Референциальный выбор в устных и письменных рассказах: сопоставительное исследование на материале корпуса «Веселые истории из жизни»

Марина Андреевна Шумилина

Московский государственный университет имени М. В. Ломоносова, Москва, Россия;
Институт языкознания РАН, Москва, Россия; mari.an.shum@iling-ran.ru

Аннотация: Исследование посвящено сравнению референциального выбора в устном и письменном дискурсе на русском языке. Референциальный выбор рассматривается в виде тройного противопоставления полных именных групп, местоимений и нулевых именных групп. Работа выполнена на материале дискурсов, каждый из которых изложен рассказчиком дважды — в устной и письменной формах. В рассказах все именные группы были идентифицированы и охарактеризованы по 29 параметрам. На собранных выборках были обучены модели мультиномиальной логистической регрессии и деревья решений, на основе деревьев решений были построены диаграммы значимости факторов. Интерпретация моделей и диаграмм показывает, что некоторые факторы референциального выбора по-разному проявляют себя в устных и письменных дискурсах, например грамматическая роль, семантическая гиперроль и неполная кореферентность между анафором и антецедентом. Также модели демонстрируют, что наборы факторов, значимых для двух выборок, неодинаковы: в частности, одушевленность референта и семантическая гиперроль анафора присутствуют в дереве решений только для письменного дискурса.

Ключевые слова: дискурс, референция, русский язык, устный дискурс

Благодарности: Автор выражает искреннюю благодарность А. А. Кибрику и двум анонимным рецензентам за полезные замечания по структуре и стилю текста, а также Г. Доброву, А. Большиной и К. Студеникиной за консультации по предварительной обработке данных, подбору и применению методов машинного обучения.

Для цитирования: Шумилина М. А. Референциальный выбор в устных и письменных рассказах: сопоставительное исследование на материале корпуса «Веселые истории из жизни». *Вопросы языкознания*, 2024, 6: 133–159.

DOI: 10.31857/0373-658X.2024.6.133-159

Referential choice in spoken and written stories: A comparative study based on the corpus “Funny life stories”

Marina A. Shumilina

Lomonosov Moscow State University, Moscow, Russia;
Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia; mari.an.shum@iling-ran.ru

Abstract: The study is concerned with referential choice in spoken and written discourse in the Russian language. I consider referential choice to consist in a threefold opposition between full noun phrases, pronouns, and zero noun phrases. The study is based on discourses each of which was presented by its narrator twice, namely in the spoken and written forms. In each story, all the noun phrases were identified and described according to 29 parameters. I trained logistic regression models and decision trees on the collected samples and analyzed factor importance diagrams built on the basis of the decision trees. The interpretation of the models and diagrams shows that some factors have different impact

on referential choice in spoken and written discourses, for instance, grammatical role, semantic hyperrole and sloppy identity between the anaphor and the antecedent. Besides, the models also demonstrate that the sets of significant factors for the two samples are not identical: in particular, the referent's animacy and the anaphor's semantic hyperrole are present solely in the decision tree for written discourse.
Keywords: discourse, oral discourse, reference, Russian
Acknowledgements: I am very grateful to A. A. Kibrik and the two anonymous reviewers for their extremely helpful suggestions and remarks on this work, as well as to G. Dobrov, A. Bolshina, and K. Studenikina for their recommendations on preliminary data processing, and also choice and application of machine-learning algorithms.
For citation: Shumilina M. A. Referential choice in spoken and written stories: A comparative study based on the corpus "Funny life stories". *Voprosy Jazykoznanija*, 2024, 6: 133–159.
DOI: 10.31857/0373-658X.2024.6.133-159

Введение

Референциальный выбор — это выбор, совершаемый говорящим в процессе порождения дискурса: для каждого референта, который говорящий хочет упомянуть, он должен выбрать план выражения. Например, один и тот же объект можно назвать *розовая стеклянная ваза*, *ваза*, *она* или *это*, а в некоторых случаях упоминание может быть фонологически не выражено, как в деепричастном обороте во фразе *ваза стояла на подоконнике и сияла, пропуская сквозь себя лучи света*.

Явление референциального выбора имеет достаточно длительную историю изучения как в синтаксисе, так и в дискурсивном анализе. В качестве примеров широко известных подходов к его моделированию можно назвать теорию центрирования [Grosz et al. 1986], теорию доступности [Ariel 1990] и теорию когнитивных статусов [Gundel et al. 1993].

В 1996 году А. А. Кибрик в статье [A. A. Kibrik 1996] предложил моделировать референциальный выбор в когнитивном ключе. Это означает, что в процессе референциального выбора решающей силой признаются когнитивные процессы внимания и активации в рабочей памяти говорящего [A. A. Kibrik 2011: 377], а также их грамматические и дискурсивные соответствия, например одушевленность и линейное расстояние в абзацах. Помимо этого, в рамках данного подхода каждый фактор получает количественную оценку того, насколько существенно его влияние на результат референциального выбора.

В когнитивной количественной модели референциальный выбор представлен как классический многофакторный процесс [А. А. Кибрик 1997: 94], моделирование которого состоит из нескольких этапов:

1) Сначала необходимо сформировать набор значимых факторов и подобрать числовые веса для значений каждого фактора.
2) Далее для каждого конкретного случая референциального выбора суммируются веса его факторных значений; полученное число отражает степень активации референта в данной точке дискурса и именуется коэффициентом активации.
3) После этого следует проинтерпретировать коэффициент с помощью специальной шкалы: при малых значениях выбор производится в пользу полной именной группы, при больших — в пользу местоимения или нулевой именной группы, промежуточные значения предписывают использовать полную ИГ либо местоимение [Там же: 101].
4) Заключительный функциональный элемент модели — набор фильтров. Фильтр — это условие, которое принуждает говорящего использовать полную ИГ вне зависимости от степени активации референта (подробнее см. п. 3.2).

Настоящая работа посвящена референциальному выбору в рассказах на русском языке. Исследование выполнено на основе 11 рассказов из корпуса «Веселые истории из жизни»,

далее ВИЖ[1] (подробнее см. п. 2.1). Материал рассматривается в рамках когнитивного количественного подхода.

Цель исследования — выявить различия в процессах референциального выбора (далее РВ) в рассказах, которые относятся к разным модусам. Под модусом в данной работе понимается канал передачи информации [Fox 1995; А. А. Kibrik 2011: 11–12]. Предлагается говорить о двух основных модусах — устном и письменном. Сравнение устных и письменных дискурсов представляется интересным направлением в свете того, что в современном языкознании еще сильна идея первостепенной важности письменного модуса: в частности, закономерности референциального выбора в письменном дискурсе могут быть неосторожно экстраполированы на устный дискурс. Как показано в [Fox 1995: 136–152], тип дискурса влияет на закономерности анафоры: автор исследует спонтанные устные диалоги и короткие тексты публицистического стиля, тем самым анализируя выборки из различных модусов.

В настоящей работе выдвинуты четыре гипотезы:

1) Существует **качественное** отличие в наборе факторов, значимых для РВ в устном и письменном модусах дискурса.
2) Для некоторых факторов, значимых в обоих модусах, существует **количественное** отличие между степенью влияния на РВ в каждом из модусов.
3) Для обоих модусов верно, что риторическое расстояние оказывает большее влияние на РВ, чем линейное расстояние[2].
4) Помимо факторов, которые уже учитывались в других машинных моделях РВ (например, в [Strube, Wolters 2002] и [van Vliet 2012]), в данной работе делается попытка ввести несколько теоретически значимых факторов, которые в этих моделях представлены не были. Такие факторы предлагается называть **экспериментальными**, большинство из них помогает более детально учитывать синтаксический контекст упоминания. Предполагается, что все они продемонстрируют значимость на материале исследования. Более подробно об этих факторах рассказано в п. 3.1.

Гипотезы 1, 2 и 3 мотивированы сопоставительным характером исследования. В основе гипотезы 4 заложено предположение, что для моделирования референциального выбора может быть полезно принимать во внимание свойства локального контекста упоминания.

Все ключевые материалы работы — размеченные дискурсы, их риторические деревья, база данных в табличной форме — доступны в облачном хранилище по ссылке: https://clck.ru/3EUWN7.

Статья организована следующим образом. В разделе 1 введены и объяснены ключевые понятия работы. В разделе 2 описаны сбор и разметка данных для моделей. Раздел 3 посвящен построению и интерпретации моделей, в его рамках отдельное внимание уделяется факторам (п. 3.1), фильтрам (п. 3.2) и данным (п. 3.3), на основе которых обучаются модели; устройству и анализу моделей посвящен п. 3.4. В разделе 4 приводится обсуждение полученных результатов. Заключительный раздел содержит краткие выводы и перспективы данного направления исследований.

---

[1] Сокращение принято вслед за [Буденная 2018]. Корпус доступен по ссылке: http://spokencorpora.ru/showcorpus.py?dir=02funny.

[2] Существует большое разнообразие способов измерить линейное расстояние между рассматриваемым упоминанием и его ближайшим линейным антецедентом. Так, в исследовании [Loukachevitch et al. 2011: 522] предложены измерения в словах, предложениях, абзацах; в статье [Strube, Wolters 2002] вводится специальная единица **major clause unit**. Кроме того, в [Carlson et al. 2003] был предложен способ выделения ЭДЕ (**элементарной дискурсивной единицы**) для письменных текстов. В настоящей работе линейное расстояние измерялось в ЭДЕ по концепции, изложенной в [А. А. Кибрик и др. 2009а]; см. подробнее в п. 2.2.

1. Ключевые понятия и явления

1.1. Основные термины

Прежде всего, следует определить понятия **референта** и **референтности именной группы**. В настоящей работе приняты определения, предложенные в книге [А. А. Kibrik 2011]. Референт — это понятие в сознании говорящего и адресата, соответствующее некоторому объекту во внеязыковом мире, реальном или сконструированном. Именная группа считается референтной, если имеет один из субстантивных референциальных статусов, конкретный либо неконкретный (подробнее о референциальных статусах см. [Ibid. 32]).

В рамках данной работы рассматривается именная референция, поэтому считается, что референт не может иметь предикатный или автонимный статус. Вслед за [А. А. Кибрик 1997] референты, упомянутые более чем однократно, называются **значимыми**.

Чтобы ввести еще два термина, обратимся к рабочей классификации ИГ, которая отражает некоторые свойства референта (рис. 1). Она разработана специально для данного исследования.

