97.(=R96-96) в соответствии с главой VIII Устава ООН в предотвращении и мирном урегулировании споров и конфликтов при сохранении главной ответственности за Советом Безопасности за поддержание международного мира и безопасности; /

96-98. способствовать более конструктивной и здоровой координации деятельности неправительственных организаций, занимающихся проблемами экономического и социального развития и гуманитарной помощи, с работой ООН и ее специализированных учреждений в этих областях.

Пример 4.2. Дискурсивная структура зависимостей дискурсивных фрагментов (4.10).

Во-вторых, несовпадающая реализация дискурсивных вершин обусловлена сегментацией РСА вслед за китайскими клаузами и выравниванием дискурсивных структур. Как показывает пример 4.3, в китайской структурной паре C05-77 – C05-78 вершиной является C05-78, а в русской – R05-77. Здесь вершины и корневые узлы определяются вручную на основе их семантического «веса», с учетом нескольких факторов. В китайском языке ЭДЕ C05-78 считается вершиной, потому что все последующие ЭДЕ (C05-79, C05-80, C05-81 и C05-82) служат для объяснения содержания C05-78. Однако, причина, по которой эти ЭДЕ не напрямую связаны с ЭДЕ C05-78 в сформированной дискурсивной структуре, заключается в том, что китайская структура здесь построена в соответствии с русской, что обеспечивает выравнивание. В русском языке, с другой стороны, РСА 05-77 считается вершиной, которая связывается с другими еди-

ницами, поскольку он имеет более полную синтаксическую структуру по сравнению с PCA 05-78.

(4.11)

C05-77. 双方呼吁国际社会共同努力，/ (*дословный перевод: Стороны призывают международное сообщество к совместной работе,*)

C05-78. 建立互信、互利、平等、协作的新型安全架构。/ (*дословный перевод: создать новую архитектуру безопасности, основанную на взаимном доверии, взаимной выгоде, равноправии и взаимодействии.*)

C05-79. 此架构应以公认的国际关系准则为政治基础，/ (*дословный перевод: Эта архитектура должна быть основана в политическом плане на общепризнанных нормах международных отношений,*)

C05-80. 以互利合作和共同繁荣为经济基础，/ (*дословный перевод: основана в экономическом плане на взаимовыгодном сотрудничестве и совместном процветании,*)

C05-81. 并应建立在尊重各国平等安全权利的基础上。/ (*дословный перевод: и должно базироваться на уважении равного права всех государств на безопасность.*)

C05-82. 平等对话、协商和谈判应成为解决矛盾和维护和平的手段。(*дословный перевод: Способом разрешения противоречий и защиты мира должны быть равноправный диалог, консультации и переговоры.*)

R05-77. Стороны призывают объединить усилия международного сообщества /

R05-77. по созданию новой архитектуры безопасности, основанной на взаимном доверии, взаимной выгоде, равноправии и взаимодействии. /

R05-78. Ее политической основой должны стать общепризнанные нормы международных отношений, /

R05-79. экономической – взаимовыгодное сотрудничество и совместное процветание. /

R05-80. Новая архитектура безопасности должна базироваться на уважении равного права всех государств на безопасность. /

R05-81. Способом разрешения противоречий и защиты мира должны быть равноправный диалог, консультации и переговоры.

Пример 4.3. Дискурсивная структура зависимостей дискурсивных фрагментов (4.11).

Следует отметить, что такие случаи в корпусе не являются типичными и встречаются крайне редко – возможно, потому, что привлеченные документы составлялись обеими сторонами одновременно, что снизило вероятность семантических несоответствий.

Выводы по четвертой главе

В данной главе на основе корпусной статистики были сопоставлены дискурсивные характеристики китайских и русских официально-деловых текстов (совместных деклараций и заявлений), а также была дана интерпретация значимых показателей в контексте существующих лингвистических знаний.

В корпусе аннотированы 12 параллельных текстов, состоящих из 858 китайских и русских абзацев. Величина абзаца (количество ЭДЕ в абзаце) варьируется от 1 до 13 ЭДЕ, при этом большинство абзацев (89,0 %) состоит менее чем из 6 ЭДЕ. Это говорит о том, что объем абзацев в официально-деловых текстах, как правило, небольшой.

Китайским клаузам в большинстве случаев соответствуют монопредикативные РСА (61,6 %), которые имеют аналогичные синтаксические параметры с исходными ЭДЕ. В корпусе также заметна доля РСА без предиката (19,4 %), полипредикативных РСА (13,7 %) и полупредикативных РСА (5,3 %). Анализ показал, что «синтаксическая плотность» дискурсивных единиц китайского и русского языков в разной степени увеличивается или уменьшается при передаче одинакового объема информации.

В параллельном дискурсивном корпусе для китайского и русского языков по отдельности было построено 1252 структурные пары, из них абсолютное большинство было выравнено. Более 60 % китайских и русских структурных пар образуются из двух ЭДЕ внутри одного предложения, причем внутри сложных предложений размечены 62,5 % китайских и только 29,2 % русских структурных пар; 32,4 % русских структурных пар размечены внутри простого предложения. Синтаксические варианты структурных пар совпадают в 59,5 % и не совпадают в 40,5 %; это означает, что во многих случаях структурные пары китайского и русского языков формируются из ЭДЕ одного и того же синтаксического уровня. Среди всех несовпадающих реализаций синтаксических вариантов наибольшую долю имеет соответствие двух китайских клауз в составе одного сложного предложения двум РСА внутри одного простого предложе-

ния, что определяется исходной установкой при сегментации параллельного корпуса.

Из двенадцати размеченных дискурсивных отношений в официально-деловых текстах наиболее частотными являются пояснение, соединение, причина – следствие и цель, в то время как остальные девять отношений встречаются гораздо реже. Подавляющее большинство ДО в корпусе встречаются с очень похожей частотой в двух языках.

Большинство ДО между ЭДЕ являются однозначно определяемыми, и только небольшое количество дискурсивных отношений нуждается в разметке вторичных типов. Дискурсивные отношения, обладающие наиболее ярким семантическим оттенком (противопоставление, сопоставление и уступка) идентифицированы без добавления вторичных типов. В добавлении вторичного типа чаще всего нуждаются ДО «причина – следствие», «соединение» и «пояснение», а наиболее часто используемым вторичным типом является «соединение».

Большинство ДО являются имплицитными, то есть реализованы без дискурсивного коннектора, при этом русские ДК используются относительно чаще, чем китайские. Большинство ДК в корпусе официально-деловых текстов являются однозначными, а многозначность проявляют в первую очередь соединительные союзы.

Дискурсивные вершины в большинстве случаев соответствуют первой ЭДЕ в структурной паре, и частота встречаемости вершинных ЭДЕ-1 в китайских и русских текстах в целом совпадает. Однако дискурсивные вершины двух языков далеко не всегда идентичны. Несовпадение китайских и русских дискурсивных вершин в основном вызвано несовпадением порядка следования ЭДЕ в двух текстах и сегментацией РСА вслед за китайскими клаузами.

ЗАКЛЮЧЕНИЕ

В диссертационном исследовании была апробирована универсальная схема формирования дискурсивной структуры зависимостей, и на ее основе построен китайско-русский параллельный дискурсивный корпус официально-деловых текстов. Исследование, проведенное на базе созданного корпуса, позволило выявить значимые сходства и различия в дискурсивных структурах китайских и русских официально-деловых текстов.

Опыт создания корпуса показал, что бинарная структура зависимостей отличается простотой формы и функциональной гибкостью, а также демонстрирует высокую адаптивность при применении к описанию сложных дискурсивных явлений. Дискурсивный структурный анализ имеет в первую очередь семантический характер. Дерево зависимостей при анализе письменного текста строится на основе абзаца как единого целого; дискурсивные структурные пары состоят из двух элементарных дискурсивных единиц (китайских клауз и русских синтаксических аналогов), связанных логико-семантическими отношениями, а направление зависимостей определяется семантическим «весом» ЭДЕ.

Создание параллельного дискурсивного корпуса официально-деловых текстов было осуществлено поэтапно: 1) этап отбора материала; 2) деление параллельных текстов на элементарные дискурсивные единицы и выравнивание текстов; 3) формирование дискурсивных структурных пар, оптимизация дискурсивной структуры зависимостей и выравнивание дискурсивных структур; 4) разметка дискурсивных структур с указанием дискурсивных отношений, дискурсивных коннекторов и дискурсивных вершин; 5) ручная проверка.

Отобранные для корпуса параллельные тексты – межправительственные двусторонние документы РФ и КНР – изначально подготовлены и согласованы обеими сторонами, поэтому они семантически эквивалентны и не являются оригиналами и переводами в классическом понимании, что исключает проблемы, возникающие в связи с качеством перевода. В то же время сегментация и выравнивание корпуса по ЭДЕ проводились в направлении от китайских тек-

стов к русским, что обусловило специфику выделяемых русских синтаксических аналогов китайских клауз.

Были выделены и сопоставлены структурные и семантические характеристики китайских и русских дискурсивных единиц. Типичны для китайских клауз простые предложения, простые предложения в составе сложных и протяженные обстоятельства причины, цели и условия, находящиеся в начале предложения. Русские синтаксические аналоги подразделяются на монопредикативные, полупредикативные, полипредикативные и РСА без предиката. Китайским клаузам в большинстве случаев соответствуют монопредикативные РСА, а остальные типы демонстрируют увеличение или уменьшение «синтаксической плотности» при передаче одной и той же информации.

Из структурных пар ЭДЕ при разметке корпуса сформированы дискурсивные деревья зависимостей, отвечающие общим ограничениям, то есть единственности корневого узла, связанности, ацикличности и проективности. Абсолютное большинство структурных пар были успешно выравнены.

Сопоставление дискурсивных отношений, дискурсивных коннекторов и дискурсивных вершин в китайской и русской части корпуса показало значимые сходства и различия в структурах официально-деловых текстов. Часть различий обусловлена понятной разницей между синтаксисом двух языков. Происхождение других было связано со спецификой построения и выравнивания китайско-русского корпуса. Этими основными причинами обусловлена и разница в наборе дискурсивных коннекторов и в порядке следования дискурсивных вершин. С другой стороны, реализация дискурсивных отношений в китайских и русских текстах оказалась достаточно близкой, а сами отношения – преимущественно однозначно определяемыми. При этом большинство ДО являются имплицитными, что потребовало при разметке восстановления дискурсивных коннекторов.

Таким образом, корпусная разметка и проведенное межъязыковое сопоставление формируют динамическую модель структурного дискурсивного ана-

лиза, а созданный китайско-русский параллельный дискурсивный корпус официально-деловых текстов открыт для дальнейших синтаксических, дискурсивных, семантических исследований, а также для решения различных лингвистических и междисциплинарных задач, включая переводоведение, компьютерную лингвистику и сравнительное изучение языков.

