

Автоматизация разметки имплицитных логико-семантических отношений: возможности и ограничения*

А. А. Гончаров¹, П. В. Ярошенко²

Аннотация: В статье описаны основные проблемы, возникающие в ходе автоматизации разметки имплицитных логико-семантических отношений, проанализированы причины этих проблем, а также предложены подходы к их решению. Последовательно рассмотрены основные этапы этого процесса: 1) поиск примеров с имплицитными логико-семантическими отношениями; 2) выявление границ аргументов отношения; 3) выбор признаков размечаемых фрагментов. Проанализированы результаты применения метода поиска с исключением в параллельных текстах и отмечены ограничения этого метода. Рассмотрены два фактора, затрудняющих автоматизацию разметки аргументов логико-семантических отношений: длина аргументов может сильно варьироваться и аргумент не обязательно включает последовательность соседних токенов, а может прерываться. Детально проанализированы пути автоматизации выбора признаков размечаемых фрагментов. Продемонстрировано, что даже обработка формальных признаков может требовать привлечения экспертов. Что же касается семантических признаков, то среди них есть как признаки, обработка которых может быть частично автоматизирована, так и те, которые на данном этапе могут размечаться только вручную. Выводы, сделанные в статье, иллюстрируются примерами из корпуса.

Ключевые слова: лингвистическое аннотирование; дискурсивные отношения; логико-семантические отношения; имплицитность; параллельные тексты

1. Введение

В настоящее время значительная часть задач, связанных с обработкой естественного языка, с неизбежностью требует привлечения материала лингвистических корпусов. В зависимости от специфики задачи тексты корпуса должны снабжаться необходимой разметкой. Особую актуальность представляет собой создание дискурсивно аннотированных корпусов. Эти корпуса снабжены не только морфологической (формальные характеристики вхождений отдельных слов) и, возможно, синтаксической (формальные характеристики предложений) разметкой, но и разметкой логико-семантической структуры текста (семантика связей между частями текста, в том числе более протяженными, чем предложение).

* Исследование выполнено за счет гранта Российского научного фонда № 24-28-00527, <https://rscf.ru/project/24-28-00527/>.

¹ Федеральный исследовательский центр «Информатика и управление» Российской академии наук, a.gonch48@gmail.com

² Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Научно-исследовательский вычислительный центр МГУ им. М. В. Ломоносова, polina.iaroshenko@yandex.ru

Определение, с одной стороны, факта наличия логической связи между фрагментами текста, а с другой стороны, семантики этой связи может способствовать улучшению результатов во многих приложениях компьютерной лингвистики: так, при составлении реферата на основе исходного документа (summarization) фактор связности порожденного текста становится одним из ключевых [1]. Кроме того, в настоящее время приобрело актуальность направление, в рамках которого рассматриваются возможности оценивать текстовую связность с помощью больших языковых моделей (например, [2]).

Важную роль играет наличие достаточного количества размеченных данных. Значительная часть имеющихся ресурсов англоязычная (речь прежде всего о Пенсильванском дискурсивно аннотированном корпусе [3, 4]), в то время как для других языков, в частности для русского, размеченных данных не хватает [5].

Логико-семантические отношения (ЛСО) – это отношения, которые устанавливаются между отрезками текста³ и описывают, как именно эти отрезки связаны друг с другом по смыслу. В англоязычной литературе для их обозначения по большей части используется термин *discourse relations* – дискурсивные отношения. В данной статье вслед за [6] используется термин *логико-семантические отношения*. Распознавание ЛСО (*discourse relation recognition*) – одна из актуальных задач в области обработки естественного языка⁴.

С точки зрения маркированности ЛСО могут быть эксплицитными – выраженными при помощи языкового средства (коннектора или иного показателя, явно указывающего на семантику отношения, см. [10]), и

³ Как правило, имеющими пропозициональное содержание.

⁴ При этом отдельной проблемой оказывается классификация ЛСО [6–7] и составление списка возможных ЛСО, где каждое ЛСО было бы снабжено структурированным определением [8–9].

