«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы конференции молодых ученых
Иркутск, 14–16 ноября 2022 г.

ISBN 978-5-9624-2180-3

УДК 004(063)
ББК 32.973л0
Ц17

Публикуется по решению ученого совета ИФИЯМ ИГУ

«Цифра» в социально-гуманитарных исследованиях: метод, поле, реальность : материалы конференции молодых ученых. Иркутск, 14–16 ноября 2022 г. / [редкол.: С. Н. Гафарова [и др.]]. – Иркутск : Издательство ИГУ, 2023. – 1 электронный оптический диск (CD-ROM). – Заглавие с этикетки диска.

https://doi.org/10.26516/ 978-5-9624-2180-3.2023.1-73

ISBN 978-5-9624-2180-3

В материалах конференции отражены вопросы, связанные с ролью и местом цифровых технологий в современной гуманитаристике, прежде всего – лингвистике и социальной антропологии. В центре внимания находятся проблемы автоматической обработки текста, исследование новых форм реализации социальных интеракций, влияние информационных технологий на традиционные методы анализа текстовых массивов, изучение роли исследователя при работе с большими данными.

Предназначено для лингвистов, социологов, психологов, антропологов, литературоведов, культурологов и других специалистов в разных областях гуманитарного знания.

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Иркутский государственный университет»
664003, г. Иркутск, ул. К. Маркса, 1; тел. +7 (3952) 51-19-00
Издательство ИГУ, 664082, г. Иркутск, ул. Лермонтова, 124
тел. +7 (3952) 52-18-53; e-mail: izdat@lawinstitut.ru
Подписано к использованию 04.09.2023. Тираж 15 экз. Объем 9,3 Мб.

Тип компьютера, процессор, частота:	32-разрядный процессор, 1 ГГц или выше
Оперативная память (RAM):	256 МБ
Необходимо на винчестере:	320 МБ
Операционные системы:	ОС Microsoft® Windows® XP, 7, 8 или 8.1. ОС Mac OS X
Видеосистема:	Разрешение экрана 1024x768
Акустическая система:	Не требуется
Дополнительное оборудование:	Не требуется
Дополнительные программные средства:	Adobe Reader 6 или выше

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы конференции молодых ученых
Иркутск, 14–16 ноября 2022 г.

ISBN 978-5-9624-2180-3

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Иркутский государственный университет»
Институт филологии, иностранных языков и медиакоммуникации

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы конференции молодых ученых
Иркутск, 14–16 ноября 2022 г.

ISBN 978-5-9624-2180-3

Содержание

Предисловие

Межвузовская молодежная конференция молодых ученых «*Цифра* в социально-гуманитарных исследованиях: метод, поле, реальность» состоялась в Иркутском государственном университете 14–16 ноября 2022 г.

В конференции приняли участие молодые ученые и специалисты Института русского языка им. В. В. Виноградова РАН, Института системного программирования им. В. П. Иванникова РАН, Дипломатической академии МИД России, Научно-исследовательского университета «Высшая школа экономики», Московского государственного университета им. М. В. Ломоносова, Санкт-Петербургского государственного университета, Воронежского государственного университета, Новгородского государственного университета им. Ярослава Мудрого, Петрозаводского государственного университета, Тихоокеанского государственного университета, Иркутского государственного университета.

В рамках конференции обсуждался широкий круг вопросов, связанных с ролью и местом цифровых технологий в современной гуманитаристике, прежде всего – лингвистике и социальной антропологии.

- Что такое «цифра» для социально-гуманитарных наук сегодня?
- Остаются ли цифровые технологии лишь продвинутыми инструментами для ученых-гуманитариев, обеспечивающими эффективную обработку больших массивов данных, или «цифра» стала чем-то большим?
- Остается ли «цифра» лишь медиатором физического и социального пространства или она порождает новую реальность, в которой социальные интеракции реализуются в новых формах и с новым результатом?
- Какие подходы к анализу текста предлагают нам цифровые технологии? Привносят ли они нечто качественно новое в традиционную методику анализа текста?
- Какие принципиально новые типы информации о языковом знаке мы можем получать с помощью цифровых технологий?
- Как общедоступность цифровых технологий меняет архитектуру исследований / повышает верифицируемость исследований?
- Каковы пределы автоматической обработки текста? Какая роль отводится исследователю в работе с большими данными?
- Может ли «цифра» стать базой для интеграции гуманитарных наук?

Обсуждение названных вопросов происходило в ходе работы нескольких секций:

- «Цифровые технологии как поле гуманитарных исследований: проблемы описания и интерпретации цифровой среды»;
- «Цифровые технологии как инструмент гуманитарных исследований: использование различных компьютерных программ для анализа языковых данных»;

• «Точные гуманитарные науки: роль количественных данных в современном гуманитарном исследовании».

Самым дискуссионным проблемам были посвящены круглые столы «Новое время – новые методы, или о соотношении традиционных и современных методик анализа данных в гуманитарных науках»; «*Цифра* как основа интеграции гуманитарных наук: ДА или ДА, НО?..»

В настоящий сборник вошли избранные материалы докладов, отражающие все перечисленные аспекты работы конференции.

Корпусные методы в современной лексикографии получили разностороннее освещение в докладах Е. Э. Базарова «Корпус как инструмент для наблюдений лексикографа», А. Р. Пестовой «Корпусные методы в лексикографической стилистике», А. С. Кулевой «О корпусных методах в лексикографии: проблемы и перспективы». В статье **А. С. Кулевой** осмысляются результаты работы, направленной на словарное представление названий растений (фитонимов); это позволяет автору наглядно представить не только высокий потенциал корпусных данных, обеспечивающих исследователя большими массивами языкового материала, но и продемонстрировать ряд проблем, порождаемых широтой этого материала, а также наметить пути решения этих проблем.

В статьях А. Ю. Зюрик «Функции нефонологической долготы в устной речи», Н. А. Кошелюк «Цифровые методы сохранения и лингвистического анализа миноритарных языков России», А. Б. Чимитовой «Использование корпусных технологий для повышения верифицируемости исследования английских фразовых глаголов», А. А. Шаляпиной «Определение близкого родства дикторов посредством проведения фоноскопической экспертизы» демонстрируются возможности использования цифровых технологий как **инструмента** гуманитарных исследований.

А. Ю. Зюрик предпринимает попытку систематизировать фрагменты устной неподготовленной речи, в которых с помощью программы Praat зафиксирована нефонологическая долгота звуков, и предложить критерии для разграничения ее функций. **Н. А. Кошелюк** сосредоточивается на характеристике нескольких опций лингвистической платформы «ЛингвоДок», позволяющих повысить верифицируемость исследований и выйти на новый уровень фонетических, фонологических, ареальных и междисциплинарых исследований различных языков и диалектов. **А. Б. Чимитова** обосновывает эффективность использования корпусных технологий для уточнения представлений о достаточно подробно описанном классе лексических единиц – английских фразовых глаголов. **А. А. Шаляпина** обращается к одной из ключевых задач прикладной лингвистики – идентификации личности по голосу и речи, представляет результаты анализа функционально-диагностических комплексов устно-речевых навыков близкородственных говорящих (родных сестер), осуществленного с помощью комплексного применения методов формантного анализа и получения проекций на латентные структуры.

Статьи Э. Е. Бурова «Выявление сочетаемостных свойств слова с помощью статистических мер ассоциации», Т. И. Давидюк «Предикативное согласование с сочиненным подлежащим в русском языке: порядок слов, взаимное расположение конъюнктов и характеристики предиката», М. Н. Лошаниной «Анализ маркера противоречивости *все-таки* в устном дискурсе с применением программы MAXQDA», У. Э. Чекмез «Индивидуальные предпочтения говорящего как фактор, влияющий на выбор леммы для маркирования обратной связи (на примере единиц *ауа* и *мщм*)» посвящены вопросам, рассмотренным в рамках секции «Точные гуманитарные науки: роль количественных данных в современном гуманитарном исследовании».

Э. Е. Буров демонстрирует значимость использования различных статистических мер (MI-score, T-score, logDice) для выявления силы синтагматической связи между членами коллокации, характеризует достоинства и недостатки различных корпусных менеджеров, позволяющих автоматически выделять в корпусе (тексте) коллокации с заданным стержневым словом. **М. Н. Лошанина** представляет возможности программы MAXQDA, разработанной для анализа неструктурированных письменных и устных данных, для изучения дискурсивных маркеров, однако недостаточно хорошо известной отечественным лингвистам. **У. Э. Чекмез** сопоставляет количественное распределение маркеров обратной связи *ауа* и *мщм* в рамках исследовательского корпуса; с помощью статистических методов выявляет значимые отличия в имеющемся распределении и формулирует отдельные гипотезы, касающиеся факторов, которые обусловливают эти отличия.

Т. И. Давидюк рассматривает стратегии предикативного согласования с сочиненным подлежащим, содержащим личное местоимение *я*; с помощью ряда экспериментальных методик осуществляет факторный анализ приемлемости и дистрибуции различных вариантов согласования и выявляет значимые переменные, обусловливающие выбор той или иной стратегии. Следует подчеркнуть, что статья Т. И. Давидюк – это одно из пяти исследований, представленных на конференции коллективом авторов из МГУ, входящих в Московскую группу экспериментального синтаксиса под руководством Е. А. Лютиковой и А. А. Герасимовой (https://expsynt.com/team/). Для полноты картины назовем темы остальных докладов: А. А. Герасимова «Языковая вариативность в свете экспериментального синтаксиса», Д. Д. Врубель «Эффект синкретизма при предикативном согласовании с сочиненными конструкциями с повторяющимся союзом *и*: экспериментальное исследование»; Л. И. Паско «Интроспекция и эксперимент: возможно ли частичное согласование симметричных предикатов?»; К. А. Студеникина «Иерархия согласования при сочинении русских именных групп: экспериментальное исследование».

Работы, выполненные на разном языковом материале, но в рамках одной исследовательской парадигмы, позволили представить актуальное исследовательское направление, которое позволяет (в том числе) решить одну из актуальных задач современной теоретической лингвистики, предъявляющей «повышен-

ные требования к качеству содержательных обобщений над языковым материалом, которые служат основой для построения теоретической модели. В частности, важными оказываются надежность суждений о приемлемости или неприемлемости языковых выражений, построенных с учетом определенных грамматических оппозиций, и обнаружение систематического варьирования между носителями языка в этой области. В этой связи задача контролируемого сбора больших объемов языковых данных и их интерпретации становится все более актуальной не только для прикладных, но и для теоретических исследований. Решение этой задачи при построении синтаксической модели языка входит в компетенцию экспериментального синтаксиса – области исследований, ставящей перед собой задачу получения объективных данных о (не)приемлемости языковых выражений»[1].

В завершение конференции состоялась Школа молодых ученых «*Цифра* в социально-гуманитарных исследованиях: методы и технологии обработки информации». Мастер-классы «R для анализа данных в лингвистике», «Использование языка Pyton для обработки лингвистических данных» провели доцент факультета гуманитарных наук, заведующий Международной лабораторией языковой конвергенции НИУ ВШЭ Г. А. Мороз и доцент кафедры математической лингвистики филологического факультета СПбГУ О. В. Митренина.

Статья О. В. Митрениной «Нейросетевая классификация данных для диалектологов и социологов» завершает сборник, выполняя функцию своеобразного послесловия, которое не закрывает тему, а открывает новые перспективы, связанные с использованием нейронных сетей для решения разнообразных задач по классификации и анализу текстов. Предложенный автором перечень шагов, которые должны осуществить гуманитарии для входа в новую для них IT-область, делает статью О. В. Митрениной не эпилогом, но прологом – следующей конференции и следующего сборника статей.

Редколлегия

[1] Герасимова А. А., Лютикова Е. А. Лингвистический эксперимент на платформе Яндекс. Толока: оценка исследовательских возможностей // Zeitschrift für Slavische Philologie. 2022. Vol. 78, N 1. P. 175–206. URL: https://agerasimova.com/wp-content/uploads/Gerasimova_Lyutikova_ 2022_Linguistic_experiment_on_the_Yandex.Toloka_crowdsourcing_platform.pdf

УДК 81'139

Э. Е. Буров

Иркутский государственный университет, Иркутск, Россия

Выявление сочетаемостных свойств слова с помощью статистических мер ассоциации

Рассматриваются статистические методы выявления сочетаемостных свойств слова. Раскрывается статистическое понятие коллокации, демонстрируется важность этого понятия для теоретической и прикладной лингвистики. Описываются широко используемые в статистических исследованиях меры ассоциации (числовые показатели силы синтагматической связи между членами коллокации): MI-score, T-score, logDice. Подчеркивается нетождественность понятия «мера ассоциации» понятию «частота совместной встречаемости». Перечисляются различные корпусные менеджеры, позволяющие автоматически выделять в корпусе (тексте) коллокации с заданным стержневым словом, анализируются их преимущества и недостатки. С помощью корпусного менеджера CoCoCo анализируются особенности сочетаемости неопределенных местоимений *что-то* и *кое-что*. Делается вывод о том, что в наборе коллокатов этих местоимений отражаются их семантико-референциальные свойства.

Ключевые слова: коллокация, корпусная лингвистика, мера ассоциации, сочетаемость, MI-score, T-score, logDice.

Identification of Words' Syntagmatic Properties by Means of Statistical Measures of Association

The article discusses some statistical methods for identifying the syntagmatic properties of words. It explains the statistical notion of collocation and demonstrates the importance of this notion for theoretical and applied linguistics. Different measures of association (numbers that indicate the strength of the syntagmatic connection between the members of a collocation) such as MI-score, T-score and logDice are described. It is emphasized that the term «measure of association» is not equivalent of the term "frequency of co-occurrence". Various corpus managers that make it possible to identify collocations with a given core word automatically are mentioned. It is pointed out that each of these managers has its advantages and disadvantages. By means of the CoCoCo corpus manager, the syntagmatic properties of Russian indefinite pronouns *chto-to* 'something' and *koe-chto* 'something' are analyzed. In conclusion, it is stated that the set of these pronouns' collocates reflects their referential properties.

Keywords: collocation, corpus linguistics, measure of association, co-occurrence, MI-score, T-score, logDice.

В современной корпусной лингвистике широко используются статистические методы выявления сочетаемостных свойств слов. Рассмотрим некоторые из этих методов. Начнем с понятия «коллокация». В статистически ориентированных лингвистических исследованиях под коллокацией принято понимать комбинацию двух или более слов, имеющих тенденцию к совместной встречаемости [Захаров, Хохлова, 2010, с. 137]. Коллокация – это сочетание некоторых лексических единиц, которое встречается в текстах чаще, чем оно встречалось бы при случайном соединении этих единиц.

Что значит «неслучайно частая совместная встречаемость»? Пусть имеется лексическая единица *А* с частотой встречаемости 25 ipm (употреблений на миллион слов) и лексическая единица *Б* с частотой встречаемости 12 ipm. При случайном соединении слов сочетание *А* + *Б* должно иметь частоту встречаемости 0,000025 × 0,000012 = 0,0000000003, т. е. 0,0003 ipm. Если фактическая частота встречаемости сочетания *А* + *Б* выше 0,0003 ipm, считается, что единицы *А* и *Б* совместно встречаются неслучайно часто, а значит, сочетание *А* + *Б* является коллокацией.