Перспективы дальнейшей разработки темы

Предложенная корпусная методика предполагает сегментацию русских дискурсивных единиц вслед за китайскими клаузами. Это не препятствует сопоставлению китайских и русских структурных единиц на уровне дискурса и сопоставлению самих дискурсивных структур. Однако из-за очередности в сегментации корпус на текущий момент имеет ограниченные возможности для изучения дискурсивных особенностей русских официально-деловых текстов. Для решения этой проблемы предполагается в перспективе создать дискурсивный корпус с разметкой зависимостей на основе исходного русского материала.

СПИСОК СОКРАЩЕНИЙ И ТЕРМИНОВ

ЭДЕ – элементарная дискурсивная единица

PDTB – Пенсильванский дискурсивный древовидный банк

БЯМ – большая языковая модель

ДК – дискурсивный коннектор

ДО – дискурсивное отношение

ИИ – искусственный интеллект

НКРЯ – Национальный корпус русского языка

РСА – соответствующие китайским клаузам русские синтаксические аналоги

ССЦ – сложное синтаксическое целое

ТРС – теория риторических структур

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Азарова И. В., Митрофанова О. А., Синопальникова А. А. Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог», 2003. С. 43–50.

2. Ананьева М. И., Кобозева М. В. Дискурсивный анализ в задачах обработки естественного языка // Труды IV Всероссийской научной конференции молодых ученых с международным участием. Тверь: Тверской государственный технический университет, 2016a. С. 138–148.

3. Ананьева М. И., Кобозева М. В. Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // Труды международного семинара «Диалог'2016» по компьютерной лингвистике и ее приложениям. Москва: Издательство РГГУ, 2016b. С. 22–28.

4. Арутюнова Н. Д. Дискурс. Речь // Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой. 2-е изд., доп. Москва: Большая российская энциклопедия, 2002.

5. Бакиева А. М. Методы автоматического анализа текстов на казахском языке. Новосибирск: Новосибирский национальный исследовательский государственный университет, 2017. С. 151.

6. Бакиева А. М., Батура Т. В. Исследование применимости теории риторических структур для автоматической обработки научно-технических текстов // Cloud of science. 2017a. Т. 4. № 3. С. 450–462.

7. Бакиева А. М., Батура Т. В. Применение теории риторических структур в системах автоматической обработки текстов // Семантические технологии. 2017b. Т. 1. С. 19–29.

8. Баранов А. Н. Введение в прикладную лингвистику. 6-е изд. Москва: URSS, 2021. 368 с.

9. Баркович А. А. Корпусная лингвистика: специфика современных метаописаний языка // Вестник Томского государственного университета. 2016. № 406. С. 5–13.

10. Батура Т. В., Бакиева А. М. Создание системы автоматического реферирования научных текстов // Вестник НГУ. Серия: Информационные технологии. 2018. Т. 16. № 3. С. 74–86.

11. Бекяшев К. А. Международное право. Москва: Проспект, 2020. 896 с.

12. Белошапкова В. А. Современный русский язык. Синтаксис. Москва: Высшая школа, 1977. 248 с.

13. Бенвенист Э. Уровни лингвистического анализа // Общая лингвистика. Москва, 1962. С. 129–140.

14. Бондарко А. В. Теория функциональной грамматики: Введение. Аспектуальность. Временная локализованность. Таксис. Ленинград: Наука, 1987. 347 с.

15. Борискина О. О. Корпусное исследование языка: мода или необходимость? // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2015. № 3. С. 24–27.

16. Валгина Н. С. Теория текста. Москва: Логос, 2003. 173 с.

17. Васильев Л. М. Системный семантический словарь русского языка: Предикаты свойства, поведения и звучания: учеб. пособие для студентов и аспирантов. Уфа: Башкирский университет, 2000. 146 с.

18. Величко М. А. Когезия и когерентность: особенности разграничения и определения понятий // Вестник Адыгейского государственного университета. Серия 2: Филология и искусствоведение. 2016. № 2 (177). С. 39–43.

19. Водясова Л. П. Сложное синтаксическое целое как основная структурная единица микротекста в прозе К. Г. Абрамова. Саранск: Мор-

довский государственный педагогический институт имени М. Е. Евсевьева, 2013. 115 с.

20. Гальперин И. Р. Текст как объект лингвистического исследования. Москва: Наука, 1981. 140 с.

21. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общений. Москва: Наука, 1985. 143 с.

22. Гладкий А. В. Формальные грамматики и языки. Москва: Наука, 1973. 368 с.

23. Гладкий А. В., Мельчук И. А. Элементы математической лингвистики. Москва: Наука, 1969. 192 с.

24. Горелов В. И. Грамматика китайского языка. 2-е изд. Москва: Просвещение, 1982. 278 с.

25. Дейк Т. А. ван. Язык. Познание. Коммуникация. Москва: Прогресс, 1989. 312 с.

26. Добровольский Д. О. Корпусный подход к исследованию фразеологии: новые результаты по данным параллельных корпусов // Вестник Санкт-Петербургского университета. Язык и литература. 2020. Т. 17. № 3. С. 398–411.

27. Добровольский Д. О. Корпус параллельных текстов и сопоставительная лексикология // Труды Института русского языка им. В. В. Виноградова. 2015. № 3 (6). С. 413–448.

28. Добровольский Д. О. Использование корпусов текстов в двуязычной лексикографии // Среди нехоженых путей: сборник научных статей к юбилею А. А. Кретова. Воронеж: НАУКА-ЮНИПРЕСС, 2012. С. 14–25.

29. Дудчук Ф. И., Подобряев А. В. Предикатно-аргументная структура // Введение в структурную лингвистику. Москва: Отделение теоретической и прикладной лингвистики МГУ, 2004. С. 13–26.

30. Дурново А. А., Зацман И. М., Лощилова Е. Ю. Кросслингвистическая база данных для аннотирования логико-семантических отношений в тексте // Системы и средства информатики. 2016. Т. 26. № 4. С. 124–137.

31. Дяченко П. В., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Подлесская О. Ю., Сизов В. Г., Фролова Т. И., Цинман Л. Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 272–300.

32. Зализняк А. А. Грамматический словарь русского языка: Словоизменение: Ок. 100000 слов. 4-е изд. Москва: Русские словари, 2003. 800 с.

33. Зализняк Анна А., Зацман И. М., Инькова О. Ю., Кружков М. Г. Надкорпусные базы данных как лингвистический ресурс // Труды международной конференции «Корпусная лингвистика – 2015». Санкт-Петербург, 2015. С. 211–218.

34. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. 3-е изд., перераб. Санкт-Петербург: Издательство Санкт-Петербургского государственного университета, 2020. 235 с.

35. Земичева С. С., Иванцова Е. В. Диалектный корпус как новый ресурс областной лексикографии // Вестник Томского государственного университета. 2019. № 446. С. 15–22.

36. Золотова Е. А. Коммуникативные аспекты русского синтаксиса. Москва: Наука, 1973.

37. Зубов А. В., Зубова И. И. Информационные технологии в лингвистике. Москва: Академия, 2004. 208 с.

38. Инькова О., Манзотти Э. Связность текста: мереологические логико-семантические отношения. Москва: Языки славянских культур, 2019. 376 с.

39. Иомдин Б. Л., Лопухина А. А., Носырев Г. В. К созданию частотного словаря значений слов // Компьютерная лингвистика и интеллектуаль-

ные технологии: по материалам ежегод. Междунар. конф. «Диалог», 2014. С. 199–212.

40. Иомдин Л. Л. В глубинах микросинтаксиса: один лексический класс синтаксических фразем // Компьютерная лингвистика и интеллектуальные технологии. 2008. С. 178–184.

41. Йоргенсен М. В., Филлипс Л. Дискурс-анализ. Теория и метод. 2-е изд. Харьков: Гуманитарный центр, 2008. 352 с.

42. Кибрик А. А. Дискурс // Введение в науку о языке. Москва: Буки Веди, 2019. С. 126–163.

43. Кибрик А. А. Анализ дискурса в когнитивной перспективе: автореферат дис. ... доктора филологических наук / Институт языкознания РАН. Москва, 2003. 90 с.

44. Кибрик А. А. Когнитивные исследования по дискурсу // Вопросы языкознания. 1994. № 5. С. 126–139.

45. Кибрик А. А., Плунгян В. А. Функционализм: дискурс как структура (У. Манн и С. Томпсон) // Современная американская лингвистика: фундаментальные направления. 2-е изд., испр. и доп. Москва: УРСС, 2002. С. 276–339.

46. Кибрик А. А., Подлесская В. И. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. Москва: ЛитРес, 2009. 736 с.

47. Кибрик А. А., Подлесская В. И. Проблема сегментации устного дискурса и когнитивная система говорящего // Когнитивные исследования. 2006. С. 138–158.

48. Кибрик А. А., Подлесская В. И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация (серия 2). 2003. Т. 6. С. 5–11.

49. Китайско-русский параллельный дискурсивный корпус официально-деловых текстов: интерактивная визуализация [Электронный ре-

курс] // URL: [https://www.crpardt.cn/index.html?lang=ru] (дата обращения: 21.10.2023).

50. Кобозева И. М., Сердобольская Н. В. Источники грамматикализации коннекторов русского языка (на материале базы Рускон) // Ученые записки Петрозаводского государственного университета. 2024. Т. 46. № 7. С. 66–74.

51. Ковальчук Н. В., Володина М. С. Теория риторических структур как прагматическая концепция анализа текста // Вестник Северного (Арктического) федерального университета. Серия: Гуманитарные и социальные науки. 2016. № 3. С. 107–113.

52. Копотев М. В. О некоторых следствиях корпусной лингвистики для общей теории языка // Филологический класс. 2021. Т. 26. № 2. С. 90–102.

53. Копотев М. В. Введение в корпусную лингвистику. Прага: Animedia Company, 2014. 194 с.

54. Копотев М. В., Мустайоки А. С. Современная корпусная русистика // Инструментарий русистики: корпусные подходы. Хельсинки: Slavica Helsingiensia, 2008. С. 7–24.

55. Кубрякова Е. С. О некоторых особенностях развития языка в свете его определения как сложнодинамической системы // Общее языкознание: формы существования, функции, история языка. Москва: Наука, 1970. С. 211–216.

56. Литвиненко А. О. Описание структуры дискурса в рамках Теории Риторической Структуры: применение на русском материале // Труды международного семинара «Диалог’2001» по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т. 1. С. 159–168.

57. Литвиненко А. О. Предикативное обстоятельство времени в устном детском нарративе // Компьютерная лингвистика и интеллектуальные технологии. Москва, 2002. Т. 1. С. 252–260.

58. Лосева Л. М. Как строится текст: пособие для учителей. Москва: Просвещение, 1980. 94 с.

