УДК 81’322.2

ИСПОЛЬЗОВАНИЕ КОРПУСОВ В ИССЛЕДОВАНИЯХ ДИСКУРСА

К. М. Шилихина

Воронежский государственный университет

Поступила в редакцию 29 апреля 2014 г.

Аннотация: *в статье обсуждается вопрос о привлечении методов корпусной лингвистики для изучения дискурсивных явлений. Несмотря на то, что возможности существующих корпусов накладывают существенные ограничения на способы их использования в анализе дискурса, ориентация на узус сближает интересы двух исследовательских направлений. В работе дается краткий обзор исследований, в которых дискурсивные явления изучаются на материале специализированных корпусов. На примере исследования вербальной иронии в статье демонстрируются способы использования данных Национального корпуса русского языка, а также обсуждаются принципы формирования специализированного корпуса и связанные с ними проблемы создания схемы дискурсивной аннотации текстов.*
Ключевые слова: *корпусная лингвистика, анализ дискурса, узус, дискурсивная разметка, вербальная ирония.*

Abstract: *the paper addresses the issue of integrating methods of corpus linguistics into discourse analysis. While modern corpora are widely used in lexical and grammatical studies, their application to the research of various discourse phenomena seems to be limited. However, both corpus linguistics and discourse analysis are interested in usage rather than the norm, and this shared interest allows for application of corpus methods in discourse studies. The paper presents a brief overview of discourse analyses based on specialized corpora. The study of verbal irony illustrates how specialized and general language corpora can be used in discourse analysis. It also serves as an example that demonstrates the development of a discourse annotation scheme.*
Key words: *corpus linguistics, discourse analysis, usage, discursive annotation, verbal irony.*

1. Анализ дискурса и корпусная лингвистика: есть ли точки пересечения?

Проблему выбора исследовательского материала можно считать одной из наиболее острых для современной лингвистики. Казалось бы, имея в распоряжении многочисленные словари и грамматики, тексты, возможности записи устной речи или прямых контактов с носителями языка, а также собственную интуицию, лингвисты не должны испытывать трудностей в выборе источника данных для анализа. Однако можно указать на две существенные проблемы, с которыми сталкиваются исследователи: во-первых, каждый из названных источников отражает только определенные аспекты существования языка. Во-вторых, не всякий источник может предоставить достаточный объем информации для конкретного исследования. Особенно трудно в этом смысле исследователям, в сферу интересов которых попадают не отдельные явления, относящиеся к определенному уровню языковой системы, но и использование языка в коммуникации, т.е. дискурс. Одна из трудностей заключается в необходимости искусственно ограничивать круг внешних факторов, которые потенциально могут оказать воздействие на речевую деятельность носителей языка. Как осуществляется это ограничение на практике?

В современном анализе дискурса очень часто текст – та единственная сущность, которая дана лингвистам в прямом наблюдении, – используется как источник иллюстративных примеров для подтверждения обсуждаемой теории [1]. Также в подобных работах важную роль играет собственная интуиция исследователя как носителя языка. Интуитивное знание языка вместе с профессиональным знанием о языке позволяет лингвистам конструировать примеры, а вместе с ними и потенциальные контексты для их употребления. Результатом такого отношения к тексту является ситуация параллельного сосуществования множества моделей и теорий, которые описывают дискурс фрагментарно, а в ряде случаев могут противоречить друг другу.

На протяжении последних пятидесяти лет снизить уровень субъективности в отборе и анализе материала в «уровневой» лингвистике помогает обращение к языковым корпусам. Важным достоинством корпуса является возможность контролировать различные «переменные» коммуникации: тексты в корпусах снабжены метаразметкой – дополнительной информацией об авторах, времени и месте создания, жанровой принадлежности и т.д. Благодаря ей пользователь получает возможность работать с текстами, характеристики которых оказываются релевантными для проводимого исследования. Казалось бы, корпус – практически идеальный источник данных и для