Коллокатами некоторой лексемы называются слова, образующие с ней коллокации. Сама лексема называется стержневым (основным, главным) словом коллокации, или ключом. Коллокаты лексемы различаются по силе их синтагматической связи со стержневым словом. Так, согласно исследованию [Захаров, Хохлова, 2010], существительное *работа* имеет следующие коллокаты (среди прочих): *совместная*, *научная*, *режиссерская*. При этом с первым прилагательным оно связано наиболее сильно, со вторым – слабее, с третьим – еще слабее.

Считается, что по коллокатам лексемы можно ее «вычислить». Знаменитый исследователь коллокаций Дж. Ферс говорил: “You shall know a word by the company it keeps” [Firth, 1957]. Даже члены синонимического ряда различаются своими коллокатами, причем не только наборами коллокатов, но и силой синтагматической связи с одними и теми же коллокатами. Так, прилагательное *алый* синтагматически гораздо сильнее связано с существительным *парус*, чем синонимичное ему прилагательное *красный*.

Как можно измерить силу синтагматической связи между лексемой и ее коллокатами? Статистическим показателем силы коллокации выступает так называемая мера ассоциации. Это число, указывающее на величину синтагматической связи между словами. Если мера ассоциации для слов *А* и *Б* больше некоторого порогового значения, сочетание *А* + *Б* считается коллокацией.

Благодаря мерам ассоциации оказывается возможным выявлять коллокаты слов, сравнивать коллокаты одного и того же слова по силе их синтагматической связи с этим словом и получать таким образом объективную информацию о сочетаемостных свойствах лексических единиц языка. Такая информация важна не только специалистам, занимающимся теоретическим описанием языка, но и тем, кто занят в прикладных областях лингвистики: лексикографам, компьютерным лингвистам (например, она используется в программах предиктивного набора текста).

Отметим, что мера ассоциации между элементами сочетания – это не то же самое, что частота встречаемости этого сочетания. Высокая мера ассоциации наблюдается не тогда, когда сочетание встречается часто, а тогда, когда оно встречается нетривиально часто, т. е. чаще, чем оно встречалось бы при случайном соединении его компонентов.

Существует множество мер ассоциации, высчитываемых по разным формулам[1]. Самая простая мера ассоциации – мера взаимной информации (Mutual

[1] См. сравнение нескольких наиболее известных метрик в [Хохлова, 2017].

Information Score, MI-score). Показатель MI-score зависит от частоты встречаемости сочетания слов, независимых частот его членов, а также от общего числа словоформ в исследуемом корпусе (тексте):

$$\text{MI-score} = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}, \quad (1)$$

где n – стержневое слово; c – коллокат; f(n,c) – частота встречаемости стержневого слова n в паре с коллокатом c; f(n), f(c) – независимые (абсолютные) частоты стержневого слова n и слова c в корпусе (тексте); N – общее число словоформ в корпусе (тексте).

Если показатель MI-score больше 3, сочетание слов считается коллокацией.

Другая часто используемая в статистике мера ассоциации – T-score. Показатель T-score высчитывается по следующей формуле:

$$\text{T-score} = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}. \quad (2)$$

Если показатель T-score больше 3, сочетание слов считается коллокацией.

Для больших корпусов чешскими лингвистами была специально разработана мера ассоциации logDice, см. [Rychlý, 2008]. Показатель logDice не зависит от объема корпуса, поэтому при использовании logDice оказывается возможным сравнивать показатели, полученные при работе с разными корпусами. Метрика высчитывается так:

$$\text{logDice} = 14 + \log_2 \frac{2 \times f(n,c)}{f(n) + f(c)}. \quad (3)$$

Максимальное значение logDice – 14 (оно достигается, если стержневое слово и его коллокат всегда употребляются совместно). Если показатель logDice больше нуля, сочетание слов считается коллокацией.

Существуют корпусные менеджеры, которые способны, анализируя загруженный в них корпус текстов, на основе приведенных выше формул автоматически составлять списки наиболее сильных коллокатов для заданных слов – так называемые скетчи. Примером может служить разработанный чешскими специалистами менеджер Sketch Engine (https://www.sketchengine.eu/). Sketch Engine работает с корпусами самых разных языков, в том числе с русским. Из сервисов, использующих только русские корпусные материалы, можно выделить ресурс CoCoCo (Collocations, Colligations and Corpora) (https://cococo.cosyco.ru/). Возможность автоматического извлечения коллокаций появилась в ноябре 2022 г. в корпусе региональных СМИ Национального корпуса русского языка (https://ruscorpora.ru/search/regional/). Sketch Engine платный и требует регистрации, в то время как CoCoCo и Национальный корпус русского языка бесплатны и регистрации не предполагают. У Sketch Engine, однако, есть свои преимущества. Например, объем веб-корпуса Russian Web 2011 (ruTenTen11), предустановленного в менеджере Sketch Engine, превышает 18 млрд словоупотреблений,

в то время как самый большой по объему корпус из предустановленных в CoCoCo – интернет-корпус Taiga – содержит лишь около 6 млрд словоупотреблений, а корпус региональных СМИ Национального корпуса русского языка – 24,5 млн словоупотреблений.

Приведем пример использования корпусного менеджера для автоматического извлечения коллокаций. С помощью менеджера CoCoCo можно выделить наиболее сильные постпозитивные адъективные коллокаты местоимения *что-то*, т. е. те постпозитивные прилагательные, с которыми *что-то* наиболее сильно связано: *новый*, *похожий*, *важный*, *странный*, *подобный* (в интернет-корпусе I-RU объемом 140 млн словоупотреблений); *новый*, *другой*, *подобный*, *важный*, *больший* (в интернет-корпусе Taiga объемом 6 млрд словоупотреблений)[1].

Интересно, что указанные коллокаты отличаются от наиболее сильных постпозитивных адъективных коллокатов местоимения *кое-что*: *интересный*, *новый*, *новенький*, *полезный* (в интернет-корпусе I-RU); *интересный*, *другой*, *важный*, *хороший*, *общий* (в интернет-корпусе Taiga). Можно заметить, что *кое-что* склонно сочетаться с прилагательными, предполагающими некоторое знакомство с объектом (*интересный*, *полезный*). Вероятно, в особенностях сочетаемости этого местоимения отразился его слабоопределенно-референтный статус, т. е. его свойство обозначать объект, (предположительно) неизвестный слушающему, но известный говорящему.

Таким образом, корпусные методы в современной лингвистике имеют большой эвристический потенциал. Компьютерные технологии открывают новые возможности для количественно-статистического анализа языкового материала. Результаты этого анализа могут быть подвергнуты лингвистической интерпретации.

Литература

Захаров В. П., Хохлова М. В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. М., 2010. Вып. 9 (16). С. 137–143.

Хохлова М. В. Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных // Компьютерная лингвистика и вычислительные онтологии. СПб., 2017. Вып. 1. С. 166–171.

Firth J. R. A synopsis of linguistic theory, 1930–1955 // Studies in Linguistic Analysis [Special volume of the Philological Society]. Oxford : Blackwell, 1957. P. 1–32.

Rychlý P. A lexicographer-friendly association score // Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008. Brno : Masaryk University, 2008. P. 6–9.

[1] Прилагательные расположены в порядке убывания показателя T-score.

УДК 811.1/.2

Т. И. Давидюк

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Предикативное согласование с сочиненным подлежащим в русском языке: порядок слов, взаимное расположение конъюнктов и характеристики предиката[1]

Рассматривается предикативное согласование с сочиненным подлежащим, содержащим личное местоимение *я*. В русском языке наблюдается зависимость использования стратегий предикативного согласования от порядка слов. Кроме того, некоторые исследователи указывают характеристики предиката как фактор, способный повлиять на выбор стратегии согласования с сочиненным подлежащим. Нами было проведено экспериментальное исследование, направленное на выявление уровня приемлемости различных стратегий согласования в зависимости от следующих факторов: порядка слов (SV, VS), взаимного расположения конъюнктов (*я и X*, *X и я*), характеристик предиката (временны́х, видовых характеристик, аргументной структуры). Результаты экспериментов показали, что в русском языке с сочиненным подлежащим, одним из конъюнктов которого является личное местоимение *я*, возможны три стратегии согласования: согласование по правилам разрешения, согласование с ближайшим конъюнктом и дефолтное согласование по 3-му лицу, множественному числу (для предикатов в непрошедшем времени). Было выявлено, что выбор перечисленных стратегий зависит от ряда факторов. Значимыми факторами оказались порядок слов и взаимное расположение конъюнктов, однако их влияние зависит от временны́х характеристик предиката. Видовая характеристика предиката, а также его аргументная структура (является он неэргативным или неаккузативным) значимо не повлияли на уровень приемлемости стратегий согласования.

Ключевые слова: предикативное согласование, сочиненное подлежащее, русский язык, иерархия лица, порядок слов.

Predicative Agreement With the Composed Subject in Russian: Word Order, Mutual Arrangement of Conjuncts and Predicate Characteristics

The article deals with predicate agreement with a coordinate subject containing the personal pronoun *я* ‘I’. In Russian, there is a dependence of the use of certain strategies of predicative agreement on word order. In addition, some researchers indicate the characteristics of the predicate as a factor that can influence the choice of agreement strategy with a coordinate subject. I present an experimental study aimed at identifying the level of acceptability of various agreement strategies depending on the following factors: word order (SV, VS), position of conjuncts (*я и X* ‘I and X’, *X и я* ‘X and I’), characteristics of the predicate (temporal and aspect characteristics, argument structure). The results of the experiments showed that in the Russian constructions with coordinate subjects containing a personal pronoun, three agreement strategies are possible – resolution agreement, closest conjunct agreement, and default 3rd plural agreement (in the non-past tense). It was found that the choice of these strategies depends on a number of factors. Word order and position of conjuncts turned

[1] *Исследование выполнено за счет гранта Российского научного фонда № 22-18-00037, реализуемого в МГУ имени М. В. Ломоносова. URL: https://rscf.ru/project/22-18-00037/.*

out to be significant factors, but their influence depends on the temporal characteristics of the predicate. The aspect characteristic of the predicate, as well as its argument structure (whether it is unergative or unaccusative) did not significantly affect the level of acceptability of agreement strategies.
Keywords: predicate agreement, coordinate subject, the Russian language, person hierarchy, word order.

Введение

Данная статья посвящена предикативному согласованию с сочиненным подлежащим, одним из конъюнктов которого является в русском языке личное местоимение *я*. Начиная с [Chomsky, 2000] согласование в минималистском синтаксисе связывается с работой Agree – операции, отвечающей за перемещение значений признаков с одной единицы на другую. В настоящее время не существует консенсуса относительно квалификации данной операции. В частности, нерешенным остается вопрос, является ли Agree синтаксической или постсинтаксической операцией. Для решения обозначенной проблемы и привлекаются данные по согласованию с сочиненными именными группами, так как во многих языках такое согласование оказывается чувствительным к линейной позиции конъюнктов и мишени согласования, следовательно, как предполагается, согласование происходит после осуществления процесса линеаризации и без учета с-командования.

Предикативное согласование с сочиненным подлежащим в русском языке демонстрирует значительную вариативность. Для личного согласования с сочиненным подлежащим нормативные грамматики (см., в частности, [Русская грамматика, 1980, с. 243–244]) указывают на действие так называемой иерархии лиц. Если в состав сочиненного подлежащего входит личное местоимение 1-го лица, то согласование будет происходить по 1-му лицу (1), если местоимение 2-го лица – то по 2-му лицу (2); при сочинении личных местоимений 1-го и 2-го лица грамматично согласование по 1-му лицу (3). В [Corbett, 1983a] показывается, что действующее в русском языке правило согласования по лицу с сочиненным подлежащим – Г. Корбетт называет его и подобные правила для других признаков правилами разрешения (англ. *resolution rules*) – справедливо и для ряда других языков.

(1) *Я и он* ***придём***.

(2) *Ты и твои родные не* ***едете***.

(3) *Ни я, ни ты не* ***едем***. [Русская грамматика, 1980, с. 244]

Тем не менее иерархия согласования по лицу в русском языке может быть нарушена. Очевидный случай возможности нарушения иерархии представляют собой конструкции с порядком слов VS (4). Кроме того, в [Пекелис, 2013; Санников, 2008, с. 161] указывается, что нарушению иерархии могут способствовать такие факторы, как использование некоторых двойных сочинительных союзов (5) и семантическая разнородность конъюнктов (6).

(4) ***Остаются*** *учителя старших классов и ты* [Пекелис, 2013].

(5) *Но гораздо лучше, когда не столько ты, сколько другие* ***считают*** / ****считаете*** *тебя красивой* [Пекелис, 2013].

(6) *Его* ***погубит*** / ****погубите*** *любовь к выпивке и ты* [Санников, 2008, с. 161].

Что касается согласования по числу, то на него оказывает влияние бóльшее число факторов. Помимо порядка слов (см. корпусное исследование в [Corbett, 1983b]), значимым фактором является одушевленность конъюнктов, ср. (7а–в).

(7) а. *Во всем был* ***виден*** / ??***видны*** *точный расчет и удивительнаяцеле устремленность.*
б. *Отсюда мне* ***виден*** / ?***видны*** *дом и опушка леса.*
в. *Отсюда мне* ***видны*** / ??***виден*** *Коля и Маша* [Санников, 2008, 157–158].

Из других свойств самих конъюнктов отмечается влияние совпадения или несовпадения рода, ср. (8а–б). Оказываются значимыми и характеристики предиката: так, как утверждается в [Санников, 2008, с. 160], при порядке слов VS согласование с ближайшим конъюнктом более предпочтительно при предикате непрошедшего времени, чем при предикате прошедшего времени, ср. (9а–б).

(8) а. *Белена и крапива* ***росла*** *прямо под окнами.*
б. ?*Бурьян и крапива* ***росла*** *прямо под окнами* [Санников, 2008, с. 158].

(9) а. *Прямо у крыльца* ***растет*** *большая сосна и дуб.*
б. ?*Прямо у крыльца* ***росла*** *большая сосна и дуб* [Санников, 2008, с. 160].

Для изучения предикативного согласования с сочиненным подлежащим экспериментальные методы выглядят достаточно перспективно, так как они позволяют зафиксировать значимые переменные и проводить факторный анализ приемлемости и дистрибуции различных вариантов согласования. Мы сосредоточимся на предикативном согласовании с подлежащим вида *я и X* и *X и я*, где X – имя собственное мужского рода. В Национальном корпусе русского языка (НКРЯ) в подавляющем большинстве предложений с подобными сочиненными подлежащими нами зафиксировано согласование по 1-му лицу, множественному числу (в непрошедшем времени) или согласование по множественному числу (в прошедшем времени). Согласование с ближайшим конъюнктом при порядке слов VS отмечается в сравнительно небольшом количестве предложений. В нескольких предложениях из корпуса нами зафиксировано согласование по 3-му лицу, множественному числу – (10), (11). Такое согласование не отмечается в нормативных грамматиках русского языка. В работе [Санников, 2008, с. 161] согласование по 3-му лицу, множественному числу признается устаревшим, однако в НКРЯ подобные примеры ((10), (11)) встречаются в текстах, созданных относительно недавно. Так как согласование по правилам разрешения (согласование по 1-му лицу, множественному числу для непрошедшего времени и согласование по множественному числу для прошедшего времени) наиболее частотно в НКРЯ, по сравнению с другими вариантами согласования, задача определения уровня приемлемости возможных стратегий согласования представляется довольно значимой.