59. Лук А. Н. Психология творчества. Москва: Наука, 1978. 128 с.

60. Макаров М. Л. Основы теории дискурса. Москва: Гнозис, 2003. 278 с.

61. Маник С. А. Параллельный корпус в практике перевода общественно-политических текстов (с английского на русский) // Современные исследования социальных проблем. 2019. Т. 11. № 4. С. 225–245.

62. Мельчук И. А., Жолковский А. К. Толково-комбинаторный словарь русского языка // Опыты семантико-синтаксического описания русской лексики. Москва: Языки славянской культуры. 2016.

63. Мельчук И. А. Русский язык в модели «Смысл – Текст». Москва: Языки русской культуры, 1995. 714 с.

64. Мельчук И. А. Опыт теории лингвистических моделей «Смысл – Текст». Семантика, синтаксис. Москва: Языки славянской культуры, 1974. 316 с.

65. Мельчук И. А. Автоматический синтаксический анализ: Общие принципы. Внутрисегментный синтаксический анализ. Москва: Наука, 1964. 370 с.

66. Мухин М. Ю., Ян И. Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2016. Т. 13. № 4. С. 23–31.

67. Олешков М. Ю. Основы функциональной лингвистики: дискурсивный аспект: учеб. пособие для студентов фак. рус. яз. и лит. Нижний Тагил: Нижнетагильская государственная социально-педагогическая академия, 2006. 146 с.

68. Падучева Е. В. О способах представления синтаксической структуры предложения // Вопросы языкознания. 1964. № 2. С. 99–113.

69. Падучева Е. В. О структуре абзаца // Ученые записки Тартусского университета. Вып. 181. Т. 2. Труды по знаковым системам. 1965. С. 284–292.

70. Пешковский А. М. Русский синтаксис в научном освещении. 8-е изд., доп. Москва: Языки славянской культуры, 2001. 510 с.

71. Плунгян В. А., Стойнова Н. М., Добрушина Е. Р. Материалы к Корпусной грамматике русского языка. Ч. I. Глагол. Санкт-Петербург: Нестор-История, 2016. 472 с.

72. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. Т. 2. № 16. С. 7–20.

73. Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. Москва: Индрик, 2005. С. 6–20.

74. Поспелов Н. С. Сложное синтаксическое целое и основные особенности его структуры // Доклады и сообщения Института русского языка АН СССР. Москва: Издательство АН СССР, 1948. С. 43–68.

75. Птухин А. А. Машинное обучение в обработке и анализе текстов // Язык в сфере профессиональной коммуникации. Екатеринбург: Ажур, 2019. С. 517–523.

76. Резникова Т. И., Копотев М. В. Лингвистически аннотированные корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. Москва: Индрик, 2005. С. 31–61.

77. Реферовская Е. А., Десницкая А. В. Лингвистические исследования структуры текста. Ленинград: Наука, 1983. 215 с.

78. Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. Москва: Просвещение, 1985. 399 с.

79. Русская грамматика: синтаксис / под ред. Н. Ю. Шведовой. Москва: Наука, 1980. 709 с.

80. Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.

81. Савчук С. О. Национальный корпус русского языка: перспективы использования в лингвистических исследованиях и в преподавании // Вестник Азиатско-Тихоокеанской ассоциации преподавателей русского языка и литературы. 2011. № 2–3. С. 62–67.

82. Семенов К. И., Дурнева С. П., Кузнецова Ю. Н. Русско-китайский параллельный корпус НКРЯ: проблемы и перспективы. Благовещенский государственный педагогический университет, 2020. С. 633–640.

83. Серио П. Анализ советского политического дискурса // Квадратура смысла. Французская школа анализа дискурса. Москва: Прогресс, 1999a. С. 337–383.

84. Серио П. Как читают тексты во Франции // Квадратура смысла. Французская школа анализа дискурса. Москва: Прогресс, 1999b. С. 12–53.

85. Син Фуи. Грамматика китайского языка. Санкт-Петербург: Издательство Санкт-Петербургского университета, 2020. 764 с.

86. Сичинава Д. В. Параллельные тексты в составе национального корпуса русского языка: новые направления развития и результаты // Труды института русского языка им. В. В. Виноградова. 2015. Т. 6. С. 194–235.

87. Сичинава Д. В., Шведова М. А. Параллельные корпуса в составе Национального корпуса русского языка: технологии и решаемые задачи // Компьютерная лингвистика: научное направление и учебная дисциплина: сб. научных статей. 2010. № 1. С. 30–34.

88. Сичинава Д. В. Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. Москва: Индрик, 2005. С. 21–30.

89. Солганик Г. Я. Синтаксическая стилистика. 5-е изд-е. Москва: ЛКИ, 2013. 232 с.

90. Степанов Ю. С. Альтернативный мир, дискурс, факт и принцип причинности // Язык и наука конца. 1995. Т. 20. С. 35–73.

91. Сусов А. А. Моделирование дискурса в терминах теории риторической структуры // Вестник Воронежского государственного университета. Серия: Филология. Журналистика. 2006. № 2. С. 133–138.

92. Сысоев П. В. Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. 2010. № 1 (9). С. 99–111.

93. Тао Ю. Создание и использование параллельного корпуса русского и китайского языков // Вестник Московского городского педагогического университета. Серия: Филология. Теория языка. Языковое образование. 2015. № 3 (19). С. 76–82.

94. Тао Ю., Захаров В. П. Разработка и использование параллельного корпуса русского и китайского языков // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2015. № 4. С. 18–29.

95. Теньер Л. Основы структурного синтаксиса. Москва: Прогресс, 1988. 653 с.

96. Тестелец Я. Г. Введение в общий синтаксис. Москва: Российский государственный гуманитарный университет, 2001. 796 с.

97. Толковый словарь русских глаголов / под ред. Л. Г. Бабенко и др. Москва: АСТ-Пресс, 1999. 702 с.

98. Фуко М. Порядок дискурса // Воля к истине: по ту сторону знания, власти и сексуальности. Работы разных лет. Москва: Касталь, 1996. 448 с.

99. Хомский Н. Синтаксические структуры // Новое в лингвистике. Вып. 2. Москва: Издательство иностранной литературы, 1962. С. 412–526.

100. Храпченко М. Б. Текст и его свойства // Контекст: литературно-теоретические исследования. 1986. С. 3–14.

101. Хурматуллин А. К. Понятие дискурса в современной лингвистике // Ученые записки Казанского университета. Серия Гуманитарные науки. 2009. Т. 151. № 6. С. 31–37.

102. Чернякова Т. А. Использование лингвистического корпуса в обучении иностранному языку // Язык и культура. 2011. № 4 (16). С. 127–132.

103. Чэнь С., Кукушкина О. В. О параллельных корпусах русских и китайских текстов // Вестник Московского университета. Серия 9. Филология. 2018. № 2. С. 170–197.

104. Шведова Н. Ю. Русская грамматика: фонетика, фонология, ударение, интонация, словообразование, морфология. Москва: Наука, 1980. 783 с.

105. Andor J. The master and his performance: An interview with Noam Chomsky // Intercultural Pragmatics. 2004. P. 93–111.

106. Asher N. Troubles on right frontier // Constraints in Discourse Pragmatics & Beyond New Series / Ed. by A. Benz, P. Kühnlein. John Benjamins Publishing Company, 2008. P. 29–52.

107. Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. Russnet: Building a lexical database for the russian language // Proceedings of the Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas, 2002. P. 60–64.

108. Bloomfield L. An introduction to the study of language // An Introduction to the Study of Language. 1983. P. 1–383.

109. Bloomfield L. Language. New York: H. Holt and Company, 1933. 584 p.

110. Böhmová A., Hajič J., Hajičová E., Hladká B. The Prague Dependency Treebank // Treebanks: Building and Using Parsed Corpora Text, Speech and Language Technology / Ed. by A. Abeillé. Dordrecht: Springer Netherlands, 2003. P. 103–127.

111. Bontempi G., Flauder M. From Dependency to Causality: A Machine Learning Approach // Cause Effect Pairs in Machine Learning The Springer Series on Challenges in Machine Learning / Ed. by I. Guyon, A. Statnikov, B. B. Batu. Cham: Springer International Publishing, 2019. P. 301–320.

112. Borge-Holthoefer J., Arenas A. Semantic networks: Structure and dynamics // Entropy. 2010. Vol. 12. № 5. P. 1264–1302.

113. Brown G., Yule G. Discourse analysis. Cambridge: Cambridge University Press, 1983. 288 p.

114. Cao S., Cunha I. da, Iruskieta M. The RST spanish-Chinese treebank // Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). New Mexico, USA: Association for Computational Linguistics, 2018. C. 156–166.

115. Carlson L., Marcu D., Okurowski M. E. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory // Current and New Directions in Discourse and Dialogue / Ed. by J. van Kuppevelt, R. W. Smith. Dordrecht: Springer Netherlands, 2003. P. 85–112.

116. Carlson L., Marcu D., Okurowski M. E. RST Discourse Treebank. Philadelphia: Linguistic Data Consortium, 2002.

117. Carlson L., Marcu D., Okurovsky M. E. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory // Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 2001.

118. Chafe W. Discourse: Overview // International encyclopedia of linguistics / Ed. by W. Bright. New York: Oxford University Press, 1992.

119. Chan K. Y., Vitevitch M. S. The influence of the phonological neighborhood clustering coefficient on spoken word recognition // Journal of Experimental Psychology: Human Perception and Performance. 2009. Vol. 35. № 6. P. 19–34.

120. Chistova E., Shelmanov A., Pisarevskaya D., Kobozeva M., Isakov V., Panchenko A., Toldova S., Smirnov I. RST Discourse Parser for Russian: An Ex-

perimental Study of Deep Learning Models // Analysis of Images, Social Networks and Texts Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021. P. 105–119.

121. Chomsky N. Syntactic structures. The Hague: Mouton Publ, 1957. 117 p.

122. Chu C. C. A Discourse Grammar of Mandarin Chinese. New York: P. Lang, 1998. 484 p.

123. Church K. Char_align: A program for aligning parallel texts at the character level // 31st Annual Meeting of the Association for Computational Linguistics. 1993. P. 1–8.

124. Cook G. W. D. A theory of discourse deviation: the application of schema theory to the analysis of literary discourse: submitted in accordance with the requirements for the degree of PhD. University of Leeds. School of English, 1990.

125. Cook G. W. D. Discourse. Oxford: Oxford University Press, 1989. 165 p.

126. Coquenet D., Chatelain C., Paquet T. SPAN: A Simple Predict & Align Network for Handwritten Paragraph Recognition // Document Analysis and Recognition – ICDAR 2021 Lecture Notes in Computer Science / Ed. by J. Lladós, D. Lopresti, S. Uchida. Cham: Springer International Publishing, 2021. P. 70–84.