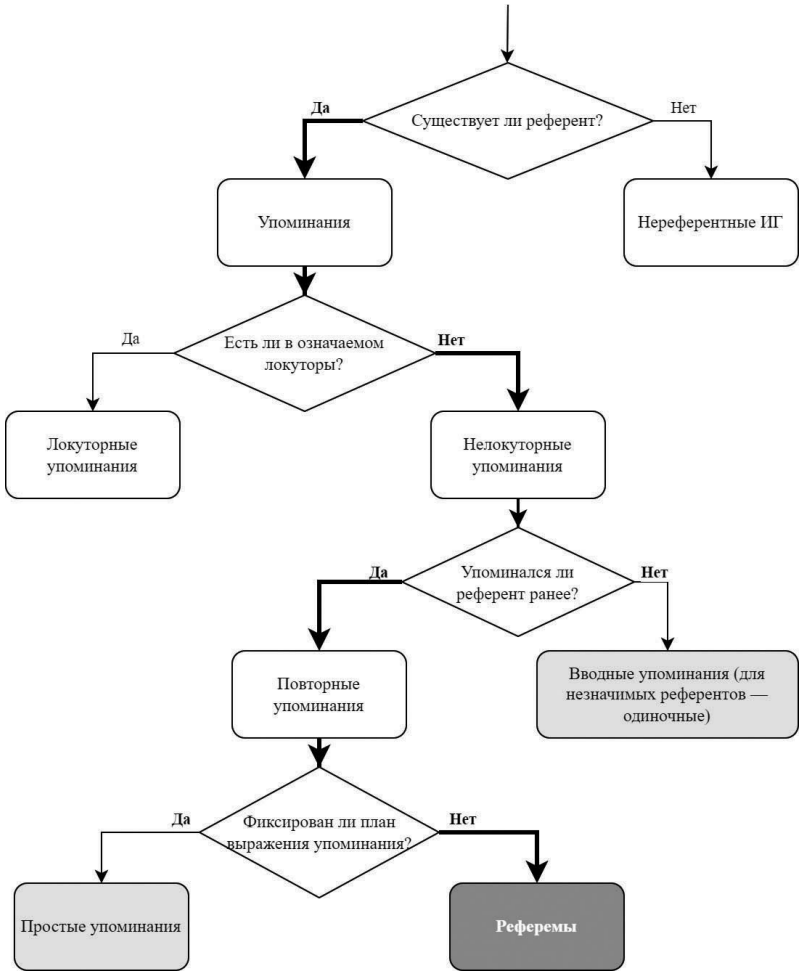

Рис. 1. Классификация именных групп

Все референтные ИГ предлагается называть **упоминаниями**. Если референт включает кого-либо из коммуникантов, то упоминание называется **локуторным** (*locutor reference* в [A. A. Kibrik 2011: 43]). **Нелокуторные** упоминания подразделяются на **вводные** (с ними также группируются **одиночные**) и **повторные**. Если говорящий должен самостоятельно выбрать план выражения для повторного упоминания, то оно является **референтом**.

Референта — одно из центральных понятий для данной работы. В качестве примера рассмотрим фрагмент устного рассказа из ВИЖ[3], сегментированный на элементарные дискурсивные единицы (далее ЭДЕ):

(1) FS_04-f_sp // № 4sp[4]:
27. моя подружка —
28. Кристина,
29. смотрит,
30. а там у него ***цветы всякие в горшках****,*
31. ну и мы спросили 0,
32. мол это е= ‖ он сам сажает ***0****.*

В ЭДЕ 32 при помощи референциального нуля повторно упоминаются цветы. В данном примере можно постулировать нуль, так как глагол «сажать» является переходным и предполагает наличие прямого объекта, при этом объект в предложении не выражен. Рассказчица имела возможность упомянуть их с помощью другого **референциального средства** (далее **РС**). В (2) показано несколько сконструированных альтернативных вариантов этой референты:

(2) *32(а). мол это е= ‖ он сам* ***их*** *сажает,*
32(б). мол это е= ‖ он сам сажает ***их****,*
32(в). мол это е= ‖ он сам ***эти цветы*** *сажает,*
32(г). мол это е= ‖ он сам сажает ***эти цветы****.*

Таким образом, **референта** — это нелокуторное повторное упоминание, для которого говорящий должен был выбрать план выражения.

Если упоминание обладает фиксированным планом выражения, то оно называется **простым**[5]. Другими словами, говорящий вынужден использовать то или иное РС в силу недискурсивных факторов (которые часто относятся к сфере грамматики). В качестве примера нулевого простого упоминания можно назвать подлежащее при инфинитивном обороте:

(3) FS_17-f_sp // № 9sp:
7. Как потом оказалось,
8. в день свадьбы —
9. нашей с ним,
10. — ***его мама*** *хотела* ***0*** *взять <его | с собой> паспорт в ЗАГС.*

---

[3] Здесь и далее каждый фрагмент из ВИЖ начинается с идентификатора рассказа, откуда он был взят. Слева от двух косых черт указан идентификатор в корпусе, справа — идентификатор в выборке в рамках исследования. Полный список дискурсов корпуса с их идентификаторами для настоящей работы приведен в табл. 1.

Каждая новая строка в примере — самостоятельная ЭДЕ. Нумерация ЭДЕ сохраняется от начала целого рассказа. Сохраняется исходная дискурсивная разметка. Кроме того, во всех примерах (не только из ВИЖ) ИГ, представляющие интерес, отмечены жирным шрифтом, а нулевые ИГ обозначаются как 0.

[4] Сохранена транскрипция, представленная в корпусе.

[5] Отсутствие стандартной процедуры референциального выбора сближает простые упоминания с ограничениями — классом явлений, который рассмотрен в статье [Залманов, Кибрик 2023].

Вводное упоминание и кореферентные ему референты вместе составляют **референциальную цепочку**, далее **РЦ** (термин «referential chain» широко используется в исследованиях референции, в частности в [Loukachevitch et al. 2011]).

1.2. Референты и простые упоминания

В работе уделено особое внимание проведению границы между референтами и простыми упоминаниями — это было необходимо, чтобы данные в выборках были однородными, что важно для качества машинных моделей. Отдельную сложность представляла работа с нулевыми повторными упоминаниями. С опорой на классификацию, изложенную в [А. А. Kibrik 1996: 262], для работы был составлен следующий список нулевых референт:

— Подлежащие:
  — нулевое подлежащее независимой клаузы;
  — нулевое подлежащее придаточного финитного предложения, например *Он позвонит вам, когда **0** закончит работу*.
— Прочие:
  — нулевое прямое дополнение (аккузативное или генитивное при отрицании);
  — нулевой дативный принципал (подробнее о системе семантических ролей см. п. 3.1) при категории состояния и модальных предикатах;
  — нулевая вершина с адъективными модификаторами (включая числительные).

Одно из ключевых отличий простых упоминаний от референт — изначальная безальтернативность референциального средства. Типы нулевых ИГ, не попавшие в список выше, были недостаточно представлены в проанализированных данных, поэтому не было возможности проверить их на наличие вариативности РС (то есть определить, возможно ли отнести их к референтам). По этой причине в данной работе они отнесены к простым упоминаниям.

Кроме того, некоторые референтные упоминания полностью исключаются из рассмотрения. Во-первых, личные местоимения локуторов, которые затруднительно считать анафорическими потому, что РС упоминаний говорящего и адресата регулируются не их антецедентами, а степенью влияния на дискурс каждого из коммуникантов [А. А. Kibrik 2011: 43]. Разумно предположить, что вариативность локуторных упоминаний регулируется иными механизмами, чем вариативность прототипических референт. Во-вторых, это референция непредметного типа. Состояния, события, а также локативная и темпоральная референция обладают особыми наборами РС[6]:

(4) Локация: {*На улице Менжинского* / *Там* / *В том месте*} *никогда не было театра*.

(5) Время: {*В 2020-м году* / *Тогда* / *В то время*} *занятия в школах проходили дистанционно*.

(6) Событие: {*Сборка мебели* / *Она* / *Это*} *не займет много времени*.

Похожая специфика отмечается у количественных групп: они обладают расширенным набором РС, что затрудняет их анализ. Разнообразие способов упомянуть такой референт показано в примерах (7)–(10):

(7) *Я училась в этой школе **пять лет**. **Они** прошли незаметно.*

(8) ***Десять лет** назад я училась в этой школе. **Столько лет** прошло с тех пор!*

[6] Примеры, при которых не указан источник, являются сконструированными.

(9) *Для исследования было нужно **тридцать участников**, но **столько их** было найти негде, поэтому пришлось опрашивать знакомых по несколько раз.*

(10) *За нарушение штраф **пять тысяч рублей**. Но у нас не было с собой **столько 0**.*

В рассмотренных дискурсах количественные группы встречаются достаточно редко. Имена существительные в составе предложений, маркирующих начало и конец рассказа, не расцениваются как упоминания, ср.: *…у меня тут на работе такая странная **история** произошла* (FS_13-f_sp // № 6sp); *В **этой истории** я была главная звезда* (FS_16-f_sp // № 9sp); *Вот такая вот **история*** (FS_02-f_sp // № 2sp).

Исключаются любые местоимения, кроме личных, указательных, притяжательных и возвратных. В прямой речи не рассматриваются обращения. Как уже было сказано в п. 1.1, исключены из рассмотрения ИГ, референты которых обладают предикатным и автонимным денотативными статусами.

Кроме того, несколько классов упоминаний рассматриваются исключительно как антецеденты реферем:

— возвратные местоимения;

— ИГ, при которых применяется один из фильтров (см. п. 3.2);

— вынесенные топики;

— нули при сочинительном сокращении;

— локуторные упоминания референтов в цитации, которые вне цитации не являются локуторами, например:

(11) FS_5-m_sp // № 11sp:
*22. пятого числа **папа** доложил,*
*23. и шестого **0** пошел **0** бегать по всяким авиакассам,*
*24. **0** объяснять,*
25. что «Вот,
26. тыры-пыры,
*27. **мне** надо 0 уехать.»,*
*28. ну **ему** говорят*

В процитированном фрагменте диалога (ЭДЕ 25–28) папа рассказчика выступает как говорящий, и в разборе рассказов подобные случаи учитываются как простые упоминания. Такое решение принято в силу эмпирической закономерности: если считать линейным антецедентом реферемы в ЭДЕ 28 упоминание в ЭДЕ 24 (то есть ближайшее упоминание вне прямой речи), то линейное расстояние между анафором и антецедентом не вполне объясняет фактический выбор РС (по данным устной выборки, при линейном расстоянии 4 полные ИГ более частотны, чем местоимения).

