Известия Воронежского государственного педагогического университета. 2022. № 1 (294). С. 167–173.
Izvestia Voronezh State Pedagogical University. 2022. (1), 167–173.

Научная статья
УДК 81'33
DOI 10.47438/2309-7078_2022_1_167

МОДЕЛИРОВАНИЕ ЗНАЧЕНИЯ СЛОВА В ТЕКСТАХ МУЖЧИН И ЖЕНЩИН МЕТОДАМИ ДИСТРИБУТИВНОЙ СЕМАНТИКИ

Татьяна Александровна Литвинова[1]

Воронежский государственный педагогический университет[1]
Воронеж, Россия

[1]*Кандидат филологических наук, научный сотрудник Регионального центра русского языка, e-mail: centr_rus_yaz@mail.ru*

Аннотация. Моделирование значения слова в индивидуальном языковом сознании является одной из актуальных задач современной психолингвистики. Одним из важнейших аспектов названной задачи является установление влияния групповых (в том числе социодемографических) характеристик говорящего на особенности семантики слова в его языковом сознании. Для решения названной задачи используется ряд экспериментальных методов (ассоциативный эксперимент, семантическое шкалирование и т.д.). Однако проведение подобных экспериментов связано со значительными временными и трудовыми затратами, в связи с чем в современной науке активно развиваются корпусные исследования с использованием алгоритмов и инструментов дистрибутивной семантики, направленные на установление особенностей семантики слова в текстах лиц с различными социодемографическими характеристиками (прежде всего мужчин и женщин). Такие исследования, однако, до сих пор немногочисленны, выполняются преимущественно на материале английского языка и требуют большого объема корпусных данных. Кроме того, в них не уделяется внимания жанровым характеристикам исследуемых текстов. В настоящем исследовании представлены результаты пилотного эксперимента, направленного на оценку влияния гендера автора на особенности семантики слова в текстах разных жанров с использованием новейших методов дистрибутивной семантики, реализованных в пакете conText. Показана необходимость учета жанра при моделировании значения слова в языковом сознании лиц с разными демографическими характеристиками.

Ключевые слова: дистрибутивная семантика, семасиология, корпус текстов, идиолект, демографические эмбеддинги.

Для цитирования: *Литвинова Т.А.* Моделирование значения слова в текстах мужчин и женщин методами дистрибутивной семантики // Известия Воронежского государственного педагогического университета. 2022. № 1 (294). С. 167–173. DOI: 10.47438/2309-7078_2022_1_167

Введение

Как известно, «традиционные исследования языкового сознания в психолингвистике основываются на свободном ассоциативном эксперименте, при котором группе испытуемых предъявляются слова-стимулы и фиксируются реакции» [1, с. 77]. Полученные данные являются основой ассоциативных словарей, являющихся отражением языкового сознания «типичного» носителя языка, средством доступа к его ментальному лексикону. Предпринимаются попытки создания ассоциативных словарей лиц, объединенных по полу, возрастной группе и т.д. (см., например, [2; 4]). Число респондентов, ассоциативные реакции которых анализируются в такого рода работах, как правило, невелико и составляет несколько сот человек (что объясняется трудоемкостью работ по сбору подобного материала), и подобные источники (безусловно, ценного) лингвистического материала существуют только в печатном виде, что затрудняет работу исследователей с ними.

Трендом мировой лингвистики является создание баз данных ассоциаций, доступных онлайн (см., например, активно развивающийся проект Small World of Words[1], содержащий данные для 14 языков, например, ассоциации для 12 000 английских слов, что является наиболее крупной базой данных ассоциаций для английского языка; наибольшее число данных собрано для голландского языка – ассоциативные реакции на 16 000 стимулов; база дает возможность визуализировать полученные результаты в виде семантической сети) [6]. Данные о характеристиках респондентов, однако, в существующих проектах отсутствуют, в связи с чем рядом исследователей ставятся задачи по сбору баз данных ассоциаций, содержащих подобные характеристики (см., например, работу [7, с. 79], в которой

[1] URL: https://smallworldofwords.org/en/project/home (дата обращения: 01.11.2021).

представлена БД ассоциаций с демографическими характеристиками авторов). Так, названная работа содержала около 300 стимулов и 800 реакций на каждый стимул, и в ходе специальных исследований было установлено, что существуют достоверные различия в ассоциациях мужчин и женщин для некоторых слов.