127. Coulthard M. An introduction to discourse analysis. 2ed new edition. London: Longman, 1985. 216 p.

128. Daneš F. Functional sentence perspective and the organization of the text // Papers on functional sentence perspective. Prague: Academia, 1974. P. 106–128.

129. Danlos L. Comparing RST and SDRT discourse structures through dependency graphs // Proceedings of Constraints in Discourse. 2005. P. 53–77.

130. Danlos L. Discourse dependency structures as constrained DAGs // 5th SIGDIAL Workshop on Discourse and Dialogue. 2004. P. 127–133.

131. Danlos L. Strong generative capacity of RST, SDRT and discourse dependency DAGSs // Constraints in Discourse Pragmatics & Beyond New Series / Ed. by A. Benz, P. Kühnlein: John Benjamins Publishing Company, 2008. P. 69–95.

132. De Beaugrande R., Dressler W. U. Introduction to Text Linguistics. London: Longman, 1981. 286 p.

133. De Marneffe M.-C., Nivre J. Dependency grammar // The Annual Review Linquist. 2019. P. 197–218.

134. Debusmann R. Extensible dependency grammar: A modular grammar formalism based on multigraphs. Verlag nicht ermittelbar, 2006. 245 p.

135. Dependency in linguistic description / Ed. by A. Polguère, I. A. Mel′čuk. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2009. 281 p.

136. Dijk T. A. van. From Text Grammar to Critical Discourse Analysis // Semantic Scholar. 2004.

137. Dijk T. A. van. Principles of Critical Discourse Analysis // Discourse & Society. 1993. Vol. 4. № 2. P. 249–283.

138. Dijk T. A. van, Kintsch W. Strategies of discourse comprehension. New York: Academic Press, 1983. 418 p.

139. Dijk T. A. van. Text and context: Explorations in the semantics and pragmatics of discourse. London and New York: Longman, 1977. 261 p.

140. Dijk T. A. van. Some aspects of text grammars. The Hague: Mouton, 1972. 375 p.

141. Dong Z., Dong Q. HowNet-a hybrid language and knowledge resource // International conference on natural language processing and knowledge engineering, 2003. P. 820–824.

142. Duchier D., Gardent C. Tree Descriptions, Constraints and Incrementality // Computing Meaning: Vol. 2. Studies in Linguistics and Philosophy / Ed. by H. Bunt, R. Muskens, E. Thijsse. Dordrecht: Springer Netherlands, 2001. P. 205–227.

143. Egg M., Redeker G. How complex is discourse structure? // Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010. P. 1619–1623.

144. Egg M., Redeker G. Underspecified discourse representation. 2008. P. 117–138.

145. Eguíluz V. M., Chialvo D. R., Cecchi G. A., Baliki M., Apkarian A. V. Scale-Free Brain Functional Networks // Physical Review Letters. 2005. Vol. 94. № 1. P. 018102.

146. Eisner J. Three new probabilistic models for dependency parsing: An exploration // Proceedings of the 16th international conference on computational linguistics (COLING). Copenhagen, Denmark, 1996. P. 340–345.

147. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, Mass: MIT Press, 1998. 222 p.

148. Feng W., Ren H., Li X., Guo H. Building a parallel corpus with bilingual discourse alignment Lecture Notes in Computer Science / Ed. by Y. Wu, J.-F. Hong, Q. Su. Cham: Springer International Publishing, 2018. P. 374–382.

149. Ferrer-i-Cancho R. The structure of syntactic dependency networks: insights from recent advances in network theory // Problems of quantitative linguistics. 2005. P. 60–75.

150. Ferrer-i-Cancho R., Solé R. V., Köhler R. Patterns in syntactic dependency networks // Physical Review E. 2004. Vol. 69. № 5. P. 051915.

151. Firth J. R. Modes of meaning // Papers in Linguistics. London: Oxford University Press, 1957. P. 190–215.

152. Forbes-Riley K., Webber B., Joshi A. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG // Journal of Semantics. 2006. Vol. 23. № 1. P. 55–106.

153. Francis W. N., Kucera H. Brown corpus manual // Letters to the Editor. 1979. Vol. 5. № 2. P. 7.

154. Fraser B. What are discourse markers? // Journal of pragmatics. 1999. Vol. 31. № 7. P. 931–952.

155. Fraser N. M. Dependency parsing. University College London (United Kingdom), 1993. P. 296–319.

156. Gaifman H. Dependency systems and phrase-structure systems // Information and control. 1965. Vol. 8. № 3. P. 304–337.

157. Gee J. P. An introduction to discourse analysis: Theory & method. London and New York: Routledge, 1999. 184 p.

158. Gelbukh A., Sidorov G. Alignment of Paragraphs in Bilingual Texts Using Bilingual Dictionaries and Dynamic Programming // Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes in Computer Science / Ed. by J. F. Martínez-Trinidad, J. A. Carrasco Ochoa, J. Kittler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006a. P. 824–833.

159. Gelbukh A., Sidorov G., Vera-Félix J. Á. Paragraph-Level Alignment of an English-Spanish Parallel Corpus of Fiction Texts Using Bilingual Dictionaries // Text, Speech and Dialogue Lecture Notes in Computer Science / Ed. by P. Sojka, I. Kopeček, K. Pala. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006b. P. 61–67.

160. Givón T. Topic Continuity in Discourse: A quantitative cross-language study. John Benjamins Publishing Company, 1983. 492 p.

161. Goldberg Y. Neural network methods for natural language processing. Springer Nature, 2022. 287 p.

162. Grimes J. E. The Thread of Discourse. Walter de Gruyter, 1975. 408 p.

163. Grosz B. J., Sidner C. L. Attention, intentions, and the structure of discourse // Computational linguistics. 1986. Vol. 12. № 3. P. 175–204.

164. Grosz B. J., Sidner C. L. Plans for discourse // Intentions in Communication. 1990. P. 417–444.

165. Gupta A., Pala K. A generic and robust algorithm for paragraph alignment and its impact on sentence alignment in parallel corpora // Proc. of the Workshop on Indian Language Data. Resource and Evaluation. 2012. P. 18–27.

166. Haiman J., Thompson S. A. Clause Combining in Grammar and Discourse. John Benjamins Publishing Company, 1988. 428 p.

167. Hajič J., Smrz O., Zemánek P., Šnaidauf J., Beška E. Prague Arabic dependency treebank: Development in data and tools // Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools. 2004. P. 110–117.

168. Hajičová E. Markedness in Synchrony and Diachrony. Trends in Linguistics // Studies in Language / Ed. by O. M. Tomic. Amsterdam/Philadephia: John Benjamins Publishing Company, 1991. P. 262–264.

169. Halliday M. A. K., Hasan R. Cohesion in English. London: Longman Group, 1976. 374 p.

170. Halliday M. A. K., McIntosh A., Strevens P. The linguistic sciences and language teaching. Longmans, Green and Co., Ltd., 1964. 352 p.

171. Halliday M. A. K. The place of ’functional sentence perspective’ in the system of linguistic description // Papers on functional sentence perspective / Ed. by F. Danes. DE GRUYTER, 1974. P. 43–53.

172. Harris Z. Discourse analysis // Language. 1952. № 28. P. 1–30.

173. Hays D. G. Dependency Grammar // Encyclopedia of Computer Science and Technology / Ed. by J. Belzer, A. G. Holzman, A. Kent. New York: Marcel Dekkerp, 1977. Vol. 7. P. 213–227.

174. Hays D. G. Dependency theory: A formalism and some observations // Language. 1964. Vol. 40. № 4. P. 511–525.

175. Hills T. T., Maouene M., Maouene J., Sheya A., Smith L. Longitudinal Analysis of Early Semantic Networks: Preferential Attachment or Preferential Acquisition? // Psychological Science. 2009. Vol. 20. № 6. P. 729–739.

176. Hirschberg J., Litman D. Now Let's Talk About Now; Identifying Cue Phrases Intonationally // 25th Annual Meeting of the Association for Computational Linguistics. Stanford, California, USA: Association for Computational Linguistics, 1987. P. 163–171.

177. Hobbs J. R. Coherence and Coreference // Cognitive Science. 1979. Vol. 3. № 1. P. 67–90.

178. Hobbs J. R. On the coherence and structure of discourse. CSLI Stanford, CA, 1985.

179. Hoey M., Mahlberg M., Stubbs M., Teubert W. Text, discourse and corpora: theory and analysis. London; New York: Continuum, 2007. 253 p.

180. Hoey M. Textual interaction: An introduction to written discourse analysis. London: Routledge, 2001. 224 p.

181. Hoey M. Patterns of lexis in text. Oxford: Oxford University Press, 1991. 276 p.

182. Hoey M. On the surface of discourse. London: Unwin Hyman, 1983. 210 p.

183. Hovy E. H. Automated discourse generation using discourse structure relations // Artificial intelligence. 1993. № 63. P. 341–385.

184. Hovy E. H., Maier E. Parsimonious or profligate: how many and which discourse structure relations? University of Southern California, Information Sciences Institute, 1992.

185. Howard J. R. A Critial Book Review of On the Surface of Discourse // ResearchGate. 2024.

186. Huang C.-R., Hsieh S.-K. Infrastructure for cross-lingual knowledge representation-towards multilingualism in linguistic studies // Taiwan NSC-granted Research Project. 2010.

187. Hudson R. Language Networks: The New Word Grammar: Oxford University Press, U.S.A., 2007. 288 p.

188. Hudson R. English Word Grammar. Oxford: Basil Blackwood, 1990. 445 p.

189. Hudson R. A. Zwicky on heads // Journal of Linguistics. 1987. Vol. 23. № 1. P. 109–132.

190. Hudson R. A. Word grammar. Oxford: Basil Blackwell, 1984. 267 p.

191. Hudson R. Word grammar // The Routledge Handbook of Cognitive Linguistics: Routledge, 1982. P. 111–126.

192. James C. Contrastive analysis. Longman, Inc, 1980. 208 p.

193. Jin M., Kim M.-Y., Kim D., Lee J.-H. Segmentation of Chinese long sentences using commas // Proceedings of the Third SIGHAN Workshop on Chinese Language Processing. 2004. P. 1–8.

194. Johansson S., Leech G., Goodluck H. Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers: Univ., Department of English, 1978. 147 p.

195. Johnstone B. Discourse analysis. Berlin: Wiley-Blackwell, 2001. 269 p.

196. Joshi A. K. Natural Language Processing // Science. 1991. Vol. 253. № 5025. P. 1242–1249.

197. Kingsbury P., Palmer M. From TreeBank to PropBank: Citeseer, 2002. P. 1989–1993.

198. Kromann M. T., Mikkelsen L., Lynge S. K. Danish dependency treebank // Proc. TLT. Citeseer, 2003. P. 217–220.

199. Krzeszowski T. P. Tertium comparationis // Contrastive linguistics: prospects and problems / Ed. by J. Fisiak. Berlin: Mouton de Gruyter, 1984. P. 301–312.