имплицитными – не выраженными ни одним из таких средств⁵. Особенно остро проблема автоматизации обработки стоит для имплицитных ЛСО. Тогда как при обработке эксплицитных ЛСО можно ориентироваться на наличие их показателей (например, *потому что* – коннектор, выражающий отношение причины), имплицитные ЛСО требуют более сложного подхода [11–13].

Цель статьи – описать основные проблемы, возникающие в ходе автоматизации разметки имплицитных ЛСО, проанализировать причины этих проблем, а также предложить подходы к их решению. В ходе исследования использовался материал параллельных текстов (для пар языков «русский–английский» и «русский–французский»), причем рассматривались только такие примеры, где ЛСО имплицитно в русскоязычном фрагменте⁶. Поиск и аннотирование примеров проводились при помощи действующей информационной системы, разрабатываемой для анализа ЛСО. Приводимые далее рисунки представляют собой скриншоты пользовательского интерфейса данной системы.

2. Автоматизация поиска имплицитных ЛСО

Сложности работы с имплицитными ЛСО начинаются уже на этапе поиска. Как было показано в [14], моноязычные тексты не позволяют целенаправленно отбирать примеры с имплицитными ЛСО. Эта возможность появляется только при работе с параллельными текстами, где можно применить метод поиска с исключением [15, 16]. Результатом применения метода становится поисковая выдача, где пары фрагментов параллельных текстов таковы, что фрагмент текста на одном из языков содержит показатель искомого ЛСО, тогда как фрагмент текста на другом языке, напротив, не содержит заданных показателей этого же ЛСО.

⁵ Так, в примере «*я не был на совещании, потому что заболел*» ЛСО причины является эксплицитным (выражено при помощи *потому что*), а в примере «*я не был на совещании, заболел*» – имплицитным (не выражено).

⁶ Несмотря на то что все примеры, использованные в статье, взяты из параллельного корпуса, в ряде случаев они даются в сокращенном виде – либо только оригинал, либо только перевод, – если этого достаточно для иллюстрации описываемых положений.

Хотя метод значительно ускоряет формирование выборки примеров с имплицитными ЛСО, он имеет как минимум три ограничения. Во-первых, сами по себе параллельные корпуса намного менее репрезентативны и объемны, чем моноязычные. Во-вторых, метод по определению позволяет обнаруживать только те примеры имплицитных ЛСО, где в другом языке это ЛСО имеет показатель (случаи, где ЛСО имплицитно в оригинале и в переводе, в поисковую выдачу не попадают). В-третьих, даже среди оставшихся примеров обнаруживается немало нерелевантных.

В качестве иллюстрации рассмотрим задачу поиска примеров имплицитного ЛСО причины в русском языке (использовались параллельные тексты для языковой пары «русский–английский»). Для решения этой задачи был разработан, в частности, следующий запрос:

- 1) англоязычный фрагмент должен содержать единицу *because*;
- 2) русскоязычный фрагмент не должен содержать ни одну из следующих единиц: *ибо, потому что / потому как, поскольку, так как, ведь, оттого, из-за, за, от, ввиду, вследствие, затем что, тем более что, в связи с.*

Затем в ходе экспертного анализа пары фрагментов делились на две группы: релевантные и нерелевантные. В данном случае к релевантным относились те пары фрагментов, где в русском языке действительно имеется имплицитное ЛСО причины; все остальные считались нерелевантными. Затем проводился лингвистический анализ релевантных пар, а результат этого анализа записывался в надкорпусную базу данных [17] в форме аннотированных переводных соответствий (АПС) [18, 19]. Оказалось, что для первых 300 пар, попавших в поисковую выдачу для направлений перевода «русский–английский» и «английский–русский», было сформировано 134 и 141 АПС соответственно, то есть доля релевантных примеров составила около 47%.

Для разных языковых пар и разных запросов доля таких примеров будет различаться, но едва ли можно представить себе ситуацию, где релевантными

будут 100% примеров. Во-первых, в запросе могут не быть учтены какие-либо способы эксплицитного выражения ЛСО (например, *по тому одному что* в отличие от *потому что* для ЛСО причины). Во-вторых, при переводе семантика ЛСО может быть изменена (например, одно ЛСО заменяется другим). В-третьих, при переводе могут добавляться или опускаться фрагменты текста. Любое из таких изменений приводит к тому, что вместо имплицитного ЛСО с заданной семантикой фрагмент текста либо содержит другое ЛСО, либо не содержит никакого.