2. Сбор материала

2.1. Корпусные дискурсы

Корпус «Веселые истории из жизни» уже применялся для исследования РВ, в частности в диссертации [Буденная 2018]. Его особенность заключается в том, что каждый информант рассказывал одну и ту же историю дважды: сначала производилась запись рассказа в устном формате, затем рассказчик излагал тот же сюжет письменно. Устные представления рассказов имеют разметку в трех степенях подробности, для данной работы взят

минимальный вариант. Проанализировано 11 сюжетов, каждый рассмотрен в обоих модусах; извлечено и рассмотрено 211 референ из письменных текстов и 280 из устных. Как представляется, количество референ, собранное из такого объема текста, составляет минимум для статистических подсчетов.

В настоящей работе не ставилась цель получить выборку, сбалансированную по полу и/или возрасту рассказчиков, поэтому тексты рассматривались сплошным образом, однако некоторые были исключены за счет большого присутствия пространственной, временной и/или количественной референции (FS_07-f и FS_09-f); большого присутствия событийной референции (FS_10-m и FS_11-m); малого числа референ при достаточно большом количестве ЭДЕ (FS_08-f и FS_12-f); большого числа хезитаций и других речевых сбоев (FS_14-f).

Все рассмотренные рассказы получили новые идентификаторы. В табл. 1 приводится соответствие этих идентификаторам исходным.

Таблица 1

Полный список проанализированных дискурсов

Новый ID	Исходный ID	Новый ID	Исходный ID
1	FS_01-f	6	FS_13-f
2	FS_02-f	7	FS_15-m
3	FS_03-m	8	FS_16-f
4	FS_04-f	9	FS_17-f
5	FS_06-f	10	FS_18-f
		11	FS_05-m

Среди 11 рассказчиков восемь женщин и трое мужчин, возраст рассказчиков — от 18 до 22 лет (средний возраст 20,4 года).

В дальнейшем для каждого текстового примера из корпуса приводится только **новый** идентификатор дискурса-источника.

2.2. Линейная сегментация дискурса

Обсудим основные термины анализа дискурсивной структуры, принятые в работе. Для локальной сегментации используется понятие ЭДЕ — элементарной дискурсивной единицы: «EDUs are quanta, or moments, of discourse time: discourse progresses in steps equalling EDUs» [A. A. Kibrik 2011: 377]. Преимущество такой сегментации заключается в ее комплексном характере [А. А. Кибрик и др. 2009а: 57]:

— с точки зрения координации речепроизводства и дыхания ЭДЕ произносится за один выдох;

— с когнитивной точки зрения ЭДЕ содержит один «фокус сознания» [Chafe 1994];

— с позиции семантики ЭДЕ описывает одно событие или одно состояние;

— с позиции синтаксиса ЭДЕ часто является клаузой;

— с точки зрения просодии ЭДЕ имеет все признаки самостоятельной «произносительной конфигурации» (единый интонационный контур, характерная динамика темпа и громкости, паузация).

Понятие ЭДЕ применимо не только к устному дискурсу, но и к письменному; один из способов делить письменные предложения на ЭДЕ представлен в [Carlson et al. 2003: 87–89]: как правило, ЭДЕ совпадает с клаузой, однако есть три особых случая.

1) Некоторые клаузы не составляют отдельных ЭДЕ — это сентенциальные подлежащие, дополнения и комплементы клауз.
2) Некоторые неклаузальные элементы размечаются как ЭДЕ — это предложные группы, которые начинаются с сильного «дискурсивного маркера», см. [Ibid.: 98–99], например *because*, *in spite of*, *according to*.
3) Авторы вводят понятие вложенной ЭДЕ: таким образом размечаются относительные придаточные предложения, именные модификаторы в постпозиции (nominal post-modifiers), а также любая другая клауза, которая располагается внутри полноценной ЭДЕ. Если в полноценную ЭДЕ вложена клауза, то фрагмент ЭДЕ до этой клаузы и фрагмент после нее рассматриваются как отдельные ЭДЕ, единство которых на уровне риторической структуры восстанавливается с помощью специального отношения.

Важно отметить, что ЭДЕ в [Carlson et al. 2003] и ЭДЕ в [А. А. Кибрик и др. 2009а] соответствуют разным явлениям, поскольку эти понятия изначально разработаны для разных модусов дискурса.

2.3. Риторическая структура

Для описания смыслового устройства рассказов применяется теория риторической структуры [Mann, Thompson 1987]. В этой теории при помощи фиксированного набора смысловых отношений (например, «Обстоятельство», «Уступка», «Последовательность») устанавливаются связи между единицами дискурса:

— самой малой единицей в оригинальной теории является клауза (в настоящей работе — ЭДЕ);

— предложения соединяются смысловыми отношениями и объединяются в группы;

— группы можно соединять с единицами и с другими группами;

— самой большой группой является целый дискурс.

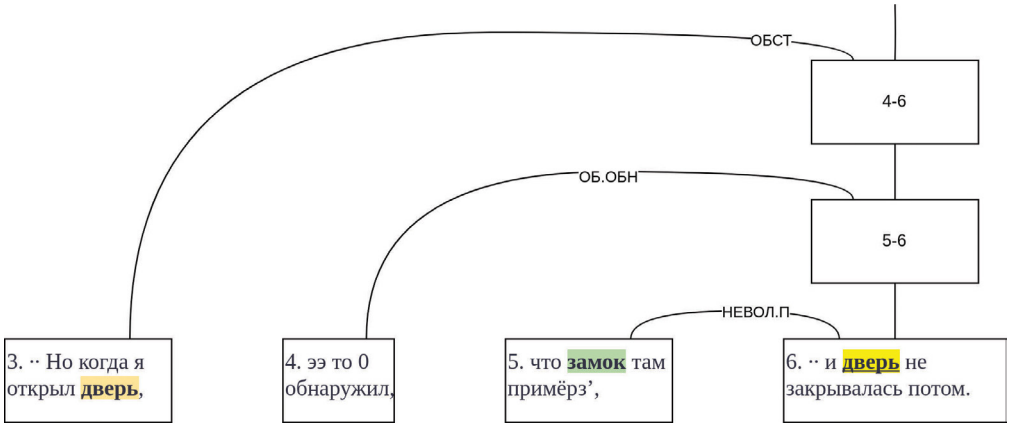

Рис. 2. Пример риторической структуры (часть рассказа № 7sp)[7]

[7] Обозначения отношений: НЕВОЛ.П — неволитивная причина; ОБСТ — обстоятельство; ОБ.ОБН — обстоятельство обнаружения. На рис. 2 и 3 для наглядности из дерева убрана дополнительная разметка, однако она присутствует в дереве данного рассказа в облачном хранилище.

Результат анализа представляет собой дерево, в котором учтены все ЭДЕ данного дискурса. Фрагмент такого дерева показан на рис. 2 (с. 141). Все диаграммы были построены вручную в draw.io — приложении для создания схем, графиков, деревьев решений и других логических и структурных визуализаций.

Различается два типа смысловых отношений.

1. Симметричное, или многоядерное, отношение соединяет два и более элементов структуры. Соединенные ЭДЕ являются одинаково значимыми в дискурсе. В дереве элементы соединены дугами между собой и косыми чертами с общим узлом группы. Примеры отношений: конъюнкция, последовательность. Также симметричное отношение показано на рис. 3.
2. Асимметричное отношение соединяет два элемента. Один из них является главным и называется ядром, другой второстепенным и называется сателлитом. Ядро соединено дугой с сателлитом и прямой чертой с общим обозначением группы. Примеры отношений: фон, обстоятельство, уступка. Такие отношения фигурируют на рис. 2 и 3.

Инвентарь риторических отношений может варьироваться в зависимости от типа рассматриваемых дискурсов. В настоящей работе задействована существенная часть перечня, который использовался в [Литвиненко и др. 2009], с добавлением трех новых отношений.

1. **Вставка**: симметрично, используется для зависимой клаузы, вложенной в главную, исключительно в письменных текстах; это отношение соединяет в один узел все три компонента, которые получаются в результате вложения клаузы[8]. Пример применения этого отношения изображен на рис. 3.

Рис. 3. Использование риторического отношения «Вставка» (обозначено как «ВСТАВ», соединяет ЭДЕ 9, 10 и 11) во фрагменте рассказа № 7wr

2. **Поиск**: симметрично, соединяет маркер поиска с полнозначными элементами дискурса до маркера и после него. Используется, когда говорящий останавливается, чтобы вспомнить нужное слово, и в процессе вспоминания порождает регуляторные ЭДЕ[9],

[8] Другой способ учитывать вложенные клаузы — специальное отношение SAME-UNIT [Carlson et al. 2003], но при таком подходе две несоседние единицы — части главной клаузы — оказываются соединенными «в обход» вложенной. С другой стороны, возможно присоединение вложенной клаузы к одной из частей главной клаузы: это помогает сохранить смысловые взаимоотношения между двумя клаузами, однако не всегда бывает легко выбрать часть главной клаузы, к которой будет естественно присоединить зависимую. Предложенное в настоящей статье решение можно назвать упрощенным в том отношении, что оно не отражает смысловых связей между клаузами, однако его преимущество состоит в том, что оно находится в рамках классической концепции риторического отношения.

[9] Термин «регуляторные ЭДЕ» используется в понимании, изложенном в [А. А. Кибрик и др. 2009б].

например *как его*[10]. Данное отношение имеет компонент разрывности («фокус сознания» уже оформлен, возникает лишь проблема с подбором конкретного слова), поэтому представляется возможным считать его симметричным. Пример использования такого отношения показан на рис. 4[11].

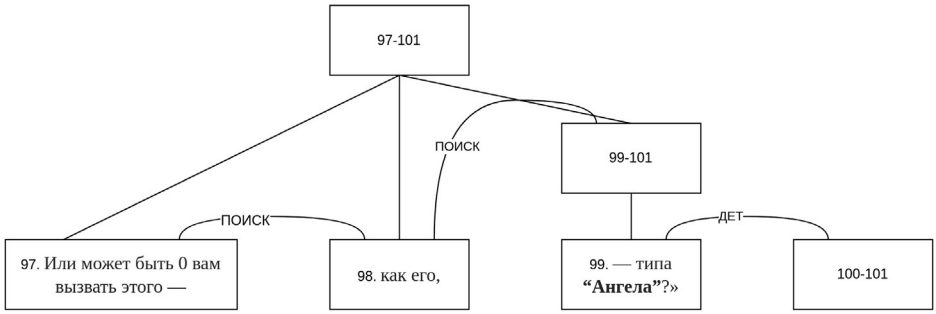

Рис. 4. Фрагмент рассказа № 2sp с применением отношения «Поиск» (соединяет ЭДЕ 97, 98 и группу 99–101)

3. **Обращение**: асимметрично, в сателлите вводит обращение в рамках прямой речи, в качестве ядра присоединяет всю остальную прямую речь.