200. Kuhlmann M., Nivre J. Mildly non-projective dependency structures // Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia: Association for Computational Linguistics, 2006. P. 507–514.

201. Lai T. B., Huang C. Functional constraints in dependency grammar // GLDV. 1999. Vol. 99. P. 235–244.

202. Lascarides A., Asher N. Temporal interpretation, discourse relations and commonsense entailment // Linguistics and Philosophy. 1993. Vol. 16. № 5. P. 437–493.

203. Le Q. H., Nguyen D. C., Pham D. H., Le A. C., Huynh V. N. Paragraph Alignment for English-Vietnamese Parallel E-Books // Knowledge and Systems Engineering Advances in Intelligent Systems and Computing / Ed. by V. N. Huynh, T. Denoeux, D. H. Tran, A. C. Le, S. B. Pham. Cham: Springer International Publishing, 2014. P. 251–259.

204. Lee A., Prasad R., Joshi A., Webber B. Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. 2008.

205. Leech G., Johansson S. The coming of ICAME // ICAME Journal. 2009. Vol. 33. P. 5–20.

206. Leech G. Corpus annotation schemes // Literary and linguistic computing. 1993. Vol. 8. № 4. P. 275–281.

207. Leech G. 100 million words of english: the british national corpus (BNC) // Second Language Research. 1992. Vol. 28. P. 1–13.

208. Li S., Wang L., Cao Z., Li W. Text-level Discourse Dependency Parsing // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014. P. 25–35.

209. Li Y., Feng J., Lai C., Feng H. Research on the construction of Chinese-English discourse cohesion alignment corpus // Proceedings of the 19th Chinese National Conference on Computational Linguistics / Ed. by Sun M.,

Li S., Zhang Y., Liu Y. Haikou, China: Chinese Information Processing Society of China, 2020. P. 795–806.

210. Li Y., Feng H., Feng W. Chinese discourse segmentation based on punctuation marks // International Journal of Signal Processing. 2015. Vol. 8. № 3. P. 177–186.

211. Li Y., Feng W., Sun J., Kong F., Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: 2014. P. 2105–2114.

212. Liu H. Statistical properties of Chinese semantic networks // Chinese Science Bulletin. 2009. Vol. 54. № 16. P. 2781–2785.

213. Liu H. The complexity of Chinese syntactic dependency networks // Physica A: Statistical Mechanics and its Applications. 2008. Vol. 387. № 12. P. 3048–3058.

214. Longacre R. E. The Grammar of Discourse. New York: Plenum Press, 1983. 423 p.

215. Lu J., Liu H. Do English noun phrases tend to minimize dependency distance? // Australian Journal of Linguistics. 2020. Vol. 40. № 2. P. 246–262.

216. Lyu C., Feng W. Analyzing Chinese text with clause relevance structure // Neurocomputing. 2023. Vol. 519. P. 82–93.

217. Mann W. C., Matthiessen C., Thompson S. A. Rhetorical structure theory and text analysis. University of Southern California, 1989.

218. Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization // Text. 1988. Vol. 8. № 3. P. 243–281.

219. Mann W. C., Thompson S. A. Rhetorical structure theory: A theory of text organization. University of Southern California: Information Sciences Institute, 1987.

220. Mann W. C. Discourse Structures for Text Generation // 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the

Association for Computational Linguistics. Stanford, California, USA: Association for Computational Linguistics, 1984. P. 367–375.

221. Marcu D. The theory and practice of discourse parsing and summarization. Cambridge, Mass: MIT Press, 2000. 248 p.

222. Marcu D. From discourse structures to text summaries // Intelligent Scalable Text Summarization: Association for Computational Linguistics, 1997a. P. 82–88.

223. Marcu D. The rhetorical parsing of natural language texts // Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics: Association for Computational Linguistics, 1997b. P. 96–103.

224. Marcu D. The rhetorical parsing, summarization, and generation of natural language texts. Department of Computer Science University of Toronto. 1997c. 331 p.

225. Marcu D. Building Up Rhetorical Structure Trees // Proceedings of the AAAI Conference on Artificial Intelligence. 1996. Vol. 13.

226. Matthews P. H. Syntax. Cambridge: Cambridge University Press, 1981. 306 p.

227. Mccarthy M., Carter R. Language as discourse: Perspectives for language teaching. London: Longman, 1994. 248 p.

228. McCulloch W. S., Pitts W. A logical calculus of the ideas immanent in nervous activity // Bulletin of Mathematical Biophysics. 1943. Vol. 5. № 4. P. 115–133.

229. McDonald R. et al. Universal Dependency Annotation for Multilingual Parsing // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers) / Ed. by H. Schuetze, P. Fung, M. Poesio. Sofia, Bulgaria: Association for Computational Linguistics, 2013. P. 92–97.

230. McEnery A., Xiao R. Parallel and comparable corpora: What are they up to? 2007.

231. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice: Cambridge University Press, 2012. 311 p.

232. Mel'čuk I. Dependency in Language // Dependency Linguistics: Recent advances in linguistic theory using dependency structures Linguistik Aktuell / Linguistics Today / Ed. by K. Gerdes, E. Hajičová, L. Wanner: John Benjamins Publishing Company, 2014. P. 1–32.

233. Mel'čuk I. Dependency Syntax: Theory and Practice. Albany, 1988. 448 p.

234. Miller G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information // Psychological Review. 1956. Vol. 63. № 2. P. 81–97.

235. Miltsakaki E., Joshi A., Prasad R., Webber B. Annotating discourse connectives and their arguments // Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation. Boston, 2004a. P. 9–16.

236. Miltsakaki E., Prasad R., Joshi A., Webber B. The Penn Discourse Treebank // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). 2004b. P. 2961–2968.

237. Miltsakaki E., Robaldo L., Lee A., Joshi A. Sense Annotation in the Penn Discourse Treebank Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. P. 275–286.

238. Moser M., Moore J. D. Toward a synthesis of two accounts of discourse structure // Computational linguistics. 1996. Vol. 22. № 3. P. 409–419.

239. Mukherjee A., Choudhury M., Basu A., Ganguly N. Self-organization of the sound inventories: Analysis and synthesis of the occurrence and co-occurrence networks of consonants // Journal of Quantitative Linguistics. 2008. Vol. 16. № 2. P. 157–184.

240. Müller S., Abeille A., Borsley R., Koenig J.-P. Head-Driven Phrase Structure Grammar: The handbook. 2021.

241. Nadkarni P. M., Ohno-Machado L., Chapman W. W. Natural language processing: an introduction // Journal of the American Medical Informatics Association. 2011. Vol. 18. № 5. P. 544–551.

242. Newman M. E. J. The structure and function of complex networks // SIAM Rev. 2003. Vol. 45. № 2. P. 167–256.

243. Nicholas N. Problems in the application of Rhetorical Structure Theory to text generation: University of Melbourne, 1994. 101 p.

244. Nivre J., Marneffe M.-C. de, Ginter F., Hajič J., Manning C. D., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference / Ed. by N. Calzolari et al. Marseille, France: European Language Resources Association, 2020. P. 4034–4043.

245. Nivre J. et al. Universal Dependencies v1: A Multilingual Treebank Collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) / Ed. by N. Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. P. 1659–1666.

246. Nivre J. Towards a universal grammar for natural language processing // Computational Linguistics and Intelligent Text Processing / Ed. by A. Gelbukh. Cham: Springer International Publishing, 2015. P. 3–16.

247. Nivre J. Dependency grammar and dependency parsing. Växjö: Växjö University, 2005a.

248. Nivre J., Nilsson J. Pseudo-projective dependency parsing. 2005b. P. 99–106.

249. Nivre J. An efficient algorithm for projective dependency parsing // Proceedings of the 8th international workshop on parsing technologies (IWPT). Nancy, France, 2003. P. 149–160.

250. Osborne T. A Dependency Grammar of English: John Benjamins Publishing, 2019. 436 p.

251. PDTB-Group. The penn discourse treebank 2.0 annotation manual. Technical report IRCS-08-01: Institute for Research in Cognitive Science, University of Pennsylvania, 2008a.

252. Peng S., Liu Y. J., Zeldes A. GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. 2022.

253. Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference “Dialogue 2017”. P. 194–204.

254. Poláková L., Mírovský J., Nedoluzhko A., Jínová P., Zikánová Š., Hajičová E. Introducing the prague discourse treebank 1.0 // Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013. P. 91–99.

255. Polanyi L. Discourse structure and discourse interpretation // General Session and Parasession on Pragmatics and Grammatical Structure. 1997. P. 492–503.

256. Polanyi L. A formal model of the structure of discourse // Journal of Pragmatics. 1988a. Vol. 12. № 5–6. P. 601–638.

257. Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. The Penn Discourse TreeBank 2.0 // Proceedings of the Sixth International Language Resources and Evaluation (LREC’08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008. P. 2961–2968.

258. Prasad R., Joshi A., Dinesh N., Lee A., Miltsakaki E., Webber B. The Penn Discourse TreeBank as a resource for natural language generation // Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation. Birmingham, U.K, 2005. P. 25–32.

259. Prasad R., Miltsakaki E., Joshi A., Webber B. Annotation and data mining of the Penn Discourse TreeBank // Proceedings of the 2004 ACL Workshop on Discourse Annotation – DiscAnnotation ’04. Barcelona, Spain: Association for Computational Linguistics, 2004. P. 88–97.

260. Quirk R. On English Usage // Journal of the Royal Society of Arts. 1966. Vol. 114. № 5122. P. 837–851.

261. Quirk R. S., Greenbaum G., Leech G., Svartvik J. A comprehensive grammar of the English language. London: Longman, 1985. 1779 p.

262. Renkema J., Schubert C. Introduction to discourse studies: new edition. Amsterdam: John Benjamins, 2018. 453 p.

263. Renkema J. Introduction to discourse studies. Amsterdam: John Benjamins, 2004. 363 p.

264. Richard-Zappella J. Lucien Tesnière aujourd’hui: actes du colloque international CNRS URA 1164 [S.U.D.L.A.]. Peeters Publishers, 1995. 436 p.

265. Robinson J. J. Dependency Structures and Transformational Rules // Language. 1970. Vol. 46. № 2. P. 259–285.

266. Samuel A., Strogatz S. H., Vitevitch M. S. Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish // Entropy. 2010. Vol. 12 (3). P. 327–337.

267. Scheffler T., Stede M. Mapping PDTB-style connective annotation to RST-style discourse annotation // Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). Bochum, Germany, 2016. P. 242–247.

268. Schiffrin D. Approaches to discourse. Oxford: Wiley-Blackwell, 1994. 480 p.

269. Schiffrin D., Tannen D., Hamilton H. E. The handbook of discourse analysis. Malden, Mass: Blackwell Publishers, 2001. 851 p.

