DOI: 10.37892/978-5-6049527-1-9-11

МОДЕЛЬ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА В УСТНОМ РУССКОМ ДИСКУРСЕ, ОСНОВАННАЯ НА МУЛЬТИНОМИАЛЬНОЙ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Марина Андреевна Шумилина
МГУ имени М. В. Ломоносова
(студент)

Аннотация. Исследование построено на корпусном материале. В статье описаны разработка и апробация модели референциального выбора в устных рассказах на русском языке. Представленная модель является опытом применения когнитивного количественного подхода ([Kibrik 1996, Кибрик 1997]) к новому материалу. В ходе разработки был получен список факторов, обладающих двойной значимостью (теоретической и математической), и были выведены веса факторных значений, также являющиеся теоретически и математически значимыми.

В статье изложен алгоритм апробации модели. При помощи модели из материала контрольной выборки были получены прогнозы референциального выбора. Была произведена оценка их точности: было выяснено, что 94% прогнозов являются верными.

Ключевые слова: русский язык, дискурсивный анализ, референциальный выбор, мультиномиальная логистическая регрессия, модель, референт, корпус, риторическая структура.

A REFERENTIAL CHOICE MODEL FOR SPOKEN NARRATIVES IN RUSSIAN BASED ON MULTINOMIAL LOGISTIC REGRESSION

Marina Shumilina
Lomonosov Moscow State University
(student)

Abstract. The research is built on corpus material. In the article, the development and approbation of a referential choice model for Russian spoken narratives is described. The presented model is an application of the Cognitive Calculative approach ([Kibrik 1996, Kibrik 1997]) to the new dataset. During the development, a set of activation factors, which possess both theoretical and mathematical significance, was obtained. The factorial weights, both theoretically and mathematically significant, were also calculated.

The article contains the model approbation algorithm. The forecasts were received by means of the model applied to the control sample. The accuracy of the forecasts was estimated: it was found that 94% of them were correct.

Key words: Russian language, discourse analysis, referential choice, multinomial logistic regression, model, referent, corpus, rhetorical structure.

1. Проблема исследования

Предметом настоящего исследования является референциальный выбор (далее РВ) — процесс, в ходе которого говорящий выбирает план выражения для некоторого заранее определенного им объекта действительности. Другими словами, говорящий принимает решение, **каким образом** упомянуть целевого референта — при помощи полной именной

группы, местоимения либо референциального нуля. Референциальные средства (далее РС) всех трех типов представлены в (1).

(1) Студент увидел книгу, взял ее и положил Ø на стол.

Варьирование РС — неотъемлемая черта естественной речи. Оно позволяет поддерживать связность дискурса: «Центральным способом выражения отношений грамматической когезии <...> является местоименная референция» [Милевская 2003: 6]. Современные автоматические диалоговые системы зачастую синтезируют **недостаточно связный** дискурс. Необходима обученная модель, которая позволит машинному алгоритму принимать правильные решения об употреблении нулей и местоимений; одно из возможных решений данной **проблемы** предлагается в настоящей статье.

Разработанная в ходе исследования модель, помимо прикладных возможностей, также обладает экспликаторным свойством и стремится объяснить устройство РВ в устных рассказах на русском языке. В этих двух аспектах — теоретическом и прикладном — состоит **актуальность** проведенной работы.

Большинство современных подходов к моделированию РВ базируется на активированности референта в сознании говорящего. Впервые мысль о прочной взаимосвязи между РВ и статусом «данное/новое» была высказана У. Чейфом в статье [Чейф 1982: 282-283]: автор отмечает, что концепты со статусом «данное» «могут подвергаться прономинализации», в отличие от концептов со статусом «новое». Активацию референта, как правило, невозможно измерить напрямую, однако ее оценку можно построить на ряде наблюдаемых факторов. Первый вариант набора значимых факторов предложил А. А. Кибрик в рамках когнитивной количественной модели РВ в статье [Kibrik 1996]: 7 значимых факторов были выявлены на материале письменных рассказов на русском языке. Многофакторный подход к анализу РВ позволил перевести дальнейшие исследования в плоскость машинного моделирования (исследования [Loukachevitch et al. 2011, Khudyakova et al. 2011, Кибрик и др. 2012]).

Модель, разработанная в настоящем исследовании, является **опытом** применения когнитивной количественной концепции к новому материалу. В ее основу положены выводы, полученные в ходе исследований в русле теоретической лингвистики. В то же время, для построения модели были привлечены методы математической статистики. Другими словами, она обладает лингвистической (= теоретической) и статистической (= математической) значимостью. В этом состоит **новизна** полученной модели.

Цель исследования состояла в разработке модели РВ и последующей ее апробации. Для достижения цели были решены следующие **задачи**:

- Был собран и обработан материал для двух выборок, тренировочной и контрольной.
- Были подобраны теоретически и математически значимые факторы.
- С помощью статистических инструментов из тренировочной выборки были получены веса факторных значений (далее ВФЗ).
- Модель, выполненная в когнитивной количественной парадигме, была апробирована на контрольной выборке.

В работе было использовано несколько **гипотез**:

1. Если референт в данной точке может быть выражен полной ИГ и нулём, то он непременно может быть выражен и местоимением.
2. Как и в письменном модусе, в устных рассказах факторы «грамматическая роль риторического антецедента» и «семантическая роль риторического антецедента» являются значимыми.
3. Помимо уже открытых факторов, значимыми являются фактор «референт / образ[31]» и фактор «совпадение линейного и риторического антецедентов».

