РАЗРАБОТКА КОРПУСА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ С РАЗМЕТКОЙ НА ОСНОВЕ ТЕОРИИ РИТОРИЧЕСКИХ СТРУКТУР

Ананьева М.И.[1] *(*ananyeva@isa.ru*)*
Кобозева М.В.[1,2]*(*mvharitonova@yandex.ru*)*

[1]Институт системного анализа ФИЦ ИУ РАН, Москва, Россия
[2]Кафедра компьютерной лингвистики Института лингвистики РГГУ, Москва, Россия

This paper presents an adaptation of the Rhetorical Structure Theory to the Russian language and the development of an RST-corpus that will be used for training of an automatic discourse parser in the future. Authors' survey shows that discourse analysis improves performance of systems for machine translation, automatic summarization, author identification etc. At the time of writing, ten texts from the SynTagRus-treebank had been annotated. A list of discourse relations (proposed by W. Mann and S. Thompson) has been modified and a list of Russian discourse markers has been made. Besides, authors present some preliminary discourse-structure statistics on the basis of this annotation.

Key words: Rhetorical Structure Theory, discourse analysis, Russian corpus, discourse relations.

Введение

Основные разделы лингвистики выделяются в соответствии с языковым уровнем, которому они посвящены: фонетика, морфология, синтаксис. Эти уровни образуют иерархическую структуру, в которой объем единиц увеличивается с повышением уровня – от звуков, фонем и морфем к словосочетаниям и предложениям. За рамки предложения выходит анализ дискурсивной (или риторической) структуры текста.

Существует несколько различных подходов к дискурсивной разметке текстов. В основном все они оперируют понятиями дискурсивных единиц (сегментов текста) и тех или иных семантических отношений между ними. К подобным подходам относятся, например, следующие: Теория риторических структур (Rhetorical Structure Theory) [1], Теория общей структуры документов (Cross-document Structure Theory) [2], модель Penn Discourse Treebank [3], Теория сегментной репрезентации дискурса (Segmented Discourse Representation Theory) [4] и другие.

Мы остановимся подробнее на наиболее популярном из перечисленных подходов – **Теории риторических структур** (ТРС) В. Манна и С. Томпсон [1]. В ней предлагается описание организации текста как древовидной иерархической структуры, элементарные единицы которой объединены логическими отношениями, образуя более крупные единицы, которые в свою очередь связаны между собой теми же самыми отношениями, и так далее. Авторы ТРС вводят 23 дискурсивных отношения: *одноядерные* (бинарные отношения, в которых один элемент играет роль ядра, а другой - сателлита) и *мультиядерные* (отношения между двумя и более равнозначными элементами), однако в литературе встречаются различные модификации ТРС, в которых количество отношений может доходить до 80. Более подробно об элементарных дискурсивных единицах (ЭДЕ) и риторических отношениях речь пойдет в следующем разделе.

Начиная с 1990-х годов дискурсивный анализ широко применяется при решении многих актуальных задач в области компьютерной лингвистики: машинного перевода, автоматического реферирования текстов, анализа тональности, жанровой классификации, вопросно-ответного поиска и прочих. Так, системы машинного перевода показывают качество на 2,5% - 9,5% выше базовых систем, в которых не учитывается дискурсивная структура [5, 6, 7]. Качество автоматического реферирования в разных исследованиях

повысилось на 5,3%, а в исследованиях по реферированию коллекций документов даже на 44% [8, 9, 10]. В задаче анализа тональности (в частности, определения полярности сообщения) дискурсивный анализ улучшает качество на 4% [11, 12]; а в задаче определения связности текста – на 4,2% на уровне предложений и на 21,3% на уровне текста [13, 14]. Более подробный обзор литературы, представленный в работе [15], показывает, что в большинстве случаев учет дискурсивной структуры текста повышает качество решения задач обработки естественного языка. В то время как для английского языка разработки в данной области выходят на достаточно высокий уровень, для русского языка подобных исследований крайне мало, поскольку даже отечественные ученые предпочитают работать с английским. Так, А.А. Кибрик и его коллеги предложили модель референциального выбора с учетом дискурсивной структуры текста, которая показала аккуратность предсказания на 14,8% выше, чем у аналогичных систем. За основу был взят уже размеченный англоязычный корпус RST-DT [16].