2.4. Разметка дискурсов для исследования

2.4.1. Общая и сегментная разметка

Каждый дискурс в выборке был скопирован в отдельный docx-файл, размечен и снабжен анкетой, которая отражает состав его референтов и распределение РС. Это позволило лучше контролировать последовательность и качество сбора данных. Пример анкеты показан в табл. 2 (с. 144):

Дополнительная разметка референциальных элементов состоит из следующих действий. Письменные дискурсы разбиты на ЭДЕ вручную в соответствии с методикой, предложенной в [Carlson et al. 2003]. Устные рассказы дополнительно размечены по эпизодам (с опорой на деление их письменных аналогов на абзацы). В каждом тексте отмечены все нулевые упоминания, также жирным шрифтом размечены все нелокуторные упоминания и разными цветами — все РЦ. Внутри РЦ реферемы отмечаются более ярким цветом и дополнительно подчеркиваются, а простые упоминания окрашиваются более бледным оттенком того же цвета.

[10] Такие ЭДЕ также называются плейсхолдерами (placeholders) [Podlesskaya 2010].

[11] Как справедливо отметил рецензент, существует другой вариант риторического анализа таких сегментов — маркер поиска можно считать сателлитом при одном из полнозначных элементов. Аналогичный подход предлагается, например, в [Литвиненко и др. 2009] для фальстартов. Однако в случае с плейсхолдерами, как и при отношении «Вставка», полнозначные элементы присутствуют по обе стороны от маркера речевого сбоя, а значит, возникает вопрос, к какому из них будет более правильно присоединять маркер в качестве сателлита. Симметричное трехместное отношение «Поиск», предлагаемое в данной статье, снимает эту проблему.

Таблица 2

Анкета рассказа № 8wr

Общее число референтов:	6
Из них значимых:	4
На них приходится упоминаний:	23
— в том числе вводных упоминаний:	4
— в том числе простых упоминаний:	5
— в том числе референ:	14
Перечисление значимых референтов (в порядке появления):	мальчик, сухари, актер, билетерши
Среди референ значимых референтов:	
Число полных ИГ [12]:	5
— из них дескрипций с местоимением *этот*:	1
— из них прочих распространенных дескрипций:	0
— из них имен собственных:	0
Число местоимений:	9
Число нулей:	0
Комментарий (опционально)	—

Затем для каждого рассказа в каждом его варианте было вручную построено дерево риторической структуры. Часто в риторических деревьях горизонтально выравниваются промежуточные узлы, ср. пример (7) из [Mann, Thompson 1987: 25]:

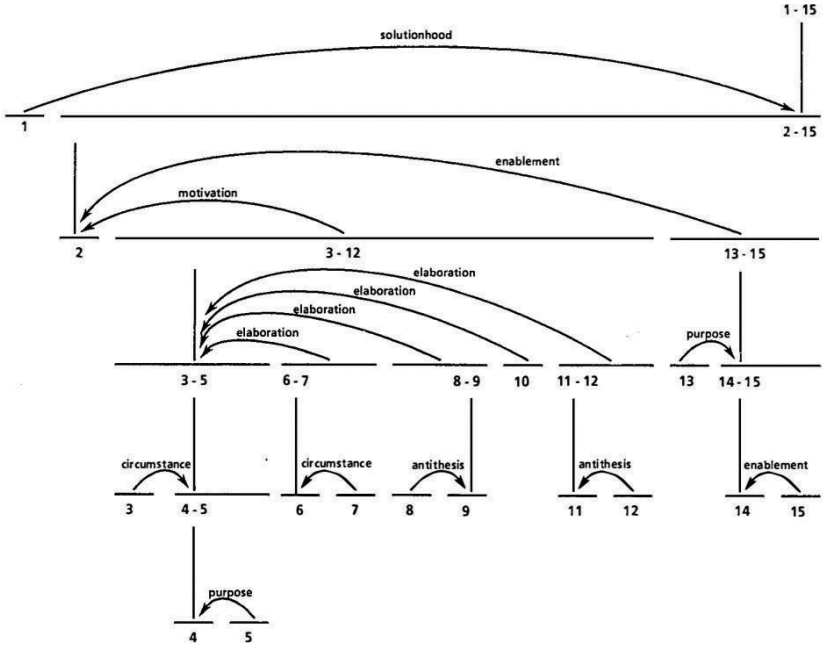

Рис. 5. Классическое дерево риторической структуры

[12] Разновидности полных ИГ подсчитывались для общей картины, итоговые показатели не были использованы для задач данной работы.

В настоящей работе, напротив, выравниванию подлежал самый нижний уровень структуры (см. рис. 2–4), уровень ЭДЕ — так было сделано для удобства чтения текста по диаграмме. Вся разметка упоминаний и референ была перенесена из docx-файлов в риторические деревья для удобства подсчета расстояний.

2.4.2. Факторная разметка

Разметка референ и их антецедентов по факторам выполнялась не в файлах с текстами, а в специальных xlsx-книгах. Данные распределяются в две книги.

1. Каждая референа в дискурсе получает уникальный идентификатор — трехчисловой код: первое число отражает номер дискурса в выборке, второе — номер референта по порядку появления, третье — номер ЭДЕ в референциальной цепочке данного референта. Перечисление дискурсов, референтов и референ, присвоение ID референам и контрольные подсчеты были произведены в **таблицах ключей** (по одной для каждого модуса). Таблицы объединены в **книгу ключей**.

2. Факторная разметка каждой референы производилась в таблице, где первый столбец содержит ID всех референ, далее следуют столбцы для 29 факторов (о них см. в следующем разделе), а также столбец, где отмечается наличие действия фильтров на данную референу. Такая таблица называется **таблицей данных**, для каждого модуса предусмотрена своя таблица с таким содержанием.

Вместе две таблицы составляют **базу данных**.

3. Моделирование

3.1. Факторы

В качестве источников факторов, влияющих на референциальный выбор, привлекались исследования [Чейф 1982; A. A. Kibrik 1996; Strube, Wolters 2002; Arnold 2008; Loukachevitch et al. 2011; van Vliet 2012; Same, van Deemter 2020] и др. Всего отобрано 29 факторов, в том числе пять экспериментальных, значимость которых ранее не проверялась методами машинного обучения. Факторы сгруппированы в факторные блоки по аналогии с [Loukachevitch et al. 2011]:

Факторы референта:

1. Конкретность референта (см. п. 1.1),
2. Одушевленность,
3. Число,
4. Род,
5. Объединенное свойство род-число,
6. Протагонизм 1,
7. Протагонизм 2.

Факторы референы:

8. Грамматическая роль,
9. Подлежащность,
10. Семантическая гиперроль,
11. Порядковый номер в РЦ.

Факторы антецедентов:

Линейный антецедент:

12. РС,
13. Грамматическая роль,
14. Подлежащность,
15. Семантическая гиперроль,
16. Неполная кореферентность с анафором (sloppy identity) [Dahl 1973],

Риторический антецедент:

17. РС,
18. Грамматическая роль,
19. Подлежащность,
20. Семантическая гиперроль,
21. Неполная кореферентность с анафором [Ibid.].

Дистанции:

22. Линейное расстояние в ЭДЕ,
23. Линейное расстояние в референтах — число референт между анафором и антецедентом (все РЦ),
24. Расстояние в эпизодах для устного дискурса, в абзацах — для письменного,
25. Риторическое расстояние,

Факторы клаузы:

26. Наличие параллелизма,
27. Наличие контрастивности,

Референциальная информативность предиката:

28. Отражает ли предикат род референта,
29. Отражает ли предикат число референта.

Остановимся подробнее на некоторых факторах.

Протагонизм-1 и протагонизм-2 (факторы 6 и 7). Алгоритмы подсчета этих мер протагонизма были представлены в [Loukachevitch et al. 2011: 523]: в первом случае протагонизм — это отношение длины данной РЦ к длине самой большой РЦ в дискурсе, во втором случае — отношение длины данной РЦ к сумме длин всех РЦ в дискурсе.

Синтаксическая роль анафора, линейного и риторического антецедентов (факторы 8, 13 и 18). Каждый из этих факторов является категориальным и допускает три значения: «подлежащее», «прямое дополнение» и «прочее». Каждый фактор имеет бинарный аналог (9, 14 и 19 соответственно), в котором подлежащее противопоставлено прочим ролям.

Семантические гиперроли анафора, линейного и риторического антецедентов (факторы 10, 15 и 20). В качестве рабочего набора гиперролей приняты принципал, пациентив и датив по определениям в статье [A. E. Kibrik 1997], а также инструмент по определению в статье [Fillmore 1968: 46].

Линейное расстояние в референтах (фактор 23) отражает количество референт других РЦ между анафором и антецедентом. Так, в примере (1), повторенном в (12), расстояние между упоминанием цветов в ЭДЕ 32 и ЭДЕ 30 равно 2, так как между ними присутствуют два упоминания того человека, который эти цветы посадил (нуль в 31 и «он» в 32):

(12) FS_04-f_sp // № 4sp:
27. моя подружка —
28. Кристина,

29. смотрит,
30. а там у него ***цветы всякие в горшках****,*
31. ну и мы спросили 0,
32. мол это е= ‖ он сам сажает ***0****,*

Этот фактор устроен аналогично фактору «Расстояние в маркабулах» (*distance in markables*) [Loukachevitch et al. 2011: 522], однако понятия референты и маркабулы не тождественны[13].

Риторическое расстояние (фактор 25). Подсчет был произведен в соответствии с системой, предложенной в [A. A. Kibrik, Krasavina 2005]):

1) за каждый горизонтальный переход по дуге между ядром и сателлитом (либо наоборот) начисляется единица;
2) однако за переход между клаузами, тесно связанными синтаксически (*syntactically tightly knit clauses* [Ibid.: 568]), начисляется 0,5 балла; столько же начисляется за переход от основного нарратива к цитированию (отношение «Содержание»);
3) за переход по симметричному отношению между сестринскими узлами начисляется 1 балл;
4) за подъем или спуск по симметричному отношению (переход от родительского узла к дочернему или наоборот) добавляется 0,5 балла.

Перейдем к рассмотрению экспериментальных факторов. Как уже было сказано во введении, все эти факторы представлены в теоретических исследованиях РВ, но все они, за исключением параллелизма, не были ранее проверены при помощи машинных моделей.

Конкретность референта (фактор 1) является обобщением над возможными денотативными статусами. Это бинарный фактор: конкретные статусы (определенный и неопределенный) противопоставлены неконкретным (экзистенциальному, универсальному, атрибутивному и родовому).