[31] Повторные упоминания могут происходить даже тогда, когда упомянутый говорящим концепт не обладает референтом. Например, в предложении «если у птиц киви нет крыльев, то они не птицы» дескрипция «птицы киви» нереферентна, однако в последующей клаузе происходит повторное обращение к

В рамках исследования вводится новый термин — «единица референциального выбора» (далее ЕРВ)[32]. Это речевая единица с фиксированным планом содержания (означает целевого референта) и параметрическим планом выражения, который детерминируется в процессе референциального выбора.

2. Материал исследования

2.1. Источник данных

Исследование проводилось на материале рассказов из устного корпуса «Веселые истории из жизни» [эл. ресурс]. Ранее этот корпус уже был использован для анализа РВ в современном русском языке в [Будённая 2018].

Каждый рассказ в корпусе размещен в двух представлениях, письменном и устном (в трех вариантах подробности).

Для настоящего исследования были привлечены дискурсы в устном представлении в минимальной транскрипции: в них размечены только разделение вербального компонента на ЭДЕ (элементарные дискурсивные единицы, понятие подробно обсуждается в [Кибрик и др. 2009]), хезитации и неясно произнесенные фрагменты. Для подсчета линейных и риторических расстояний было необходимо сохранить деление на ЭДЕ.

Из всего объема корпуса (40 рассказов) было отобрано 12 дискурсов, это количество детерминировано целевым объемом выборок. Отбор происходил по порядку следования дискурсов в корпусе. Для рассмотрения отбирались все дискурсы, в которых содержалось не менее девяти ЕРВ.

2.2. Разметка данных

В каждом из отобранных дискурсов были размечены не только ЕРВ, но также и первые упоминания значимых[33] референтов (т.е. референтов, упомянутых более одного раза). Каждая ЕРВ была охарактеризована по ряду факторов, результаты разметки фиксировались в базе данных. База данных для тренировочных дискурсов состоит из 39 столбцов, которые разделены на 4 группы:

- Группа данных A: нумерация дискурса и референта. Содержит 4 столбца: код дискурса в корпусе, номер дискурса в выборке, имя референта совместно с его порядковым номером (по первому упоминанию), а также номер ЭДЕ, в которой референт был впервые упомянут.
- Группа данных B: общие сведения о ЕРВ. Содержит 2 столбца: код ЕРВ в формате «номер дискурса – номер референта – номер ЭДЕ» и план выражения ЕРВ.
- Группа данных C делится на 4 подгруппы в соответствии с классификацией факторов, приведенной в [Loukachevitch et al. 2011: 522]: свойства референта (подгруппа C-1), дистанции (подгруппа C-2), свойства анафора (подгруппа C-3), свойства антецедентов (подгруппы C-4а и C-4б).
- Группа данных D: контрольные столбцы, по которым производится учет ЕРВ. Содержит 4 столбца: подсчет ЕРВ по референтам, подсчет ЕРВ по дискурсам, подсчет всех упоминаний по дискурсам и столбец для подтверждения, что все подсчеты по дискурсам совпадают с соответствующими числами в анкетах.

Контрольная база отличается от тренировочной меньшим числом факторов, поскольку ее разметка производится только по математически значимым параметрам. Деление столбцов на группы идентично тренировочной базе.

тому же концепту. Семантическое наполнение дескрипции, замещающее референт — денотат / сигнификат / экстенсионал соответствующей лексемы — предлагается называть «образом».

[32] Этот термин является более удачной заменой слову «маркабула» (встречено в [Loukachevitch 2011]), которое можно было встретить в докладе.

[33] Термин «значимый референт» встречен в [Кибрик 1997: 96].

Для измерения риторических расстояний потребовалось построить деревья риторической структуры для каждого дискурса. Это было сделано в соответствии с оригинальной РТС У. Манна и С. Томпсон [Mann, Thompson 1987]; для адаптации этой теории под устный модус дискурса был применен опыт проекта «Рассказы о сновидениях» [Литвиненко и др. 2009].

С полной таблицей, а также с размеченными дискурсами и их риторическими деревьями можно ознакомиться в облачном хранилище по адресу https://drive.google.com/drive/folders/1eLHvqutqOMPhHUV8qsY76Q69PqR2TbhL?usp=sharing

3. Методология исследования

3.1. Теоретическое моделирование

С лингвистической точки зрения, представляемая модель выполнена в русле когнитивного количественного подхода, автором которого является А. А. Кибрик. Название теории хорошо соотносится с одним из ее ключевых исходных допущений: «референциальный выбор не может быть понят в рамках внутритекстового подхода; он непосредственно предопределяется **когнитивным состоянием говорящего** в данный момент» [Кибрик 1997: 94]. Основная теоретическая предпосылка модели — зависимость РВ от активации референта в рабочей памяти говорящего. Модель базируется на следующем принципе: необходимо получить количественную оценку активации, а затем конвертировать ее в доступные референциальные средства. При этом референциальный выбор не является категориальным, т. е. план выражения большинства ЕРВ может варьироваться.

Ниже представлена оригинальная схема модели (рис. 1). Для удобства различения будем называть **компонентным** деление модели на значимые концепты (представлено на схеме), а **блочным** — деление, при котором каждой части модели соответствует одна значимая процедура (приводится на рис. 2).