На сегодняшний день едва ли не единственным размеченным дискурсивным корпусом на русском языке является корпус устных текстов, созданный А.А. Кибриком и его коллегами (корпус «Рассказы о сновидениях») [17]. Для его разметки список отношений был пополнен 20 новыми, описывающими “коммуникативные связи и оформительные компоненты структуры дискурса” [18]. Данный корпус создавался для исследований особенностей устного дискурса у здоровых детей и детей с неврозами, и не может быть использован для разработки автоматических систем анализа текстов.

Принципы разметки дискурсивных структур в текстах на русском языке

В ходе нашего исследования мы разработали принципы дискурсивной разметки текстов на русском языке в рамках Теории риторических структур и оформили их в виде руководства по разметке, с которым можно ознакомиться на сайте лаборатории "Компьютерной лингвистики и интеллектуального анализа информации" (ФИЦ ИУ РАН)[1]. В общем варианте разметка риторической структуры текста делится на два этапа: а) выделение элементарных дискурсивных единиц (в базовом варианте – клауз, см. работы [1], [19]) и б) установление отношений между ними. Поэтому, прежде всего, мы должны выявить критерии выделения ЭДЕ. Следует отметить, что для этого учитываются как формальные грамматические, так и семантические характеристики сегментов. Итак, в качестве отдельных ЭДЕ мы выделяем:

- Финитные клаузы (кроме клауз, являющихся сентенциальными актантами и при этом не входящих в косвенную речь):
  a) *[При изменении объемов экологических ниш этот фактор срабатывает автоматически,] [и ниши постоянно оказываются предельно заполненными,] [так что значительная часть их обитателей балансирует на грани физического выживания.]*
  b) *[Поэтому уже не удивительно, когда данные таможни не совпадают с реальными объемами перевозки в 5 и более раз.]*
- Деепричастные обороты с причинно-следственным и уточняющим значением: [*Из этого многие уже сделали соответствующие выводы,] [конвертировав часть сбережений в самую молодую валюту Старого Света.]*
- Описательные (нерестриктивные) причастные обороты: *[Здесь, в особом сарае, царь Иван и поместил ученого слона,] [обученного становиться перед самодержцем на колени.]*
- Предложные группы со значением причины, следствия, уступки и контраста: *[Несмотря на свойственную возрасту впечатлительность,] [Прокофьев не злоупотребляет инициалами.]*

[1] http://nlp.isa.ru/projects/discourse/manual.v1.pdf

Следующей задачей нашего исследования является определение списка отношений. Опираясь на базовую работу Теории риторических структур В. Манна и С. Томпсон [1] и на инструкцию к разметке Л. Карлсона и Д.Марку [20], а также отталкиваясь от собственного опыта пробной разметки, мы выделили предварительный список отношений, который в ходе дальнейшего исследования может быть модифицирован (см. *Таблица 1*).

Таблица 1. Список риторических отношений

Одноядерные отношения:	**Мультиядерные отношения:**
1. Background (Фон)	1. Contrast (Контраст)
2. Evidence (Обоснование)	2. Restatement (Переформулировка)
3. Cause (Причина)	3. Sequence (Последовательность)
4. Effect (Следствие)	4. Joint (Конъюнкция)
5. Condition (Условие)	5. Comparison (Сравнение)
6. Purpose (Цель)	6. Same-unit (Прерывающаяся единица)
7. Concession (Уступка)	
8. Preparation (Подготовка)	
9. Conclusion (Вывод)	
10. Elaboration (Детализация)	
11. Antithesis (Антитезис)	
12. Solutionhood (Решение)	
13. Motivation (Мотивация)	
14. Evaluation (Оценка)	
15. Attribution (Источник)	

Мы изменили изначальный набор риторических отношений, предложенный В. Манном и С. Томпсон [1]. Так, было добавлено отношение Conclusion для оформления заключительных сегментов всего текста или некоторых его частей (это отношение используется, например, в [20]), а также зеркальное для него отношение – Preparation (см., [21] [22]). Так как они пересекаются с отношением Summary (Резюме), последнее мы исключили. Отношения Otherwise ("В противном случае"), Interpretation (Интерпретация), Enablement (Возможность) – мы убрали, так как в списке содержатся родственные или синонимичные к ним отношения: Antithesis и Contrast для Otherwise, Evaluation для Interpretation, и Condition для Enablement. Волитивные и неволитивные типы некоторых отношений (Cause и Result) были объединены, поскольку на данном этапе исследования мы решили сосредоточиться на семантике текста, пока не затрагивая такую широкую область как прагматический аспект.