Таким образом, хотя метод поиска с исключением в параллельных текстах позволяет значительно облегчить процесс формирования набора данных, даже этот метод не позволяет полностью автоматизировать поиск целевых примеров так, чтобы не требовалась ручная постобработка.

3. Автоматизация разметки аргументов

Разметка примеров имплицитных ЛСО предполагает, среди прочего, выявление границ отрезков текста, непосредственно связанных размечаемым ЛСО. Такие отрезки будем называть аргументами ЛСО. Их объем должен быть необходимым и достаточным для анализа отношения: сокращать аргументы так, чтобы семантика ЛСО осталась неясной, или включать в состав аргументов то, что не требуется для анализа ЛСО, не следует. На рисунках ниже аргументы отношения выделены цветом.

Автоматизация разметки аргументов ЛСО на данном этапе исследования затруднительна, что обусловлено двумя факторами. Во-первых, длина аргументов может сильно варьироваться; во-вторых, аргумент не обязательно включает в себя исключительно соседние токены, а может прерываться.

Основную проблему представляет вариативность аргументов с точки зрения их длины. Аргументы отношения могут включать одно слово, словосочетание, предложение и даже несколько предложений. Такой разброс объясняется природой ЛСО, поскольку отношение одного и того же типа теоретически может связывать отдельные слова (рис. 1), части одного

предложения, самостоятельные предложения или даже целые абзацы (рис. 2) в зависимости от описываемой ситуации и детальности ее описания.

Новиков промолчал, хотя он любил поговорить с Макаровым о жутких, чем-то влекущих первых днях войны.	Novikov hadn't said anything, though he enjoyed talking to Makarov about that terrible but fascinating time.
NoMarker	but

Рис. 1 Пример однословного аргумента ЛСО противопоставления⁷

Стараясь держаться прямо, он сделал несколько движений в сторону и повалился на бок. Так падал и вскакивал он до вечера и вернулся домой с катка, к огорчению матери, весь в снегу, с подкашивающимися от усталости ногами. На следующее утро он опять был на катке.	Trying to keep his body upright, he took several steps sideways and fell on his side; and so he fell and got up over and over again until sunset, and, to his mother's vexation, returned home all covered with snow, his legs trembling with weariness. But next morning he was back on the ice again.
NoMarker	but

Рис. 2 Пример аргументов ЛСО противопоставления, представляющих собой одно или несколько предложений⁸

Пример на рис. 2 демонстрирует еще одну сложность автоматизации разметки параллельного корпуса – деление текста на сегменты может различаться в оригинале и переводе. Это может быть связано как со спецификой синтаксиса языка, так и с частными решениями переводчика. Так, в примере на рис. 2 в русскоязычном оригинале аргумент 1 включает два предложения, в то время как в переводе на английский – одно сложное предложение.

Разрыв аргументов зачастую связан с наличием в размечаемых фрагментах прямой речи. В примере на рис. 3 в русскоязычном блоке текста разорваны оба аргумента, а в переводе – только аргумент 2.

⁷ Источник примера: [В. С. Гроссман. Жизнь и судьба (ч. 1) (1960) | Life and Fate (pt. 1) (пер. R. Chandler; 1985)].

⁸ Источник примера: [Б. Н. Полевой. Повесть о настоящем человеке (1946) | A Story about a Real Man (пер. J. Fineberg; 1950)].

-- Ну, ну! ну ставь еще четыре тысячи гульденов на красную! Вот бумажник, бери. -- Она вынула из кармана и подала мне бумажник. -- Ну, бери скорей, тут двадцать тысяч рублей чистыми деньгами. -- Бабушка, -- прошептал я, -- такие куши... Жива не хочу быть -- отыграюсь. Ставь!	"Very well. Stake another four thousand gulden upon the red. Take this banknote to do it with." "I have still got twenty thousand roubles in actual cash." "But," I whispered, "such a quantity of money —" "Never mind. I cannot rest until I have won back my losses. Stake!"
NoMarker	but

Рис. 3 Пример разрывных аргументов ЛСО противопоставления⁹

Таким образом, вариативная структура аргументов, их возможная разрывность, а также специфика размечаемого материала – параллельные тексты, где разбиение оригинала и перевода на части может различаться – не позволяют на данном этапе автоматизировать разметку границ аргументов.