270. Sgall P., Hajicová E., Panevová J. The meaning of the sentence in its semantic and pragmatic aspects. Springer Science & Business Media, 1986. 353 p.

271. Sharma N., Sharma R., Biswas K. K. Recognizing Textual Entailment using Dependency Analysis and Machine Learning // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop / Ed. by D. Inkpen, S. Muresan, S. Lahiri, K. Mazidi, A. Zhila. Denver, Colorado: Association for Computational Linguistics, 2015. P. 147–153.

272. Sinclair J., Coulthard M. Towards an analysis of discourse // Advances in spoken discourse analysis. Routledge, 2013. P. 7–40.

273. Stede M. The Potsdam Commentary Corpus // Proceedings of the Workshop on Discourse Annotation. Barcelona, Spain: Association for Computational Linguistics, 2004. P. 96–102.

274. Stede M., Neumann A. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. P. 925–929.

275. Steyvers M., Tenenbaum J. B. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth // Cognitive Science. 2005. Vol. 29. № 1. P. 41–78.

276. Stubbs M. Discourse analysis: The sociolinguistic analysis of natural language. Chicago: University of Chicago Press, 1983. 279 p.

277. Taboada M., Mann W. C. Rhetorical structure theory: looking back and moving ahead // Discourse Stud. 2006. Vol. 8. № 3. P. 423–459.

278. Taylor A., Marcus M., Santorini B. The Penn Treebank: An Overview // Treebanks Text, Speech and Language Technology / Ed. by A. Abeillé. Dordrecht: Springer Netherlands, 2003. P. 5–22.

279. Tesnière L. Éléments de syntaxe structurale. Paris: Klincksieck, 1959. 690 p.

280. Tiedemann J. Bitext alignment. Morgan & Claypool, 2011. 165 p.

281. Tognini-Bonelli E. Corpus linguistics at work. John Benjamins Publishing, 2001. 224 p.

282. Toldova S., Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. Rhetorical relations markers in Russian RST Treebank // Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms. 2017. P. 29–33.

283. Webber B. D-LTAG: extending lexicalized TAG to discourse // Cognitive Science. 2004. Vol. 28. № 5. P. 751–779.

284. Webber B. L., Joshi A. K. Anchoring a lexicalized tree-adjoining grammar for discourse. 1998. P. 86–92.

285. Webber B., Egg M., Kordoni V. Discourse structure and language technology // Natural Language Engineering. 2012. Vol. 18. № 4. P. 437–490.

286. Webber B., Knott A., Stone M., Joshi A. Discourse relations: A structural and presuppositional account using lexicalised TAG // Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Maryland: Association for Computational Linguistics, 1999. P. 41–48.

287. Webber B., Prasad R., Lee A., Joshi A. The penn discourse treebank 3.0 annotation manual. 2019.

288. Wiersma W. Research methods in education: An introduction. Boston: Allyn & Bacon, 1999. 476 p.

289. Wolf F., Gibson E. Coherence in Natural Language: Data Stuctures and Applications. Cambridge: MIT Press, 2006. 137 p.

290. Wolf F., Gibson E. Representing discourse coherence: A corpus-based study // Computational linguistics. 2005. Vol. 31. № 2. P. 249–287.

291. Xing F. Modern Chinese Grammar: a Clause-Pivot Approach. London & New York: Routledge, 2017. 639 p.

292. Xue N., Xia F., Chiou F.-D., Palmer M. The Penn Chinese Treebank: Phrase structure annotation of a large corpus // Natural Language Engineering. 2005. Vol. 11. № 2. P. 207–238.

293. Xue N., Yang Y. Chinese sentence segmentation as comma classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, 2011. P. 631–635.

294. Yang A., Li S. SciDTB: Discourse Dependency TreeBank for Scientific Abstracts // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Melbourne, Australia, 2018. P. 444–449.

295. Yoshida Y., Suzuki J., Hirao T., Nagata M. Dependency-based Discourse Parser for Single-Document Summarization // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) / Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha, Qatar: Association for Computational Linguistics, 2014. P. 1834–1839.

296. Zhang M., Qin B., Liu T. Chinese discourse relation semantic taxonomy and annotation // Journal of Chinese Information Processing. 2014. Vol. 28. № 2. P. 28–36.

297. Zhou Y., Xue N. PDTB-style discourse annotation of Chinese text // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea: Association for Computational Linguistics, 2012. P. 69–77.

298. Zhou Y., Xue N. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations // Language Resources and Evaluation. 2015. Vol. 49. № 2. P. 397–431.

299. Zipf G. K. Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA.: Harvard University Press, 1932. 51 p.

300. Zwicky A. M. Heads // Journal of Linguistics. 1985. Vol. 21. № 1. P. 1–29.

301. 柏晓静, 常宝宝, 詹卫东, 吴拥华. 构建大规模的汉英双语平行语料库 // 机器翻译研究进展——2002 年全国机器翻译研讨会论文集. 2002.

302. 陈洁. 俄汉超句统一体对比与翻译: 上海外语教育出版社, 2007. 309 p.

303. 陈洁. 论超句统一体的本质特征 // 外国语(上海外国语大学学报). 1997. № 04. P. 59–62.

304. 陈莉萍. 汉语篇章结构标注的理论支撑 // 南京航空航天大学学报(社会科学版). 2008. № 03. P. 68–71.

305. 陈莉萍. 英语语篇结构标注研究综述 // 外语与外语教学. 2007. № 7. P. 9–10.

306. 陈芯莹, 刘海涛. 汉语句法网络的中心节点研究 // 科学通报. 2011. Vol. 56. № 10. P. 726–731.

307. 崔卫, 李峰. 俄汉-汉俄平行语料库的构建设想与应用展望 // 中国俄语教学. 2014. Vol. 33. № 01. P. 1–5.

308. 崔卫, 张岚. 俄汉翻译平行语料库及其应用研究 // 解放军外国语学院学报. 2014. Vol. 37. № 01. P. 81–87.

309. 董振东, 董强, 郝长伶. 知网的理论发现 // 中文信息学报. 2007. № 4. P. 3–9.

310. 董振东, 董强. 知网和汉语研究 // 当代语言学. 2001. № 1. P. 33– 44, 77.

311. 董振东. 语义关系的表达和知识系统的建造 // 语言文字应用. 1998. № 3. P. 79–85.

312. 冯文贺, 陈伊琳, 任亚峰, 任函. 汉语篇章小句关联结构的表示与识别 // 北京大学学报(自然科学版). 2020. Vol. 56. № 01. P. 23–30.

313. 冯文贺, 李青青. 汉语复句的成分共享与英译断句 // 外语教学与研究. 2022. Vol. 54. № 05. P. 762–772, 801.

314. 冯文贺, 徐钰仪, 李青春. 汉语篇章依存结构的标注难点与处理 // 中文信息学报. 2020. Vol. 34. № 10. P. 19–26.

315. 冯文贺. 汉英篇章结构平行语料库的对齐标注研究 // 中文信息学报. 2013. Vol. 27. № 6. P. 158–165.

316. 冯文贺. 汉英篇章结构平行语料库构建与应用研究. 北京: 科学出版社, 2019.

317. 冯志伟. 机器翻译研究. 北京: 中国对外翻译出版公司, 2004. 841 p.

318. 冯志伟. 自然语言的计算复杂性研究 // 外语教学与研究. 2015. Vol. 47. № 5. P. 659–672, 799.

319. 黄伯荣, 廖序东. 现代汉语（增订三版）下册. 北京: 高等教育出版社, 2002. 337 p.

320. 黄伯荣、廖序东. 现代汉语（下册）（增订六版）: 高等教育出版社, 2017. 245 p.

321. 黄橙紫. 科技英语词汇的统计特征 // 同济大学学报(社会科学版). 2003. № 02. P. 97–101.

322. 蒋玉茹, 宋柔. 基于广义话题理论的话题句识别 // 中文信息学报. 2012. Vol. 26. № 5. P. 114–119, 128.

323. 靳光瑾, 肖航, 富丽, 章云帆. 现代汉语语料库建设及深加工 // 语言文字应用. 2005. № 02. P. 111–120.

324. 孔芳, 王红玲, 周国栋. 汉语篇章理解研究综述 // 软件学报. 2019. Vol. 30. № 07. P. 2052–2072.

325. 乐明, 冯志伟. 汉语财经评论的修辞结构标注及篇章研究 // 2006.

326. 乐明. 汉语篇章修辞结构的标注研究 // 中文信息学报. 2008. Vol. 22. № 4. P. 19–23, 42.

327. 李维刚, 刘挺, 王震, 李生. 双语语料库段落重组对齐方法研究 // 孙茂松, 陈群秀. 语言计算与基于内容的文本处理. 2003. P. 332–338.

328. 李文中, 濮建忠. 语料库索引在外语教学中的应用 // 解放军外国语学院学报. 2001. Vol. 24. № 2. P. 20–25.

329. 李艳翠, 冯继克, 来纯晓, 冯洪玉, 冯文贺. 汉英篇章衔接对齐语料库构建研究 // 中文信息学报. 2022. Vol. 36. № 04. P. 39–47, 56.

330. 李艳翠, 冯文贺, 周国栋, 朱坤华. 基于逗号的汉语子句识别研究 // 北京大学学报(自然科学版). 2013. Vol. 49. № 01. P. 7–14.

331. 李艳翠, 孙静, 周国栋. 汉语篇章连接词识别与分类 // 北京大学学报(自然科学版). 2015. Vol. 51. № 02. P. 307–314.

332. 李艳翠, 周国栋. 汉语篇章结构表示体系及资源构建研究 // 2015.

333. 李正华. 汉语依存句法分析关键技术研究 // 2014.

334. 连淑能. 英汉对比研究: 高等教育出版社, 1993. 363 p.

335. 梁茂成. 利用 WordPilot 在外语教学中自建小型语料库 // 外语电化教学. 2003. № 06. P. 42–45.

336. 廖秋忠. 篇章与语用和句法研究 // 语言教学与研究. 1991. № 04. P. 16–44.

337. 刘辰诞, 赵秀凤. 什么是篇章语言学. 上海: 上海外语教育出版社, 2011.

338. 刘复. 中国文法通论. 长沙: 岳麓书社, 1920. 91 p.

339. 刘海涛. 计量语言学导论. 北京: 商务印书馆, 2017. 219 p.

340. 刘海涛. 依存语法的理论与实践: 科学出版社, 1991. 319 p.

341. 刘淼, 邵青. 俄汉文学翻译语料库的创建——基于契诃夫小说平行语料库的设计与建构 // 外语学刊. 2016. № 01. P. 154–158.

342. 刘挺, 马金山. 汉语自动句法分析的理论与方法 // 当代语言学. 2009. Vol. 11. № 2. P. 100–112, 189.