Кроме того, следуя примеру Л. Карлсона и Д.Марку [20], мы добавили связь Same-unit, которая по сути не является риторическим отношением, однако соединяет составные части одной ЭДЕ, разделенные другим сегментом текста. Для каждого риторического отношения в нашем руководстве по разметке приводится количество ядер в нем, описание его значения, возможные дискурсивные маркеры, примеры; для одноядерных отношений указывается, какая семантическая составляющая содержится в ядре, а какая в сателлите (в примерах ядро выделяется курсивом). Например:

> **Cause** (Причина). Одноядерное отношение. Ситуация ядра является причиной ситуации сателлита. Сателлит представляет собой результат действия. Причина важнее, чем следствие; цель автора – выделить причину. Если результат в ядре, то следует выбрать отношение **Effect** (Следствие**).** Может иметь дискурсивные маркеры: *потому что*, *причина* (в различных формах), *быть/стать/являться причиной*. Пример:

[Алгоритм А2 такой возможности не предоставляет,] *[потому что до тех пор, пока все нити не завершат очередной цикл резольвирований, управляющий процесс не получит «новые» дизъюнкты.]*

Стоит отметить, что при разметке мы опираемся на абзацное членение текста, т.е. связи сначала устанавливаются внутри и между предложениями, затем внутри абзаца и только потом между абзацами.

Опыт и результаты разметки дискурсивных структур

В качестве материала для первого варианта корпуса были выбраны 30 текстов из корпуса СинТагРус [23], которые принадлежат к публицистическому и научно-популярному жанру. В рамках других проектов, ранее проводимых в нашей лаборатории, на данные 30 текстов была нанесена семантическая разметка [24], и анафорическая [25]. Таким образом, коллектив лаборатории ставит целью создание корпуса текстов на русском языке с наиболее полной разметкой, что позволит в дальнейшем исследовать взаимодействие разных языковых уровней.

Для ручной разметки текстов дискурсивными структурами существуют готовые инструменты, такие как RSTTool [26] и rstWeb [27]. Они позволяют выделять в текстах элементарные дискурсивные единицы (ЭДЕ), объединять ЭДЕ по близости симметричными (мультиядерными) и асимметричными (одноядерными) отношениями, строить древовидную структуру. Для настоящего исследования была выбрана вторая система (rstWeb), разработанная А.Зельдесом из Джорджтаунского университета [27]. Такой выбор был сделан по многим причинам: у данного разметчика открытый исходный код, что позволяет редактировать список отношений и функционал, он понятен и удобен в использовании, а также работает на основе браузера, что позволяет использовать его на любом компьютере без установки дополнительного ПО. На рис.1 показан фрагмент дискурсивного дерева в данном разметчике (фрагмент приводится с сохранением некоторых элементов интерфейса).