анализа дискурсивных явлений, однако для тех, кто занимается анализом дискурса, возможность обращения к корпусам связана с рядом проблем. Во-первых, качество исследования очень сильно зависит от размера корпуса и времени его создания. Во-вторых, для лингвистов, занимающихся коммуникативной проблематикой, существует серьезное препятствие: современные корпусы ориентированы преимущественно на анализ лексических и/или грамматических явлений, а единицы коммуникации, которые не имеют стандартных способов выражения (например, речевые акты), в корпусах не аннотируются. В результате корпус не может предоставить пользователю возможность получения данных о сложных единицах коммуникации. Поэтому исследователи, опирающиеся на собственную интуицию и читательский опыт, отрицают пользу корпусов в исследованиях свойств текста и дискурса [2]. В качестве аргументов в поддержку данной точки зрения высказываются мысли о невозможности создания такой аннотации, которая адекватно отражала бы структурные и семантические свойства текстов. Действительно, аннотирование дискурсивных явлений остается одной из неразработанных проблем корпусной лингвистики. Кроме того, исследователи сомневаются в возможности использования статистического анализа применительно к структурно-смысловым свойствам текстов.

На первый взгляд, совместить идеологию анализа дискурса и приемы и методы корпусной лингвистики невозможно. Однако также известно и то, что у анализа дискурса и корпусной лингвистики есть общий объект исследования – употребление, т.е. узус (в отличие от ориентации на системные свойства языка, свойственные «уровневой» лингвистике). Еще один немаловажный фактор в пользу объединения усилий – это то, что корпусный анализ позволяет увидеть факты, которые оказываются недоступными при «интуитивном» подходе к отбору и анализу текстов. Названные обстоятельства побуждают лингвистов искать возможности для применения корпусных методов сбора и анализа лингвистического материала.

2. Опыт использования корпусных методов в исследованиях дискурса

Примером успешного применения методов корпусной лингвистики в исследованиях дискурсивных явлений можно считать работу Э. Семино и М. Шорта, посвященную способам представления речи и мысли в английских текстах [3]. Для этого проекта был создан корпус объемом 250 тыс. словоупотреблений. Дискурсивная аннотация текстов, т.е. приписывание дополнительной информации о структурных и содержательных свойствах текста, производилась вручную. При этом исследователи опирались на уже существовавшие классификации способов передачи речи и мысли в нарративных текстах.

Можно указать на ряд корпусных исследований устного дискурса: ярким примером является проект «Рассказы о сновидениях» А. А. Кибрика, В. И. Подлесской и др. [4], в котором записи устных нарративов (детских рассказов о снах) были транскрибированы и далее размечены в терминах Теории риторической структуры текста [5]. Риторические отношения также лежат в основе корпусного исследования М. Табоады, посвященного вопросам когерентности и когезии в диалогическом общении [6]. Материалом для изучения риторических связей послужил параллельный англо-испанский корпус записей диалогов, цель которых – договориться о времени встречи.

Еще один пример использования корпуса в изучении дискурса – исследование анафорических отношений, представленное в работе Р. Гарсайда, С. Флигельстоуна и С. Ботли [7]. Изучение анафоры – это попытка ответить на вопрос, как с помощью кореферентных групп обеспечивается смысловая связность дискурса.

Упомянутые выше исследования объединяет, во-первых, методологическое стремление лингвистов использовать корпусы для решения проблем анализа дискурса, во-вторых, желание ответить на вопрос, как именно обеспечивается смысловая целостность, т.е. когерентность дискурса.

Изучение механизмов обеспечения когерентности – это сложная исследовательская задача, имеющая как теоретическое, так и прикладное значение. Однако необходимо помнить и о том, что в реальности далеко не вся коммуникация воспринимается нами как семантически и прагматически когерентный дискурс. Нарушения смысловой структуры дискурса – это не менее интересный для изучения феномен. В частности, намеренно создаваемая говорящим некогерентность, т.е. смысловая «нестыковка» между элементами высказывания или между высказыванием и описываемой ситуацией лежит в основе вербальной иронии. В разделе 4 данной статьи мы постараемся продемонстрировать, как различные стратегии и тактики создания некогерентности в дискурсе могут быть использованы в качестве классификационной основы для дискурсивной аннотации корпуса. В разделе 3 сосредоточимся на вопросе, в какой мере может быть полезен национальный корпус для изучения вербальной иронии.