4. Автоматизация разметки признаков

Признаки, размечаемые в ходе аннотирования примеров с имплицитными ЛСО, весьма разнообразны; они подробно описаны в работах [19, 20]. В контексте же автоматизации разметки важно разделить их на две большие группы: формальные и семантические. Формальные признаки связаны прежде всего с планом выражения (пунктуация, морфология, синтаксис); семантические – с планом содержания (использование слов той или иной семантической категории, наличие заданных семантических структур).

Наиболее простой с точки зрения автоматизации оказывается разметка ряда формальных признаков, таких как наличие некоторых знаков препинания между аргументами, наличие в составе аргумента двух и более предложений. Однако даже среди формальных признаков есть те, которые требуют экспертной постобработки, не говоря уже о семантических. Далее рассмотрим примеры таких признаков.

4.1 Формальные признаки

Признак «Лексический повтор» указывает на то, что первый и второй аргументы ЛСО включают формы одних и тех же слов. На первый взгляд этот признак может быть полностью автоматизирован при наличии

⁹ Источник примера: [Ф. М. Достоевский. Игрок (1866) | The Gambler (пер. С. J. Hogarth; 1935)].

морфологической разметки, но это не так. Прежде всего необходимо исключить повтор служебных частей речи – т. н. «стоп-слов»: в рамках данной работы перечни таких слов составлялись для каждого языка (русский, английский и французский) путем расширения списков стоп-слов из библиотеки Natural Language Toolkit [21]. Кроме того, из-за неснятой омонимии к повторам могут быть отнесены те единицы, для которых имеется более одного морфологического разбора. Например, в одном из разборов слова *видимо* в качестве леммы указан глагол *видеть*. Следовательно, если в одном аргументе используется слово *видимо* в значении ‘может быть’, а в другом – какая-либо из форм глагола *видеть*, при автоматической обработке такой пример будет снабжен признаком «Лексический повтор», что некорректно.

Признак «Императив» присваивается аргументу, если в нем присутствует глагол в форме повелительного наклонения, и на первый взгляд при наличии морфологической разметки данный признак тоже достаточно легко поддается автоматизации. Однако при анализе результатов автоматической простановки было выявлено, что и в этом случае требуется ручная постобработка. Так, признак «Императив» не нужен, если словоформа в императиве присутствует, но утрачивает свое изначальное значение. К подобным «ложным» императивам можно отнести, например, устойчивое сочетание *(не) дай бог* или глаголы в повелительном наклонении, употребляемые в дискурсивном значении: *извольте, разрешите*.

4.2 Семантические признаки

При разметке примеров, содержащих имплицитные ЛСО, особенно значимыми представляются признаки, связанные с семантикой. Эти признаки можно условно разделить на «скорее конкретные» и «скорее абстрактные». В первом случае для простановки признака, как правило, требуется найти лексические единицы или их сочетания, относящиеся к заданному семантическому классу; во втором же смысловое значение может быть

настолько абстрактным и выражаться лексически настолько по-разному, что составление списков искомых единиц представляется затруднительным.

Рассмотрим некоторые из семантических признаков, которые относятся к «скорее конкретным».

Признак «Ментальный предикат» указывает на то, что предикат главной клаузы аргумента можно отнести к категории ментальных (то есть касающихся знаний, рассуждений, мышления; например, глаголы *знать*, *мечтать*, *предполагать* и др.). Чтобы найти в исследуемом материале лексические единицы, входящие в целевую группу, и присвоить соответствующий признак, были использованы данные Национального корпуса русского языка: для семантически размеченной части корпуса доступна функция поиска по лексико-семантическим признакам [22]. Таким образом, можно выгрузить¹⁰ список глаголов, которые имеют семантическую помету «ментальная сфера» (по состоянию на 30.05.2025 он включает 245 единиц).