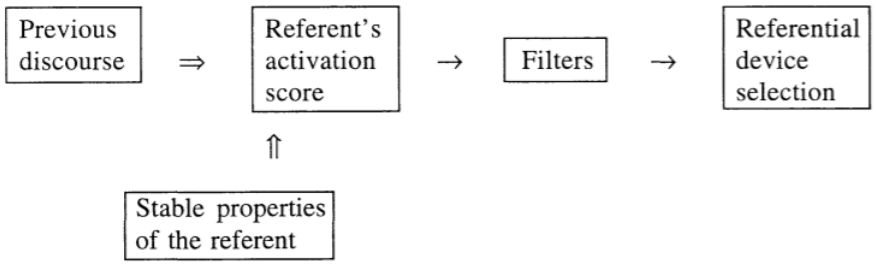

Рисунок 1. Оригинальная схема когнитивной количественной модели ([Kibrik 1996: 258])

Каждый элемент данной схемы является отдельным компонентом, важным для общего смысла модели. Автор парадигмы предлагает учитывать контекст (*previous discourse*) и внутренние свойства референта (*stable properties of the referent*), чтобы оценить активированность референта в определенный момент времени *t*. Совместно эти два типа свойств называются факторами. У каждого фактора есть набор значений, у каждого значения имеется числовой вес. В качестве примера в табл. 1 показано устройство фактора «расстояние до антецедента в абзацах» ([Кибрик 1997: 99]):

Таблица 1

Структура фактора в факторном блоке

Расстояние до антецедента в абзацах	
Значение:	Вес значения:
0	0
1	-0.2
2	-0.4

Следующий компонент модели — коэффициент активации референта (*referent's activation score*) (далее КА). Чтобы его получить, необходимо взять все факторные значения ЕРВ и суммировать их веса. КА, как правило, варьируется в пределах отрезка [0, 1]; чем он больше, тем более редуцированные РС можно использовать для данного ЕРВ. Прогноз дополняется оценкой приемлемости с точки зрения контекста, которая производится с помощью фильтров (*filters*) «референциальный конфликт» и «смена границ действительности». При срабатывании оба фильтра требуют, чтобы говорящий употребил полную ИГ вне зависимости от активации референта.

Результирующий компонент модели — выбор референциального средства (*referential device choice*). Таким образом, РВ может быть осуществлен за 4 концептуальных шага.

На рис. 2 представлено процедурное деление модели, содержащее 3 процедурных блока.

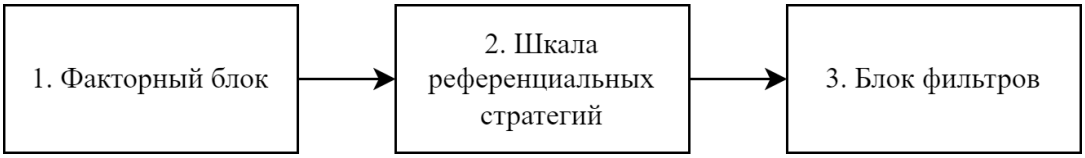

Рисунок 2. Процедурная схема когнитивной количественной модели

Блок факторов необходим для получения КА: он содержит факторы, их значения и веса этих значений.

Шкала референциальных стратегий (рис. 3) позволяет интерпретировать КА, полученный из первого блока. Она представляет собой таблицу из двух строк: в нижней строке записаны диапазоны значений КА, в верхней — средства, которые допустимы при этих диапазонах:

	Только полная ИГ	Полная ИГ наиболее вероятна, местоимение или ноль сомнительны	Полная ИГ или местоимение/ноль	Только местоимение или ноль
КА:	0 0.1 0.2 0.3	0.4 0.5 0.6	0.7 0.8 0.9	1

Рисунок 3. Шкала референциальных стратегий в оригинальной модели ([Кибрик 1997: 101])

Третий, заключительный блок модели содержит фильтры.

3.2. Статистические методы

При разработке модели для устных нарративов были использованы следующие математические методы:

- Для оценки статистической значимости факторов был использован **критерий хи-квадрат**.
- Однако критерий хи-квадрат не дает возможности сравнивать влияние различных факторов, поэтому для каждого фактора был получен его **коэффициент сопряженности** (*contingency coefficient*). Это позволило построить иерархию факторов.
- Оценки весов факторных значений были получены при помощи **мультиномиальной логистической регрессии** (МЛР). Логистическая регрессия ранее применялась в [Strube, Walters 2002] и показала эффективность 93.1 % [ibid.: 93]. Кроме того, в статье [Khudyakova et al. 2011] отмечается достаточно высокая эффективность логистической регрессии: она демонстрирует корректность 87% при двухвариантном РВ и 77.4% при трехвариантном.
- Модель должна включать факторы, разнообразные по своему способу влияния. С точки зрения математики, это означает, что никакие факторы не должны линейно зависеть друг от друга. Если такая зависимость присутствует, это означает, что два фактора влияют на РВ схожим образом. Такие пары факторов называются коллинеарными. Чтобы сохранить сбалансированность модели, необходимо отследить и ликвидировать коллинеарность факторов. Эту задачу позволяет решить **корреляционная матрица**.

4. Разработка модели

4.1. Исходные допущения

Общие исходные допущения, на которые опирается новая модель, практически полностью совпадают с общими исходными допущениями в [Кибрик 1997: 94-95]:

- «выбор референциального средства — не случайный, а **мотивированный** процесс;
- адекватная модель референциального выбора должна быть в состоянии объяснить **каждый** случай выбора в исследуемом дискурсе;
- референциальный выбор не может быть понят в рамках внутритекстового подхода; он непосредственно предопределяется **когнитивным состоянием говорящего** в данный момент;
- число факторов, влияющих на когнитивное состояние говорящего, может быть **велико**».