(10) *Я и мои товарищи от всего сердца* ***благодарят*** *Вас за достигнутые успехи* (НКРЯ. Сергей Довлатов. Компромисс (1981–1984 гг.)).

(11) *Секретный код* ***знают*** *только я и Сергей Петрович* (НКРЯ. Андрей Ростовский. По законам волчьей стаи (2000 г.)).

Экспериментальное исследование

Наше исследование состоит из четырех экспериментов, методика которых заключалась в оценке приемлемости предложений от 1 до 7. Эксперименты проводились нами на платформе PCIbex. Набор респондентов осуществлялся через социальную сеть «ВКонтакте» и платформу «Яндекс.Толока». Обработка и визуализация полученных результатов проводилась в программе RStudio. Оценки предложений были приведены к стандартизированным z-оценкам. Кроме того, был произведен отсев респондентов по следующим параметрам:

- отсев по отклоняющимся ответам: а) по сумме квадратов отклонений от «эталонных» оценок в 6 и 2 балла для грамматичных и неграмматичных филлеров соответственно (если данное значение у некоторого респондента отличается более чем на два стандартных отклонения, то результаты данного респондента исключаются из дальнейшего анализа); б) по стандартному отклонению от средних оценок для филлеров (если несколько оценок некоторого респондента не попали в заданный интервал, то результаты данного респондента исключаются из дальнейшего анализа);
- отсев по временно́му критерию (если у некоторого респондента было несколько быстрых ответов длительностью меньше 300 миллисекунд, то результаты данного респондента исключаются из дальнейшего анализа);
- отсев по контрольным вопросам на внимательность к некоторым стимульным предложениям (если некоторый респондент ошибся в двух или более вопросах, то результаты данного респондента исключаются из дальнейшего анализа).

В качестве метода статистического исследования влияния переменных нами был выбран регрессионный анализ с использованием смешанных линейных моделей. Смешанные линейные модели позволяют учесть влияние случайных факторов, в нашем случае это были особенности конкретной лексикализации или предпочтений респондента. Подбор смешанных линейных моделей осуществлялся нами вручную; после обнаружения модели, которая наиболее эффективно описывает данные, мы производили множественное попарное сравнение между условиями при помощи теста Тьюки.

В каждом эксперименте использовались филлерные предложения, соотношение которых к экспериментальным предложениям составляло 1:1. Филлеры делились на грамматичные и неграмматичные. В экспериментах использовались непереходные предикаты. На каждое условие приходилось по восемь лексикализаций, формирование экспериментальных листов производилось по правилу латинского квадрата.

В экспериментах № 1 и 2 мы фиксировали порядок конъюнктов (подлежащее имело вид *я и X*), но при этом варьировали порядок слов. В эксперименте № 1 были использованы предикаты непрошедшего времени, он состоял из 64

предложений. Эксперимент № 2 содержал предикаты прошедшего времени и состоял из 32 предложений. Пример одной из лексикализаций для эксперимента № 1 приведен в (12а–б), для эксперимента № 2 − в (13а–б). Грамматичные филлеры в обоих экспериментах содержали сочиненное подлежащее, состоящее из имен собственных мужского рода (например, *Вася и Петя*), согласование с которым осуществлялось по 3-му лицу, множественному числу (для эксперимента № 1) или по множественному числу (для эксперимента № 2). Неграмматичные филлеры в эксперименте № 1 содержали ошибку в предикативном согласовании: с подлежащим, состоящим из двух имен собственных мужского рода, «согласование» осуществлялось по 1-му лицу, множественному числу (*Дима и Витя убираем на кухне*). В эксперименте № 2 неграмматичные филлеры содержали ошибку в предложном управлении (например, *Кирилл и Федя ушли по киоск*). Предикаты в обоих экспериментах были сбалансированы по виду.

(12) а. *Я и Вася приду/придём/придёт/придут на вечеринку.*
б. *На вечеринку приду/придём/придёт/придут я и Вася.*

(13) а. *Я и Вася пришёл/пришли на вечеринку.*
б. *На вечеринку пришёл/пришли я и Вася.*

В эксперименте № 1 приняли участие 84 респондента (35 женщин, 49 мужчин), средний возраст респондентов составил 36 лет. На рис. 1 представлены результаты эксперимента в виде графика взаимодействия.

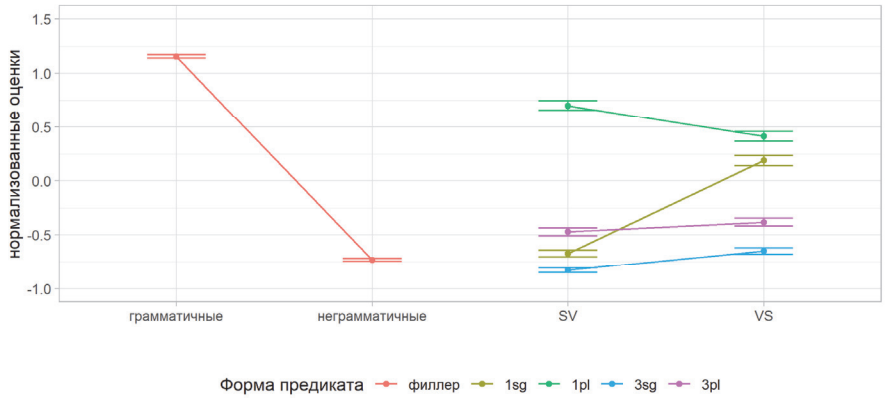


Рис. 1. График взаимодействия факторов эксперимента № 1

Наиболее высоко оценивается согласование по 1-му лицу, множественному числу, однако на него влияет порядок слов: при порядке VS оно оценивается значимо ниже ($p < 0{,}05$). При порядке слов VS также достаточно высокие оценки получило согласование по 1-му лицу, единственному числу, тем не менее его оценки значимо ниже оценок для согласования по 1-му, лицу множественному числу ($p = 0{,}01$). Оценки для согласования с первым конъюнктом при порядке слов SV, а также для согласования по 3-му лицу, единственному числу при

обоих порядках слов значимо не отличаются от оценок для неграмматичных филлеров. Интересно, что оценки для согласования по 3-му лицу, множественному числу при порядке слов SV значимо отличаются от оценок для согласования по 3-му лицу, единственному числу ($p < 0{,}05$). Глагольный вид оказался незначимым фактором для данного эксперимента.

В эксперименте № 2 поучаствовали 40 человек (23 женщины, 17 мужчин); средний возраст респондентов – 30 лет. На рис. 2 представлен график взаимодействия полученных для эксперимента № 2 оценок.

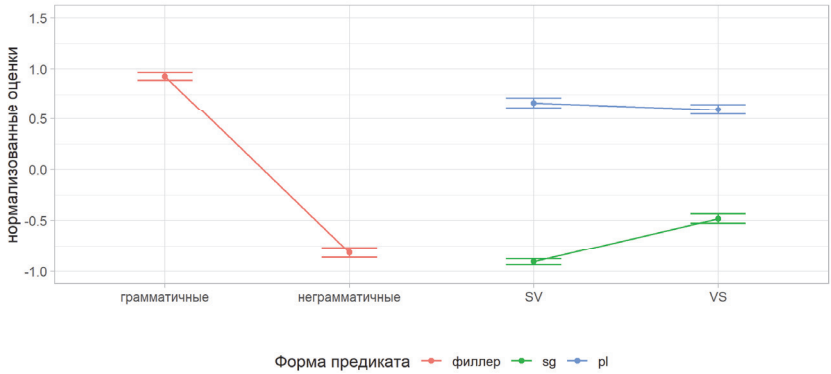


Рис. 2. График взаимодействия факторов эксперимента № 2

Наиболее высокие оценки получило согласование по множественному числу. Примечательно, что на его оценки порядок слов не оказывает существенного влияния ($p = 0{,}70$). Однако порядок слов важен для согласования по единственному числу: при порядке слов SV согласование по единственному числу оценивается на уровне оценок неграмматичных филлеров ($p = 0{,}59$), тогда как при порядке слов VS оценки для согласования по единственному числу значимо выше и оценок для неграмматичных филлеров, и оценок для согласования по единственному числу при порядке слов SV. Вид предиката, как и в эксперименте № 1, оказался незначим.

В экспериментах № 3 и 4 мы фиксировали порядок слов, взяв предложения с порядком слов VS, но при этом варьировали порядок конъюнктов: подлежащее имело вид *я и Вася* или *Вася и я*. В эксперименте № 3 были использованы предикаты непрошедшего времени, он состоял из 64 предложений. Эксперимент № 4 содержал предикаты прошедшего времени и состоял из 32 предложений. Пример одной из лексикализаций для эксперимента № 3 приведен в (14а–б), для эксперимента № 4 – в (15а–б). Конъюнкты грамматичных филлеров в обоих экспериментах не были рассогласованы по лицам, и согласование с ними было грамматично (*придут*/*пришли Петя и Вася*). В неграмматичных филлерах имелась ошибка в предикативном согласовании (*придём*/*приду Петя и Вася*) или в согласовании между прилагательным и существительным в предложной группе (*в*

правом углах). Предикаты в обоих экспериментах были сбалансированы по аргументной структуре: в одной половине предложений были представлены неэргативные предикаты (*танцевать*, *работать*, *идти* и т. д.), в другой половине − неаккузативные предикаты (*спать*, *оставаться*, *мерзнуть* и т. д.).

(14) а. *В районном центре живу/живём/живёт/живут я и Петя.*
б. *В районном центре живу/живём/живёт/живут Петя и я.*

(15) а. *В районном центре жил/жили я и Петя.*
б. *В районном центре жил/жили Петя и я.*

В эксперименте № 3 приняли участие 85 человек (34 женщины, 48 мужчин), средний возраст которых составил 38 лет. Результаты эксперимента № 3 в виде графика взаимодействия отражены на рис. 3.

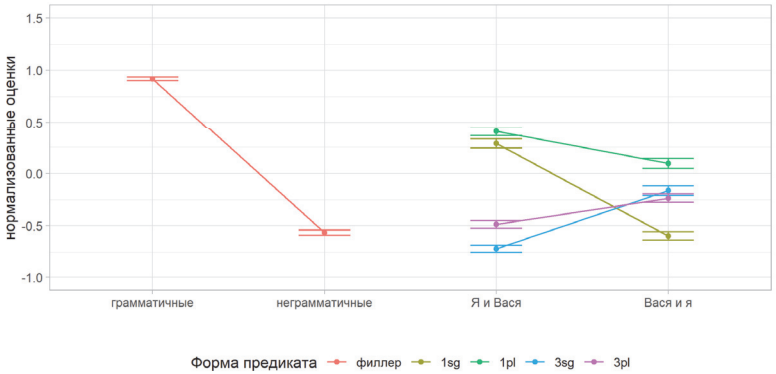


Рис. 3. График взаимодействия факторов эксперимента № 3

При порядке конъюнктов *я и Вася* наиболее высокие оценки получили согласование по 1-му лицу, множественному числу и согласование по 1-му лицу, единственному числу, разница между ними незначима ($p = 0{,}98$). При порядке *Вася и я* наиболее высоко оценивалось согласование по 1-му лицу, множественному числу. Для согласования по 1-му лицу, множественному числу значим порядок конъюнктов: при порядке *Вася и я* такое согласование оценивается значимо ниже. Для согласования с первым конъюнктом фактор порядка слов также значим: для порядка *я и Вася* оно оценивается выше, чем для порядка *Вася и я*. Таким образом, можно заключить, что личное местоимение *я* является более сильным аттрактором в роли контролера согласования по сравнению с именем собственным. Интересно, что при порядке слов *Вася и я* оценки для согласования по 3-му лицу, множественному числу значимо не отличаются от оценок для согласования с первым конъюнктом ($p = 0{,}40$). Оценки для согласования по 3-му лицу, множественному числу оказываются значимо более высокими, чем оценки для согласования со вторым конъюнктом. Фактор аргументной структуры предиката (является он неэргативным или неаккузативным) оказался незначим (см.

[Babyonyshev, 1996] о гипотезе влияния аргументной структуры предиката на согласование с первым конъюнктом в русском языке).

В эксперименте № 4 приняли участие 43 респондента (15 женщин, 28 мужчин). Результаты эксперимента № 4 представлены на рис. 4.

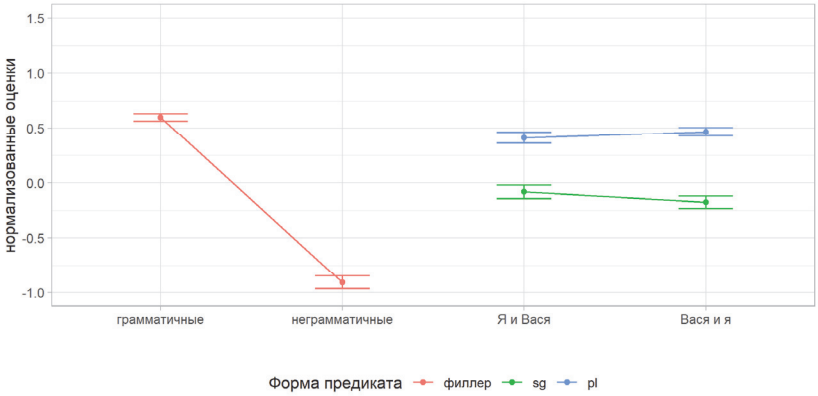


Рис. 4. График взаимодействия факторов эксперимента № 4

Наиболее высоко оценивается согласование по множественному числу. Ниже него оценивается согласование по единственному числу. Порядок конъюнктов для эксперимента № 4 оказывается незначим: у предиката в прошедшем времени отсутствует признак лица, следовательно, изменение порядка двух рассогласованных по признаку лица конъюнктов ожидаемо не повлияло на распределение стратегий согласования. Как и в эксперименте № 3, фактор аргументной структуры предиката оказался незначим.

Заключение

В данной статье мы описали результаты четырех лингвистических экспериментов, направленных на исследование предикативного согласования с сочиненным подлежащим, содержащим личное местоимение *я*. Согласно результатам экспериментов в русском языке с подобным подлежащим возможны три стратегии согласования: согласование по правилам разрешения (согласование по 1-му лицу, множественному числу в непрошедшем времени и согласование по множественному числу в прошедшем времени), согласование с первым конъюнктом при порядке слов VS и (для непрошедшего времени) согласование по 3-му лицу, множественному числу. Согласование с последним конъюнктом для исследуемых конструкций можно признать неграмматичным.