Далее проводился поиск аргументов, которые включают леммы из списка. Анализ полученных результатов позволил выявить несколько факторов, затрудняющих автоматизацию разметки. Во-первых, из-за неснятой омонимии в выдачу попадали непредикативные единицы, которые в качестве леммы при этом имели глагол (*кажущийся* по лемме *казаться*, *определенный* по лемме *определить*). Во-вторых, признак актуален только в том случае, когда к ментальным относится именно главный предикат аргумента. Ср. примеры, приведенные на рис. 4 и 5, содержащие формы глагола *знать*, который включен в список ментальных глаголов из Национального корпуса русского языка. В первом случае *знала* выступает в качестве сказуемого главной клаузы аргумента (т. е. признак нужен); во втором же эту функцию выполняет *надо* (т. е. признак не нужен).

¹⁰ <https://ruscorpora.ru/page/instruction-semantic/>

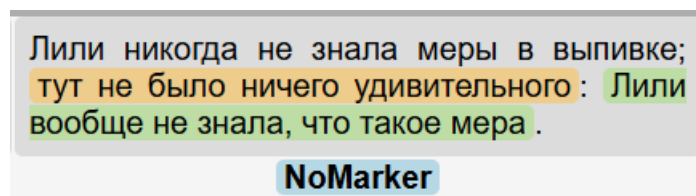


Рис. 4 Пример, где нужен признак «Ментальный предикат»¹¹

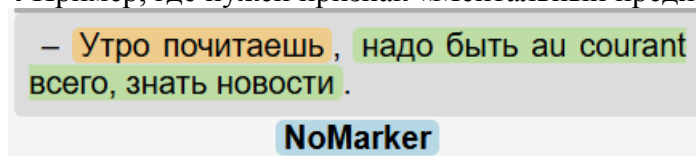


Рис. 5 Пример, где не нужен признак «Ментальный предикат»¹²

Наряду с зашумленностью выдачи при автоматическом поиске, существует также обратная проблема: ментальный предикат может быть выражен словом, которое не входит в изначальный список для поиска. В примере на рис. 6 в роли главного ментального предиката клаузы выступает глагол *чувствовать*. Сам по себе этот глагол скорее относится к семантическому классу эмоциональной лексики, но здесь он употреблен в качестве ментального предиката, сближаясь по значению с глаголом *понимать*.

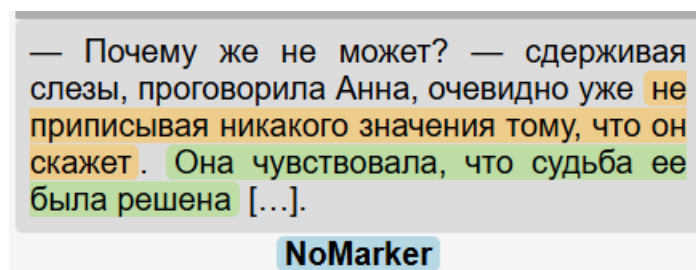


Рис. 6 Пример, где имеется ментальный предикат, не входящий в список для поиска¹³

Таким образом, даже если признак таков, что его потенциальные показатели теоретически можно задать в виде списка, полностью автоматизировать разметку не представляется возможным. Тем не менее использование списков целевой лексики позволяет значительно облегчить разметку и сделать ее более последовательной, а также проводить дополнительную проверку уже размеченных данных.

¹¹ Источник примера: [Лорен Вайсбергер. Дьявол носит Прада (пер. М. Маяков и др.; 2006)].

¹² Источник примера: [И. А. Гончаров. Обломов (1848–1859)].

¹³ Источник примера: [Л. Н. Толстой. Анна Каренина (1873–1877)].

Рассмотрим некоторые из признаков, которые относятся к «скорее абстрактным».

Признак «Смена состояния» подразумевает, что один из аргументов включает указание на состояние, а другой – указание на то, что это состояние впоследствии изменилось. В примере на рис. 7 лексическими маркерами смены состояния служат слова *проспал – проснулся*. Однако реализация этого признака может быть крайне разнообразной. Переход от сна к бодрствованию – лишь один из многих вариантов смены состояния, и заранее составить ограниченный набор целевых лексических единиц не представляется возможным.