К перечисленным пунктам добавляется два новых. Первое допущение мотивировано особенностями устного модуса в русском языке. В отличие от письменных текстов, в устной речи нулевые выражения являются самостоятельным референциальным средством, поэтому представляется актуальным рассматривать тройное противопоставление «полная ИГ – местоимение – нуль». В то же время, нет оснований считать, что одни и те же факторы одинаково влияют на выбор «полная ИГ – местоимение» и на выбор «местоимение – нуль». Допущение можно сформулировать так:

- Одни и те же факторные значения могут влиять **по-разному** на выбор между полной ИГ и местоимением, с одной стороны, и на выбор между местоимением и нулем — с другой.

В соответствии с этим допущением предлагается тройное противопоставление разбить на две дискретные пары: дихотомия полной ИГ и местоимения и дихотомия

местоимения и нуля. Для краткости дихотомии называются по номерам: «дельта»-1, «дельта»-2, или Δ1 и Δ2.

Второе допущение было обосновано в [Kibrik et al. 2016] и [Ferreira et al. 2016]:

– РВ **не является категориальным** и во многих случаях план содержания ЕРВ имеет несколько равновозможных вариантов плана выражения.

4.2. Двойная значимость факторов и весов их значений

Чтобы модель была применима в автоматическом синтезе речи, необходимо, чтобы приоритетность факторов и числовые значения их весов были **математически фундированными**. Однако у методов статистики и машинного обучения существует склонность имитировать, но не отражать реальные закономерности. Поэтому необходим дополнительный контроль за тем, чтобы в модели на новом материале был учтен **опыт предыдущих исследований**. В этих двух требованиях заключена двойная значимость модели — **математическая** и **теоретическая**. Контроль за значимостью обоих типов осуществляется на уровне факторов и на уровне весов факторных значений при помощи специальных диагностик — **индикаторов значимости** (табл. 2):

Таблица 2

Индикаторы значимости

	Математическая значимость	Теоретическая значимость
Факторы	Уровень значимости критерия χ^2 не больше 0.05.	Фактор признан значимым хотя бы в одном из ранее проведенных исследований.
Веса факторных значений	Уровень значимости ВФЗ в регрессионной модели не более 0.05; при соответствии теоретическому индикатору — не больше 0.1.	Теоретические ожидания

Остановимся подробнее на теоретическом индикаторе для весов. Для регулирования теоретической значимости ВФЗ достаточно двух теоретических ожиданий:

1) Если вес отличен от нуля, то его знак (= влияние) должен коррелировать с тем, что наблюдается в выборке. Например, если в Δ «полная ИГ – местоимение» значение «пациенс» чаще имеют местоимения, чем полные ИГ, то это значение оказывает повышающее действие на Δ1, следовательно его вес должен быть положительным.
2) ВФЗ факторов из категории «Дистанции» должны образовывать невосходящий тренд: чем больше дистанция, тем больше отрицательный модуль веса. По аналогии с фактором «Риторическое расстояние» в [Кибрик 1997], наименьшее значение фактора может иметь положительный вес.

Четыре типа индикаторов при совместном применении позволяют максимизировать значимость итоговой модели.

4.3. Алгоритм разработки

Построение модели выполнялось в соответствии со следующим планом:

1) Сбор теоретически значимых факторов;
2) Получение набора математически значимых факторов;
3) Подготовка к построению МЛР;
4) Построение МЛР и получение математически значимых весов;
5) Корректировка весов в соответствии с теоретическими ожиданиями и иерархией факторов;

6) Нормализация весов.

В табл. 3 суммирован процесс отбора факторов (деление на 4 группы произведено аналогично факторному блоку в [Loukachevitch et al. 2011: 522]). Все представленные в таблице факторы (за исключением двух, проверяемых в рамках гипотезы 3) являются теоретически значимыми.

Таблица 3

Процесс и результат отбора факторов

Фактор	**Математическая значимость**	**Прохождение теста на коллинеарность**	**Значимость для Δ1**	**Значимость для Δ2**
Свойства референта				
Референт или образ	Да	-	Да	Да
Одушевленность	Да	-	Да	Да
Число	**Нет**			
Род	Да, но менее значимо, чем род-число			
Род-число	Да	-	Да	Да
Протагонизм 1	Да	Да	Да	**Нет**
Протагонизм 2	Да, но менее значима, чем 1-я			
Свойства анафора				
Грамматическая роль	Да	-	**Нет**	**Нет**
Подлежащность	**Нет**			
Семантическая роль	Да	-	Да	Да
Порядковый номер в референциальной цепи	**Нет**			
Неполная кореферентность	Да	-	Да	Да
Свойства антецедентов				
Форма линейного антецедента	Да	-	**Нет**	**Нет**
Грамматическая роль линейного антецедента	Да	-	**Нет**	**Нет**
Семантическая роль линейного антецедента	Да	-	Да	Да
Форма риторического антецедента	Да	-	**Нет**	Да
Грамматическая	Да	-	Да	**Нет**

Фактор	**Математическая значимость**	**Прохождение теста на коллинеарность**	**Значимость для Δ1**	**Значимость для Δ2**
роль риторического антецедента				
Семантическая роль риторического антецедента	Да	-	Да	Да
Совпадение антецедентов	**Нет**			
Дистанции				
Линейное расстояние в ЭДЕ	Да	Да	Да	Да
Число ЕРВ до ближайшего антецедента (все РЦ[34])	Да	**Нет**		
Число ЕРВ до ближайшего антецедента, выраженного полной ИГ (только собственная РЦ)	Да	Да	**Нет**	Да
Риторическое расстояние	Да	Да	Да	Да
Расстояние в эпизодах	Да	**Нет**		
Итого значимых факторов:			11	11

На первом этапе был собран список теоретически значимых факторов — всего их насчитывалось 24. Источниками факторов являются ранее проведенные исследования ([Кибрик 1997, Kibrik, Grüning 2005, Loukachevitch et al. 2011, Kibrik et al. 2016, Strube, Walters 2002, Same, van Deemter 2020, Саенко, Прокопеня 2020]).