Для русского языка существуют по меньшей мере две базы данных ассоциаций, содержащие информацию о социодемографических (пол, возраст, профессия) характеристиках респондентов: Русский ассоциативный тезаурус, отражающий результаты крупнейшего ассоциативного эксперимента для русского языка (1988–1997 гг.) и содержащий свыше 1 млн ассоциаций, более 6 тыс. уникальных стимулов и 100 тыс. реакций от 11+ тыс. респондентов (доступен по адресу http://tesaurus.ru/dict/), и Русская региональная ассоциативная база данных, основанная на вербальных ассоциациях русских, проживающих на Дальнем Востоке и Сибири (http://adictru.nsu.ru/dictright#). База содержит 1000 слов-стимулов на русском языке; число респондентов – 5 011.

Таким образом, для русского языка существуют уникальные ресурсы, позволяющие исследовать ассоциации, в том числе в сопоставительном аспекте. Однако, как известно, интерпретация ассоциативных реакций связана со значительными методологическими трудностями, детально проанализированными в работе В.А. Пищальниковой [3]. Кроме того, в соответствующих исследованиях на материале русского языка не используются статистические методы анализа для оценки достоверности различий в ассоциативных реакциях лиц разных групп. Очевидно, для эффективного моделирования значения слова в языковом сознании, в том числе с учетом групповых характеристик респондентов, нельзя ограничиваться исключительно данными ассоциативного эксперимента.

В настоящее время для моделирования значения слова активно используются корпусы текстов и методы дистрибутивной семантики, которые позволяют «выявлять семантически близкие слова, основываясь на векторных представлениях слов в дистрибутивно-семантических моделях» [1, с. 77]. Предполагается, что «слова схожей семантики, встречающиеся в одних и тех же контекстах, связаны в языковом сознании человека, как и ассоциаты» [1, с. 77], в связи с чем в ряде исследований последних лет проводится сопоставительный анализ результатов моделирования значения слова с использованием методов дистрибутивной семантики и ассоциативного эксперимента, при этом в качестве аналога ассоциатов рассматриваются наиболее близкие к стимулу лексемы в дистрибутивной модели (см., например, [1], а также обзор существующих работ в статье [14]). Результаты подобных исследований показывают, что векторное представление слова является новым перспективным способом представления его значения [8, p. 38]. Предполагается, что обучение нейронных моделей может быть схоже с тем, как люди узнают значения слов, и, следовательно, подобные модели устраняют разрыв между традиционными моделями дистрибутивной семантики и психологически обоснованными принципами обучения [11, p. 57]. Высказывается интересное предположение (см. подробнее: [14]) о том, что различные оценки близости слов (в дистрибутивной модели и ассоциативном эксперименте) отражают различия между внешними и внутренними языковыми моделями: векторное представление семантики слова, извлеченное из корпуса, рассматривает язык как внешний объект, состоящий из всех высказываний, производимых данным сообществом, тогда как внутренние языковые модели (например, сети семантических ассоциаций) рассматривают язык как совокупность знаний, содержащихся в мозге говорящего. При таком понимании ассоциации отражают представления, которые не могут быть извлечены дистрибутивными моделями, поскольку последние формируются под влиянием ряда прагматических и коммуникативных факторов. Все сказанное выше, на наш взгляд, красноречиво свидетельствует о необходимости дополнения существующих методов моделирования значения слова, представленных в отечественной психолингвистике преимущественно ассоциативным экспериментом, методами дистрибутивной семантики.

В цитируемых выше работах, однако, не ставился вопрос о сравнении методов ассоциативного эксперимента и моделирования значения слова с использованием дистрибутивных семантических моделей для оценки влияния демографических характеристик авторов текстов на особенности семантики слова (в ее векторном представлении). В целом вопрос о влиянии названных характеристик пишущих на различия в векторном представлении слов на русском языке является неисследованным. В настоящей работе впервые ставится задача оценки влияния гендера автора на степень близости слов в семантической модели при контроле жанра для текстов на русском языке.

Современное состояние исследований в области моделирования значения слова методами дистрибутивной семантики в текстах мужчин и женщин

Общение между людьми требует понимания значения употребляемых ими слов, в связи с чем исследователи активно изучают феномен значения слова, причины его изменения, способы манипулирования им. Если ранее подобные исследования носили качественный характер, то в последнее время преимущественно используются методы исследования, предполагающие конструирование значения слова на основе дистрибуции лексем, которые окружают данное слово в (кон)тексте. Представляя каждое слово корпуса как вектор (эмбеддинг) и исследуя связи между ними, ученые получают новые данные о языке и людях, которые на нем говорят [5].