Распределение выявленных экспериментальным методом стратегий согласования с сочиненным подлежащим, одним из конъюнктов которого является местоимение *я*, зависит от порядка слов, порядка конъюнктов и временны́х характеристик предиката. При порядке слов VS становится возможным согласова-

ние с первым конъюнктом, однако в непрошедшем времени оно оценивается заметно более высоко, чем в прошедшем времени. Кроме того, в непрошедшем времени на уровень приемлемости оказывает влияние порядок конъюнктов, тогда как в прошедшем времени такого эффекта не наблюдается. Чувствительность согласования с первым конъюнктом при порядке слов VS к временны́м характеристикам предиката и морфосинтаксическому классу первого конъюнкта еще предстоит учесть в синтаксических моделях согласования с первым конъюнктом. Примечательно, что согласование с первым конъюнктом в непрошедшем времени оценивается примерно на том же уровне, что и согласование по 1-му лицу, множественному числу. В типологическом исследовании [Nevins, Weisser, 2019] о согласовании с ближайшим конъюнктом в языках мира было сделано эмпирическое обобщение о том, что разные признаки с различной вероятностью вызывают согласование только с одним из конъюнктов, и было показано, что признак лица реже вызывает согласование с ближайшим конъюнктом по сравнению с другими признаками. Полученные нами данные относительно согласования с первым конъюнктом при порядке слов VS несколько противоречат этому обобщению.

Представляется любопытным обнаруженное согласование по 3-му лицу, множественному лицу, так как механизм его возникновения еще предстоит смоделировать. Возможными направлениями работы по обозначенной проблематике могут стать исследования о дефолтности 3-го лица и о том, что механизм согласования в принципе расположен вне грамматики (см. [Lyskawa, 2021]), поэтому правила разрешения могут давать сбой.

Литература

Пекелис О. Е. Сочинение // Материалы для проекта корпусного описания русской грамматики. 2013. URL: http://rusgram.ru/Сочинение (дата обращения: 11.12.2022).

Русская грамматика. Т. 2. Синтаксис / Н. Ю. Шведова (гл. ред.). М. : Наука, 1980. 714 с.

Санников В. З. Русский синтаксис в семантико-прагматическом пространстве. М. : Языки славянских культур, 2008. 624 с.

Babyonyshev M. Structural connections in Syntax and Processing: Studies in Russian and Japanese. PhD thesis. Cambridge, 1996. 293 p.

Chomsky N. Minimalist inquiries: The framework // Step by step: Essays on Minimalist syntax in honor of Howard Lasnik / eds. R. Martin, D. Michaels, J. Uriagereka. Cambridge, MA : MIT Press, 2000. P. 89–155.

Corbett G. Resolution rules: agreement in person, number, and gender // Order, Concord and Constituency / eds. G. Gazdar, E. Klein, G. K. Pullum. Dordrecht : Foris, 1983a. P. 175–205.

Corbett G. Hierarchies, targets and controllers: Agreement patterns in Slavic. London : Croom Helm, 1983b. 272 p.

Lyskawa P. Coordination without grammar-internal feature resolution. PhD thesis. College Park, 2021. 329 p.

Nevins A., Weisser P. Closest conjunct agreement // Annual Review of Linguistics. 2019. N 5. P. 219–241.

УДК 8'81

А. Ю. Зюрик

Иркутский государственный университет, Иркутск, Россия

Функции нефонологической долготы в устной речи

Описывается нефонологическая долгота звуков. Рассматриваются характерные для данного средства функции: маркирование хезитации, иконическое выражение пространственной и временной величины, интенсификации признака, акцентирование значимого элемента. Особую роль в исследовании занимает выявление контекстуальных критериев, которые позволяют разграничить функции нефонологической долготы. Помимо примеров, в которых удлинения выполняют только одну функцию, выделяются также случаи совмещения нескольких функций. На основании проведенного анализа выдвигается гипотеза об уменьшении количества случаев нефонологической долготы, маркирующих хезитацию, по мере развертывания беседы. Для проверки гипотезы сравнивается количество употреблений удлинений в отрывках из начала, середины и финала глубинного интервью.

Ключевые слова: нефонологическая долгота, удлинение звука, хезитация, маркеры речевого сбоя, просодические маркеры хезитации, устный дискурс.

Functions of Non-Phonological Length in Oral Speech

This article is devoted to description of non-phonological longitude of the sounds. Some functions specific to the means: hesitation marking, iconic expression of spatial and temporal magnitude, of feature intensification, emphasizing a significant element are considered there. Identification of contextual criteria which allow to demarcate non-phonological longitude functions plays a special role in the study. In addition to examples in which the lengthening performs only one function, some cases of combining several functions were distinguished. Based on the analysis hypothesis about decrease in the number cases of non-phonological longitude marking hesitation as the conversation unfolds was put forward. To test a hypothesis the number of uses the lengthening in excerpts from the begining, the middle and the ending of in-depth interview is compared in the article.

Keywords: non-phonological length, sound lengthening, hesitation, markers of speech failure, prosodic markers of hesitation, oral discourse.

Нефонологическая долгота (далее – НФД) – удлинение звуков, не обладающее смыслоразличительной функцией. Описанию этого явления посвящены работы А. А. Кибрика и В. И. Подлесской [2009], С. В. Кодзасова [2009], В. П. Москвина [2015]. Исследователями описаны виды и различные функции нефонологической долготы. Кроме того, группа ученых под руководством А. А. Кибрика разработала процедуру транскрибирования удлиненных звуков.

Но, несмотря на некоторое количество достижений в данной области, в ней остаются лакуны. Особую сложность представляет разграничение типов НФД в зависимости от функций: не всегда удается точно определить, какую именно функцию выполняет данное средство в определенном контексте, так как в научной литературе отсутствует описание контекстуальных показателей, позволяющих квалифицировать употребления НФД.

Анализ нефонологической долготы в устной неподготовленной речи позволит нам дополнить дискурсивное описание данного явления, рассмотреть особенности его использования конкретным субъектом в своей речи.

Материалом исследования послужили фрагменты интервью[1] с Лидией Пантелеймоновной Поляничко (три отрезка по 4 минуты из начала, середины и финала интервью).

Целью данной статьи является анализ и описание типов нефонологической долготы звуков, выявление критериев для разграничения функций удлинения в устной неподготовленной речи.

Для этого определен репертуар функций нефонологической долготы. В. П. Москвиным [2015] выделяются удлинение звука как выражение различных оттенков экспрессии (*Red Bull окрыляяяяеееет*) и как прием эмфатического подчеркивания (*Все это было немножко до[с:]адно / И довольно нелепо*)[2]. Мы считаем такое разделение неточным. Так как эмфаза – это эмоциональное выделение, проводить границу между экспрессией и эмфатическим подчеркиванием нелогично. Вместо выражения различных оттенков экспрессии и эмфатического подчёркивания предложена **функция акцентирования значимого элемента**.

Группой ученых под руководством А. А. Кибрика [Кибрик, Подлесская, 2009] установлено, что нефонологическая долгота является одним из маркеров речевых сбоев, ее использование предоставляет говорящему время для формулирования мысли, подбора лексемы или грамматической формы:

Лунная доро-ожка,

(0.7) огра-ада[3].

Таким образом, может быть выделена **хезитационная функция удлинения звуков**. Кроме того, долгота помогает выражать пространственные и временные значения, а также указывает на интенсификацию признака – все это объединяется в **иконической функции**, описанной С. В. Кодзасовым [2009].

Данная функция реализуется в следующих примерах:

1) *– А он когда туда пошел?*

– Да давно-о уже.

2) *– Далеко-о полетела стрела*[4].

В результате анализа предложен следующий список функций нефонологической долготы:

- маркирование хезитации;
- иконическое выражение пространственной и временной величины, интенсификации признака;
- акцентирование значимого элемента.

[1]Интервью записано в рамках проекта «История Иркутска», хранится в архиве кафедры русского языка и общего языкознания ИФИЯМ ИГУ.

[2] Примеры из [Москвин, 2015].

[3] Примеры из [Кибрик, Подлесская, 2009].

[4] Примеры из [Кодзасов, 2009].

Представляется интересным выяснить, как в анализируемом материале распределяются примеры, в которых эти функции реализуются. Для этого в глубинном интервью с помощью программы Praat выявлено 89 случаев удлинения звуков. Рассмотрим их и представим значения количества удлинений в каждом из трёх отрезков в виде диаграммы (рис. 1).

Рис. 1. Сравнительный анализ количества употребления НФД в трех частях интервью

На диаграмме видно, что в начальном отрезке 36 употреблений НФД (из них 30 гласных звуков, 6 согласных), в серединном фрагменте также 36 (33 удлинения гласных и 3 согласных), в финальном отрывке количество долготы звуков значительно меньше – 17 случаев (13 гласных, 4 согласных). Данные примеры удлинений необходимо распределить в зависимости от выполняемой ими функции.

Для этого предложены критерии разграничения функций нефонологической долготы: близость других видов маркеров речевого сбоя, наличие фразового ударения и/или изменение интенсивности произнесения слова, наличие единиц, значение которых содержит семы величины, интенсивности и т. д. Такой выбор обусловлен тем, что данные критерии позволяют предполагать, что в примере реализуется та или иная функция НФД.

Так, **о речевом сбое** могут свидетельствовать грамматические, лексические и просодические маркеры хезитации. Нефонологическая долгота в качестве хезитационного средства может употребляться в элементарной дискурсивной единице (далее – ЭДЕ), где наблюдается перестроение синтаксической конструкции, дискурсивные слова *вот*, *ну*, маркеры препаративной подстановки, фальстарты и разные виды пауз. В приведённом примере нефонологическая долгота гласного звука употребляется совместно с фальстартом *на=*, абсолютной паузой *(0.20)*, перестройкой конструкции и длительным глоттальным скрипом *(ʔ0.78)* и, следовательно, выполняет функцию маркирования хезитации:

Дом на[ж] (0.50[вдох]) прямо ***на= (0.20)*** *б-был* ***(ʔ0.78)*** *н-на берегу рекии [0.33]*.

В следующем примере можно увидеть фальстарт *н-от*, абсолютную паузу *(1.65)*, два удлинения в дискурсивном слове *ннуу [0.28] [0.27]*, маркер *вот* и заполненную паузу *(э0.48),* что свидетельствует о том, что удлинения звуков в данном примере также выполняют функцию хезитации:

Н-от (1.65) ннуу *[0.28] [0.27] ии [0.43]* ***вот (э0.48)*** *собственно говоря.*

Падающее на слово фразовое ударение и изменение тона и интенсивности произнесения слов указывают на значимость информации для говорящего. Это позволяет предполагать, что в контекстах, где используются такие просодические средства, реализуется **функция акцентирования значимого элемента**.

В следующем примере фразовое ударение выделяет слово с НФД гласного звука, изменение тона и повышение интенсивности произнесения лексической единицы *прекрасное* свидетельствует о выражении экспрессии говорящим (рис. 2, 3):

*Детство было у нас прекр**ааа**сное [0.27] замечательное.*

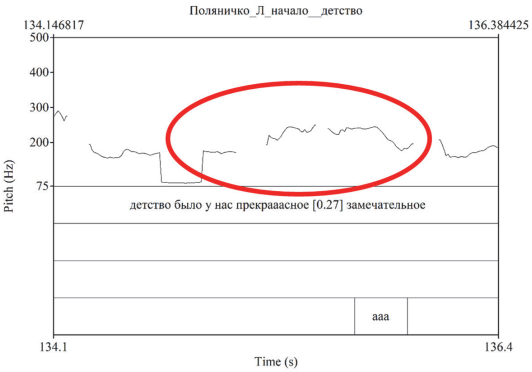


Рис. 2. Тонограмма (*Детство было у нас прекрааасное [0.27] замечательное)*

Рис. 3. Осциллограмма (*Детство было у нас прекрааасное [0.27] замечательное)*

Удлинение согласных звуков в слове также может выполнять функцию акцентирования значимого элемента, как, например, в следующей ЭДЕ. На это указывает резкое повышение тона при произнесении лексемы с НФД (рис. 4):

Мало машшин [0.29]

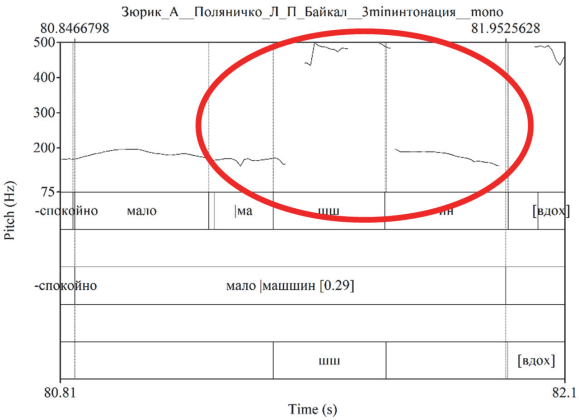


Рис. 4. Тонограмма (*Мало машшин [0.29]*)

Важно обращать внимание и на компоненты значения лексем с удлинением или лексем из той же ЭДЕ, что и лексема с удлинением – семы продолжительности, величины, интенсивности, так как это указывает на то, что в таких контекстах долгота может выполнять **иконическую функцию**.

В следующем примере наличие у нефонологической долготы иконической функции (акцентирование внимания на большом размере объекта) подтверждается близостью слова *огромные*, значение которого – ‘очень большой’:

огромные там тополяя [0.37].

С помощью удлинения согласных звуков также может выражаться значение пространственной величины: сема размера в слове *массив* указывает на иконическую функцию:

(1.01[вдох]) такой лесной масссиив [0.30][0.21].

В рассмотренных ранее примерах удлинения гласных и согласных звуков выполняют только одну функцию, однако встречаются и такие контексты, в которых **совмещается несколько функций**.

Так, в следующем примере удлинение гласного звука находится в одной ЭДЕ с заполненной паузой (э 0.38) и абсолютной паузой (0.38), что свидетельствует о речевом сбое. Однако, помимо этого, в лексеме *быыстрая* есть сема интенсивности (‘происходящий, совершающийся с большой скоростью’), что говорит об иконической функции НФД. Это позволяет заключить, что удлинение в этом случае выполняет сразу две функции:

(1.15[вдох]) (э 0.38) река э Белая очень (0.38) быыстрая [0.33]

(0.68[вдох]) (э 0.32) глубоокая [0.40].

Разработанный репертуар функций НФД и выделенные критерии их разграничения позволяют выявить несколько групп примеров: контексты, в которых НФД выступает только как маркер хезитации; контексты, в которых НФД вы-

полняет иконическую функцию; контексты, в которых НФД выполняет акцентирующую функцию; контексты, в которых НФД сочетает маркирование хезитации с другими функциями (табл. 1).