Далее факторы были проверены на математическую значимость на материале тренировочной выборки с помощью критерия χ^2. По результатам проверки, число факторов сократилось с 24 до 20. Таким образом, был сформирован набор факторов, обладающих двойной значимостью.

На третьем шаге была проведена подготовка факторного блока к построению МЛР. Она включала в себя два действия:

а) Выявление и ликвидация случаев коллинеарности факторов с помощью корреляционной матрицы (рис. 4).

[34] Референциальная цепь. Термин обнаружен в [Strube, Walters 2002].

		Число маркабул между анафором и линейным антецедентом	Число маркабул до ближайшего антецедента, выраженного полной ИГ	Линейное расстояние	Расстояние до линейного антецедента в эпизодах
Число маркабул между анафором и линейным антецедентом	Pearson's r	—			
	p-value	—			
Число маркабул до ближайшего антецедента, выраженного полной ИГ	Pearson's r	-0.087	—		
	p-value	0.189	—		
Линейное расстояние	Pearson's r	0.792	-0.065	—	
	p-value	< .001	0.321	—	
Расстояние до линейного антецедента в эпизодах	Pearson's r	0.598	-0.038	0.772	—
	p-value	< .001	0.566	< .001	—

Рисунок 4. Корреляционная матрица

Факторы коллинеарны, если их коэффициент корреляции (*Pearson's r*) близок к 0.8. По причине коллинеарности с линейным расстоянием из факторного блока были удалены расстояние в эпизодах и расстояние до ближайшего антецедента в ЕРВ (все РЦ).

b) Выбор базовых факторных значений.

При построении МЛР в рамках каждого фактора выбирается одно значение, базовое, и с его частотой сравниваются частоты всех остальных значений. Каждое такое сравнение называется отношением шансов; именно натуральный логарифм отношения шансов для некоторого факторного значения является оценкой его веса. Базовое значение всегда обладает весом, равным нулю. Базовые значения выбирались с тем расчетом, чтобы как можно больше регрессионных коэффициентов (т. е. статистических оценок ВФЗ) были статистически значимыми. У части факторов оценки весов оказались статистически не значимыми (т.е. p-level > 0.1) при любых базовых значениях. Такие факторы были исключены из блока для обеих дихотомий либо для одной из них.

После подготовительных действий была построена логистическая модель, ниже (рис. 5) приводится ее фрагмент для дихотомии «местоимение — нуль», или Δ2.

		95% Confidence Interval						95% Confidence Interval	
Predictor	Estimate	Lower	Upper	SE	Z	p	Odds ratio	Lower	Upper
Риторическое расстояние:									
1.5 – 1	1.870	-0.1228	3.8631	1.01684	1.8392	0.066	6.48935	0.88443	47.6142
2 – 1	4.292	1.5571	7.0270	1.39541	3.0758	0.002	73.11634	4.74502	1126.6555
2.5 – 1	3.988	1.0005	6.9760	1.52438	2.6163	0.009	53.95942	2.71966	1070.5838
3 – 1	5.827	2.0150	9.6391	1.94496	2.9960	0.003	339.36565	7.50099	15353.8522
3.5 – 1	3.000	-0.2598	6.2602	1.66329	1.8038	0.071	20.08936	0.77121	523.3082
5.5 – 1	-7.037	-384.9248	370.8517	192.80369	-0.0365	0.971	8.79e -4	6.75e-168	1.15e+161
Форма риторического антецедента:									
нуль – мест.	3.976	0.7552	7.1974	1.64345	2.4195	0.016	53.32043	2.12808	1335.9780
полная ИГ – мест.	2.509	-0.8492	5.8680	1.71359	1.4644	0.143	12.29737	0.42776	353.5305
Референт / образ:									
образ – референт	49.384	48.2466	50.5212	0.58028	85.1034	< .001	2.80e+21	8.98e +20	8.73e +21

Рисунок 5. Фрагмент МЛР

Все регрессионные коэффициенты с уровнем значимости не более 0.1 считались математически значимыми.

Следующий этап в построении — корректировка весов в соответствии с теоретическими ожиданиями, а также внедрение в модель иерархии факторов (показана в табл. 4), основанной на сравнении коэффициентов сопряженности. При этом допускалось только два вида изменений:

1. Смена знака веса;
2. Получение веса с помощью арифметических действий с опорой на соседние веса, значимые математически и теоретически.

Таблица 4

Иерархия факторов

Номер	Фактор	Коэффициент сопряженности
1	Линейное расстояние	0.396
2	Одушевленность	0.361
3	Семантическая роль линейного антецедента	0.356
4	Число ЕРВ до ближайшего антецедента, выраженного полной ИГ	0.341
5	Семантическая роль	0.340
6	Семантическая роль риторического антецедента	0.331
7	Риторическое расстояние	0.33
8	Форма риторического антецедента	0.314
9	Грамматическая роль линейного антецедента	0.284
10	Родо-численность	0.283
11	Форма линейного антецедента	0.264
12	Грамматическая роль линейного антецедента	0.229
13	Грамматическая роль	0.196
14	Референт или образ	0.169

Номер	Фактор	Коэффициент сопряженности
15	Неполная кореферентность	0.164
16	Протагонизм 1	-

У каждого фактора было выделено значение, вес которого обладает максимальным модулем (будет называть такой вес MaxAbs). Далее было применено 2 правила:

– Если MaxAbs нижестоящего фактора больше, чем MaxAbs вышестоящего, сделать его равным MaxAbs вышестоящего;
– Все остальные веса нижестоящего фактора уменьшить во столько же раз, во сколько уменьшился его MaxAbs.