Основная идея векторной семантики («значение слова возможно смоделировать на основе распределения слов в контексте, понимаемом как соседние слова») появилась в исследованиях 1950-х годов в трудах по лингвистике, психологии, компьютерным наукам, при этом каждая из названных областей знания внесла свой вклад в фундаментальные аспекты векторной семантики. Особенно популярной векторная семантика стала после появления таких инструментов, как word2vec [12], Glove [13], в которых были представлены предиктивные модели (в дополнение к существовавшим ранее счетным моделям), позволившие значительно увеличить скорость расчетов.

Следует, однако, отметить, что эмбеддинги, полученные при помощи word2vec либо GloVe, отражают синтаксические и семантические свойства идиолектов всех лиц, чьи тексты вошли в исследовательский корпус, при этом они не учитывают особенности индивидуального словоупотребления. Проблема влияния групповых (в том числе демографических) характеристик на векторное представление слов мало исследована. Существующие техники не позволяют ответить на вопрос о том, к примеру, существуют ли статистически значимые различия в употреблении слова «любовь» в текстах мужчин и женщин. Безусловно, особенности словоупотребления не находятся в прямой зависимости от демографических характеристик носителей язы-

ка, однако, моделируя особенности семантики слов в речи лиц разных групп, возможно более эффективно строить языковые модели и разрабатывать прикладные инструменты (например, чат-боты).

Существующие немногочисленные сравнения эмбеддингов текстов лиц разных групп (см. обзор в работе [15]) сложны для реализации и требуют больших объемов языковых данных. Для нашего пилотного исследования мы будем использовать новый пакет на языке R conText [15] и методы, разработанные его авторами и специально созданные для работы с корпусами малого объема и выявления отличий между эмбеддингами слов в разных корпусах текстов. В пакете реализован представленный в работе Khodak et al. [9] метод, названный «эмбеддинги a la carte» (ALC) и основанный на Glove-подобных эмбеддингах. Создатели пакета conText предлагают с использованием методов регрессии отвечать на вопросы вида «Используют ли люди из группы A данные слова иначе, чем люди из группы B? Если да, то в чем именно состоит такое отличие?» В работе 2021 [15] было показано, что ALC сравним по эффективности с современными методами моделирования значения слова в контекста, предназначенными для больших данных (например, BERT).

Отметим, что под значением слова в дальнейшем нами подразумевается особенности его совместной встречаемости с другими словами, то есть значение в контексте дистрибутивной семантики.

Основной используемой нами функцией пакета conText является функция *nns_ratio* (), которая вычисляет отношение мер косинусового сходства между группами и эмбеддингами, то есть сначала вычисляется косинусовая близость слова с другими словами для текста каждой группы, после чего вычисляется их соотношение, которое отражает, насколько отличителен данный признак для группы. Уникальной особенностью названного пакета, является то, что в нем реализована возможность проведения пермутационного теста. Для каждой пермутации группирующая переменная назначается случайным образом, и вычисляется абсолютное отклонение значения меры косинусовой близости.

Экспериментальное исследование

Материалом экспериментального исследования послужили тексты, входящие в созданную под нашим руководством БД RusIdiolect [10]. Мы использовали подкорпус текстов одних и тех же лиц, создавших тексты двух разных жанров, – описание картины, предложенной респондентам, и письмо другу.

Общее число текстов составило 884, из них описаний картин – 343 (247 текстов женщин, 96 текстов мужчин), 541 письмо другу (358 текстов авторов-женщин, 183 текста авторов-мужчин). Респонденты – студенты российских вузов.

Исследования проводились на лемматизированном с использованием анализатора Treetagger[1] корпусе. Знаки препинания, символы были удалены, как и служебные слова с использованием соответствующей функции пакета quanteda[2] и встроенного словаря служебных слов для русского языка, а также слова длиной менее 3 букв. Общее число уникальным лемм корпуса – 8 083. Мы рассматривали отдельно три датасета – объединенный корпус, куда вошли тексты двух жанров (1); описания картины (2); письма другу (3).

Наиболее частотные леммы анализируемых датасетов представлены в табл. 1 (жирным выделены совпадающие слова).