Параллелизм (фактор 26) как значимый фактор упоминается, в частности, в статье [Arnold 2008]. В [Strube, Wolters 2002] этот фактор присутствует, однако понимается в узком смысле — предлагается считать референту и ее антецедент параллельными в том случае, если они имеют одинаковые синтаксические роли [Ibid.: 20]. В [Fahnestock 2003: 131] представлен более широкий взгляд на это явление: параллелизм рассматривается на шести уровнях — 1) длительности, 2) интонации, 3) фонетики (аллитерация и ассонанс), 4) грамматики (повтор последовательности синтаксических элементов), 5) семантики (элементы с одинаковой синтаксической ролью относятся к одной «семантической категории»[14]),

[13] Обсудим соотношение понятий «маркабула» и «референта». В [A. A. Kibrik 2011: 472] маркабула определяется как элемент, который следует аннотировать в корпусе, после определения приводятся все классы ИГ и разряды местоимений, которые составляют множество маркабул: «The markables to be annotated include definite, demonstrative, and possessive full NPs (headed by a common noun), proper nouns, personal pronouns, and demonstratives. In addition, other NPs are annotated if they serve as antecedents of anaphors (in particular, indefinite full NPs)». Референта — это ИГ с определенным семантическим свойством (референт не локутор), повторное упоминание некоторого референта, план выражения которого не предопределен (главным образом грамматически).

Все классы ИГ и разряды местоимений, включенные во множество маркабул, могут быть референтами, однако из факторного блока [Ibid.] можно заключить, что в состав маркабул входят вводные упоминания. Кроме того, не каждая референта попадает во множество маркабул — прежде всего это касается нулевых ИГ. Таким образом, множество, которое задается определением маркабулы, пересекается со множеством, которое задается определением референты, но ни одно из них не является подмножеством другого.

[14] Исходя из примеров в данной статье, можно заключить, что автор называет семантическими категориями не семантические роли, а принадлежность слов к одному онтологическому

6) повторения. В настоящей работе клаузы считаются параллельными, если демонстрируют грамматический или семантический признак из данного перечня.

Контрастивность (фактор 27) рассматривается в статье [Чейф 1982]: исходя из эмпирических данных, контрастивные референты чаще бывают выражены полными ИГ.

Референциальная информативность предиката с точки зрения рода и числа (факторы 28 и 29). Понятие референциальной информативности глагола было введено в диссертации [Ефимова 2006] на материале японского языка. Этот фактор сообщает о наличии или отсутствии бенефактивной конструкции, которая состоит из полнозначного глагола и следующего за ним вспомогательного. З. В. Ефимова отмечает, что в японском языке бенефактивный референт часто бывает выражен нулевой ИГ. В таких случаях вспомогательный глагол может выступать единственным показателем присутствия этого референта, поэтому может быть полезно учитывать наличие такой конструкции. С опорой на идею влияния данного фактора в настоящей работе сделана попытка учесть согласование предиката с подлежащим по роду и числу:

(13) № 4wr:
*24. **Он** ответил,*
*25. что **0** купил*
*26. и **0** посадил их сам,*

(14) (Пример, сконструированный на основе (13)):
*24. **Он** ответил,*
*25. что **0** покупает*
*26. и **0** сажает их сам,*

В (13) оба нулевых упоминания являются актантами предиката в прошедшем времени: в каждом случае предикат содержит информацию о роде и о числе референта. В (14) прошедшее время предикатов в ЭДЕ 25 и 26 заменяется на настоящее: такие предикаты несут информацию о числе референта, но не передают сведений о его роде.

3.2. Фильтры

В когнитивной количественной модели **фильтр** — это проверка, которая проводится после подсчета коэффициента активации референта; фильтр имеет условия действия, и при наличии этих условий он переопределяет выбор РС. В [A. A. Kibrik 1996: 280–282] описано два фильтра: референциальный конфликт и смена границ действительности (*world boundary filter*). Оба они предписывают рефереме план выражения в виде полной ИГ вне зависимости от величины коэффициента. Первый фильтр срабатывает, когда использование редуцированного РС может спровоцировать референциальную неоднозначность. Второй фильтр действует, когда в рамках цитирования некоторого дискурса в виде прямой речи (как правило, диалога или размышлений) говорящий впервые упоминает референт, но при этом в рамках всего дискурса данное упоминание не является вводным.

классу. Семантическое сходство в статье понимается достаточно узко, например, оно постулируется в паре предложений:

(i) *During February, ice storms occur frequently, creating havoc with the traffic.*

(ii) *Around August, tornadoes happen occasionally, causing danger on the roadways.*

Из объяснений автора можно сделать вывод, что семантический параллелизм в этих предложениях представлен в виде трех пар: названия месяцев, стихийные бедствия, проблемы на дорогах.

В настоящей работе к ним добавляется третий фильтр — недопустимость нулевой предложной группы[15]. Необходимость его введения обоснована эмпирически: существует несколько классов предложений (особенно распространенных в устной речи), в которых референция происходит при помощи предложной группы и в случае замены этой предложной группы на нуль предложение становится неясным или неграмматичным, как, например, в ЭДЕ 39:

(15) № 11sp:
36. Дождался он шестого числа,
37. 0 пришел,
38. сажают на самолет.
39. Папа со ***своим билетом***[16].

3.3. Данные

Из корпусного материала были сформированы две выборки — устная и письменная, — куда вошли все референты из проанализированных текстов. Подсчеты референтов и РС приведены в табл. 3.

Таблица 3

Статистика референтов и референ по двум выборкам

	Письменная выборка	Устная выборка
Всего референтов	119	154[17]
Значимых референтов	51	58
Упоминаний значимых референтов	337	483
Референ, в том числе:	211	280
— полных ИГ	109 (52 %)	121 (43 %)
— местоимений	75 (35 %)	115 (41 %)
— нулевых ИГ	27 (13 %)	44 (16 %)

Перед построением моделей в среде R[18] был выполнен поиск отсутствующих значений — в обеих таблицах все значения заполнены. Кроме того, проверка методом nearZeroVar показала, что конкретность референта, контрастивность и параллелизм являются незначимыми на материале выборок обоих модусов. Таким образом, три из пяти экспериментальных фактора были исключены из дальнейшего рассмотрения.

При анализе выборок было выявлено, что значение «средний род» факторов «род» и «род-число», а также значение «инструмент» семантических факторов встречаются

[15] Возможно, этот фильтр было бы полезно рассматривать как ограничение референциального выбора, см. [Залманов, Кибрик 2023].

[16] В настоящей статье предлоги не включаются в упоминания, однако такое решение можно назвать вынужденным упрощением. Изучение взаимосвязей между предлогами и референциальными выражениями является темой для дальнейшего исследования.

[17] Разница между выборками в общем количестве референтов составляет 35, в количестве значимых референтов — 7. Таким образом, в устных текстах незначимых референтов на 24 больше, чем в письменных, что примечательно.

[18] Доступна по ссылке https://cran.r-project.org.

крайне редко — модели не смогли бы корректно обучиться на единицах с этими значениями. Реферемы с данными признаками были исключены из выборок.

3.4. Построение и интерпретация моделей

Для моделирования РВ были применены два метода классического машинного обучения — мультиномиальная логистическая регрессия и дерево решений. Оба метода не предъявляют жестких требований к количеству материала и могут работать с малыми выборками. Кроме того, на основе деревьев решений получены диаграммы значимости факторов (Variable Importance in Projection, сокр. VIP) — в них ранжированы по степени важности все факторы, в том числе те, которые не вошли в финальные деревья.

Регрессия построена в статистическом пакете jamovi[19], деревья решений и диаграммы — в среде R.

3.4.1. Логистическая регрессия

Данный метод машинного обучения применяется во многих исследованиях РВ, например в [Loukachevitch et al. 2011; van Vliet 2012; Rosa, Arnold 2017; Same, van Deemter 2020]. В настоящей работе использована его мультиномиальная разновидность, предназначенная для случаев, когда зависимая переменная имеет более двух значений. Ниже приведено краткое описание устройства этой модели; оно может быть полезно для понимания того, как были получены веса факторных значений, а также для объяснения принципов их интерпретации.

Прежде чем описать устройство мультиномиальной модели, рассмотрим ее бинарный прототип. Бинарная логистическая регрессия позволяет определить для каждого целевого объекта, принадлежит ли он к некоторому целевому классу или нет. В ее основе лежит линейная функция вида $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_0$, где n — это количество факторов, x_i — это некоторое факторное значение, каждый β_i из множества $\beta_1 \ldots \beta_n$ — это вес значения x_i в модели, или численное выражение его вклада в общий результат, а β_0 — свободный коэффициент. Веса также принято называть бета-коэффициентами. Входными значениями для логистической модели являются характеристики некоторого объекта; в процессе ее работы рассчитывается вероятность, что данный объект относится к целевому классу. Взвешенная сумма факторных значений может быть любым действительным числом, однако необходимо, чтобы ее результат был отображен в пределах отрезка [0; 1] — для этого она проходит через функцию сигмоиды. Полученная вероятность сравнивается с некоторым порогом, и если она превышает его, то исследуемый объект относится к целевому классу. В состав логистической модели входит много компонентов, однако для задач исследования первостепенное значение имеют веса $\beta_1 \ldots \beta_n$: именно они подвергаются интерпретации.

В качестве примера бинарной логистической регрессии рассмотрим модели, представленные в гл. 6 книги [van Vliet 2012]. Автор моделирует референциальный выбор между местоимением и именем собственным. Так как изначально логистическая модель предсказывает принадлежность/непринадлежность объекта к целевому классу, выбор между двумя РС адаптируется в виде вариантов «имя собственное» и «не имя собственное», причем второй вариант полностью совпадает с классом «местоимение», поскольку другие РС автор в своей модели не учитывает. В настоящей работе класс, не являющийся целевым, предлагается называть базовым; в представленных ниже логистических моделях в качестве базового класса выбрано местоимение, как и в моделях в [Ibid.]. Однако в настоящей

[19] Доступен по ссылке https://www.jamovi.org.

работе у зависимой переменной — референциального средства — не два возможных значения, а три. Это значит, что у нее два не-базовых значения: полная ИГ и нулевая ИГ. Следовательно, в модели отдельно вычисляется вероятность полной ИГ относительно местоимения и вероятность нулевой ИГ относительно местоимения. Таким образом, при интерпретации моделей речь будет идти не о трех значениях совместно, а о двух парах, поскольку веса для каждой пары рассчитываются отдельно. Универсальных правил выбора базового значения не существует; выбор в пользу местоимения сделан для удобства интерпретации.