343. 卢达威, 宋柔, 尚英. 从广义话题结构考察汉语篇章话题认知复杂度 // 中文信息学报. 2014. Vol. 28. № 5. P. 112–124.

344. 吕国英, 苏娜, 李茹, 王智强, 柴清华. 基于框架的汉语篇章结构生成和篇章关系识别 // 中文信息学报. 2015. Vol. 29. № 06. P. 98–109.

345. 吕叔湘. 汉语语法分析问题: 商务印书馆, 1979. 96 p.

346. 彭宣维. 语言与语言学概论: 汉语系统功能语法: 北京大学出版社, 2011. 353 p.

347. 秦洪武, 周霞. 大语言模型与语言对比研究 // 外语教学与研究. 2024. Vol. 56. № 2. P. 163–176, 318.

348. 申小龙. 普通语言学教程 // 复旦大学出版社. 上海, 2005. 336 p.

349. 宋柔. 宋柔语言工程论文集. 2012. 278 p.

350. 陶源. 人文社科学术文本俄汉平行语料库的创建与研究 // 语料库语言学. 2014. Vol. 1. № 01. P. 78–93, 112–113.

351. 王克非. 新型双语对应语料库的设计与构建 // 中国翻译. 2004. № 06. P. 75–77.

352. 王立非, 梁茂成. WordSmith 方法在外语教学研究中的应用 // 外语电化教学. 2007. № 03. P. 3–7, 12.

353. 王龙吟, 何安平. 基于语料库的外语教学与二语习得的链接 // 外语与外语教学. 2005. № 03. P. 28–32.

354. 王跃龙. 汉语交叉依存类非投射性现象 // 王跃龙：新加坡国立大学中文系: 博士学位论文, 2012. – 299 с.

355. 卫乃兴. 基于语料库的对比短语学研究 // 外国语（上海外国语大学学报）. 2011. Vol. 34. № 4. P. 32–42.

356. 吴为章, 田小琳. 汉语句群. 北京: 商务印书馆, 2000. 246 p.

357. 吴永芃, 李素建, 秦沐坤, 杨安, 王厚峰. 中英文篇章依存树库构建与分析 // 中文信息学报. 2018. Vol. 32. № 01. P. 75–82.

358. 吴云芳, 徐艺峰, 王恺然. 汉语篇章级小句关系的标注体系 // 中文信息学报. 2015. Vol. 29. № 03. P. 71–81.

359. 奚雪峰, 褚晓敏, 孙庆英, 周国栋. 汉语篇章微观话题结构建模与语料库构建 // 计算机研究与发展. 2017. Vol. 54. № 08. P. 1833–1852.

360. 肖维青; 自建语料库与翻译批评 // 外语研究. 2005. № 04. P. 60–65.

361. 谢家成. 小型英汉平行语料库的建立与运用 // 解放军外国语学院学报. 2004. № 03. P. 45–48.

362. 谢元花. 语料库与词汇研究 // 外语教学. 2002. № 03. P. 70–76.

363. 邢福义. 汉语复句研究: 商务印书馆, 2001. 695 p.

364. 邢福义. 汉语语法学（修订本）: 商务印书馆, 2016. 517 p.

365. 邢福义. 小句中枢说 // 中国语文. 1995. № 06. P. 420–428.

366. 徐赳赳. 话语分析在中国 // 外语教学与研究. 1997. № 04. P. 21– 25, 81.

367. 许余龙, 刘海涛, 刘正光. 关于语言研究的理论与方法 // 外语教学与研究. 2020. Vol. 52. № 01. P. 3–11.

368. 许余龙. 定量对比研究的方法问题 // 外国语（上海外国语大学学报）. 2001. № 4. P. 1–7.

369. 荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 大数据背景下 BCC 语料库的研制 // 语料库语言学. 2016. Vol. 3. № 1. P. 93–118.

370. 杨惠中, 黄人杰. JDEST 科技英语计算机语料库 // 外语教学与研究. 1982. № 4. P. 60–62.

371. 叶常青. 自建语料库在翻译教学中的应用——《红楼梦》中英文本用于翻译教学的课堂设计 // 外国语言文学. 2003. № 03. P. 41–44.

372. 詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. 北京大学 CCL 语料库的研制 // 语料库语言学. 2019. Vol. 6. № 01. P. 71–86, 116.

373. 张凤珍, 陈洁. 俄语超句统一体语言学分析 // 2005.

374. 张俐, 李晶皎, 胡明涵, 姚天顺. 中文 WordNet 的研究及实现 // 东北大学学报. 2003. № 4. P. 327–329.

375. 张牧宇, 宋原, 秦兵, 刘挺. 中文篇章级句间语义关系识别 // 中文信息学报. 2013. Vol. 27. № 06. P. 51–57.

376. 张培佳, 冯德正. 基于修辞结构理论的多模态语料库研究 // 当代修辞学. 2018. № 2/206. P. 71–81.

377. 张仕仁. 汉语复句的结构分析 // 中文信息学报. 1994. Vol. 8. № 4. P. 43–54.

378. 赵宏展; 对小型语料库的初步研究 // 辽宁行政学院学报. 2006. № 12. P. 214–215.

379. 朱德熙. 语法讲义: 商务印书馆, 1982. 231 p.

ПРИЛОЖЕНИЯ

Приложение 1. Перечень документов, размеченных в корпусе

1994. Совместная российско-китайская декларация (3 сентября 1994 г.)

1996. Совместная китайско-российская декларация (25 апреля 1996 г.)

2000-1. Совместное российско-китайское заявление по вопросам противоракетной обороны (17 июля 2000 г.)

2000-2. Пекинская декларация Российской Федерации и Китайской Народной Республики (18 июля 2000 г.)

2001. Московское совместное заявление глав государств России и Китая (16 июля 2001 г.)

2002. Совместная декларация Российской Федерации и Китайской Народной Республики (2 декабря 2002 г.)

2004. Совместная декларация Российской Федерации и Китайской Народной Республики (14 октября 2004 г.)

2005. Совместная декларация Российской Федерации и Китайской Народной Республики о международном порядке в XXI веке (1 июля 2005 г.)

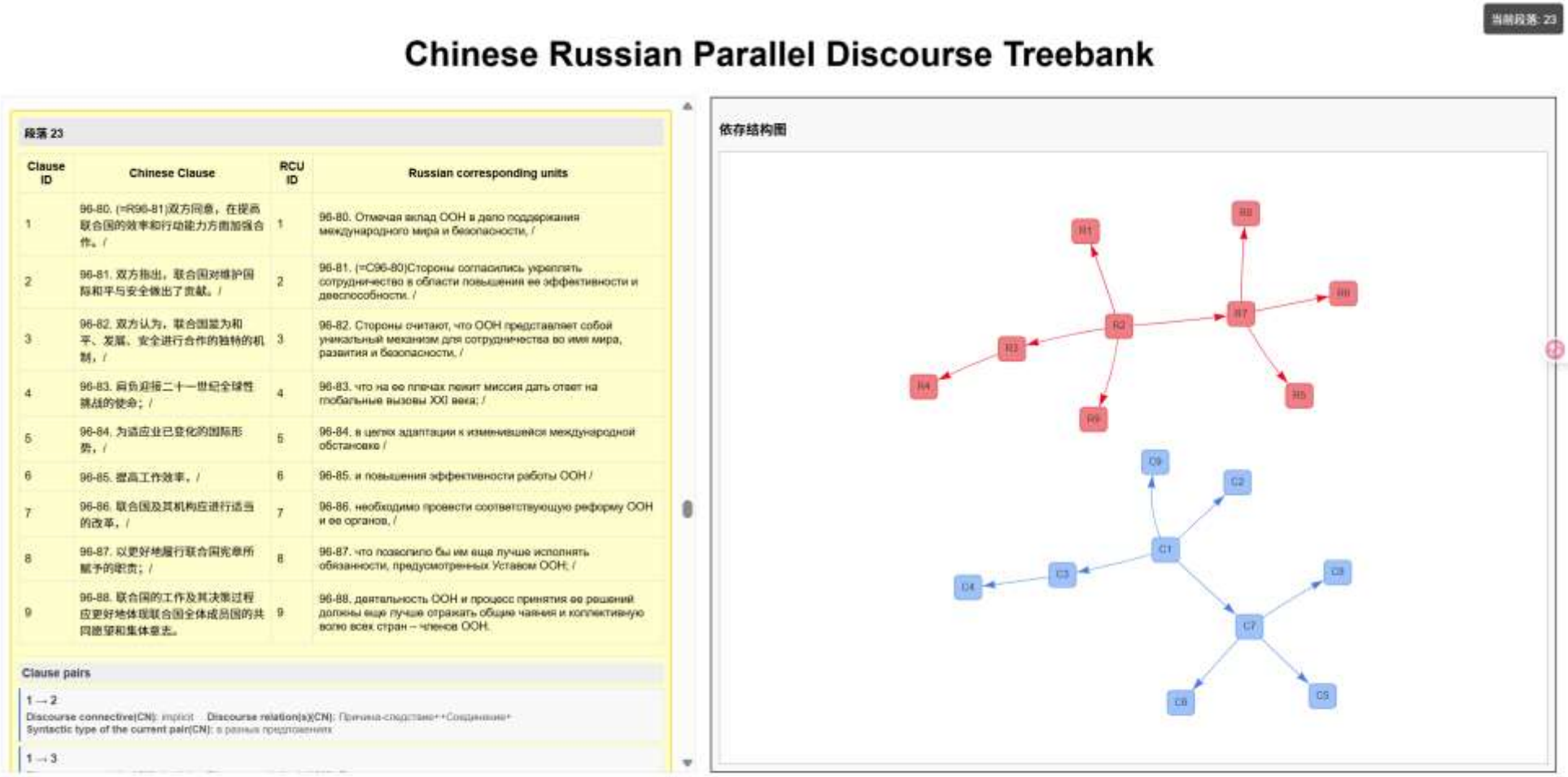
2006. Совместная декларация Китайской Народной Республики и Российской Федерации (21 марта 2006 г.)

2007. Совместная декларация Российской Федерации и Китайской Народной Республики (26 марта 2007 г.)

2008. Совместная декларация Российской Федерации и Китайской Народной Республики по основным международным вопросам (23 мая 2008 г.)

2010. Совместное заявление Российской Федерации и Китайской Народной Республики о всестороннем углублении российско-китайских отношений партнерства и стратегического взаимодействия (27 сентября 2010 г.)

Приложение 2. Визуализация Китайско-русского параллельного дискурсивного корпуса официально-деловых текстов.