Таблица 1

Функция	Начало (36)	Середина (36)	Конец (17)
Хезитация (40)	18	13	9
Иконическая (13)	4	6	3
Акцентирование (23)	9	9	5
Совмещение (13)	5	8	0

На основании анализа примеров выдвинута гипотеза о том, что в процессе развёртывания беседы коммуникативный барьер между говорящими заметно ослабевает и, как кажется, это приводит к уменьшению количества используемых средств хезитации, в том числе нефонологической долготы. Чтобы проверить выдвигаемую гипотезу, предлагается рассмотреть распределение случаев удлинений, выполняющих различные функции, в разных отрывках интервью.

Диаграмма (рис. 5) позволяет увидеть, что в начальном отрезке 18 примеров, в которых НФД выполняет функцию хезитации, в серединном 13, а в финальном 9. К ним также следует добавить случаи совмещения функций (во всех присутствует хезитация): 5 из первого фрагмента и 8 из второго. Следовательно, к концу интервью количество удлинений, выполняющих функцию маркера речевых сбоев, уменьшается.

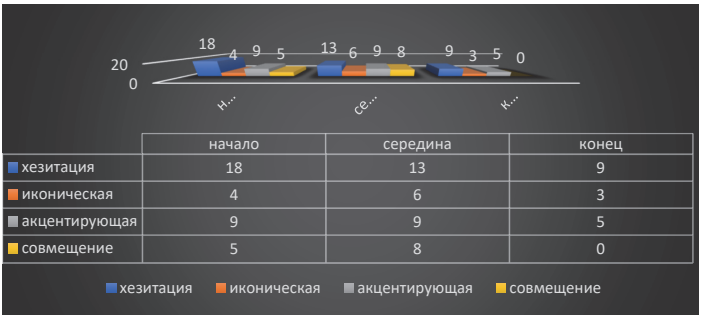


Рис. 5. Распределение функций нефонологической долготы по фрагментам интервью

Таким образом, можно сделать вывод о том, что разработанные критерии разграничения функций нефонологической долготы позволяют определить, какую функцию выполняет удлинение в том или ином случае. Выявлены случаи совмещения двух функций в пределах одного примера. На основании проведенного анализа установлено, что в анализируемом интервью с развертыванием беседы уменьшается количество речевых сбоев, которые маркирует нефонологическая долгота.

Литература

Кибрик А. А., Подлесская В И. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М. : Языки славян. культур, 2009. 736 с.

Кодзасов С. В. Исследования в области русской просодии. М. : Языки славян. культур, 2009. 496 с.

Москвин В. П. Удлинение звука в русском слове: риторический и орфоэпический аспекты // Лингвистические заметки. 2015. С. 54–60.

УДК 81’33

Н. А. Кошелюк

Институт системного программирования РАН, Москва, Россия

И. В. Федотова

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
Институт системного программирования РАН, Москва, Россия

Цифровые методы сохранения и лингвистического анализа миноритарных языков России

Рассматриваются возможности цифрового анализа языкового материала, которые предлагаются лингвистической платформой «ЛингвоДок». Описывается работа опций *cognate analysis*, позволяющая выявлять общие связи исследуемого словаря с другими диалектами и языками; *phonemic analysis* – инструмента, служащего выявлению набора фонем и аллофонов в современных полевых данных; *phonology* – опции для определения звукового строя языков и диалектов, *search и maps* – опции, предназначенной для поиска по любому лингвистическому признаку или языку и диалекту всего множества словарей и корпусов, размещенных на платформе, после чего становится возможным строить ареалы конкретных языковых явлений и отображать территорию их распространения. Опыт реализации этих опций на примерах словарей и корпусов доступных языков России, имеющихся на «ЛингвоДок», позволяет сделать вывод о том, что использование платформы при работе с языковым материалом повышает верифицируемость исследований, делает возможной комфортную автономную и командную работу и позволяет выйти на новый уровень фонетических, фонологических, ареальных и междисциплинарных исследований, значительно сокращая время работы исследователей над поставленными задачами.

Ключевые слова: миноритарные языки, лингвистическая платформа, анализ когнатов, фонетический анализ, фонологический анализ, контактные языковые изменения.

Digital Methods of Preservation and Linguistic Analysis of Minority Languages of Russia

The article discusses the possibilities of digital analysis of language material, which are offered by the linguistic platform “LingvoDoc”. We describe the operation of the cognate analysis options, which allows to identify common connections of the dictionary under study with other dialects and languages; phonemic analysis – a tool that serves to identify a set of phonemes and allophones in modern field data; phonology is an option for determining the sound structure of languages and dialects, search and maps, designed to search for any linguistic feature or language and dialect of the entire set of dictionaries and corpora placed on the platform, after which it becomes possible to build areas of specific linguistic phenomena and display the territory of their distribution. The experience of implementing these options using the examples of dictionaries and corpora of the languages of Russia available on «LingvoDoc» allows us to conclude that using the platform when working with language material increases the verifiability of research, makes possible comfortable autonomous and team work and allows you to reach a new level of phonetic, phonological, areal and interdisciplinary research, significantly reducing the time of the researchers' work on the assigned tasks.

Keywords: minority languages, linguistic platform, cognate analysis, phonetic analysis, phonological analysis, contact-induced language change.

Введение

Исследования миноритарных языков России в настоящее время выходят на новый уровень: создаются виртуальные исследовательские лаборатории, специальные программы и платформы, позволяющие обрабатывать лингвистический материал более совершенным способом. Так, например, в 2009 г. в Германии был запущен проект по изучению родственных хантыйского и мансийского языков Ob-Ugric languages: conceptual structures, lexicon, constructions, categories (OUL), а в 2014 г. – проект Ob-Ugric Database: analyzed text corpora and dictionaries for less described Ob-Ugric dialects (OUDB). За несколько лет лингвистам из Германии (Е. К. Скрибник), Австрии (Т. Ризе), Финляндии и Венгрии удалось организовать виртуальную исследовательскую среду, до недавних пор не имеющую мировых аналогов. Задачей проекта было создание инновационного описательного ресурса, включая базы данных, мультимедийные библиотеки, ссылки на всевозможные ресурсы и опции, и лингвистического анализа, основанного на базе корпуса современных полевых данных и данных экспедиций Мункачи, Каннисто и других исследователей [Кошелюк, 2020, с. 757].

В России стоит отметить создание в 2012 г. виртуальной научной лаборатории LingvoDoc – лингвистической платформы, предназначенной для хранения, составления и анализа словарей, корпусов и конкордансов различных языков и диалектов[1]. Описание программного обеспечения и принципа устройства платформы подробно описаны в [Программная система … , 2022]. Работа на платформе предполагает следующие возможности осуществления научной деятельности:

1) коллективную и обособленную от других исследователей работу в системе;

2) составление собственных словарей и корпусов;

3) предоставление прав на свои словари другим пользователям платформы;

4) сортировку материалов по языкам, грантам, институтам;

5) создание РИДов для отчетности и закрепления прав на свои данные на законодательном уровне;

6) проведение полного анализа языкового материала с помощью предложенных платформой опций с возможностью проверки ошибок в обработанных данных.

На сегодняшний день база этого ресурса составляет более 1000 этимологических диалектных аудиословарей и 300 корпусов на разных языках народов России. Значительный объем этих данных – уникальный лингвистический материал по редким языкам, носителей которых осталось не более 10 человек старше 60 лет или по уже исчезнувшим языкам/диалектам (камасинский, орочский языки, западные, восточные и южные мансийские диалекты и др.). Необходимо отметить, что большинство этих источников были лишь недавно обнаружены в раз-

[1] См. ЛингвиДок 3.0. URL: http://lingvodoc.ispras.ru/

личных архивах России и Европы и ранее не вводились в научный оборот. Анализ этих словарей с помощью опций, предлагаемых платформой «ЛингвоДок», позволил существенно обогатить имеющиеся знания о языках.

Для понимания того, какую исследовательскую деятельность можно осуществлять на «ЛингвоДок» и какие примерные результаты можно получить, ниже мы рассмотрим несколько основных опций платформы: *cognate analysis*, *phonemic analysis*, *phonology*, *search и maps*.

1. Cognate analysis

Анализ когнатов, осуществляемый в рамках этой опции, является лишь одним из возможных направлений работы со словарями и корпусами на платформе[1]. *Cognate analysis* строится на возможности алгоритма осуществлять анализ этимологических связей: выявлять ряды соответствий, показывать стандартные соответствия и соответствия, основанные на сходстве переходов. Анализ когнатов словаря проводится по гласным и согласным звукам 1-го и 2-го слогов и по соответствиям сочетаний и позволяет определять близость между языками и диалектами, после чего выстраивать генетические деревья (рис. 1).

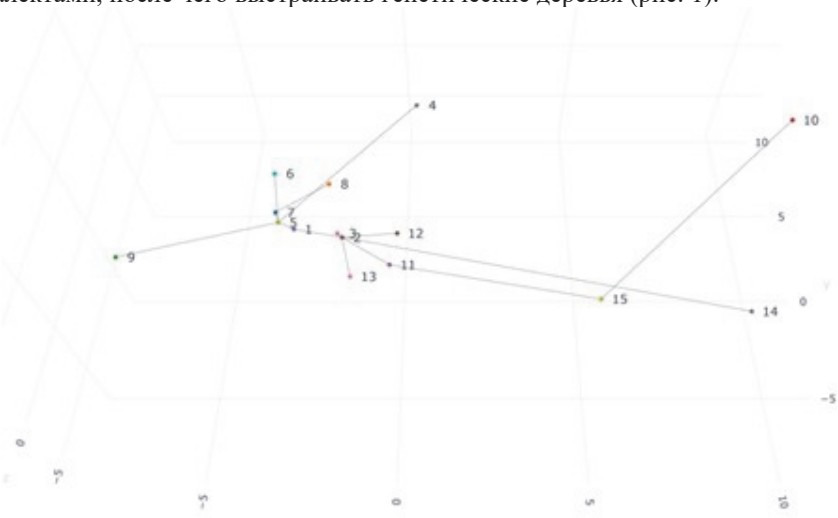


Рис. 1. 3D-модель близости мансийских диалектов на основе результатов обсчета опции *cognate analysis*

2. Phonemic analysis

С помощью этой опции происходит определение фонематического инвентаря исследуемого языка или диалекта. Алгоритм построен таким образом, что позволяет выводить на экран полный набор гласных и согласных фонем и аллофонов, показывая позиции их появления в лексемах (рис. 2).

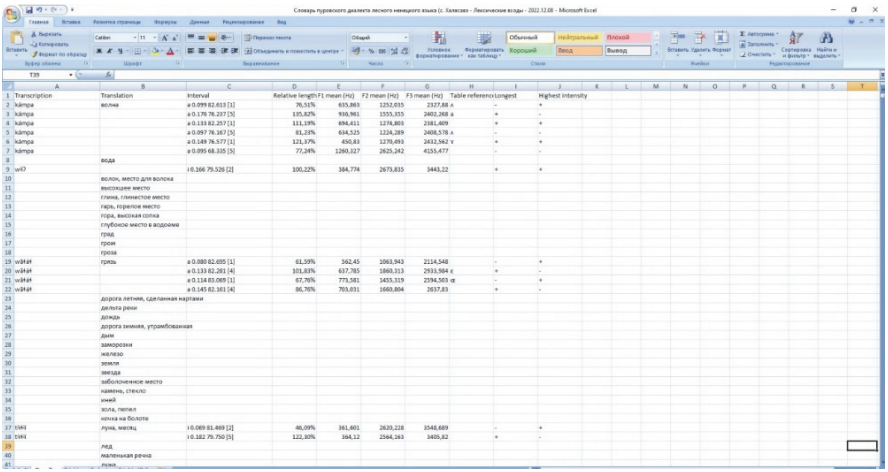
[1] Подробнее см. ЛингвиДок 3.0. URL: http://lingvodoc.ispras.ru/tools

Рис. 2. Пример использования опции *phonemic analysis*

3. Phonology

Phonology – опция, позволяющая строить и верифицировать фонологические системы для любого языка или диалекта. В ее основе лежат результаты, изначально полученные с помощью фонетической программы Praat, в которой для каждого звука определяются физические характеристики звука: форманты, длительность, интенсивность, и впоследствии подгруженные исследователем в соответствующий словарь на платформе «ЛингвоДок». Алгоритм работы этой опции строится по принципу сбора физических характеристик каждого гласного звука во всех произнесениях и обработки спектрограмм всего фонетического материала исследуемого словаря. На выходе пользователь платформы получает таблицу Excel со списком всех физических характеристик каждого звука всего множества лексем анализируемого языка или диалекта (рис. 3).

В результате такой обработки все данные таблицы Excel с указаниями физических параметров каждого звука преобразуются в 3D-график, построенный на основании показателей трех формант (F1-F2-F3) для каждого гласного звука (рис. 4). О правильности выделения той или иной фонемы говорит слабое пересечение облаков формант – менее 30 %, и наоборот – их пересечение более чем на 31 % свидетельствует об ошибках исследователя в первоначальной разметке данных в программе Praat либо о разном способе образования выявленных алгоритмом фонем. Важно, что на каждую выявленную фонему *phonology* выводит все примеры из анализируемого словаря, что позволяет быстро выполнить его фонологическое описание и произвести проверку транскрипций, таким образом осуществив верификацию данных.