Заключительный шаг в разработке — нормализация весов с сохранением их порядка и иерархии. Эта мера позволит сделать так, чтобы в большинстве случаев КА не превышал единицу.

В результате был получен новый набор весов, который с точки зрения индикаторов значимости является теоретически и математически значимым. На основе тренировочной выборки были также получены шкалы референциальных стратегий и выведен новый фильтр — «наличие адъективных зависимых».

5. Апробация

Перед апробацией модели были выполнены подготовительные действия:

- Каждая ЕРВ в контрольной выборке была размечена по значимым факторам.
- Каждое факторное значение было конвертировано в два числовых веса, по Δ1 и Δ2.
- При помощи шкал референциальных стратегий для каждой ЕРВ были выведены прогнозы РВ.

Проверка эффективности модели состояла в оценке качества полученных прогнозов. Ключевыми для оценивания были два требования:

1) В прогнозе должно содержаться то средство, которое реально было употреблено говорящим.
2) Все остальные спрогнозированные средства должны быть приемлемыми в данном контексте. Оценка приемлемости является бинарной и производится с опорой на интроспекцию.

Каждый прогноз мог получить один из трех статусов:

- Верно: соответствие обоим требованиям.
- Частичное соответствие: выполнено первое требование, но не выполнено второе.
- Неверно: не выполнено первое требование.

Результаты апробации представлены в Таблице 5.

Таблица 5

Статистика по результатам апробации

	Верных прогнозов	Частичных соответствий	Неверных прогнозов	Всего
Абсолютные показатели	37	9	3	49
Относительные показатели	76%	18%	6%	

Модель сделала ошибочные прогнозы в 6% случаев, как правило, она не предсказывала полные ИГ там, где они были употреблены говорящими. Это может быть связано с переоцененностью семантической роли анафора «агенс». В 94% случаев модель дала верные либо приемлемые прогнозы.

Чтобы проиллюстрировать работу модели, рассмотрим фрагмент контрольного дискурса[35]:

(2) 14.83 16. А у нас **мальчик** один,
16.03 17. **он** всё время жевал сухари.
16.98 18. Мы ещё так и звали **его**:
17.71 19. **Сухарь**.

Кореферентные ЕРВ отмечены жирным шрифтом, единица, план выражения которой нужно спрогнозировать, подчеркнута. Для целевой ЕРВ были получены и количественно оценены (табл. 6) релевантные факторные значения:

Таблица 6

Пример использования модели

Фактор	**Значение**	**Вес значения по Δ1**	**Вес значения по Δ2**
Референт или образ?	референт	0	0
Одушевленность	Одушевленное[36]	0,3530	0,6670
Род-число	Мужской	0,0830	0
Протагонизм	1	0,0830	0,0830
Линейное расстояние	1	0,0730	0,0830
Число маркабул до ближайшего антецедента, выраженного полной ИГ	0	0	0
Риторическое расстояние	1	0	0
Семантическая роль	агенс	0,1410	-0,0200
Неполная кореферентность	нет	0	0
Семантическая роль линейного антецедента	агенс	0,0950	0,3500
Форма риторического антецедента	полная ИГ	0	0
Грамматическая роль риторического антецедента	подлежащее	0,0830	-0,1770
Семантическая роль риторического антецедента	агенс	0,1750	0
Коэффициент активации		1,086	0,9860
Коэффициент активации		1	0,9860

[35] Рассказ 13 из корпуса «Веселые истории из жизни» [эл. ресурс], код рассказа в исследовании К2.
[36] В оригинальной таблице используются сокращения «одуш.», «неодуш», «муж» и т.д.

Фактор	Значение	Вес значения по Δ1	Вес значения по Δ2
(нормализованный)			
Прогноз		Только местоимение	Местоимение или нуль

Итоговый прогноз модели включает реально примененное средство, а также содержит альтернативное — нуль. По данным интроспекции, нуль в рассмотренном контексте представляется приемлемым.

6. Выводы и перспективы

В рамках исследования была создана модель РВ в устных рассказах на русском языке. Она продемонстрировала достаточно высокую эффективность на контрольной выборке. В ходе исследования нашли подтверждение гипотеза №2. а также первая часть гипотезы №3: критерий Пирсона подтвердил значимость нового фактора «референт или образ» в устном дискурсе. Остальные два предположения не нашли подтверждения.

Опровержение гипотезы №1: существуют ЕРВ, которые могут быть выражены полной ИГ и нулем, но не могут быть выражены местоимением. Это все ЕРВ, находящиеся под действием фильтра «наличие адъективных зависимых». Например, ЕРВ, подчеркнутая во фрагменте ниже, не может быть выражена местоимением по той причине, что имеет зависимое прилагательное (дискурс 8, в выборке — К1):

(3) 1.38 2. У меня была ужасная **причёска**,
2.68 3. и я была просто как парень,
3.70 4. вообще настолько ужасная **0**,
5.32 5. что и ==

Еще два примера действия этого фильтра содержатся в рассказе 9 (по нумерации в выборке — 8):

(4) 4.11 3. ·· так как я свой **день варенья** не устраивала,
6.37 4. то 0 решили его отметить в этой **компании**,

У обеих подчеркнутых ЕРВ есть зависимые адъективные местоимения, поэтому невозможно сказать **свой его* или **в этой ней*; единственные альтернативы — *свой 0* и *в этой 0*.