当前段落: 1

Chinese Russian Parallel Discourse Treebank

段落 1

Clause ID	Chinese Clause	RCU ID	Russian corresponding units
1	07-1. 应俄罗斯联邦总统普京邀请，中华人民共和国主席胡锦涛于2007年3月26日至28日对俄罗斯联邦进行了国事访问。/	1	07-1. По приглашению Президента Российской Федерации В.В.Путина 26–28 марта 2007 года Председатель Китайской Народной Республики Ху Цзиньтао совершит официальный визит в Российскую Федерацию. /
2	07-2. 两国元首在莫斯科举行正式会谈，/	2	07-2. Главы двух государств провели переговоры в Москве, /
3	07-3. 并出席了"中国年"开幕式和中国国家展开幕式。	3	07-3. участвовали в церемонии официального открытия Года Китайской Народной Республики в Российской Федерации и в открытии Национальной выставки КНР.

Clause pairs

1 → 2
Discourse connective(CN): implicit Discourse relation(s)(CN): Пояснение++
Syntactic type of the current pair(CN): в разных предложениях

2 → 3
Discourse connective(CN): implicit Discourse relation(s)(CN): Дополнение++
Syntactic type of the current pair(CN): в одном сложном предложении

RCU pairs

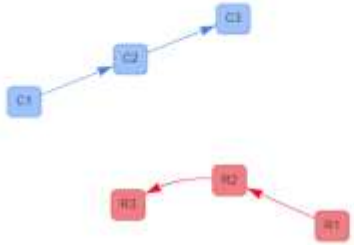
1 → 2
Discourse connective(Rus): implicit Discourse relation(s)(Rus) Пояснение++
Syntactic type of the current pair(Rus): в разных предложениях

2 → 3
Discourse connective(Rus): implicit Discourse relation(s)(Rus) Дополнение++
Syntactic type of the current pair(Rus): в одном сложном предложении

段落 2

Clause	Chinese Clause	RCU	Russian corresponding units

依存结构图



当前段落: 3

Chinese Russian Parallel Discourse Treebank

段落 3

Clause ID	Chinese Clause	RCU ID	Russian corresponding units
1	00-1-4. 两国元首全面回顾了近10年来中俄两国关系的发展历程。/	1	00-1-4. Главы двух государств, всесторонне рассмотрев развитие связей между Россией и Китаем на протяжении последнего десятилетия, /
2	00-1-5. 满意地指出，1996年宣布建立的中俄平等信任、面向21世纪的战略协作伙伴关系完全符合两国人民的根本利益。/	2	00-1-5. с удовлетворением отметили, что провозглашенное в 1996 году установление отношений равноправного доверительного партнерства, направленного на стратегическое взаимодействие в XXI веке, полностью отвечает коренным интересам народов двух стран. /
3	00-1-6. 强调发展平等信任的战略协作伙伴关系对加强中华人民共和国和俄罗斯联邦的全面合作、巩固中俄两国人民的友谊、推动世界多极化和建立公正合理的国际新秩序具有重要意义和作用。	3	00-1-6. Подчеркнуто, что развитие отношений равноправного доверительного партнерства и стратегического взаимодействия имеет важное значение для укрепления всестороннего сотрудничества между Российской Федерацией и Китайской Народной Республикой, упрочения дружбы народов России и Китая, способствует формированию многополярного мира и нового справедливого и рационального международного порядка.

Clause pairs

2 → 1
Discourse connective(CN): implicit **Discourse relation(s)(CN)**: Время++Соединение+
Syntactic type of the current pair(CN): в одном сложном предложении

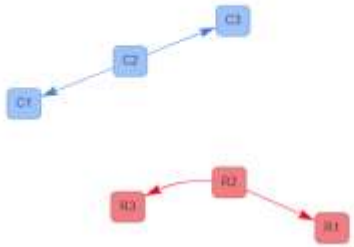
2 → 3
Discourse connective(CN): implicit **Discourse relation(s)(CN)**: Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

RCU pairs

2 → 1
Discourse connective(Rus): implicit **Discourse relation(s)(Rus)** Время++
Syntactic type of the current pair(Rus): РСА внутри простого предложения с деепричастным оборотом

2 → 3
Discourse connective(Rus): implicit **Discourse relation(s)(Rus)** Соединение++
Syntactic type of the current pair(Rus): в разных предложениях

依存结构图



当前段落: 7

Chinese Russian Parallel Discourse Treebank

段落 7

Clause ID	Chinese Clause	RCU ID	Russian corresponding units
1	00-1-17. 二中国和俄罗斯将继续保持两国高层领导人之间的密切接触和经常交往，/	1	00-1-17. II. Россия и Китай будут и впредь поддерживать постоянные и тесные контакты между высшими руководителями двух стран. /
2	00-1-18. 以多种方式就双边关系和国际形势的重大问题交换意见。/	2	00-1-18. используя различные каналы, вести обмен мнениями по важнейшим вопросам двусторонних отношений и международной обстановки. /
3	00-1-19. 两国外交、国防、执法、经济、科技等部门将保持经常和密切的交流。/	3	00-1-19. Внешнеполитические, оборонные, правоохранительные, экономические и научно-технические ведомства двух стран будут поддерживать регулярные и тесные контакты. /
4	00-1-20. 加强协调与合作。/	4	00-1-20. усиливать координацию и укреплять сотрудничество. /
5	00-1-21. 这有利于增进两国间的相互理解和信任。/	5	00-1-21. Это способствует дальнейшему углублению взаимопонимания и доверия между двумя странами /
6	00-1-22. 加强中俄的全面战略协作。	6	00-1-22. и укреплению всестороннего российско-китайского стратегического взаимодействия.

Clause pairs

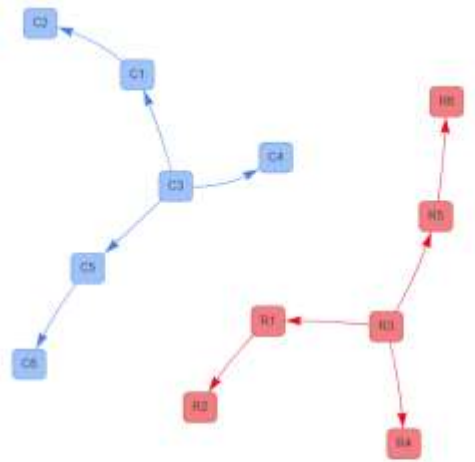
1 → 2
Discourse connective(CN): implicit Discourse relation(s)(CN): Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

3 → 1
Discourse connective(CN): implicit Discourse relation(s)(CN): Пояснение++
Syntactic type of the current pair(CN): в разных предложениях

3 → 4
Discourse connective(CN): implicit Discourse relation(s)(CN): Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

3 → 5
Discourse connective(CN): implicit Discourse relation(s)(CN): Оценка++
Syntactic type of the current pair(CN): в разных предложениях

依存结构图



当前段落: 12

Chinese Russian Parallel Discourse Treebank

段落 12

Clause ID	Chinese Clause	RCU ID	Russian corresponding units
1	00-1-33. 四此次点峰会期间签署《中华人民共和国主席和俄罗斯联邦总统关于反导问题的联合声明》的目的，是巩固全球和地区战略稳定。/	1	00-1-33. IV. Целью подписанного в ходе настоящей встречи на высшем уровне Совместного заявления Президента Российской Федерации и Председателя Китайской Народной Республики по вопросам противоракетной обороны является укрепление глобальной и региональной стратегической стабильности. /
2	00-1-34. 维护现有军控与裁军条约体系。/	2	00-1-34. сохранение существующей системы договоров в области контроля за вооружениями и разоружения. /
3	00-1-35. 推动防止大规模杀伤性武器及其运载工具扩散的进程。/	3	00-1-35. продвижение процесса нераспространения оружия массового уничтожения и средств его доставки. /
4	00-1-36. 并无一例外地确保所有国家的安全利益。/	4	00-1-36. гарантирование интересов безопасности всех без исключения государств. /
5	00-1-37. 对当前国际现实的分析表明，以某些国家的所谓导弹威胁作为要求修改"反导条约"的借口是根本站不住脚的。	5	00-1-37. Анализ нынешних международных реалий показывает полную несостоятельность использования так называемой ракетной угрозы со стороны некоторых государств в качестве предлога для обоснования требований об изменении Договора по ПРО.

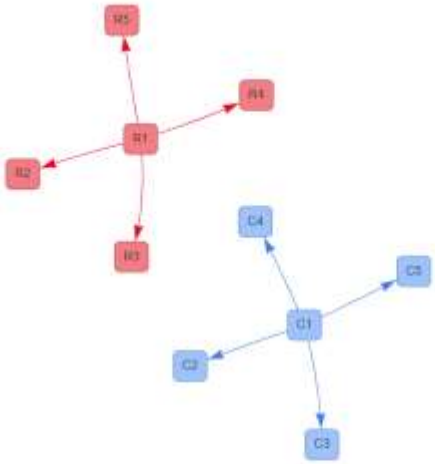
Clause pairs

1 → 2
Discourse connective(CN): implicit Discourse relation(s)(CN): Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

1 → 3
Discourse connective(CN): implicit Discourse relation(s)(CN): Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

1 → 4
Discourse connective(CN): 并 Discourse relation(s)(CN): Соединение++
Syntactic type of the current pair(CN): в одном сложном предложении

依存结构图



当前段落: 3

Chinese Russian Parallel Discourse Treebank

段落 3

Clause ID	Chinese Clause	RCU ID	Russian corresponding units
1	10-10. 两国元首回顾了中俄战略协作伙伴关系的发展状况。/	1	10-10. Главы государств рассмотрели вопросы развития российско-китайского партнерства и стратегического взаимодействия /.
2	10-11. 高度评价近年来两国各领域合作取得的重大进展。/	2	10-11. и дали высокую оценку значительному прогрессу, достигнутому за последние годы во всех сферах сотрудничества. /.
3	10-12. 双方满意地指出，近年来中俄政治互信不断增强。/	3	10-12. Стороны с удовлетворением отмечают, что в эти годы усиливалось политическое взаимодоверие, /.
4	10-13. 务实合作稳步扩大。/	4	10-13. стабильно расширялись двусторонние связи в практических областях, /.
5	10-14. 在国际和地区事务中保持密切沟通和协调。/	5	10-14. поддерживались тесные контакты и координация в международных и региональных делах, /.
6	10-15. 两国人民相互了解和友谊不断巩固。/	6	10-15. крепли взаимопонимание и дружба между народами. /.
7	10-16. 中俄关系具有战略性和长期性，成为当今国际关系中的重要稳定因素。	7	10-16. Благодаря своему стратегическому и долговременному характеру российско-китайские отношения стали важным стабилизирующим фактором современной международной политики.

Clause pairs

2 → 1
Discourse connective(CN): implicit Discourse relation(s)(CN): Время**Соединение*
Syntactic type of the current pair(CN): в одном сложном предложении

2 → 3
Discourse connective(CN): implicit Discourse relation(s)(CN): Пояснение**
Syntactic type of the current pair(CN): в разных предложениях

2 → 4

依存结构图