	A	B	C	D	E	F	G	H	I	J
1	Transcription	Translation	Interval	Relative length	F1 mean (Hz)	F2 mean (Hz)	F3 mean (Hz)	Table reference	Longest	Highest intensity
2	kämpa	волна	a 0.099 82.613 [1]	76,51%	615,863	1252,035	2127,88	a	-	+
3	kämpa		a 0.176 76.237 [3]	135,82%	936,961	1555,355	2402,268	a	+	-
4	kämpa		a 0.133 82.257 [1]	111,19%	694,411	1274,803	2381,409		+	+
5	kämpa		a 0.997 76.167 [5]	81,22%	634,525	1224,389	2408,578	a	-	-
6	kämpa		a 0.149 76.577 [1]	121,37%	450,83	1270,493	2432,562	y	+	+
7	kämpa		a 0.095 68.335 [5]	77,24%	1260,327	2625,242	4155,477		-	-
8		вода								
9	wʲɑ		i 0.166 75.526 [2]	100,22%	384,774	2673,835	3443,22		+	+
10		волок, место для волока								
11		высокое место								
12		глина, глинистое место								
13		гарь, горелое место								
14		гора, высокая сопка								
15		глубокое место в водоеме								
16		град								
17		гроза								
18		гроза								
19	wʲɑʲɑ	грязь	a 0.080 82.695 [1]	61,19%	362,45	1063,943	2114,548		-	+
20	wʲɑʲɑ		a 0.133 82.281 [4]	101,83%	637,785	1860,313	2931,564	a	+	-
21	wʲɑʲɑ		a 0.114 83.069 [1]	67,76%	773,581	1455,319	2594,503	ɑ	-	+
22	wʲɑʲɑ		a 0.145 82.161 [4]	86,76%	763,031	1660,804	2637,83		+	-
23		дорога летняя, сделанная картами								
24		дельта реки								
25		дождь								
26		дорога зимняя, утрамбованная								
27		дым								
28		заморозки								
29		железо								
30		земля								
31		звезда								
32		заболоченное место								
33		камень, стекло								
34		иней								
35		зола, пепел								
36		ночка на болота								
37	tʲiʲi	луна, месяц	i 0.069 81.469 [2]	46,09%	361,401	2620,218	3548,689		-	+
38	tʲiʲi		i 0.182 79.750 [5]	122,30%	384,12	2564,383	3405,82		+	-
39		лед								
40		маленькая речка								
41		река								

Рис. 3. Итоговая таблица Excel с фонологическими данными, сформированная с помощью опции *phonology*

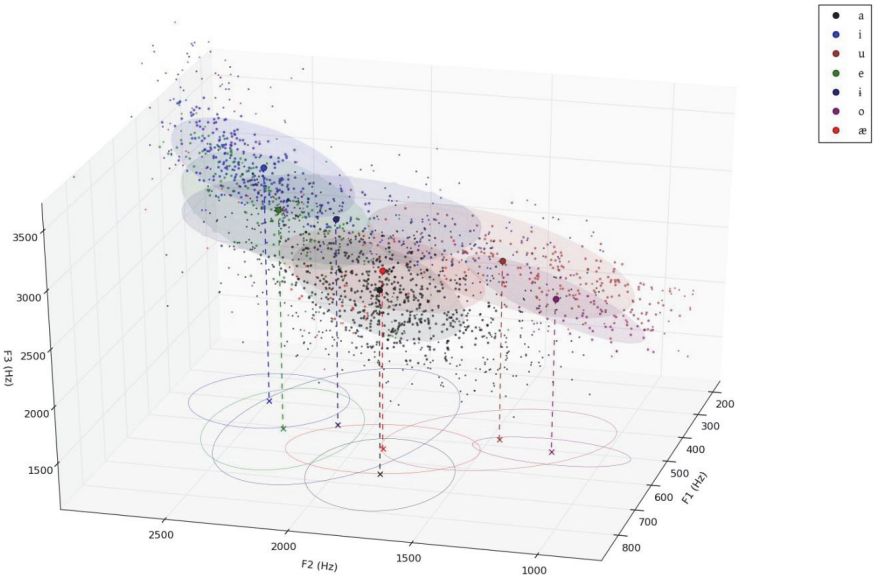


Рис. 4. Пример облаков формант гласных первого слога, сформированных опцией *phonology*

4. Search and Maps

Опция поиска и построения карт на платформе «ЛингвоДок» предлагает принципиально новый способ визуализации лингвистических явлений относительно существующих на сегодняшний день диалектологических атласов. В таких атласах (например, см.: [Диалектологический атлас русского … ; Лексический атлас русских … , 2017; Общеславянский лингвистический … , 2015]) на картах отмечаются дистрибуция лексем с определенным значением, семантические и мотивационные карты, но строятся они только на материале родственных языков или диалектов одного языка. Аналогичный принцип на неродственных языках применен в [Atlas Linguarum Europae, 1997] (ALE, «Лингвистический атлас Европы»): здесь представлены языки из 6 неродственных семей, ономасиологические, семасиологические и мотивационные карты; однако выпуски этого атласа – результат многолетнего труда большого коллектива ученых. Преимущество и новизна подхода построения карт на «ЛингвоДок» заключается в том, что их может строить любой лингвист, зарегистрированный на платформе, используя любые хранящиеся там и постоянно пополняемые данные.

С помощью поиска можно показывать ареалы распространения определенных семантических моделей, семантических переходов. Отметим, что такой же принцип применяется в [Database of Semantic Shifts, 2006–2022] (рис. 5): визуализируются семантические переходы. Но по сравнению с «ЛингвоДок» (рис. 6) пользователи не могут сами добавлять материал на платформу DatSemShift, чтобы потом по ним строить карты, и там невозможно показывать географический ареал распространения – в Database of Semantic Shifts одна точка равна одному языку и нельзя отобразить весь ареал, который занимает язык.

Выводы

Таким образом, краткий обзор основных опций «ЛингвоДок» для исследования различных языков и диалектов позволяет выделить следующие преимущества работы на этой платформе:

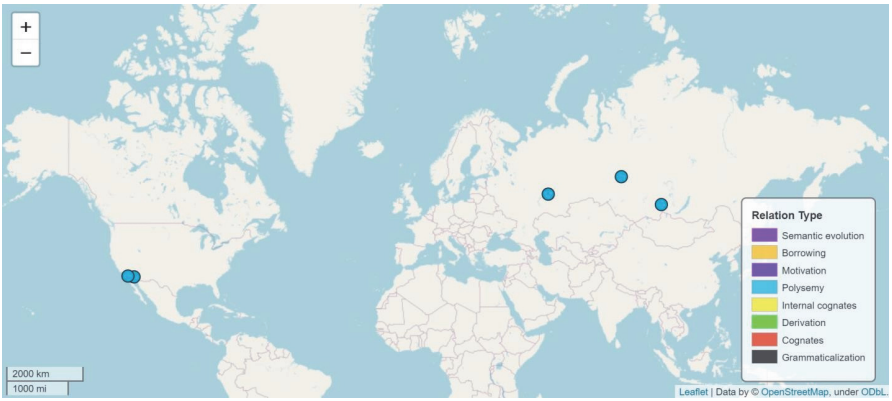


Рис. 5. Пример построения карты семантического перехода *earth, soil > clay* с помощью программы Database of Semantic Shifts

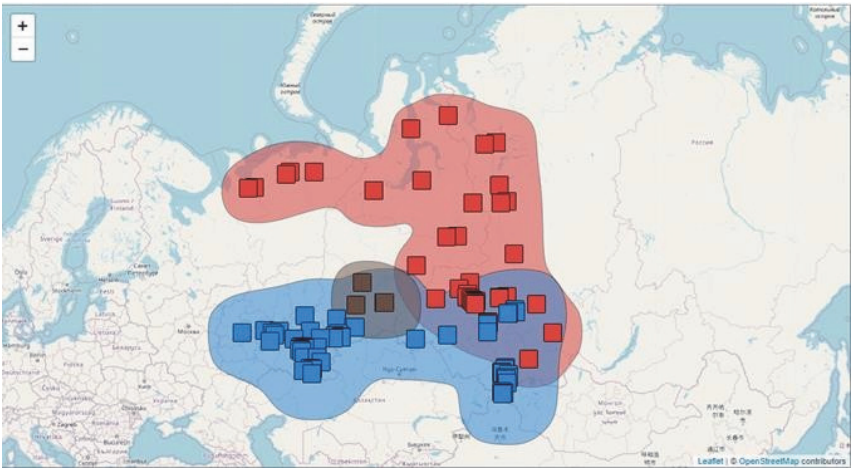


Рис. 6. Пример построения карты семантического перехода *earth, soil > clay* с помощью опций Search and Maps платформы «Лингводок»

1) повышение **верифицируемости** исследований с помощью автоматических программ (фонетические, этимологические, морфологические);

2) вывод на **новый уровень** ареальных исследований благодаря большому массиву данных по языкам разных семей и возможностям визуализации полученных результатов исследования;

3) создание базы для **интеграции отдельных направлений** лингвистики, истории и археологии под зонтичным проектом исследований миноритарных языков России;

4) открытость и **доступность** данных, возможность получения принципиально нового знания в команде или при индивидуальной работе.

Литература

Диалектологический атлас русского языка. М. : Ин-т рус. языка им. В. В. Виноградова РАН. https://da.ruslang.ru/

Лексический атлас русских народных говоров (ЛАРНГ): Т. 1. Растительный мир. М. ; СПб. : Нестор-История, 2017. 736 с.

Программная система LingvoDoc и возможности, которые она предлагает для документирования и анализа обско-угорских языков / Ю. В. Норманская, О. Д. Борисенко, И. Б. Белобородов, А. И. Аветисян // Доклады Российской академии наук. Математика, информатика, процессы управления. 2002. Т. 504. С. 51–73.

Общеславянский лингвистический атлас: Серия лексико-словообразовательная (ОЛА): Выпуск 10. Народные обычаи. М. ; СПб. : Нестор-История, 2015. 276 с.

Кошелюк Н. А. Мансийские исследования: от истоков к современности // Oriental Studies. 2020. Т. 13, № 3. С. 743-765.

Atlas Linguarum Europae Perspectives nouvelles en géolinguistique. Roma, 1997.

Database of Semantic Shifts / A. Zalizniak [et al.]. Moscow : Institute of Linguistics RAS, 2016–2022. URL: https://datsemshift.ru/ (дата обращения: 29.04.2022).

УДК 81’33

А. С. Кулева

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

О корпусных методах в лексикографии: проблемы и перспективы

Рассматриваются проблемы, возникающие при использовании корпусных методов исследования, рассматриваются на примерах из опыта работы над различными лексикографическими проектами. Анализируется словарное представление одной тематической группы лексики – названий растений (фитонимов). Отмечается, что составители современных толковых словарей опираются на корпус при формировании словника, определении употребительности и описании функционирования лексемы, подборе иллюстративного материала, однако существующие корпуса продолжают пополняться и совершенствоваться. Доказывается, что незначительная представленность фитонимов в материалах Национального корпуса русского языка связана с недостаточным привлечением соответствующих источников – научной и научно-популярной литературы. Когда при дальнейшем пополнении корпуса такой дисбаланс будет преодолен, представления о языке, формируемые корпусными данными, также изменятся. Приводятся примеры словарных статей, подтверждающие необходимость критического отношения к материалу и привлечения дополнительных источников для адекватного лексикографического описания современного состояния языка. Важность фитонимической лексики для языкового сознания иллюстрируется данными поэтического корпуса НКРЯ в соотнесении с материалами «Словаря языка русской поэзии XX века». Показано, что названия растений, в том числе редкие и специальные, широко используются в поэтическом языке, что говорит о значимости этого тематического пласта в русской языковой картине мира. Приводятся количественные данные, доказывающие, что отдельные недостатки корпусных методов опосредованы несовершенством существующих словарей. Так, семантическая разметка корпуса включает тематическую группу «растения», в которую собственно названия растений входят как одно из подмножеств; с другой стороны, соответствующий семантический признак не приписывается фитонимам, не отраженным существующими словарями. Таким образом, корпусные методы дают лексикографам новые возможности, а практическая работа над словарями позволяет не только выявить возникающие проблемы, но и наметить пути их решения.

Ключевые слова: авторские словари, лексикография, названия растений, Национальный корпус русского языка, толковый словарь, фитонимы.

On Corpus Methods in Lexicography: Problems and Prospects

The paper is devoted to corpus methods in lexicography and exploring of related problems. The plant names (phytonyms) representation in dictionaries is analyzed. It is noted that the authors of modern explanatory dictionaries rely on the corpus when forming a dictionary, determining the use and description of the functioning of the lexeme, selecting illustrative material, but the existing corpus continues to be replenished and improved. It is proved that the lack of representation of phytonyms in the materials of the Russian National Corpus is associated with insufficient involvement of relevant sources – academic literature and non-fiction. When such an imbalance is overcome with further replenishment of the corpus, the ideas about the language formed by the corpus data will also change. The article provides examples of dictionary entries confirming the need for a critical attitude to the material and the involvement of additional sources for an adequate lexicographic description of the current state of the language. The importance of phytonymic vocabulary is illustrated by the data of

the poetry corpus of the Russian National Corpus in comparison with the materials of the “Dictionary of the language of Russian poetry (20th century)”. It is shown that the names of plants, including rare and special ones, are widely used in the poetic language. The semantic markup of the corpus includes the thematic group “plants”, in which the actual names of plants are included as one of the subsets; on the other hand, the corresponding semantic feature is not attributed to phytonyms that are not reflected in existing dictionaries. Thus, corpus methods give lexicographers new opportunities, and practical work on dictionaries allows not only to identify emerging problems, but also to find ways to solve them.

Keywords: author dictionary, explanatory dictionary, phytonyms, lexicography, plant names, Russian National Corpus.

Лексикографическая работа сейчас во многом переходит в электронную форму, базируется на корпусных методах и инструментах. Все больше проектов изначально создается в электронной форме, переводится в формат баз данных и т. п. Важнейшим достоинством такой формы работы над толковым словарем становится возможность постоянного пополнения и совершенствования словаря, увеличение иллюстративной зоны, в определенной перспективе – подключение обратной связи. Тем не менее классический жанр бумажного словаря не теряет своей актуальности. Удачным примером сочетания старых и новых форматов работы можно назвать [РОС].

Но и работа в традиционном формате существенно отличается от труда предшественников, вынужденных создавать многотомные словари в алфавитном порядке на основе ограниченного фактического материала и собственной интуиции. Зачастую работа над новым толковым словарем ярко показывает всю глубину проблем, стоявших перед нашими великими предшественниками, и заставляет восхититься их достижениями, несмотря на все имеющиеся недостатки, противоречия или даже ошибки.

С одной стороны, мы сейчас можем опираться на такие объемы информации, которые раньше невозможно было представить. Корпусные источники (прежде всего [НКРЯ]) позволяют обрабатывать большие массивы языкового материала, дают более объективные количественные и статистические данные, показывают динамику употребления лексической единицы, позволяют проверить языковую интуицию лексикографа. Кроме того, интернет выступает и как источник специальных сведений, и как своего рода расширенный корпус языкового материала.

С другой стороны, составитель словаря в ходе своей работы снова и снова сталкивается с проблемами, вызванными уже не ограниченностью, а широтой материала. В какой-то степени ограничивающие факторы, игравшие значительную роль при создании классических толковых словарей, имели свое преимущество: лексикографы работали на основе картотеки примеров, тщательно подготовленных и выверенных, привлекали к своей работе консультантов-специалистов, ответственных за достоверность и системность соответствующих групп лексики. Сейчас в качестве такого сдерживающего фактора составитель словаря может использовать главным образом свое критическое мышление.

Как уже упоминалось, важной особенностью работы современного лексикографа становится возможность оперировать большими объемами информации. Хотя в работе над многотомным словарем по-прежнему приходится зависеть от алфавитной последовательности, но на предварительных этапах работы появляется возможность заранее прорабатывать отдельные группы лексики, выделяемые на самых разных основаниях. К такому подходу толковая лексикография обращается все чаще: можно вспомнить опыт [РСС], а из новых проектов следует назвать [ТСРРР].

«Академический толковый словарь русского языка» [АТоС] создается в Институте русского языка им. В. В. Виноградова как толковый словарь среднего типа (по классификации [Ожегов, 1952]). Параллельно с работой над томами АТоСа коллектив разрабатывал проект «Стратификация лексики современного русского языка и ее отражение в толковом словаре», поддержанный грантом РФФИ, по системному анализу и представлению нескольких лексических групп. Подобная работа позволяет более глубоко и системно представить соответствующие лексемы в словаре: оценить полноту словника словаря (включить или исключить те или иные единицы), проанализировать правомерность постановки помет, соотнося единицы в ряду им подобных. Помимо традиционной системы отсылочных толкований, указания синонимов и антонимов, полезным становится приведение при толковании синонимического ряда, указание на гипо-гиперонимические связи [Крысин, 2010; ТСРРР].