Таблица 1 – Самые частотные леммы датасетов

Самые частотные леммы объединенного датасета	Самые частотные леммы жанра «описание картины»	Самые частотные леммы жанра «письмо другу»	Леммы, встречающиеся в текстах обоих жанров (первые 50 по абсолютной частотности), ранг (картина/письмо)
очень, это, свой, время, хотеть, весь, женщина, картина, думать	женщина, картина, **свой**, молодой, **это**, мать, изобразить, который, мочь, **думать**	**очень**, **это**, **время**, твой, **весь**, **хотеть**, привет, новый, **свой**, **друг**	очень (12/1), это (5/2), весь (19/5), хотеть (17/6), свой (3/9), друг (38/10), время (31/3), хороший (46/12), знать (40/14), думать (10/27), видеть (25/43), жить (37/38)

Нами были построены контексты выделенных в столбце 4 табл. 1 лемм отдельно для трех датасетов со длиной окна, равной 6. Модель Glove была обучена на объединенном датасете с использованием настроек, приведенных в тьюториале к пакету[3].

В табл. 2 приведены для совпадающих лемм двух корпусов их «ближайшие соседи», значимо различающие группы по гендеру авторов ($p < 0{,}1$), а также «общие» соседи.

Таблица 2 – Данные о «ближайших соседях» лемм по датасетам

Лемма	Датасет								
	Объединенный корпус			Картина			Письмо		
	Общие	**Муж.**	**Жен.**	**Общие**	**Муж.**	**Жен.**	**Общие**	**Муж.**	**Жен.**
Очень	время, очень, новый, дело	–	скучать, надеяться, твой	–	–	–	время, твой, очень, дело	–	скучать, привет, давно, весь
Это	–	–	–	–	–	молодой, казаться	–	–	группа, говорить, это, надеяться,

[1] URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (дата обращения: 01.11.2021).
[2] URL: https://quanteda.io (дата обращения: 01.11.2021).
[3] URL: https://cran.r-project.org/web/packages/conText/vignettes/quickstart.html (дата обращения: 01.11.2021).

Лемма	Датасет								
	Объединенный корпус			Картина			Письмо		
	Общие	**Муж.**	**Жен.**	**Общие**	**Муж.**	**Жен.**	**Общие**	**Муж.**	**Жен.**
									пока, знать
Весь	весь, знать, время	работа, работать	очень, это, твой, год, хороший, рассказать, свой	–	–	казаться, девушка, взгляд, старушка	наш, время, весь	ждать, знать, работа, пока, дом	твой, это, новый, очень, дело, год, хороший
Хотеть	очень, знать, это, хотеть	–	весь надеяться твой	–	–	возможно	знать, очень весь	–	это, твой, весь, надеяться, хотеть
Свой	мочь, свой, это	–	который, весь, мать	женщина, молодой картина	–	дочь	–	–	писать
Друг	очень, время, друг	твой, привет, писать, давно	знать, весь, это, хотеть, год	знать, друг	план	свой, год	новый, видеться, друг, очень	твой	время, хотеть, знать, весь
Время	очень, время, хотеть, весь, знать	новый, свой, хотеться, год, работа	твой, надеяться, хороший, это, дело	время	изобразить, девушка	мать, который, это план молодой, женщина	время, новый, знать очень дело весь	учеба, работа, хотеть хотеться	рассказать, хороший, твой надеяться
Хороший	время, хороший, очень	–	новый, знать, хотеться, год	–	–	лицо	очень, хороший, учиться, хотеться	–	надеяться, учеба, новый, время, весь
Знать	знать	–	друг, это, весь	–	–	картина	знать	–	год, друг, весь
Думать	время, думать, свой, хотеть, это, год	жить, который	женщина, очень, мочь, знать	женщина	–	молодой, изобразить	скоро, работа, новый, хороший	надеяться, лето, знать	дело, хотеть, время, год, твой, очень
Видеть	–	–	–	–	–	девушка, взгляд	–	–	–
Жить	жить	–	это, дом, свой, год	–	–	–	хороший, весь, жить	время, дело	год, это, очень, хотеть, учиться, свой, новый

Пример визуализации результатов анализа для леммы «Новый», выполненный с использованием названного пакета, представлен на рис. 1, где звездочками обозначены статистически значимые отличия.