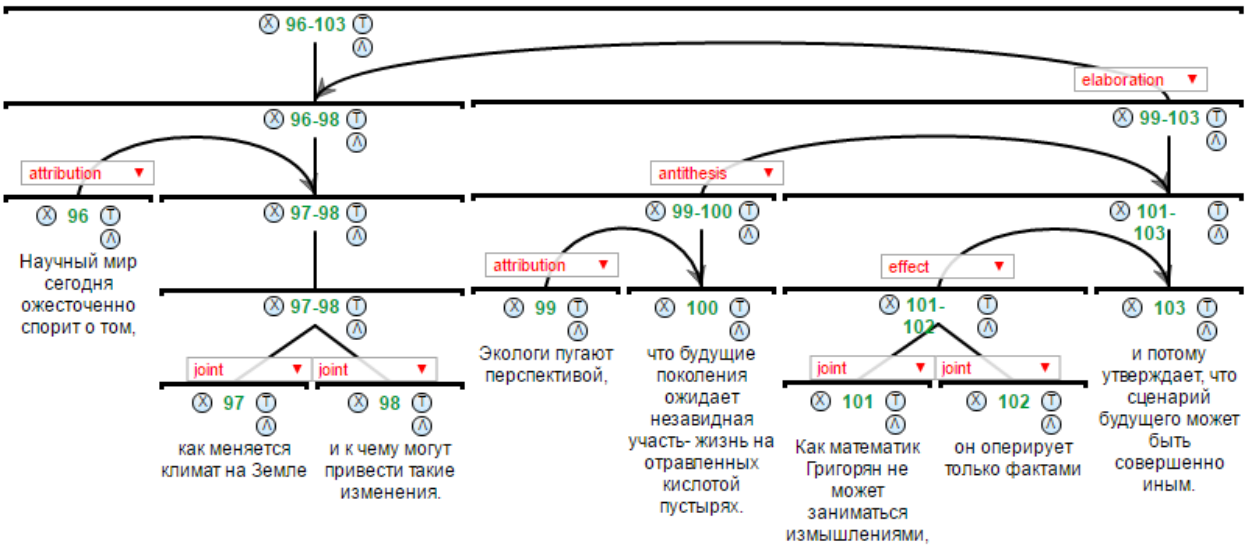

Рис. 1. Фрагмент дискурсивной структуры текста

На рисунке показана разметка трех предложений, каждое из которых разбито на несколько ЭДЕ, являющиеся листьями дерева. ЭДЕ объединяются отношениями внутри предложений, а предложения – друг с другом. Стрелками маркируются одноядерные отношения (стрелка направлена от сателлита к ядру), а единицы симметричных отношений объединяются линиями в одной точке (см. сегменты 97 и 98, отношение Конъюнкция).

На текущий момент мы разметили 10 текстов, что составляет более 1200 ЭДЕ и 1484 риторических отношения. На данном материале был проведен предварительный анализ частоты встречаемости отношений в публицистических текстах на русском языке (Рис. 2). Очевидно, что в текстах других жанров, а, возможно, и на большем корпусе публицистических текстов, данное распределение может измениться. Однако обзор аналогичных исследований показал, что и в других корпусах на разных европейских языках самыми частотными являются отношения со значением перечисления и пояснения [28, 29] – в нашем корпусе это Конъюнкция и Детализация.