Разрабатывать таким образом можно любые лексические группы. Остановимся на такой группе, как названия растений (фитонимы).

С одной стороны, кажется, что это достаточно простая и очевидная лексико-семантическая группа, однако при ближайшем рассмотрении оказывается, что работа с ней позволяет выявить серьезные и системные лексикографические проблемы, существенные и для других групп, причем связанные как с традиционным лексикографическим описанием русского языка, так и с новыми цифровыми методами.

Прежде всего возникает проблема, рассматривать ли названия растений как общеязыковые единицы или как терминологическую или терминоподобную группу, подраздел номенклатуры. От этого зависит и объем этой группы, и выбор толкования. Сопоставление толковых словарей [СУ; БАС-1; МАС] показывает, что в толковой лексикографии был сделан упор на терминологическом понимании.

Это привело к тому, что из словника стали выводиться единицы, стилистически окрашенные, устаревшие, областные, неконкретные, дублетные и пр. (*желтоцвет*, *желтофиоль*, *конопель*, *смородина*, *шпажник*), а толкования стали по возможности унифицироваться и приближаться к научному описанию. Ср. в МАС: АСТРА ‘травянистое декоративное растение сем. сложноцветных, с цветками различной окраски без запаха’; ДЕВЯСИЛ ‘травянистое растение или кустарник сем. сложноцветных’; ЗОЛОТАРНИК ‘травянистое растение сем. сложноцветных, с желтыми цветками, собранными в соцветия’; ЦИНЕРАРИЯ ‘декоративное травянистое растение сем. сложноцветных, с яркими соцветиями’; ЦИННИЯ ‘травянистое декоративное растение сем. сложноцветных’.

Как представляется, это привело к сомнительным результатам: исключались вполне употребительные единицы, встречающиеся в языке литературы лексемы стали переходить в разряд агнонимов, унифицированные толкования становились малоинформативными, сочетание в толковании научного и наивноязыкового компонента входило в противоречие с установкой на научность.

Легко заметить, что от словаря к словарю состав этой лексической группы сокращается, тем более что при исключении «неподходящих» единиц новые и более актуальные добавляются в незначительных количествах.

Так, [СУ] (85 тыс. слов), словарь среднего типа, отражающий лексику первой половины XX в., содержит 581 словарную статью, где описываются названия растений; [МАС] (90 тыс. слов), словарь среднего типа, отражающий лексику второй половины XX в., содержит 616 статей, однотомный [СШ] (82 тыс. слов), который можно признать наиболее современным лексическим трудом в своем жанре, включает 391 статью. Сводный словник, уточненный по РСС, состоит из 738 позиций. Для сравнения: указатель русских названий растений только в классическом труде «Флора средней полосы европейской части России» [Маевский, 2014] включает 1270 названий.

Установка на терминологичность ставит еще одну проблему: нужно ли включать единицы такого рода в словарь среднего типа – АТоС – и тем более в однотомные словари. Кажется, что в этих словарях упор должен быть сделан на общеязыковую широкоупотребительную лексику, тогда как специальная лексика должна описываться в специальных словарях.

С другой стороны, специальных словарей, описывающих фитонимы, не существует. Не углубляясь в проблему, достаточно отметить, что с собственно научной точки зрения фитонимы должны быть предметом рассмотрения естественно-научных дисциплин, прежде всего систематики как специального подраздела ботаники. Но с соответствующей научной позиции в роли «настоящих» названий растений выступает латинская номенклатура, которая к тому же постоянно меняется и совершенствуется, а национальные (в нашем случае – русские) названия растений остаются на далекой периферии.

Если говорить о словарях русского языка, то полные, большие академические словари [БАС-1; БАС-3] принципиально отражают язык прошлого, язык классической литературы, рассматриваемая лексическая группа в них достаточно обширна, но недостаточно актуальна. Относительно новые единицы системно включаются в РОС. Ряд названий фиксируется словарями иноязычных слов, но при этом за их пределами оказывается большое количество вполне употребительных обиходных именований (типа *декабрист*) или актуальных и при этом не иноязычных (типа *борщевик*).

Ср. употребительные сейчас названия, не представленные в толковых словарях: растения как продукты питания – *брокколи*, *лайм*, *рукола*, *цукини*; декоративные (садовые) – *агератум*, *арония* (*черноплодка*), *астильба*, *бузульник*, *вейгела*, *волжанка*, *гейхера*, *дельфиниум*, *дербенник*, *дицентра* (*разбитое сердце*), *ирга*, *клематис*, *колеус*, *космея* (*немецкая ромашка*), *котовник*, *купена*, *лилейник*,

лунник, *маттиола*, *монарда*, *морозник*, *мускари*, *посконник*, *пролеска*, *пузыреплодник*, *рудбекия*, *снежноягодник*, *форзиция*, *хоста*; декоративные (срезочные, букетные) – *гербера*, *калла*, *лиатрис*, *фрезия*; комнатные (оранжерейные): *бромелия*, *декабрист* (*шлюмбергера*), *сенполия* (*узамбарская фиалка*), *хлорофитум*.

Кажется, что ответы на многие возникающие вопросы можно найти с помощью уже привычного и прекрасно разработанного инструмента – на корпусном материале. Безусловно, современную исследовательскую работу, прежде всего лексикографическую, без корпусных методов уже невозможно представить.

Обращение к НКРЯ показывает, что большинство фитонимов, особенно из числа новых кандидатов на включение в словник, действительно малоупотребительно, что, казалось бы, подтверждает мнение о том, что в однотомных и средних словарях наличие таких единиц необязательно. Этой позиции противоречит языковая интуиция составителя словаря: названия растений, это важная часть языковой картины мира, они употребляются в речи, в языке литературы, их отсутствие в толковом словаре очень огорчает.

Противоречие между данными корпуса и языковой интуицией во многом объясняется, если критически взглянуть на источники корпуса. Безусловно, НКРЯ сейчас аккумулирует огромный языковой материал, он постоянно пополняется, но при этом нельзя говорить о его всеохватности, это пока по совершенно объективным причинам лишь отдаленная перспектива.

Действительно, корпус постоянно пополняется, и не только более новыми источниками, но и более широким привлечением старых, но остается дискуссионным вопрос о выборе таких источников.

Особенно остро сказанное относится к названиям растений. Поиск по НКРЯ показывает, что некоторые названия, особенно вошедшие в язык недавно, употребляются очень редко и ограниченно. Так, поиск названия популярного сейчас садового растения *гейхера* дает скромный результат: 9 документов, 19 примеров. Эти примеры взяты из узкоспециальных популярных источников и относятся к одному временному отрезку (2001–2009 гг. и одиночный пример 1988 г.). Ср.: *Ирисы, флоксы, примулы, хризантемы, лилейники, гейхеры, медуницы, анемоны, баданы, пиретрумы, ромашки похолоданий не боятся* (Марина Шалавеене «Ранняя рассада», 2009 г.); *Здесь в легкой полутени можно посадить гейхеру гибридную* (Валерия Иршенкова «Горный ландшафт в миниатюре», 2001 г.); *Мы не представляем себе сад без оранжевых и желтых пятен купальницы, красных гравилата, гейхеры, агристемы, изящных аквилегий, гайлярдий, разнообразных ирисов, лилейников* («Осенние заботы», журнал «Работница», 1988 г.).

Само это название было зафиксировано раньше: ге́йхера (лат. Heuchera), причем в авторитетной специальной литературе, включая [БЭС], ср. [Головкин, Китаева, Немченко, 1986, с. 161–162]. Однако в орфографическом словаре до недавнего времени слово отсутствовало, сейчас в [РОС]: «ге́йхера, -ы». В популярной литературе по садоводству название *гейхера* употреблялось [Воронцов, 2013] наряду с вариантами *геухера* и *хеухера* (ср.: Хессайон Д. Г. «Все о хвойных и вечнозеленых растениях», 2014 г.). На данный момент при поиске в интернете

эти варианты встречаются преимущественно на малоавторитетных вторичных сайтах, в НКРЯ они отсутствуют.

Похожие результаты дает и поиск по ряду других фитонимов.

Таким образом, в данном случае мы видим не столько динамику употребления лексемы в языке, сколько отсутствие в НКРЯ источников по теме, в особенности отражающих язык позднее 2009 г. Действительно, специальная и научно-популярная литература пока представлена в НКРЯ очень избирательно, с пополнением этого круга источников взгляд на употребительность фитонимов может существенно измениться.

Пока же составителю словарных статей следует опираться не только на НКРЯ, но и на широкий круг специальной литературы, а также язык интернета.

Еще один очень интересный аспект работы с корпусом обнаруживается при постановке другой исследовательской задачи. Так, появление в инструментарии НКРЯ семантической разметки позволило исследователям существенно расширить свои возможности. Поиск в основном корпусе лексем со значением «растение» (gr:S & r:concr & t:plant) дает внушительные результаты: 51 317 документов, 956 582 примера. Обработать вручную такой массив затруднительно.

Здесь хотелось бы обратиться к другому лексикографическому жанру. В отличие от толковой лексикографии, описывающей общеупотребительный язык, авторская лексикография позволяет взглянуть на общеупотребительный язык через призму художественного текста. В особенности интересно это по отношению к языку поэзии. Безусловно, словарь поэтической речи позволяет составить более объективное представление об авторском идеостиле. Однако с другой стороны, через отражение в языке поэзии мы можем по-новому взглянуть на динамику общеязыковых процессов, прослеживать тенденции в словообразовании или в синтаксисе, анализировать механизмы словотворчества.

Если сравнить выборку из [СЯРП] и поиск по семантическому признаку, можно увидеть существенное различие результатов. Традиционная ручная выборка из опубликованных томов и материалов СЯРП дала 365 словарных статей (или гнезд с объединением однокоренных единиц) – собственно названий растений. Это как широко употребительные единицы, представленные многочисленными контекстами (иногда сотнями примеров) – *береза*, *дуб*, *ива*, *роза*, *ромашка,* так и редкие: устаревшие, областные, терминологичные, индивидуально-авторские. В ПК НКРЯ был осуществлен поиск по семантическому параметру «растения» по пользовательскому подкорпусу, соотнесенному с материалом СЯРП (поэзия Серебряного века), количество найденных примеров – 43 491.

Поиск на ограниченном материале позволил обнаружить 293 лексемы, из которых было исключено 48: родовые понятия и их производные (*древо*, *деревце*; *кустарник, кустик; травка*, *травушка*, *мурава*, *былинка*; *цветик*, *цветочек*; *мох*, *лишайник*, *водоросль*), названия грибов, названия частей или совокупностей растений и др. Гораздо сложнее оказалась ситуация с омонимами. Так, лексемы *марь*, *череда* практически не встречаются реально, а многочисленные соответствующие употребления относятся к омонимичным лексемам: *Мария* (имя), *че-*

реда ‘ряд, очередь’ и *чередой*/*чередом* (нареч.). Учитывая несоразмерность реальной употребительности омонимов, возможно, такие единицы следовало бы лишить маркера «растение». Вручную было выбрано еще 113 лексем, обозначающих растения, которые не вошли в выборку по семантическому признаку. Одной из причин их отсутствия в выборке можно назвать невозможность маркирования индивидуально-авторских, редких, несловарных единиц. Но и вполне системные названия растений не получили семантического маркирования, если соответствующих словарных статей не было в лексикографических источниках, на которых семантическая разметка была основана.

Полученный сводный список был сопоставлен с выборкой из СЯРП (365 единиц) и толковых словарей (738 позиций) и дополнен лексемами, обнаруженными в поэтических текстах вручную (немаркированными в корпусе). Обобщенный словник был заново проверен поиском по каждой лексеме по ПК НКРЯ. Итоговый словник включает 825 позиций – существенно больше, чем в каждом из рассмотренных толковых словарей.

Если взглянуть на функционирование названий растений в языке поэзии – с точки зрения словаря и с точки зрения корпуса – мы можем заметить много интересного. Прежде всего, такая исследовательская задача позволяет не только ограничить объем материала, но и еще раз поставить вопрос о степени терминологичности фитонимов. Думается, если название растения встречается не только в специальной литературе, но и в стихотворном тексте, это должно говорить о том, что эта лексема вполне освоена языком, даже если в количественном отношении ее представленность невелика. Ср. примеры: *букс* (Сологуб, 1894 г.), *буксус* (Волошин, 1913 г.), *самшит* (Брюсов, 1916 г.); *гелиотроп* (Мережковский, 1885 г.); *глициния* (Бальмонт, 1900 г.); *далия* (Пастернак, 1917 г.); *дурро* (Гумилев, 1912 г.); *княженика* (Пастернак, 1917 г.); *крокус* (Кузмин, 1908 г.); *люпин* (Мартынов, 1967 г.); *манго* (Цветаева, 1925 г.); *пижма* (Хлебников, 1913 г.); *пьяника = голубика* (Клычков, 1922 г.); *флокс* (Брюсов, 1913 г.); *шпажник = гладиолус* (Бунин, 1903).

С одной стороны, добавление в инструментарий корпуса семантической разметки, включающей опцию «растения», стало серьезным шагом вперед в корпусной обработке такого сложного материала, как поэтический текст. С другой стороны, неудовлетворительность получаемых результатов напрямую связана с неадекватностью современного словарного описания этой группы лексики. Данные поиска в ПК НКРЯ подчеркивают те особенности описания фитонимов в толковых словарях, которые ярко свидетельствуют об их недостаточном осмыслении.

Таким образом, названия растений, как представляется, должны быть специально исследованы, а их основной массив должен включаться в толковые словари с учетом их хронологической и стилистической стратификации, неоднозначности и вариативности.

Использованные словари

АТоС – Академический толковый словарь русского языка. Т. 1. А – Вилять. Т. 2. Вина – Гяур / отв. ред. Л. П. Крысин. М. : Языки славян. культуры, 2016.

БАС-1 – Словарь современного русского литературного языка. Т. 1–17. М., 1948–1965. 17 т.

БАС-3 – Большой академический словарь русского языка. Т. 1. СПб. : Наука, 2004.

БЭС – Большая советская энциклопедия : [в 30 т.] / гл. ред. А. М. Прохоров. 3-е изд. М. : Сов. энциклопедия, 1969–1978.

МАС – Словарь русского языка. Т. 1–4 / под ред. А. П. Евгеньевой. 2-е изд. М., 1981–1984. 4 т.

НКРЯ – Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата обращения: 10.10.2022).

РОС – Русский орфографический словарь / Лопатин В. В., Иванова О. Е. (ред.). АКАДЕМОС. URL: http://orfo.ruslang.ru/ (дата обращения: 10.10.2022).

РСС – Русский семантический словарь / под общ. ред. акад. Шведовой Н. Ю. Т. 1. М. : Азбуковник, 2002.

СУ – Толковый словарь русского языка : в 4 т. / под ред. Д. Н. Ушакова. М., 1935–1940. 4 т.

СШ –Толковый словарь русского языка с включением сведений о происхождении слов / Шведова Н. Ю. (ред.). М. : Азбуковник, 2007. 1175 с.