Опровержение гипотезы №3б: по результатам применения критерия Пирсона, фактор «совпадение антецедентов» не является значимым для РВ.

У исследования есть три основные перспективы. Первая — расширение тренировочной выборки и увеличение точности ВФЗ. Вторая — расширение факторного блока, выявление новых значимых факторов. Третья — проведение аналогичного исследования на материале письменного нарративного дискурса, что позволило бы сделать кроссмодальное сравнение когнитивных систем референциального выбора.

Список сокращений

РВ — референциальный выбор
РС — референциальное средство (либо средства)
ЕРВ — единица референциального выбора
КА — коэффициент активации
МЛР — мультиномиальная логистическая регрессия
MaxAbs — вес, модуль которого в рамках данного фактора максимален

Литература

Будённая Е. В. Эволюция субъектной референции в языках балтийского ареала: Дис. … канд. филол. наук. – Москва, 2018.

Кибрик А. А. Моделирование многофакторного процесса: выбор референциального средства в русском дискурсе // Вестник Московского университета. — 1997. — Серия 9: Филология. № 4. — С. 94-105.

Кибрик А. А., Подлесская В. И., Коротаев Н. А. Структура устного дискурса: основные элементы и канонические явления // Кибрик, А. А., Подлесская, В. И. (ред.). Рассказы о сновидениях: корпусное исследование устного русского дискурса / Институт языкознания РАН, Российский государственный гуманитарный университет. — Москва: Издательство «Языки славянских культур», 2009. — С. 55-101.

Кибрик А. А., Линник А. С., Добров Г. Б., Худякова М. В. Оптимизация модели референциального выбора, основанной на машинном обучении // Компьютерная лингвистика и интеллектуальные технологии: труды XVIII Международной конференции «Диалог 2012»: в 2-х томах, Бекасово, 30 мая — 03 2012 года. — Бекасово: Российский государственный гуманитарный университет, 2012. — С. 237-246.

Корпус «Веселые истории из жизни». Эл. ресурс: http://spokencorpora.ru/showcorpus.py?dir=02funny

Литвиненко А. О., Подлесская В. И., Кибрик А. А. Анализ рассказов о сновидениях с точки зрения иерархической структуры дискурса // Кибрик А. А., Подлесская В. И. (ред.). Рассказы о сновидениях: корпусное исследование устного русского дискурса / Институт языкознания РАН, Российский государственный гуманитарный университет. — Москва: Издательство «Языки славянских культур», 2009. — С. 431-463.

Милевская Т. В. Связность как категория дискурса и текста (Когнитивно-функциональный и коммуникативно-прагматический аспекты): Автореф. дисс. канд. докт. фил. наук. — Ростов-на-Дону, 2003.

Саенко Е., Прокопеня В. К. Референциальный выбор при описании видеоролика в режиме реального времени // Студенческие Смольные чтения, Санкт-Петербург, 01 октября 2015 года — 31 2017 года / Санкт-Петербургский государственный университет. — Санкт-Петербург: Центр научно-информационных технологий «Астерион», 2020. – С. 214-223.

Чейф У. Данное, контрастивность, определенность, подлежащее, топики и точка зрения // Кибрик А. Е. (ред.). Новое в зарубежной лингвистике. – Москва: ПРОГРЕСС, 1982. – С. 277-317.

Ferreira, T. C., Krahmer, E., Wubben, S. Individual Variation in the Choice of Referential Form // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — [https://www.researchgate.net/publication/305334477_Individual_Variation_in_the_Choice_of_Referential_Form], 2016. — P. 423-427.

Kibrik, A. A. Anaphora in Russian Narrative Prose: A Cognitive Calculative Account // Fox B. A. (ed.). Studies in Anaphora. — Amsterdam, Philadelphia: John Benjamins Publishing Company, 1996. — P. 255-286.

Kibrik, A. A., Grüning, A. Modelling Referential Choice in Discourse: A Cognitive Calculative Approach and a Neutral Network Approach // Anaphora Processing. Linguistic, cognitive and computational modelling. – Lisbon : John Benjamins Publishing, 2005. – P. 163-198.

Kibrik, A. A., Zalmanov, D. A., Khudyakova, M. V. [et al.] Referential choice: Predictability and its limits // Frontiers in Psychology. – 2016. – Vol. 7. – No SEP. – P. 1429. – DOI 10.3389/fpsyg.2016.01429.

Khudyakova, M. V, Dobrov, G.B., Kibrik, A. A., Loukachevitch, N.V. Computational modeling of referential choice: Major and minor referential options, 2011.

[https://www.researchgate.net/publication/247768213_Computational_modeling_of_referential_choice_Major_and_minor_referential_options]. — 5 p.

Loukachevitch, N. V., Dobrov, G. B., Kibrik, A. A., Khudyakova, M. V., and Linnik, A. S. Factors of referential choice: computational modeling // A. E. Kibrik (ed.). In Proceedings of the Papers from the Annual International Conference «Dialogue» (2011): Computational Linguistics and Intellectual Technologies. — Moscow, 2011. — P. 518–528.

Mann, W., Thompson, S. Rhetorical Structure Theory: A Theory of Text Organization. — Marina del Rey: Information Sciences Institute, 1987. — 92 p.