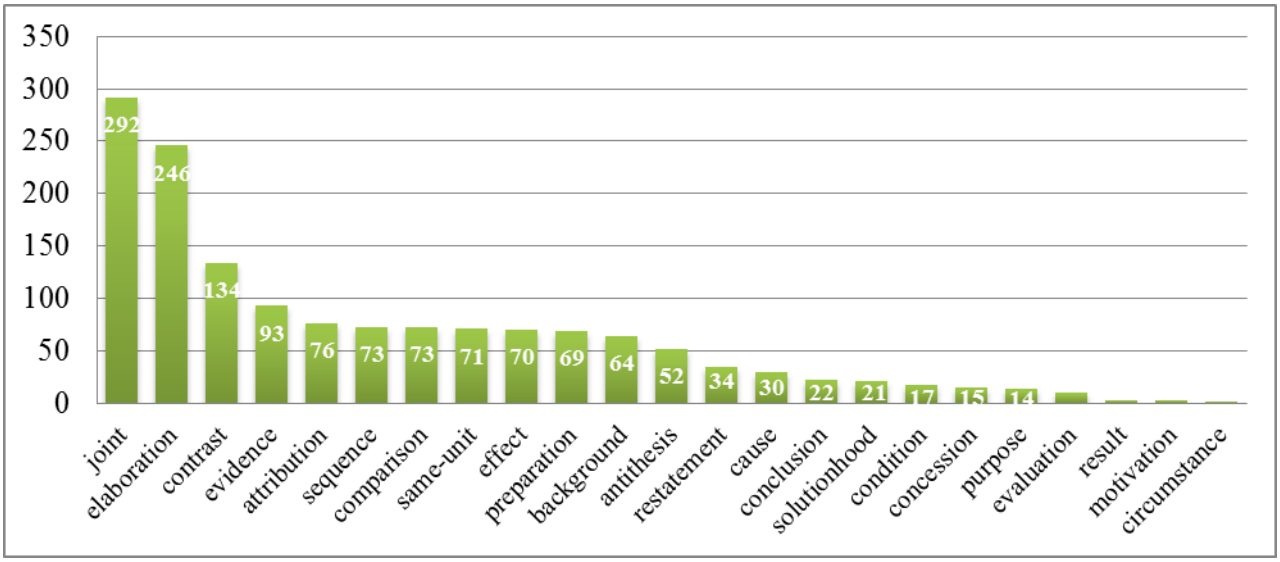

Рис. 2. Распределение отношений по частоте. (Количество как одноядерных, так и мультиядерных связей соответствует количеству вхождений данного отношения, т.е. при подсчете мультиядерных отношений не учитывается, сколько именно ЭДЕ включает в себя данное отношение.)

Кроме того, в рамках данного исследования ведется работа по созданию списка лексических маркеров риторических отношений (на данный момент выделено 50 маркеров), а также по выявлению их корреляции с конкретными риторическими отношениями. После завершения разметки данного корпуса планируется определить список сильных маркеров, которые однозначно коррелируют с определенным отношением и впоследствии могут быть использованы при автоматизации процесса разметки, и слабых маркеров, связь которых с тем или иным отношением прослеживается не так явно. Аналогичную классификацию маркеров можно найти во многих работах для разных языков (см., например [29, 30, 31, 32, 33]). Обобщая результаты, приведенные в данных исследованиях, можно сделать вывод, что наиболее однозначна корреляция между отношением *Concession (Уступка)* и маркером *'although'*, а также между *Condition (Условие)* и *'if'*. Кроме того, достаточно просто формализуются другие контрастные отношения (*Contrast, Antithesis*), а также различные причинно-следственные отношения. Соответственно, к наиболее сильным маркерам обычно относят *'but', 'however'* и *'since', 'because', 'due to'*. Можно предположить, что для русского языка будет наблюдаться сходная картина, однако все же со своей спецификой. Так, например, мы выяснили, что конструкция с союзом 'если…, то...' только в 70% случаев содержит в себе отношение Условия: в остальных же - либо отношение Контраста, либо только одну ЭДЕ (20% и 10% соответственно).

Заключение

Создание русскоязычного корпуса текстов с риторической разметкой является очень важной и актуальной задачей. На данном этапе работы необходима доработка инструкции по дискурсивной разметке, уточнение состава риторических отношений, а также продолжение разметки и увеличение объема корпуса – первый вариант корпуса будет содержать 30 текстов из СинТагРус. В дальнейшем этот корпус будет использован при разработке дискурсивного парсера для русского языка. Предполагается, что данный парсер будет работать на основе вручную составленных правил, учитывающих в том числе и дискурсивные маркеры, а с увеличением объема корпуса и на основе методов машинного обучения. Такой дискурсивный парсер будет использован для решения множества задач обработки текстов на русском языке.