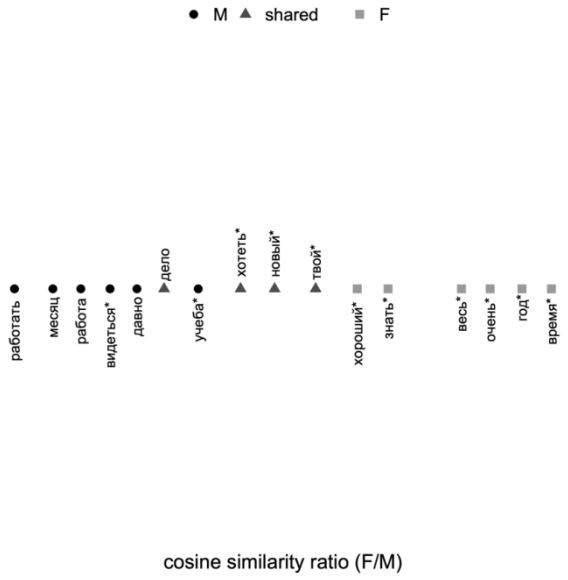

Рисунок 1 – Пример визуализации результатов сравнительного анализа «ближайших соседей» леммы «новый» в текстах мужчин и женщин

Нами был рассчитан индекс гендерной дифференциальности дистрибутивной семантики леммы как соотношение числа дифференцирующих элементов («ближайших соседей») к числу совпадающих элементов (табл. 3). Как видно из табл. 3, исследуемые леммы имеют более высокий индекс гендерной дифференциальности в датасете писем другу.

Таблица 3 – Индекс дифференциальности самых частотных лемм по датасетам

Лемма	Датасет		
	Объединенный	Картина	Письмо
очень	0,75	**0**	**1**
это	**0**	2	**6**
весь	**3**	**4**	**4**
хотеть	0,75	**1**	**1,67**
свой	**1**	0,33	**1**
друг	**3**	1,5	1
время	2	**8**	1,33
хороший	**1,33**	1	1,25
знать	3	**1**	**3**
думать	1	**2**	**2**
видеть	**0**	**1**	**0**
жить	**4**	**0**	1,4

В табл. 4–5 представлены результаты анализа гендерной дифференциальности значений самых частотных несовпадающих по жанрам лемм.

Таблица 4 – Гендерная специфика значений самых частотных лемм датасета «Письма другу»

Лемма	**Общие соседи**	**Мужские**	**Женские**	**Индекс гендерной дифференциальности**
Твой	ждать, рассказать, очень, твой, писать	–	ответ, скучать, новый, написать	0,8
Привет	привет, хороший, давно, твой, дело, видеться	здравствуй, друг	соскучиться, время, новый, очень	1,17
Новый	твой, знать, новый, дело	давно, видеться, учеба, хотеть, работа, кстати	год очень весь время хороший скоро	3
Дело	твой, очень, соскучиться, дело, видеться, давно	работа, поживать	новый, рассказать, хороший	1
День	весь, очень, это	–	наш, собираться, часто, давно	1,33
Год	год, очень, хороший, новый	–	скоро, хотеть	0,5
Учеба	учеба, дело, время, хороший, новый, лето	заниматься, хотеть	твой, очень, учиться	0,83

Таблица 5 – Гендерная специфика значений самых частотных лемм датасета «Описание картины»

Лемма	**Общие «соседи»**	**Мужские**	**Женские**	**Индекс гендерной дифференциальности**
Женщина	молодой, женщина	–	–	0
Картина	–	–	–	0
Молодой	молодой, взгляд	–	возможно	0,5
Мать	сын, мать, дочь	–	–	0
Изобразить	молодой	–	–	0
Мочь	мать, женщина, картина, мочь	–	сын, молодой, казаться	0,75
Взгляд	молодой, женщина	–	лицо	0,5

Таким образом, как показал проведенный эксперимент, у большинства проанализированных нами лемм значение, определяемое в контексте дистрибутивной семантики, обладает гендерной спецификой, однако для датасетов разного жанрового состава наблюдаются разные значения индекса гендерной дифференциальности у одних и тех же лемм. Наибольшим индексом гендерной дифференциальности обладают леммы в текстах жанра «письма другу», что, как представляется, объясняется тем, что названный жанр дает наибольшие возможности для конструирования гендерной идентичности в сравнении с жанром описание картины, а также тем, что контроль жанра дает возможность алгоритму обучиться более эффективно.