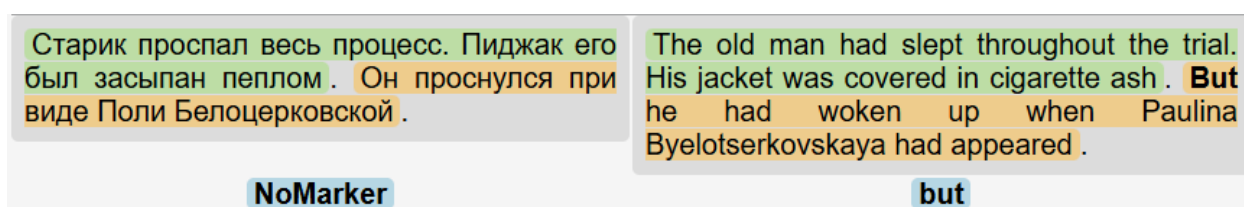


Рис. 7 Пример, где необходим признак «Смена состояния»¹⁴

Еще один признак – «Действие – результат» – присваивается блоку текста в случае, если один из аргументов – указание на действие, а другой – указание на результат действия¹⁵. В примере на рис. 8 описана ситуация, где действующее лицо ищет некий предмет (действие), но не находит его (результат). Лексическими маркерами признака в русском языке служат слова *обошел – не нашел*. Данный пример хорошо иллюстрирует проблему, обозначенную выше: ввиду разнообразия лексических маркеров автоматизировать разметку затруднительно. Так, более ожидаемой парой глаголов для описания ситуации безрезультатного поиска было бы *искать – не найти*, а не *обойти – не найти*.

¹⁴ Источник примера: [И. Э. Бабель. Карл-Янкель (1931) | Karl-Yankel (пер. P. Constantine; 2002)].

¹⁵ В случае с ЛСО противопоставления результат, как правило, не соответствует ожидаемому.

Just before dinner two telegrams were handed in. The first ran: -- "Have just heard that Barrymore is at the Hall. -- BASKERVILLE." The second: -- "Visited twenty-three hotels as directed , but sorry, to report unable to trace cut sheet of Times . -- CARTWRIGHT."	Перед самым обедом нам подали две телеграммы. Первая гласила: "Только что сообщил Бэрримор дома Баскервиль". Вторая: "Обошел двадцать три гостиницы сожалению изрезанной страницы "Таймса" не нашел . Картрайт".
but	NoMarker

Рис. 8 Пример, где необходим признак «Действие – результат»¹⁶

Таким образом, разметка семантических признаков, которые относятся к «скорее абстрактным», сложнее всего поддается автоматизации.

5. Заключение

В статье были рассмотрены основные этапы разметки имплицитных логико-семантических отношений: поиск целевых фрагментов в корпусном материале, выделение аргументов логико-семантического отношения, а также присвоение признаков целевым фрагментам текста.

Показано, что разнообразие семантических критериев, которым должны удовлетворять размеченные данные, приводит к необходимости выходить далеко за пределы собственно анализа имплицитных ЛСО: это и вопросы анализа синтаксической структуры текста, и проблемы снятия омонимии, и задачи составления списков лексики, относящейся к заданному семантическому классу. Полная автоматизация на данном этапе не представляется возможной, сохраняется необходимость привлекать эксперта по крайней мере для постобработки, однако представленные подходы к автоматизации разметки способны значительно ускорить создание наборов данных с имплицитными ЛСО и сделать разметку более согласованной.

Информационная система для анализа ЛСО разрабатывается авторами с учетом всех вышеописанных сложностей: она позволяет не только создавать АПС в ручном режиме, но и обрабатывать ранее сформированные АПС, при необходимости корректируя разметку в автоматизированном режиме. Таким

¹⁶ Источник примера: [Arthur Conan Doyle. The Hound of the Baskervilles (1902) | Собака Баскервилей (пер. Н. Волжина; 1948)].

образом, система представляет собой эффективный инструмент для работы разметчика и эксперта.