3. Использование Национального корпуса русского языка в исследовании вербальной иронии

Поскольку ирония не имеет постоянных способов выражения, имеющиеся корпусы оказываются малоинформативными: морфологической, синтаксической

и семантической разметки не хватает для обнаружения иронии. Так, в Национальном корпусе русского языка (далее – НКРЯ) можно обнаружить только 6 случаев указания на иронию говорящего с помощью пометы *[с иронией]* в устном подкорпусе:

Ну / дать ему еще денег побольше / у него мало денег (сказано ***с иронией****).*

Возникает вопрос: означает ли факт отсутствия дискурсивной аннотации в национальном языковом корпусе невозможность его использования в исследовании иронии?

Возможным вариантом применения такого корпуса можно считать исследование метапрагматических маркеров модуса коммуникации (bona fide или non-bona fide), в том числе и иронии как одного из режимов модуса non bona fide. Национальный корпус позволяет формировать конкордансы контекстов, в которых ироническая интенция маркирована эксплицитно, например, с помощью глагола *иронизировать* или *говорить с иронией.*

С семиотической точки зрения эксплицитные маркеры модуса коммуникации (*я говорю серьезно, я иронизирую* и т.д.) – это индексы, с помощью которых участники дискурса «увязывают» ситуацию и высказывание. Их появление в речи объясняется потребностью в обеспечении когерентности дискурса в тех случаях, когда возможна двойственная интерпретация текста/высказывания.

Для изучения метапрагматической деятельности носителей русского языка был сформирован конкорданс из 675 контекстов, в которых глагол *иронизировать* или предложная группа *с иронией* функционируют как металингвистические комментарии. Анализ конкорданса показал, что необходимость снятия иллокутивной неоднозначности может быть вызвана несколькими причинами:

1) говорящий опасается неверной интерпретации его намерений адресатом; чтобы избежать непонимания, он эксплицитно обозначает свои коммуникативные намерения:

Автор не не умеет говорить «нет», как Вы деликатно (на Вас не похоже) выразились, она ищет повод остаться хорошей и при этом соблюсти шкурный интерес. Это и называется двойной моралью. Я это так называю: -). И именно над этим я ***иронизирую****. Теперь понятно? Для «особо одаренных» комментирую: про мебель из Лувра, «Три медведя», «Ералаш» и «ЦЕЛЫЙ МЕСЯЦ любви и счастья» – это была ирония. Да, недобрая.*

2) говорящий хочет, чтобы адресат правильно понял, как соотносятся высказывание и реальность: в случае bona fide коммуникации высказывание соответствует реальности, в случае non-bona fide общения соответствие нарушается; именно об этом сообщают маркеры:

Но, человек важный и значительный, придя в цирк, он без сомнений взялся за дело. И уже через неделю выглядел «крупным специалистом» в цирковом деле. Я не зря говорю об этом ***с иронией****.*

3) адресат хочет проверить, правильно ли он/она понимает интенции собеседника:

Вот это цифрища! Просто кошмар какой-то! Как ты после этого можешь смотреть в глаза своим коллегам? – ***Ты иронизируешь****? А у тебя сколько нераскрытых дел? – Десять. На мне висит целая десятка.*

4) адресат отказывается принять предлагаемый модус non-bona fide общения и возвращает диалог в исходный модус:

– Как интересно, – вздохнула Валя. – Вот выпила я стакан воды и… стала умной. – ***Напрасно ты иронизируешь****, – сурово начал Арсений Никитич, но Валя его перебила: – Дед, а правда, почему ты не хочешь переехать ко мне?*

5) наблюдатель (повествователь) оценивает высказывание как ироническое, задавая способ интерпретации текста читателем:

Милюкову принесли заявление четырех великих князей, соглашавшихся на ответственное министерство. «Интересный исторический документ», – ***с иронией заметил*** *он, убирая бумажку в портфель.*

Таким образом, НКРЯ оказывается полезным ресурсом для изучения металингвистической деятельности: в исследовании вербальной иронии корпус дает возможность выяснить, какие высказывания классифицируются носителями русского языка как ирония.