Логистическая регрессия может принимать факторы нескольких типов: категориальные (например, семантическую гиперроль), количественные дискретные (например, расстояние в ЭДЕ или в абзацах) и количественные непрерывные (к этому типу в моделях относится только протагонизм). Их веса устроены следующим образом:

— у категориального фактора одно из значений выбирается базовым, его вес фиксирован и составляет 0. Остальные значения получают свои веса в процессе обучения модели. Если не-базовых значений больше одного, то каждое получает свой собственный вес;

— количественный фактор имеет один вес, который умножается на его величину; базового значения для нулевой величины не предусмотрено.

В [Molnar 2022] приведены рекомендации, как следует интерпретировать веса логистической регрессии:

— для категориальных факторов:
  — базовое значение не вносит вклад в результат работы модели;
  — не-базовое значение, имеющее вес *n*, повышает вероятность не-базового значения зависимой переменной в *exp(n)* раз;
— для количественных факторов:
  — нулевое значение величины не вносит вклад в результат работы модели;
  — если фактор имеет вес *n* и величину *t*, то вероятность не-базового значения зависимой переменной возрастает в *t*exp(n)* раз;
— свободный коэффициент β_0 не подвергается интерпретации.

Если вес *n* факторного значения отрицательный, то *exp(n)* меньше единицы, и тогда, наоборот, повышается вероятность базового значения зависимой переменной.

Программа jamovi, в которой были получены модели регрессии на материале выборок, рассчитывает не только сами веса, но также и их p-уровни значимости — в первом приближении, степени их недостоверности: это значит, что чем меньше p-уровень, тем достовернее величина. Анализу были подвергнуты только веса с p-уровнем значимости не более 0,1. Важно отметить, что уровень значимости выше 0,1 говорит не об отсутствии влияния фактора на результат, а о том, что на материале выборки невозможно точно вычислить его вес.

Прежде чем перейти к интерпретации весов, рассмотрим процесс выбора базовых значений для категориальных факторов. Как и в случае с зависимой переменной, универсальной процедуры нет, поэтому везде, где это возможно, был сделан обоснованный лингвистически выбор. Например, для фактора одушевленности базовым значением выбрана неодушевленность: таким образом в модель закладывается идея о повышении активации в случае, если референт одушевленный [A. A. Kibrik 2011: 413]. Несколько сложнее было выбрать базовое значение для числа, однако есть основания предполагать, что множественное число усложняет РВ и определенным образом влияет на него ([Ibid.: 554–555]; наблюдения за дистрибуцией РС в зависимости от числа референта также представлены в статье [Zalmanov, A. A. Kibrik 2021]), поэтому в качестве базы было выбрано единственное число.

Достаточно сложно было выбрать базовое значение для грамматической роли: известно, что «подлежащее» повышает активацию, но каково влияние значений «прямое дополнение» и «прочее»? Представляется важным, чтобы под не-базовым факторным значением объединялись однородные опции, которые единообразно влияют на РВ; если категория «прочее» окажется не-базовой, то ее интерпретация будет затруднена в силу разнообразия значений, которые в нее входят. Так, у фактора грамматической роли в состав «прочего» включены дативные дополнения и несогласованные определения; представляется более разумным принять его как базовое. Таким же образом была выбрана категория «прочее» как базовое значение для всех семантических факторов.

Для факторов рода и рода-числа выбрать базовое значение обоснованным образом оказалось невозможно, поэтому они не вошли в регрессии.

Перейдем к анализу весов логистических моделей. В табл. 4 и 5 представлено сравнение вероятностей РС в парах «полная ИГ — местоимение» и «нуль — местоимение» соответственно. Третий столбец таблицы — «Значение» — заполнен только для категориальных факторов, так как у них каждое значение имеет свой собственный вес; модуль коэффициента соответствует силе влияния данного факторного значения на РВ. Также для каждого категориального фактора в скобках приведено его базовое значение, которое не вносит вклад в работу модели. В столбцах «РС» указывается средство, вероятность которого повышается за счет данного фактора или факторного значения. В таблицах размещены только коэффициенты, уровень значимости которых оказался приемлемым в обоих модусах.

Таблица 4

Значимые коэффициенты регрессий в паре «полная ИГ — местоимение»

Группа	Фактор	Значение	Письм. модус		Уст. модус	
			Коэфф.	РС	Коэфф.	РС
Референт	протагонизм-1		−4.64	мест.	−2.28	мест.
	протагонизм-2		15.53	полная ИГ	7.84	полная ИГ
Референта	грамматическая роль (прочее)	подлеж.	−4.8	мест.	1.08	полная ИГ
	подлежащность (нет)	да	−4.8	мест.	1.08	полная ИГ
Лин. ант.	неполная кореферентность (нет)	да	−18.7	мест.	28.81	полная ИГ
Рит. ант.	неполная кореферентность (нет)	да	35	полная ИГ	−27.59	мест.
Дист.	риторическое расстояние		0.83	полная ИГ	0.74	полная ИГ
Клауза	отражает ли предикат число референта (нет)	да	−3.38	мест.	−1.31	мест.

Таблица 5

Значимые коэффициенты регрессий в паре «нуль — местоимение»

Группа	Фактор	Значение	Письм. модус		Уст. модус	
			Коэфф.	РС	Коэфф.	РС
Референт	одушевленность (неодуш.)	одуш.	−2.59	мест.	−3.27	мест.
Референта	грамматическая роль (прочее)	подлеж.	16.3	нуль	−10.35	мест.
	подлежащность (нет)	да	16.3	нуль	−10.35	мест.
	семантическая гиперроль (прочее)	принципал	−15.75	мест.	23.47	нуль
Лин. ант.	неполная кореферентность (нет)	да	−4.99	мест.	−16.8	мест.
Рит. ант.	неполная кореферентность (нет)	да	−3.9	мест.	−17.54	мест.
Дист.	линейное расстояние в абзацах/эпизодах		−13.6	мест.	−27.15	мест.

В паре «полная ИГ — местоимение» в обоих модусах вероятность полной ИГ повышают протагонизм-2 и риторическое расстояние, вероятность местоимения повышают протагонизм-1 и референциальная информативность предиката. Разнонаправленное влияние демонстрируют четыре факторных значения, в том числе неполная кореферентность с линейным и риторическим антецедентами. Это может быть связано с малой представленностью обоих факторов в выборках. Интерес также представляет различие во влиянии двух вариантов протагонизма — это явление нуждается в дальнейшем исследовании.

В паре «нуль — местоимение» одушевленность, неполная кореферентность и линейное расстояние в абзацах/эпизодах повышают вероятность местоимения, три других факторных значения действуют на модусы различным образом. Примечательно, что гиперроль принципал у референты повышает вероятность местоимения в письменном дискурсе и нуля — в устном.

Таким образом, при анализе логистических моделей обнаруживаются некоторые расхождения по силе и направлению влияния факторов на два модуса. Проведенное сопоставление позволяет говорить об истинности второй гипотезы: один и тот же значимый фактор может влиять на два модуса с разной силой.

3.4.2. Дерево решений и диаграмма VIP

Данный метод выявляет наборы факторных значений, наиболее характерные для каждого РС, разбивая всю выборку на подмножества — узлы. Начальный узел, в состав которого входит вся выборка, называется корневым, конечные узлы, которые не подвергаются дальнейшему делению, — листьями.

Деревья построены в среде R при помощи пакетов rpart и rpart.plot. Они показаны на рис. 6 и 7 (с. 154).

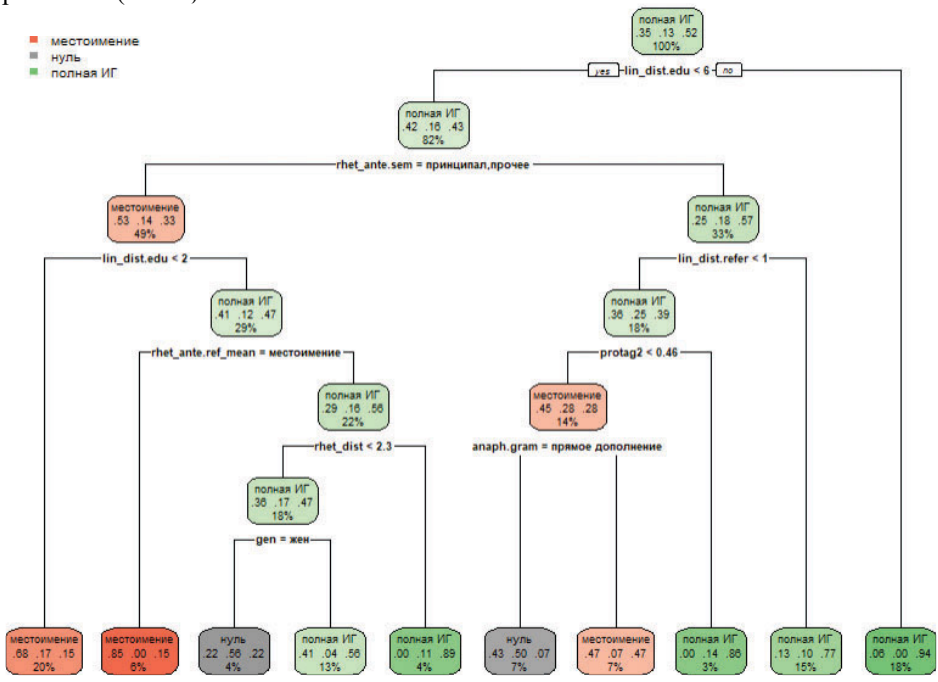

Рис. 6. Дерево решений для письменного модуса[20]

[20] Сокращения в алфавитном порядке: **anaph.gram** — грамматическая роль анафора, **gen** — род референта, **lin_dist.edu** — линейное расстояние в ЭДЕ, **lin_dist.refer** — линейное расстояние