Литература

1. *Parmar M., Deilamsalehy H., Dernoncourt F., Yoon S., Rossi R. A., Bui T.* Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – Miami, FL, USA: Association for Computational Linguistics, 2024. P. 19810–19820.
2. *Barbosa B., Campelo C.* LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis // Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. – Porto Alegre: Sociedade Brasileira de Computação, 2024. P. 278–287. <https://sol.sbc.org.br/index.php/stil/article/view/31140>.
3. *Prasad R., Webber B., Joshi A.* The Penn Discourse Treebank: An Annotated Corpus of Discourse Relations // Handbook of Linguistic Annotation. – Dordrecht: Springer Science + Business Media, 2017. P. 1197–1217.
4. *Webber B., Prasad R., Lee A., Joshi A.* The Penn Discourse Treebank 3.0: Annotation Manual, 2019. <https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf>.
5. *Jiang D., He J.* Tree Framework With BERT Word Embedding for the Recognition of Chinese Implicit Discourse Relations // IEEE Access, 2020. Vol. 8. P. 162004–162011.
6. *Инькова О. Ю.* Логико-семантические отношения: проблемы классификации // Связность текста: мереологические логико-семантические отношения. – М.: ЯСК, 2019. С. 11–98.
7. *Гончаров А. А.* Классификации внутритекстовых отношений: основания и принципы структурирования // Вопросы языкознания, 2021. № 3. С. 97–119.
8. *Инькова О. Ю., Кружков М. Г.* Структурированные определения дискурсивных отношений в Надкорпусной базе данных коннекторов // Информатика и её применения, 2021. Т. 15. Вып. 4. С. 27–32.
9. *Инькова О. Ю., Кружков М. Г.* Критерии определения семантической близости дискурсивных отношений // Информатика и её применения, 2023. Т. 17. Вып. 3. С. 100–106.
10. *Гончаров А. А., Инькова О. Ю.* Извлечение знаний о средствах выражения логико-семантических отношений при помощи Надкорпусной базы данных // Информатика и её применения, 2021. Т. 15. Вып. 2. С. 96–103.
11. *Xiang W., Liu S., Wang B.* Parsing and encoding interactive phrase structure for implicit discourse relation recognition // Neural Computing and Applications, 2024. Vol. 36. P. 13783–13797.
12. *Cai M., Yang Zh., Jian P.* Improving Implicit Discourse Relation Recognition with Semantics Confrontation // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Torino, Italia: ELRA and ICCL, 2024. P. 8828–8839.
13. *Xiang W., Wang B.* A survey of implicit discourse relation recognition // ACM Computing Surveys, 2023. Vol. 55. Iss. 12. Art. 258. P. 1–34.
14. *Гончаров А. А.* Методы поиска имплицитных логико-семантических отношений в монологических текстах // Системы и средства информатики, 2022. Т. 32. № 3. С. 92–102.

15. Гончаров А. А. Методы поиска имплицитных логико-семантических отношений в параллельных текстах // Системы и средства информатики, 2022. Т. 32. № 4. С. 32–44.
16. Гончаров А. А. Поиск с исключением в параллельных текстах // Системы и средства информатики, 2023. Т. 33. № 4. С. 102–114.
17. Кружков М. Г. Концепция построения надкорпусных баз данных // Системы и средства информатики, 2021. Т. 31. № 3. С. 101–112.
18. Гончаров А. А. Аннотирование параллельных корпусов: подходы и направления развития // Информатика и её применения, 2023. Т. 17. Вып. 4. С. 81–87.
19. Гончаров А. А., Ярошенко П. В. Принципы аннотирования имплицитных логико-семантических отношений в параллельных текстах // Информатика и её применения, 2024. Т. 18. Вып. 3. С. 106–114.
20. Goncharov A. A., Iaroshenko P. V. How to Describe Implicit Discourse Relations: The Experience of Creating a Dataset in Russian // Вестник Московского университета. Сер. 9: Филология, 2025 (в печати).
21. Bird St., Loper E., Klein E. Natural Language Processing with Python. – O'Reilly Media Inc., 2009. 502 p.
22. 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.