4. Создание собственного корпуса текстов для исследования вербальной иронии

Поскольку для изучения иронии в дискурсе поисковых возможностей НКРЯ оказывается недостаточно, решением проблемы отбора материала для исследования является создание специализированного корпуса, в котором тексты снабжены дискурсивной разметкой (аннотацией). Дискурсивная разметка позволяет выделять фрагменты текстов, содержащих вербальные сигналы иронии, а также осуществлять статистический анализ частоты появления различных стратегий и тактик создания иронии в дискурсе.

Одной из проблем, с которой сталкиваются разработчики корпусов, является проблема сбалансированности и репрезентативности корпуса. Поскольку исходной задачей нашего исследования было изучение иронии в различных сферах коммуникации, в корпус включались записи устной речи, фрагменты компьютерно-опосредованной коммуникации и письменные нехудожественные тексты, функционирующие в сфере академического и политического

дискурса, публикации различных СМИ. Благодаря разнообразию источников обеспечивается репрезентативность корпуса. Что касается сбалансированности, то она достигается, с одной стороны, благодаря жанровому разнообразию включенных текстов (например, для компьютерно-опосредованной коммуникации такими источниками послужили блоги, форумы, социальные сети, твиттер, новостные ленты и др.). С другой стороны, сбалансированность обеспечивается объемом текстового материала. Объем корпуса составляет около 2,5 млн словоупотреблений. Информация о составе корпуса обобщена в табл. 1.

Метаразметка, т.е. информация о текстах, вошедших в корпус, включает информацию о языке (русский, английский), форме коммуникации (письменные тексты, записи устной речи, фрагменты компьютерно-опосредованной коммуникации), жанре и тематике текстов. Основной проблемой была разработка схемы дискурсивной разметки, которая позволила бы аннотировать фрагменты текстов с точки зрения способа создания иронии.

Метаразметка текстов и дискурсивная аннотация осуществлялись в программе UAM CorpusTool, которая приписывает xml-тэги, содержащие дополнительную информацию о тексте или его отдельных элементах [8]. Поскольку набор тэгов устроен иерархически, сам пользователь может определить, в какой степени детальной должна быть аннотация текстового материала.

Создание схемы дискурсивной аннотации носило итеративный характер. На начальном этапе исследования в схему разметки были включены способы создания иронии, уже описанные в лингвистической литературе [9; 10]. Далее, по мере добавления текстов, схема разметки уточнялась, список способов создания иронии расширялся. Тексты, размеченные ранее, проверялись и при необходимости размечались повторно. Таким образом, создание схемы – это процесс уточнения существующих представлений о способах создания иронии в высказывании и тексте.

В окончательном варианте схемы аннотации выделенные способы создания иронии были разделены на три стратегии: вербальную, дискурсивную (риторическую) и когнитивную. Эти стратегии соответствуют трем группам факторов, которые влияют на наше восприятие дискурса.

Первую группу составляют лингвистические факторы, связанные с имеющимся у носителей языка эксплицитным и имплицитным знанием конвенций узуса и языковой нормы, а также с умением отличать нормативное и приемлемое, ожидаемое использование языка от ненормативного, необычного, неожиданного.

Во вторую группу объединены когнитивные факторы, которые связаны с имеющимися у коммуникантов знаниями о мире и представлениями о нормальном ходе событий.

Третья группа – это дискурсивные (риторические) факторы, которые проявляются в умении устанавливать смысловые связи между отдельными отрезками дискурса, в том числе и в тех случаях, когда эти отрезки не являются составными компонентами единой коммуникативной ситуации, а разделены во времени и пространстве. К этой же группе относятся и наши знания о том, как должен быть организован диалог, какими риторическими свойствами должен обладать семантически связный текст.

Схема дискурсивной разметки приведена на рисунке.

Лингвистическая стратегия проявляется в намеренном отступлении от конвенций узуса или языковой нормы, в нетривиальном использовании лексических, грамматических или стилистических языковых средств. Описание лингвистических тактик создания иронии ориентировано на традиционное для лингвистики уровневое представление о структуре языка, в соответствии с которым были выделены лексические, грамматические и стилистические средства создания иронии в высказывании/тексте.

Риторическая стратегия создания иронии основана на имеющемся у коммуникантов знании о том, как должен быть организован семантически целостный текст. Дискурсивные тактики также делятся на две группы: к первой относятся те случаи, когда говорящий намеренно отступает от канонической структуры текста. Во вторую группу включены случаи, когда для создания иронии говорящий опирается на интертекстуальные связи.