Библиография

1. Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization //Text-Interdisciplinary Journal for the Study of Discourse. – 1988. – Т. 8. – №. 3. – С. 243-281.
2. Radev D. R. A common theory of information fusion from multiple text sources step one: cross-document structure //Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10. – Association for Computational Linguistics, 2000. – С. 74-83.
3. Miltsakaki E. et al. The Penn Discourse Treebank //LREC. – 2004.
4. Lascarides A., Asher N. Segmented discourse representation theory: Dynamic semantics with discourse structure //Computing meaning. – Springer Netherlands, 2008. – С. 87-124.
5. Guzmán, F., Joty, S., Màrquez, L., & Nakov, P. Using Discourse Structure Improves Machine Translation Evaluation //ACL (1). – 2014. – С. 687-698.
6. Joty S., Nakov P. DiscoTK: Using discourse structure for machine translation evaluation //In Proceedings of the Ninth Workshop on Statistical Machine Translation. – 2014.
7. Meyer, T., Popescu-Belis, A., Hajlaoui, N., & Gesmundo, A. Machine translation of labeled discourse connectives //Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA). – 2012. – №. EPFL-CONF-192524
8. Yoshida, Y., Suzuki, J., Hirao, T., & Nagata, M. Dependency-based Discourse Parser for Single-Document Summarization //EMNLP. – 2014. – С. 1834-1839.
9. Carenini G., Cheung J. C. K., Pauls A. MULTI-DOCUMENT SUMMARIZATION OF EVALUATIVE TEXT //Computational Intelligence. – 2013. – Т. 29. – №. 4. – С. 545-576.
10. Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., & Nejat, B. Abstractive Summarization of Product Reviews Using Discourse Structure //EMNLP. – 2014. – С. 1602-1613.
11. Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. Fine-Grained Sentiment Analysis with Structural Features //IJCNLP. – 2011. – С. 336-344.
12. Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. Polarity analysis of texts using discourse structure //Proceedings of the 20th ACM international conference on Information and knowledge management. – ACM, 2011. – С. 1061-1070.
13. Feng, V. W., Lin, Z., Hirst, G., & Holdings, S. P. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence //COLING. – 2014. – С. 940-949.

14. Feng V. W. RST-style discourse parsing and its applications in discourse analysis : дис. – University of Toronto, 2015.].
15. Ананьева М.И., Кобозева М.В. Дискурсивный анализ в задачах обработки естественного языка. // Конференция «Информатика, управление и системный анализ», ИУСА, 2016, в печати.
16. https://catalog.ldc.upenn.edu/LDC2002T07
17. Кибрик А. и др. (ред.). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. – Litres, 2014.
18. Литвиненко А. О. Описание структуры дискурса в рамках Теории Риторической Структуры: применение на русском материале //Труды Международного семинара Диалог. – 2001. – С. 159-168.
19. Кибрик А. А. Анализ дискурса в когнитивной перспективе //Дисс.… докт. филол. наук. – 2003.
20. Carlson L., Marcu D. Discourse tagging reference manual //ISI Technical Report ISI-TR-545. – 2001. – Т. 54.
21. Lüngen H. et al. Discourse relations and document structure //Linguistic modeling of information and markup languages. – Springer Netherlands, 2010. – С. 97-123.
22. Taboada M., Habel C. Rhetorical relations in multimodal documents //Discourse Studies. – 2013. – Т. 15. – №. 1. – С. 65-89.
23. http://www.ruscorpora.ru/instruction-syntax.html
24. Shelmanov A. O., Smirnov I. V., Methods for Semantic Role Labeling of Russian Texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014). Issue 13 (20). – 2014. – pp. 580-592.
25. Kamenskaya M.A, Khramoin I.V., Smirnov I.V. Data-driven methods for anaphora resolution of Russian texts //Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, 2014, pp. 134-136].
26. http://www.wagsoft.com/RSTTool/
27. rstWeb - Browser Annotation of Rhetorical Structure Theory: https://corpling.uis.georgetown.edu/rstweb/info/
28. Marcu D., Amorrortu E., Romera M. Experiments in constructing a corpus of discourse trees //Proceedings of the ACL’99 Workshop on Standards and Tools for Discourse Tagging. – 1999. – С. 48-57.
29. Cardoso P. C. F. et al. CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese //the Proceedings of the 3rd RST Brazilian Meeting. – 2011. – С. 88-105.
30. Cao S., da Cunha I., Bel N. A contrastive study of Spanish-Chinese intra-sentence discourse structures based on the discourse marker “although”.
31. da Cunha I. A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers //Research in Computing Science. – 2013. – Т. 70. – С. 93-104.
32. Spenader J., Lobanova A. Reliable discourse markers for contrast relations //Proceedings of the Eighth International Conference on Computational Semantics. – Association for Computational Linguistics, 2009. – С. 210-221.
33. Taboada M. Discourse markers as signals (or not) of rhetorical relations //Journal of Pragmatics. – 2006. – Т. 38. – №. 4. – С. 567-592.