Выводы

Понимание того, как контекст влияет на значение слова, является критически важным для мно-

гих областей науки и практики, однако моделирование и оценка такого влияния представляют собой крайне сложную научную задачу, особенно в случае анализа текстов малого объема. Несмотря на стремительное развитие алгоритмов моделирования контекстного значения слова, многие вопросы в данной области остаются нерешенными, и в частности вопросы оценки влияния характеристик автора на значение слова. Названный вопрос важен для разработки более эффективных методов моделирования языка, в том числе для устранения влияния гендерных стереотипов, которые отражаются в контекстных значениях слов, как было показано в ряде новейших исследований.

Проведенный нами эксперимент показал перспективность использования основанных на регрессии методов выявления оценки влияния гендера автора на особенности контекстной семантики слова, реализованные в виде свободно распространяемых программных пакетов. Вместе с тем была показана необходимость учета жанра при проведении подобного анализа. В дальнейшем нами будет исследовано влияние совокупности факторов на контекстное значение слова, что возможно благодаря богатой метаразметке базы данных RusIdiolect. Кроме того, в дальнейших исследованиях будет проведено сравнение результатов, полученных с использованием дистрибутивных моделей, с данными ассоциативных словарей, содержащими демографическую информацию о респондентах.

Источник финансирования

Статья подготовлена при поддержке гранта Российского научного фонда № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики», выполняемого в Воронежском государственном педагогическом университете.

Конфликт интересов

Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Библиографический список

1. Антипенко А.А., Митрофанова О.А. Исследование ассоциативных связей слов в корпусе социальных сетей с помощью дистрибутивно-семантических моделей // Компьютерная лингвистика и вычислительные онтологии : труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19–22 июня 2019 г. СПб. : Университет ИТМО, 2019. Вып. 3. С. 77–91.

2. Гендерный дифференциальный психолингвистический словарь русского языка / науч. ред. И.А. Стернин, А.В. Рудакова. Воронеж : РИТМ, 2020. 198 с.

3. Пищальникова В.А. Интерпретация ассоциативных данных как проблема методологии психолингвистики // Вестник РУДН. Серия: Лингвистика. 2019. № 3.

4. Соколова Т.В. Ассоциативный словарь ребенка. Вербальные реакции детей 3–7 лет. Архангельск, 1996. Ч. 1: От стимула к реакции. 255 с.

5. Caliskan A., Bryson J., Narayanan A. Semantics derived automatically from language corpora contain human-like biases // Science. 2017. Vol. 356 (6334). P. 183–186.

6. The “Small World of Words”: English word association norms for over 12,000 cue words / S. De Deyne, D.J. Navarro, A. Perfors [et al.] // Behav Res. 2019. Vol. 51. P. 987–1006.

7. Garimella A., Banea C., Mihalcea R. Demographic-aware word associations // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. P. 2285–2295.

8. Gladkova A., Drozd A. Intrinsic evaluations of word embeddings: What can we do better? // Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, 2016. P. 36–42.

9. La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors / M. Khodak, N. Saunshi, Y. Liang [et al.] // ArXiv, abs/1805.05388.

10. Litvinova T. RusIdiolect: A New Resource for Authorship Studies // Lecture Notes in Networks and Systems. 2021. Vol. 186. P. 14–23.

11. Mandera P., Keuleers E., Brysbaert M. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation // J Mem Lang. 2017. Vol. 92. P. 57–78.

12. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, Kai Chen [et al.] // Advances in neural information processing systems. 2013. P. 3111–3119.

13. Pennington J. Glove: Global vectors for word representation / J. Pennington, R. Socher, Ch. D. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. pp. 1532-1543.

14. Pericбn-Pascual C. Measuring associational thinking through word embeddings // Artif Intell Rev. 2021. August.

15. Rodriguez P., Spirling A., Stewart B.M. Embedding Regression: Models for Context-Specific Description and Inference // Working papers Vanderbilt University, 2021.