Когнитивная стратегия «эксплуатирует» имеющиеся у коммуникантов знания о мире и о том, какое положение дел считается нормальным. Ироничные высказывания и тексты противоречат этим

Таблица 1

Состав корпуса

	Письменные тексты	Устные тексты	Компьютерно-опосредованная коммуникация
Количество текстов	500	500	500
Количество словоформ	774 000	882 000	879 000

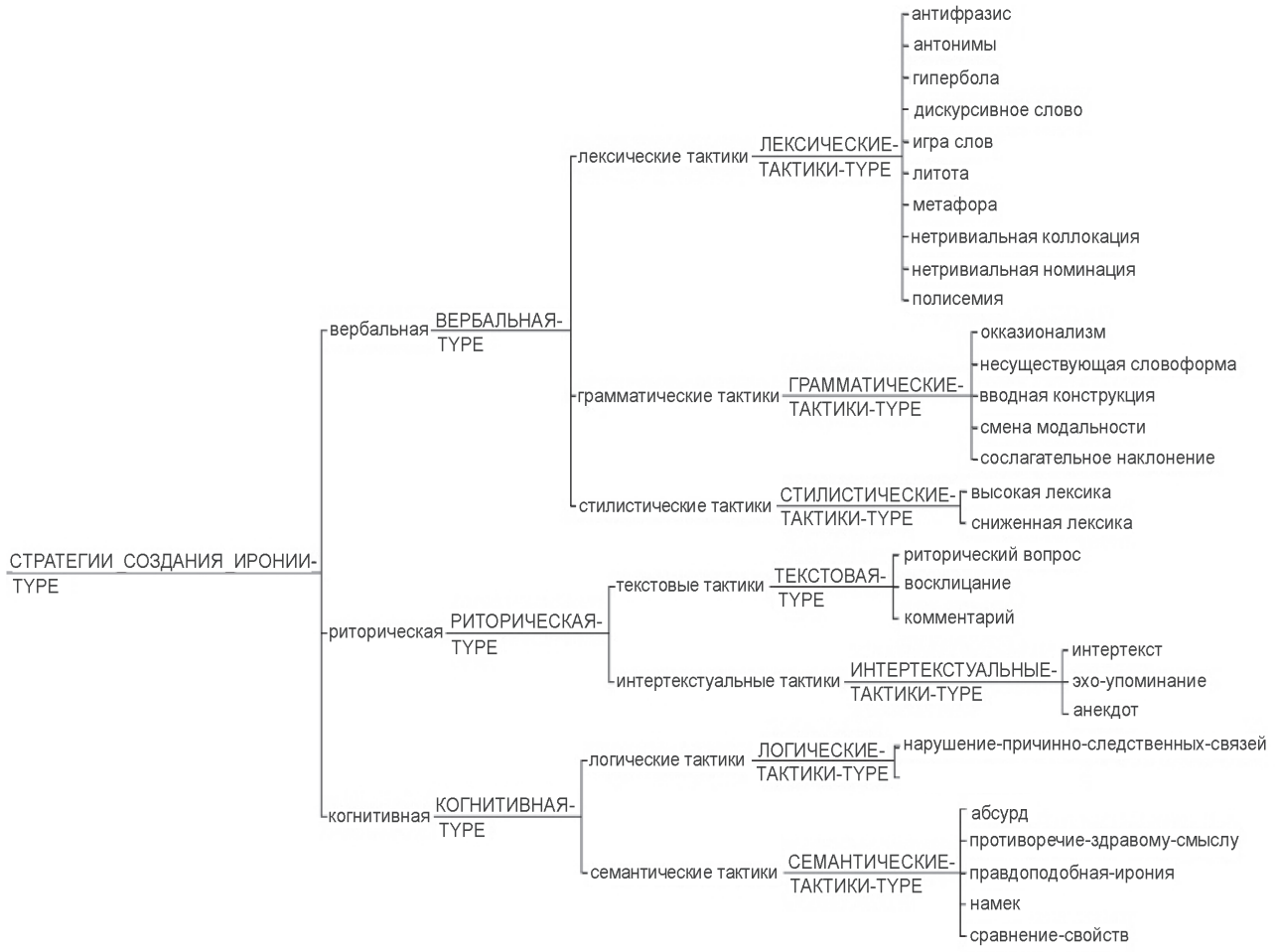

Рисунок. Схема дискурсивной разметки стратегий и тактик создания иронии

знаниям, представляя ситуацию странным или необычным образом.