Same, F., Deemter, K. A Linguistic Perspective on Reference: Choosing a Feature Set for Generating Referring Expressions in Context, 2020. https://www.researchgate.net/publication/346570847_A_Linguistic_Perspective_on_Reference_Choosing_a_Feature_Set_for_Generating_Referring_Expressions_in_Context. — 12 p.

Strube, M., Wolters, M. A Probabilistic Genre-Independent Model of Pronominalization // Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, 2002. — [https://www.researchgate.net/publication/2539169_A_Probabilistic_Genre-Independent_Model_of_Pronominalization]. — 8 p.

References

Budennaya, E. V. Subject Reference Evolution in the Kanguages of the Baltic Area. Diss. kand. fil. nauk. – Москва, 2018.

Kibrik, A. A. Multifactorial Process Modelling: Referential Choice in Russian Discourse. // Moscow State University Bulletin. — 1997. — Series 9: Philology. № 4. — P. 94-105.

Kibrik, A. A., Podlesskaya, V. I., Korotaev, N. A. Spoken Discourse Structure: Principal Elements and Canonical Phenomena // Kibrik, A. A., Podlesskaya, V. I. (eds.). Dream Stories: a Corpus Research of Russian Spoken Discourse / Institute of Linguistics of the Russian Academy of Sciences, Russian State University of the Humanities. — Moscow: «Языки славянских культур» Publishing, 2009. — P. 55-101.

Kibrik, A. A., Linnik, A. S., Dobrov, G. B., Khudyakova, M. V. Machine-Learning-Based Referencial Choice Model Optimization // Applied Linguistics and Intellectual Technologies: in Proceedings of the 18th International Conference “Dialog 2012”: in 2 volumes, Bekasovo, 30th May — 03 2012. — Bekasovo: Russian State University of the Humanities, 2012. — P. 237-246.

«Funny Life Stories» Corpus. URL: http://spokencorpora.ru/showcorpus.py?dir=02funny

Litvinenko, A. O., Kibrik, A. A., Podlesskaya, V. I. Dream Stories Analysis in Terms of the Hierarchical Structure of Discourse // Kibrik, A. A., Podlesskaya, V. I. (eds.). Dream Stories: a Corpus Research of Russian Spoken Discourse / Institute of Linguistics of the Russian Academy of Sciences, Russian State University of the Humanities. — Moscow: «Языки славянских культур» Publishing, 2009. — P. 431-463.

Milevskaya, T.V. Cohesion as a Discursive and Textual Category (Communicative-Functional and Communicative-Pragmatic Aspects): Avtoref. diss. dokt. fil. nauk. — Rostov-On-Don, 2003.

Saenko, E., Prokopenya, V. K. Referential Choice during Online Video Description // Student Smolny Conference, Saint Petersburg, 01st October 2015 — 31 2017 / Saint Petersburg State University. — Saint Petersburg: Information Technologies Center «Asterion», 2020. – P. 214-223.

Chafe, W. Givenness, contrastiveness, definiteness, subjects, topics, and point of view // Kibrik, A. E. (ed.). New in Foreign Linguistics. – Moscow: PROGRESS, 1982. – P. 277-317.

Ferreira, T. C., Krahmer, E., Wubben, S. Individual Variation in the Choice of Referential Form // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. —

[https://www.researchgate.net/publication/305334477_Individual_Variation_in_the_Choice_of_Referential_Form], 2016. — P. 423-427.

Kibrik, A.A. Anaphora in Russian Narrative Prose: A Cognitive Calculative Account // Fox B. A. (ed.). Studies in Anaphora. — Amsterdam, Philadelphia: John Benjamins Publishing Company, 1996. — P. 255-286.

Kibrik, A.A., Grüning, A. Modelling Referential Choice in Discourse: A Cognitive Calculative Approach and a Neutral Network Approach // Anaphora Processing. Linguistic, cognitive and computational modelling. – Lisbon: John Benjamins Publishing, 2005. – P. 163-198.

Kibrik, A. A., Zalmanov, D. A., Khudyakova, M. V. [et al.] Referential choice: Predictability and its limits // Frontiers in Psychology. – 2016. – Vol. 7. – No SEP. – P. 1429. – DOI 10.3389/fpsyg.2016.01429.

Khudyakova, M. V, Dobrov, G.B., Kibrik, A. A., Loukachevitch, N.V. Computational modeling of referential choice: Major and minor referential options, 2011. [https://www.researchgate.net/publication/247768213_Computational_modeling_of_referential_choice_Major_and_minor_referential_options]. — 5 p.

Loukachevitch, N. V., Dobrov, G. B., Kibrik, A. A., Khudyakova, M. V., and Linnik, A. S. Factors of referential choice: computational modeling // A. E. Kibrik (ed.). In Proceedings of the Papers from the Annual International Conference «Dialogue» (2011): Computational Linguistics and Intellectual Technologies. — Moscow, 2011. — P. 518–528.

Mann, W., Thompson, S. Rhetorical Structure Theory: A Theory of Text Organization. — Marina del Rey: Information Sciences Institute, 1987. — 92 p.

Same, F., Deemter, K. A Linguistic Perspective on Reference: Choosing a Feature Set for Generating Referring Expressions in Context, 2020. [https://www.researchgate.net/publication/346570847_A_Linguistic_Perspective_on_Reference_Choosing_a_Feature_Set_for_Generating_Referring_Expressions_in_Context]. — 12 p.

Strube, M., Wolters, M. A Probabilistic Genre-Independent Model of Pronominalization // Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, 2002. — [https://www.researchgate.net/publication/2539169_A_Probabilistic_Genre-Independent_Model_of_Pronominalization]. — 8 p.