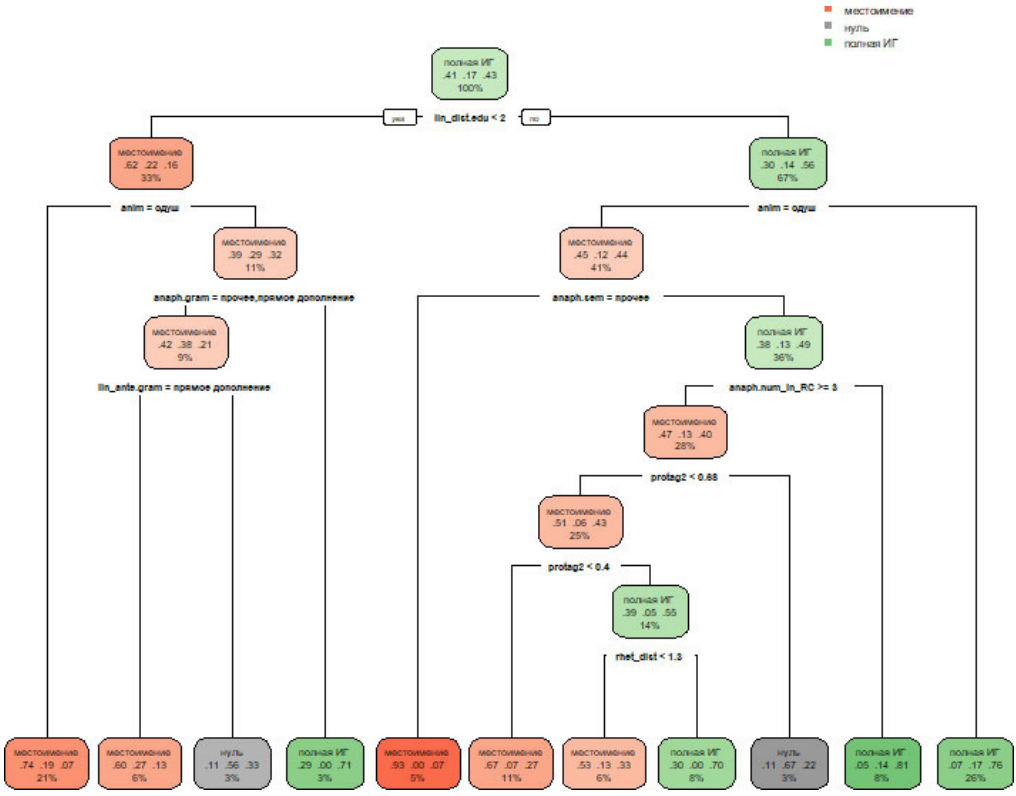

Рис. 7. Дерево решений для устного модуса[21]

В состав дерева входят:

— Узлы. Каждый узел дерева — это подмножество выборки с определенными свойствами. Для каждого узла имеется четыре показателя:

1) название, которое присваивается по самому частотному РС в данном подмножестве;
2) цвет, который означает РС, преобладающие в данном подмножестве: подмножествам, в которых большинство референт выражено полными ИГ, соответствуют зеленые листья, местоимениями — оранжевые, нулевыми ИГ — серые. Чем насыщеннее цвет, тем сильнее перевес в сторону самого частотного РС;
3) распределение референт в данном подмножестве между местоимениями, нулями и полными ИГ (процентные доли указаны именно в таком порядке);
4) процент, который данное подмножество составляет от всей выборки. Например, первый слева нижний узел на рис. 6 содержит 20 % всей выборки. 68 % его референт являются местоимениями, 17 % — нулями, 15 % — полными ИГ. Он озаглавлен как «местоимение».

в референтах, **protag2** — протагонизм-2, **rhet_ante.sem** — семантическая гиперроль риторического антецедента, **rhet_ante.ref_mean** — РС риторического антецедента, **rhet_dist** — риторическое расстояние.

[21] Сокращения, которых нет на рис. 6, в алфавитном порядке: **anaph.num_in_RC** — порядковый номер анафора в РЦ, **anaph.sem** — семантическая гиперроль анафора, **anim** — одушевленность, **lin_ante.gram** — грамматическая роль линейного антецедента.

— Условия деления (или ветвления). Каждое дерево содержит оптимальное количество и состав условий, который был выведен алгоритмом. Каждое деление бинарно.

— «Ветки», или переходы. Для корневого узла отмечено, что переход влево осуществляется при соблюдении условия деления, переход вправо — при несоблюдении. Переходы при остальных делениях устроены так же.

Первое ветвление обоих деревьев происходит по линейному расстоянию, что говорит о его высокой значимости. Также в обоих деревьях присутствуют деления по риторическому расстоянию, протагонизму-2, грамматической роли риторического антецедента. Дерево для устной выборки содержит больше семантических факторов (одушевленность и семантическая роль референты), факторный набор для письменной выборки более разнообразен.

На рис. 8 и 9 (с. 156) показаны VIP-диаграммы, построенные на базе деревьев решений.

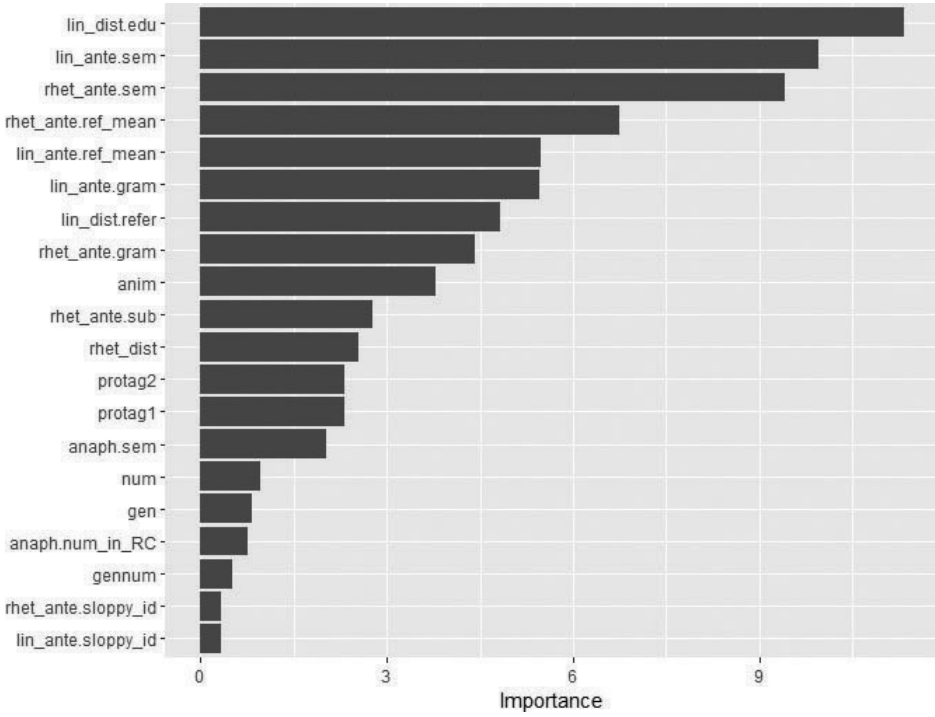

Рис. 8. VIP для письменного модуса

Для обеих моделей риторическое расстояние (обозначено как rhet_dist) является значимым. Однако на графике для письменной выборки оно занимает 7 место и уступает некоторым свойствам антецедентов; на графике для устной выборки оно занимает 8 место и уступает достаточно разнородному набору факторов, в том числе всем трем семантическим свойствам (референты и обоих антецедентов). Как представляется, имеет смысл обращать больше внимания на позицию фактора в рейтинге, чем на его величину VIP, так как эта величина, по-видимому, является безразмерной, что затрудняет ее интерпретацию. Интересно отметить, что оба экспериментальных фактора, оставшихся в выборках после фильтрации — референциальная информативность предиката по роду и по числу — фигурируют в диаграмме устного модуса и демонстрируют незначительное влияние, в то время как в письменном модусе референциальная информативность предиката по роду (pred:shows_gen) отсутствует как значимый фактор РВ.

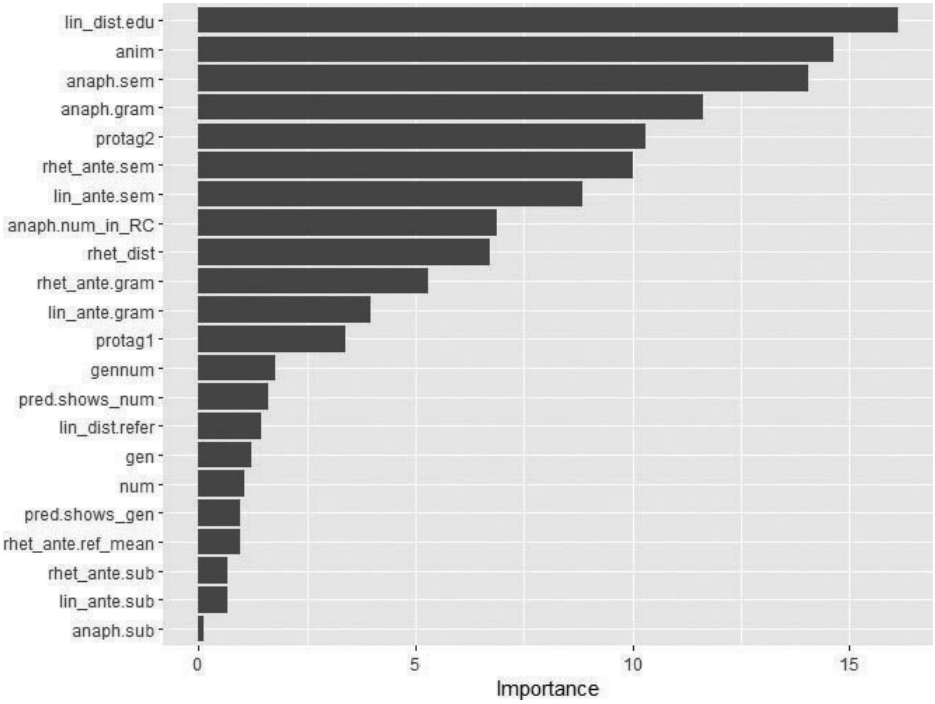

Рис. 9. VIP для устного модуса

4. Обсуждение результатов

Проверка гипотез дала следующие результаты:

Гипотеза 1: наборы факторов, значимых для РВ в двух модусах, различаются (показано с помощью деревьев решений).

Гипотеза 2: для некоторых факторов, значимых в обоих модусах, существует количественное отличие в степени влияния на каждый из модусов (наблюдается в моделях регрессии и VIP).

Гипотеза 4 подтверждена частично: модели позволяют говорить о значимости только одного экспериментального фактора — референциальной информативности предиката с точки зрения числа.

Что касается гипотезы 3 об относительной значимости линейного и риторического расстояния, примененные методы не позволяют делать однозначных выводов. Анализ бета-коэффициентов логистической регрессии не позволил оценить влияние линейного расстояния. Деревья решений демонстрируют, что для обеих выборок первым по приоритетности из всех факторов является линейное расстояние в ЭДЕ, в то время как риторическое расстояние уступает по степени влияния многим факторам. Вероятно, такой противоречивый результат был получен из-за невыясненного до конца характера связи этих факторов с остальными. Так, в когнитивной количественной модели для письменных нарративов риторическое расстояние учитывается дважды: как самостоятельный фактор и в составе фактора «одушевленность». Возможно, похожим образом эти факторы-дистанции ведут себя и в устном модусе, однако раздельного применения регрессии и деревьев решений оказалось недостаточно, чтобы выявить такое влияние.