Дискурсивная разметка не может выполняться в автоматическом режиме, поскольку нас интересуют явления, связанные не столько с формальной, сколько с содержательной стороной коммуникации. Дискурсивная аннотация текста – это результат аналитической деятельности с опорой на контекстуальные и прагматические составляющие. Кроме того, сам процесс создания схемы – это также и процесс уточнения существующих представлений о способах создания иронии в высказывании и тексте. Данные о частоте реализации выделенных стратегий приведены в табл. 2.

Корпусное исследование иронии оказалось полезным в нескольких отношениях. Во-первых, были выявлены группы факторов, влияющих на возможность иронической интерпретации. Во-вторых, удалось увидеть то общее, что объединяет все разрозненные способы создания иронии в дискурсе – этими свойствами являются намеренное нарушение смысловой целостности дискурса, игровое поведение говорящего и имплицитное выражение деонтической оценки. Кроме того, аннотирование текстов по созданной схеме позволило увидеть не только самые частотные способы создания иронии в дискурсе, но и «маргинальные» стратегии и тактики.

Таблица 2

Частота реализации различных стратегий создания иронии

	Вербальная стратегия	Риторическая стратегия	Когнитивная стратегия
Количество контекстов	1569	738	591

ЛИТЕРАТУРА

1. *Плунгян В. А.* Корпус как инструмент и как идеология : о некоторых уроках современной корпусной лингвистики / В. А. Плунгян // Русский язык в научном освещении. – 2008. – № 2(16). – С. 7–20.

2. *Fludernik M.* Towards a Natural Narratology / M. Fludernik. – London ; New York : Routledge, 2002. – 472 p.

3. *Semino E.* Corpus Stylistics : Speech, Writing and Thought Presentation in a Corpus of English Writing / E. Semino, M. Short. – London ; New York : Routledge, 2004. – 272 p.

4. Рассказы о сновидениях : корпусное исследование устного русского дискурса / под ред. А. А. Кибрика и В. И. Подлесской. – М. : Языки славянских культур, 2009. – 736 с.

5. *Mann W.* Rhetorical Structure Theory : Toward a Functional Theory of Text Organization / W. Mann, S. Thompson // Text. – 1988. – № 8. – P. 243–281.

6. *Taboada M.* Building Coherence and Cohesion : Task-oriented Dialogue in English and Spanish / M. Taboada. – Amsterdam ; Philadelphia : John Benjamins, 2004. – 261 p.

7. *Garside R.* Discourse Annotation : Anaphoric Relations in Corpora / R. Garside, S. Fligelstone & S. Botley // Corpus Annotation / ed. by R. Garside, G. Leech & T. McEnery. – London ; New York : Routledge, 1997.

8. *O'Donnell M.* Demonstration of the UAM CorpusTool for Text and Image Annotation / M. O'Donnell // Proceedings of the ACL-08: HLT Demo Session (Companion Volume). – Columbus : The Ohio State University, 2008. – P. 13–16.

9. *Ермакова О. П.* Ирония и ее роль в жизни языка / О. П. Ермакова. – Калуга : Изд-во КГПУ им. К. Э. Циолковского, 2005. – 204 с.

10. *Походня С. И.* Языковые виды и средства реализации иронии / С. И. Походня. – Киев : Наукова думка, 1989. – 126 с.

Воронежский государственный университет

Шилихина К. М., кандидат филологических наук, доцент, докторант кафедры теории перевода и межкультурной коммуникации

E-mail: shilikhina@gmail.com

Тел.: 8 (473) 220-41-49

Voronezh State University

Shilikhina K. M. Candidate of Philology, Associate Professor, Post-Doc Researcher of the Translatology and Intercultural Communication Department

E-mail: shilikhina@gmail.com

Tel.: 8 (473) 220-41-49