References

1. Antipenko, A.A., Mitrofanova, O.A. (2019) Research of associative connections of words in the corpus of social networks using distributive semantic models. In: *Computational linguistics and computational ontologies : Proceedings of the XXII International Joint Scientific Conference "Internet and Modern Society", IMS-2019, June 19–22, St. Petersburg*. St. Petersburg, Universitet ITMO publ., vol. 3, pp. 77–91. (in Russian)

2. Sternin, I.A., Rudakova, A.V. (eds.) (2020) *Gendernyi differentsial'nyii psikholingvisticheskiii slovar' russkogo yazyka* [Gender differential psycholinguistic dictionary of the Russian language]. Voronezh, RITM publ. 198 p. (in Russian)

3. Pishchal'nikova, V.A. (2019) Interpretatsiya assotsiativnykh dannykh kak problema metodologii psikholingvistiki [Interpretation of associative data as a problem of psycholinguistics methodology]. *Vestnik RUDN. Seriya: Lingvistika*. (3). (in Russian)

4. Sokolova, T.V. (1996) *Assotsiativnyi slovar' rebenka. Verbal'nye reaktsii detei 3–7 let. Ch. 1: Ot stimula k reaktsii* [Associative dictionary of the child. Verbal reactions of children 3–7 years old. Part. 1]. Arkhangelsk. 255 p. (in Russian)

5. Caliskan, A., Bryson, J., Narayanan, A. (2017) Semantics derived automatically from language corpora contain human-like biases. *Science*. 356 (6334), 183–186.

6. De Deyne, S., Navarro, D.J., Perfors, A. e.a. (2019) The "Small World of Words": English word association norms for over 12,000 cue words. *Behav Res*. (51), 987–1006.

7. Garimella, A., Banea, C., Mihalcea, R. (2017) Demographic-aware word associations. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2285–2295.

8. Gladkova, A., Drozd, A. (2016) Intrinsic evaluations of word embeddings: What can we do better?. In: *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pp. 36–42.

9. Khodak, M., Saunshi, N., Liang, Y. e.a. La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors. *ArXiv, abs/1805.05388*.

10. Litvinova, T. (2021) RusIdiolect: A New Resource for Authorship Studies. *Lecture Notes in Networks and Systems*. (186), 14–23.

11. Mandera, P., Keuleers, E., Brysbaert, M. (2017) Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *J Mem Lang*. (92), 57–78.

12. Mikolov, T., Sutskever, I., Kai, Chen e.a. (2013) Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*, pp. 3111–3119.

13. Pennington, J., Socher, R., Manning, Ch. D. (2014) Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

14. Pericбn-Pascual, C. (2021) Measuring associational thinking through word embeddings. *Artif Intell Rev*.

15. Rodriguez, P., Spirling, A., Stewart, B.M. (2021) Embedding Regression: Models for Context-Specific Description and Inference. *Working papers Vanderbilt University*.

Поступила в редакцию 15.12.2021
Подписана в печать 28.03.2022

Original article
UDC 81'33
DOI 10.47438/2309-7078_2022_1_167

MODELING THE MEANING OF A WORD IN MALE AND FEMALE TEXTS BY DISTRIBUTIONAL SEMANTICS METHODS

Tatiana A. Litvinova[1]

Voronezh State Pedagogical University[1]
Voronezh, Russia

[1]*Cand. Philolog. Sci., researcher at the Regional Center for the Russian Language, e-mail: centr_rus_yaz@mail.ru*

Abstract. Modeling the meaning of a word in individual mental lexicon is one of the urgent problems of modern psycholinguistics. One of the most important aspects of this problem is to establish the influence of the group (including sociodemographic) characteristics of the author on semantics of the word in his mental lexicon. To solve this problem, a number of experimental methods are used (associative experiment, semantic scaling, etc.). However, conducting such experiments is associated with significant time and labor costs. To overcome this problem, corpus studies using algorithms and tools of distributive semantics are being developed aimed at establishing the features of the semantics of a word in the texts of persons with different sociodemographic characteristics (primarily men and women). Such studies, however, are still few in number, are carried out mainly on the material of the English language and require a large amount of corpus data. In addition, they do not pay attention to the genre characteristics of the studied texts. This study presents the results of a pilot experiment aimed at assessing the influence of the author's gender on the semantics of the word in texts of different genres using the latest methods of distributive semantics implemented in the conText package. The necessity of taking into account the genre when modeling the meaning of a word in the linguistic consciousness of people with different demographic characteristics is shown.

Keywords: distributive semantics, semasiology, corpus of texts, idiolect, demographic embeddings.

Cite as: Litvinova, T.A. (2022) Modeling the meaning of a word in male and female texts by distributional semantics methods. *Izvestia Voronezh State Pedagogical University*. (1), 167–173. DOI 10.47438/2309-7078_2022_1_167 (In Russ., abstract in Eng.)

Received 15.12.2021
Accepted 28.03.2022