Заключение

Целью настоящей работы было изучение сходств и различий референциального выбора в устных и письменных нарративах на русском языке. Для исследования были привлечены рассказы из специального корпуса, где каждый сюжет содержится в двух изложениях — устном и письменном. Было взято 11 сюжетов, то есть 22 рассказа; все рассказы были размечены согласно принципам теории риторической структуры. В рассказах были найдены референты: собрано около 280 единиц из устных текстов и около 210 из письменных. Каждая референта получила характеристику по 29 факторам; результирующие базы данных были переданы машинным алгоритмам, отдельно для устной и письменной выборок были построены модели логистической регрессии и деревья решений. Интерпретация моделей показала, что наборы значимых факторов в устных и письменных нарративах различаются; кроме того, факторы, имеющие большую значимость в одном модусе, могут оказывать малый эффект на РВ во втором модусе. Ряд заключений на основе построенных моделей нуждается в рассмотрении на большем массиве данных.

Референциальный выбор продолжает оставаться актуальной темой не только в прикладном, но и в когнитивном ключе. Более глубокое понимание различий между механизмами РВ в устных и письменных дискурсах можно назвать промежуточным шагом к сопоставительному исследованию когнитивных процессов при их порождении. В качестве перспектив данной линии исследований можно назвать уточнение границы между референтами и простыми упоминаниями, а также более детальный учет различных видов цитирования.

СПИСОК СОКРАЩЕНИЙ

ВИЖ — корпус «Веселые истории из жизни»
РВ — референциальный выбор
РС — референциальное средство
РЦ — референциальная цепочка
ЭДЕ — элементарная дискурсивная единица
VIP — Variable Importance in Projection

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Буденная 2018 — Буденная Е. В. *Эволюция субъектной референции в языках балтийского ареала*. Дис. … канд. филол. наук. М.: ИЯз РАН, 2018. [Budennaya E. V. *Evolyutsiya sub”ektnoi referentsii v yazykakh baltiiskogo areala* [Evolution of subject reference in languages of Baltic sprachbund]. Candidate diss. Moscow: Institute of Linguistics of the Russian Academy of Sciences, 2018.]

Ефимова 2006 — Ефимова З. В. *Референциальная структура нарратива в японском языке (в сопоставлении с русским)*. Дис. … канд. филол. наук. М.: РГГУ, 2006. [Efimova Z. V. *Referentsial’naya struktura narrativa v yaponskom yazyke (v sopostavlenii s russkim)* [Referential structure of Japanese narrative (in comparison with Russian)]. Candidate diss. Moscow: Russian State Univ. for the Humanities, 2006.]

Залманов, А. А. Кибрик 2023 — Залманов Д. А., Кибрик А. А. Уровень активации и специальные ограничения при математическом моделировании референциального выбора. *Язык и искусственный интеллект: Сб. ст. по итогам конф. «Лингвистический форум 2020: Язык и искусственный интеллект», Москва, 12–14 ноября 2020 г.* Вдовиченко А. В. (ред.). М.: Языки славянских культур, 2023, 142–166. [Zalmanov D. A., Kibrik A. A. Activation level and special constraints in referential choice computational modeling. *Yazyk i iskusstvennyi intellect. Coll. of papers from the conf. “Lingvisticheskii forum 2020: Yazyk i iskusstvennyi intellekt”, Moscow, November 12–14, 2020*. Vdovichenko A. V. (ed.). Moscow: Yazyki slavyanskikh kul’tur, 2023, 142–166.]

А. А. Кибрик 1997 — Кибрик А. А. Моделирование многофакторного процесса: выбор референциального средства в русском дискурсе. *Вестник Московского университета. Сер. 9: Филология*, 1997, 4: 94–105. [Kibrik A. A. Multifactorial process modeling: Choosing referential devices in Russian discourse. *Lomonosov Philology Journal*, 1997, 4: 94–105.]

А. А. Кибрик и др. 2009а — Кибрик А. А., Подлесская В. И., Коротаев Н. А. Структура устного дискурса: основные элементы и канонические явления. *Рассказы о сновидениях: корпусное исследование устного русского дискурса*. Кибрик А. А., Подлесская В. И. (ред.). М.: Языки славянских культур, 2009, 55–101. [Kibrik A. A., Podlesskaya V. I., Korotaev N. A. Spoken discourse structure: Main elements and canonical phenomena. *Rasskazy o snovideniyakh: korpusnoe issledovanie ustnogo russkogo diskursa*. Kibrik A. A., Podlesskaya V. I. (eds.). Moscow: Yazyki slavyanskikh kul'tur, 2009, 55–101.]

А. А. Кибрик и др. 2009б — Кибрик А. А., Подлесская В. И., Коротаев Н. А. Неканонические явления. *Рассказы о сновидениях: корпусное исследование устного русского дискурса*. Кибрик А. А., Подлесская В. И. (ред.). М.: Языки славянских культур, 2009, 102–176. [Kibrik A. A., Podlesskaya V. I., Korotaev N. A. Non-canonical phenomena. *Rasskazy o snovideniyakh: korpusnoe issledovanie ustnogo russkogo diskursa*. Kibrik A. A., Podlesskaya V. I. (eds.). Moscow: Yazyki slavyanskikh kul'tur, 2009, 102–176.]

Литвиненко и др. 2009 — Литвиненко А. О., Подлесская В. И., Кибрик А. А. Анализ рассказов о сновидениях с точки зрения иерархической структуры дискурса. *Рассказы о сновидениях: корпусное исследование устного русского дискурса*. Кибрик А. А., Подлесская В. И. (ред.). М.: Языки славянских культур, 2009, 431–463. [Litvinenko A. O., Podlesskaya V. I., Kibrik A. A. Dream stories analysis from the point of hierarchical discourse structure. *Rasskazy o snovideniyakh: korpusnoe issledovanie ustnogo russkogo diskursa*. Kibrik A. A., Podlesskaya V. I. (eds.). Moscow: Yazyki slavyanskikh kul'tur, 2009, 431–463.]

Чейф 1982 — Чейф У. Данное, контрастивность, определенность, подлежащее, топики и точка зрения. *Новое в зарубежной лингвистике*. Кибрик А. Е. (ред.). М.: Прогресс, 1982, 277–316. [Chafe W. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Novoe v zarubezhnoi lingvistike*. Kibrik A. E. (ed.). Moscow: Progress, 1982, 277–316.]

Ariel 1990 — Ariel M. *Accessing noun-phrase antecedents.* London: Routledge, 1990.

Arnold 2008 — Arnold J. E. Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, 2008, 23: 495–527.

Carlson et al. 2003 — 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*. van Kuppevelt J., Smith R. (eds.). Dordrecht: Kluwer, 2003, 85–112.

Chafe 1994 — Chafe W. *Discourse, consciousness, and time*. Chicago: Univ. of Chicago Press, 1994.

Dahl 1973 — Dahl Ö. On so-called "sloppy identity". *Synthese*, 1973, 26: 81–112.

Fahnestock 2003 — Fahnestock J. Verbal and visual parallelism. *Written Communication*, 2003, 20(2): 123–152.

Fillmore 1968 — Fillmore C. The case for case. *Universals of linguistic theory*. Bach E., Harms R. T. (eds.). New York: Holt, Rinehart & Winston, 1968, 1–88.

Fox 1995 — Fox B. *Discourse structure and anaphora: Written and conversational English*. Cambridge: Cambridge Univ. Press, 1995.

Grosz et al. 1986 — Grosz B., Weinstein S., Joshi A. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 1986, 12(3): 175–204.

Gundel et al. 1993 — Gundel J. K., Hedberg N., Zacharski R. Cognitive status and the form of referring expressions in discourse. *Language*, 1993, 69: 274–307.

A. A. Kibrik 1996 — Kibrik A. A. Anaphora in Russian narrative prose: A cognitive calculative account. *Studies in anaphora*. Fox B. (ed.). Amsterdam: John Benjamins, 1996, 255–303.

A. E. Kibrik 1997 — Kibrik A. E. Beyond subject and object: Toward a comprehensive relational typology. *Linguistic Typology*, 1997, 1(3): 279–346.

A. A. Kibrik 2011 — Kibrik A. A. *Reference in discourse*. Oxford: Oxford Univ. Press, 2011.

A. A. Kibrik, Krasavina 2005 — Kibrik A. A., Krasavina O. N. A corpus study of referential choice: The role of rhetorical structure. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, 2005, 4: 561–569.

Loukachevitch et al. 2011 — Loukachevitch N. V., Dobrov G. B., Kibrik A. A., Khudyakova M. V., Linnik A. S. Factors of referential choice: computational modeling. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conf. "Dialogue"*, 2011, 10: 518–528.

Mann, Thompson 1987 — Mann W. C., Thompson S. A. *Rhetorical structure theory: A theory of text organization*. Los Angeles: Univ. of Southern California, 1987.

Molnar 2022 — Molnar C. *Interpretable machine learning. A guide for making black box models explainable*. 2022. https://christophm.github.io/interpretable-ml-book/.

Podlesskaya 2010 — Podlesskaya V. I. Parameters for typological variation of placeholders. *Fillers, pauses and placeholders*. Amiridze N., Davis B., Maclagan M. (eds.). Amsterdam: John Benjamins, 2010, 11–32.

Rosa, Arnold 2017 — Rosa E. C., Arnold J. E. Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language*, 2017, 94: 43–60.

Same, van Deemter 2020 — Same F., van Deemter K. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. *Proc. of the 28th International Conference on Computational Linguistics*. Scott D., Bel N., Zong C. Barcelona: International Committee on Computational Linguistics, 2020, 4575–4586.

Strube, Wolters 2002 — Strube M., Wolters M. A Probabilistic genre-independent model of pronominalization. *Proc. of the 1st North American Chapter of the Association for Computational Linguistics Conference*. Stroudsburg: Association for Computational Linguistics, 2002, 18–25.

van Vliet 2012 — van Vliet S. *Proper nouns and pronouns: The production of referential expressions in narrative discourse*. Utrecht: LOT, 2012.

Zalmanov, Kibrik 2021 — Zalmanov D. A., Kibrik A. A. Short definite descriptions and referent activation. *Advances in cognitive research, artificial intelligence and neuroinformatics.* Velichkovsky B. M., Balaban P. M., Ushakov V. L. (eds.). Cham: Springer, 2021, 284–292.

Получено / received 26.01.2023 *Принято / accepted 24.09.2024*