СЯРП – Словарь языка русской поэзии XX века. Т. 1 / сост.: Григорьев В. П. (отв. ред.), Шестакова Л. Л. (отв. ред.), Колодяжная Л. И. (ред.), Кулева А. С. (ред.), Бакеркина В. В., Гик А. В., Реутт Т. Е., Фатеева Н. А. М. : Языки славянской культуры, 2001.

ТСРРР – Толковый словарь русской разговорной речи. Вып. 1 / отв. ред. Л. П. Крысин. М. : Языки славян. культуры, 2014.

Литература

Воронцов В. В. Садовые растения от А до Я. М. : Фитон XXI, 2013. 368 с.

Головкин Б. Н., Китаева Л. А., Немченко Э. П. Декоративные растения СССР. М. : Мысль, 1986. 320 с.

Крысин Л. П. Толковый словарь иноязычных слов. М. : Эксмо, 2010. 939 с.

Маевский П. Ф. Флора средней полосы европейской части России. 11-е изд. М. : Товарищество научных изданий КМК, 2014. 635 с.

Ожегов С. И. О трех типах толковых словарей современного русского языка // Вопросы языкознания. 1952. № 2. С. 85–103.

УДК-81’33

М. Н. Лошанина

Иркутский государственный университет, Иркутск, Россия

Анализ маркера противоречивости *все-таки* в устном дискурсе с применением программы MAXQDA

Описаны специфические характеристики текстов устного дискурса, в которых показан некоторый внутренний конфликт (или, по крайне мере, рассогласованность) представлений говорящего о ситуации. Выделены основные дискурсивные маркеры противоречивости, в том числе единица *все-таки*. Показаны основные преимущества программы MAXQDA как инструмента для анализа дискурсивных маркеров. Представлены результаты поиска единицы *все-таки* в текстах интервью с использованием программы MAXQDA и на основании результатов соотношения данных о частотности употребления дискурсивных единиц разными говорящими, намечены пути дальнейшего анализа маркеров противоречивости в устном дискурсе.

Ключевые слова: устный дискурс, MAXQDA, дискурсивные маркеры, маркеры противоречивости, все-таки.

Analysis of the Marker of Contradiction *vse-taki* in Oral Discourse Using the MAXQDA Program

The paper is devoted to the specific characteristics of oral discourse, which presentы some internal conflict (or inconsistency) of speaker’s ideas of the situation. The main discursive markers of inconsistency are distinguished, including *vse-taki*. The main advantages of the MAXQDA program as a discursive marker analysis tool are shown. By means of the MAXQDA program, the results of the search for a unit *vse-taki* are presented, and based on the results of the correlation of data on the frequency of use of discursive units by different speakers, ways of further analysis of contradiction markers in oral discourse are outlined.

Keywords: oral discourse, MAXQDA, discourse markers, contradiction markers, vse-taki.

Объектом данной работы являются нарративы, в которых показан некоторый внутренний конфликт (или, по крайне мере, рассогласованность) представлений говорящего о ситуации. Проблемы, связанные с интерпретацией таких нарративов, все чаще становятся объектом исследования не только в лингвистической науке (см. [Дроздова, 2011; Демьянков, 2011; Ирисханова, 2014 и др.]), но и в смежных с языковедческой наукой областях [Аронсон, 1984; Фестингер, 1984; 1999; Рикер, 1995 и др.]. Так, Д. Кэмерон, характеризуя специфику работы с разговорным дискурсом в социальных исследованиях, особенно отмечает необходимость иметь в виду те «противоречия», которые встречаются в рассматриваемых текстах: “It is not uncommon for informants in the course of an interview or a group discussion to give more than one account of the same thing, and sometimes these accounts may appear to **contradict one another**” (выделено нами. – *М. Л.*). Even in a single continuous sequence of talk [Cameron, 2001, p. 156].

Такие «противоречия» широко представлены в текстах глубинных интервью, полученных в ходе реализации проекта «Устная история Иркутска и технологии oral history в междисциплинарной перспективе», выполняемого на кафедре

русского языка и общего языкознания Иркутского государственного университета. Проект ориентирован на многоаспектный анализ текстов и аудиозаписей неподготовленных полуструктурированных интервью, в ходе которых информанты отвечают на вопросы интервьюера о собственном прошлом, истории своей семьи, города и страны. Приведем один выразительный пример:

Была война, конечно, ***трудное очень время****. Но мы все молодые были. Наверное, поэтому, как говорится,* ***все равно*** *находили время и* ***повеселиться, и экзамены сдавать вовремя****. И в общем* ***голодное время было****.* ***И голодное, и холодное****.*

Как видно, «генерируемый дискурс настолько захватывает говорящего, что он оказывается неспособен отрефлексировать противоречие, возникающее между порождаемыми им смыслами и субъективными оценками (выделены содержательно контрастирующие фрагменты)» [Ташлыкова, 2013, с. 45].

В свете сказанного актуальной исследовательской задачей является выявление дискурсивных маркеров, объективирующих подобные противоречия в процессе порождения устного дискурса.

Дебора Шиффрин дает следующее определение дискурсивных маркеров: "**Discourse markers are linguistic**, paralinguistic, or non-verbal **elements that signal relations between units of talk by virtue of their semantic and syntactic properties** (if any) and, most importantly, by virtue of their sequential position as initial or terminal brackets demarcating discourse units" [Schiffrin, 1987, p. 35–40].

К традиционным маркерам, выражающим противопоставление, относятся единицы: *но*, *однако*, *хоть*, *хотя*, *зато*, *все-таки*, *все равно*, *несмотря на*, *тем не менее*, *в то же время* и т. д.

Поиск всех употреблений единиц, маркирующих противоречие, в текстах глубинных интервью, не имеющих цифрового лингвистического корпуса, был осуществлен в программе MAXQDA[1], которая была разработана для анализа неструктурированных письменных и устных данных.

Безусловным преимуществом программы MAXQDA является функция автокодирования, т. е. поиска единицы в контексте и присваивание ей определенного кода: «хотя», «все-таки», «все равно» и т. д. Заметим, что код присваивается не единице, а окружающему ее контексту, для которого на этапе разметки можно задать любые параметры. В созданном нами корпусе из 30 интервью был осуществлен поиск по параметрам «10 слов до единицы – 10 слов после единицы» (рис. 1), после чего все контексты, в которых употреблена единица, были закодированы.

Безусловно, такую разметку можно выполнить, воспользовавшись в качестве инструмента Национальным корпусом русского языка (НКРЯ) или комбинацией CTRL+F, однако функция автокодирования в программе MAXQDA поз-

[1] MAXQDA доступна как платное лицензированное приложение как для Windows, так и для Mac OS X (доступна бесплатная бета-версия для 14-дневного тестирования). Дополнительная информация о программе представлена на официальном сайте разработчика (http://www.maxqda.de/).

воляет осуществить поиск сразу нескольких единиц одновременно во всех интервью и, благодаря тому что единицы представлены в виде тегов, автоматически переключаться между разными контекстами.

Начало	Контекст	Ключевое слово	Контекст
15: 793	нии? Ээ... вот и, конечно, приходилось много чего делать, и плюсы, и минусы, вот	все-таки	организация была лучше в городе поставлена. Тогда никто не имел права
41: 1999	России. Вот... Ну и что, образования-то не было, а уже понимала, что надо же	все-таки	учиться чему-то. Так, работала в разных организациях, в том числе сначала в
46: 178	наверняка, можно было меньшим количеством жертв обойтись. Ну что	все-таки	в этой войне в страшной. А потом стройки эти. Вот все это... люди работали. Просто
48: 952	там должен и жить, потому что ты приспособлен к этим условиям. Это... эээ...	все-таки	идет: гены никуда не денешь. Приспосабливаться плохо. А так
52: 794	оказалась на Украине, выйдя замуж за инженера, который был направлен в	все-таки	здесь, в Сибири. Именно Сибирь мы считаем своей родиной и самым любимым.
52: 1051	зье, бывала на Востоке. Имела возможность остаться на юге, там мой муж-	все-таки	нас тянуло всегда сюда, в Си- бирь. Вплоть до того, что мы написали ходатайство,
52: 1277	жизнь наша была пол- на всех благ, какие многие мечтают, желая жить на юге. А	все-таки	тянуло сюда, на Байкал, в Сибирь. – А расскажите, пожалуйста, каким было ваше
57: 1896	и я посмотрела. Я много побывала везде: и на востоке, и на юге. Но всегда	все-таки	было желание вернуться опять сюда. И я считаю, что наш климат самый

Рис. 1. Контексты с маркером противоречивости *все-таки*

Единицы, преобразованные в код, автоматически формируются в список кодов (рис. 2):

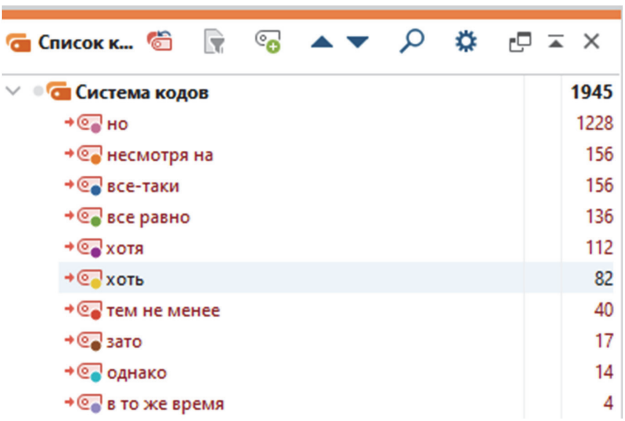


Рис. 2. Количественные корреляции кодов

Согласно результатам автокодирования видно, что единица *все-таки* вошла в тройку лидеров по частотности употребления (156 употреблений).

Прежде чем переходить к анализу специфики функционирования дискурсивного маркера противоречивости *все-таки* в текстах интервью, вкратце опишем его семантику. Результаты анализа семантики частицы *все-таки* с учетом данных «Активного словаря русского языка» и НКРЯ показали, что лексема *все-таки: 1)* используется как маркер противопоставления имеющегося положения вещей гипотетическому, но более стандартному, типичному, «естественному», ср.: *Лил дождь* (*А1*)[1]*, но мы* ***все-таки*** *пошли гулять* (*А2*); *2)* как маркер причины, обусловившей реализацию одной из двух альтернатив, ср.: *Хотя экзамен сложный* (*А1*)*, она сдала на пятерку* (*А2*) – ***все-таки*** *круглая отличница* (*А3*); *3*) как маркер эмоциональной реакции говорящего на совпадение фрагментов ментального и реального мира, ср.: *Какие* ***все-таки*** *в Москве высокие цены*.

Результаты поиска контекстов с лексемами *все-таки 1*, *все-таки 2*, *все-таки 3* представлены в виде интерактивного дерева (рис. 3). Интерактивное древо имеет следующую структуру: корень дерева образует слово, расположенное в левой части, в данном случае это единица *все-таки*; ветви дерева, расположенные правее, выстраиваются с учетом частотности появления слова в созданном корпусе. При наведении курсора мыши на слово появляется диалоговое окно, в котором отображается частотность появления этого слова в анализируемых текстах.

Рис. 3. Интерактивное дерево единицы *все-таки*

Программа MAXQDA имеет функцию визуализации данных и результатов, среди которых «Портрет документа», «Браузер кодовых связей», «Облако слов», «Диаграмма сравнения документов» и др. (рис. 4).

Визуализация данных с помощью диаграммы сравнения документов дала очень, на наш взгляд, значимый результат для дальнейшего исследования. Бросается в глаза принципиальное различие в частотности использования маркеров в начальных и финальных фрагментах интервью (рис. 5, *а, б*). Если учесть, что

[1] Здесь и далее (*А1*), (*А2*), (*А3*) обозначены сопоставляемые в высказывании ситуации (одна из которых в явном виде не названа).

интервью состоит из нескольких блоков («Детство», «Учеба», «Годы войны», «Работа», «город Иркутск»), то имеет ли значение высокая частотность появления маркеров противоречивости в первых блоках интервью о детстве и юности интервьюеров? Связаны ли социальные характеристики интервьюируемых (пол, возраст, профессия) с типом употребляемых в речи единиц, выражающих противоречие (или рассогласованность), и их частотностью? Существует ли зависимость частотности маркеров, выражающих противоречие, от типа нарратива и его тематики?

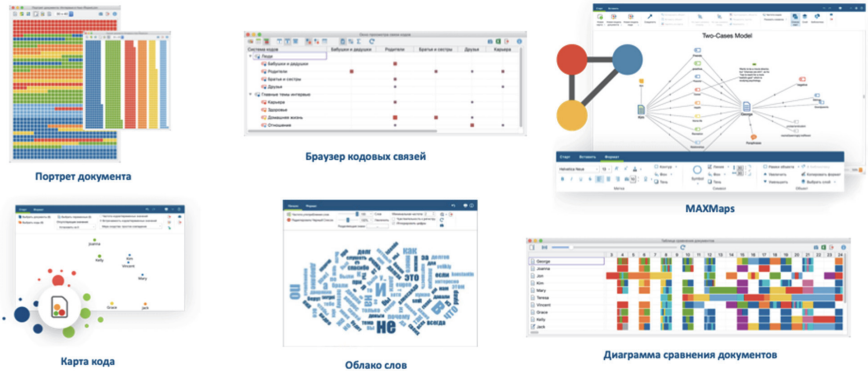


Рис. 4. Инструменты визуализации данных и результатов в MAXQDA

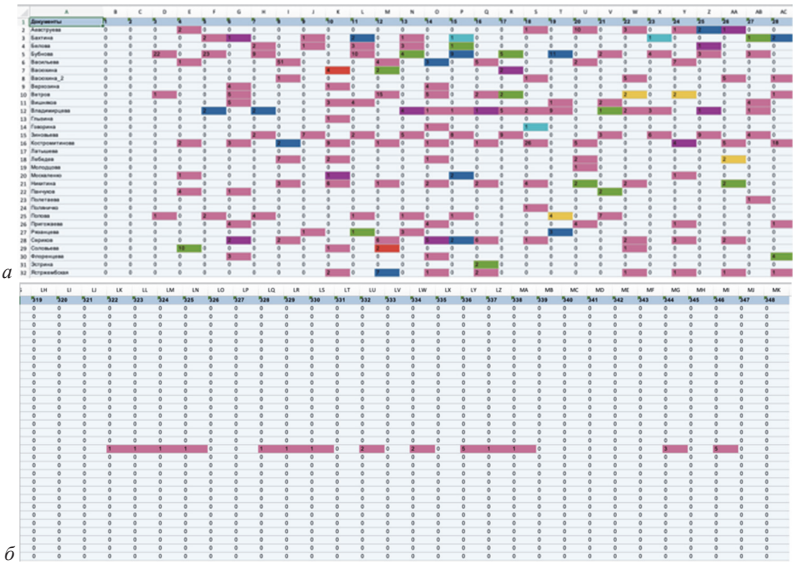


Рис. 5. Матрица сравнения кодов в программе MAXQDA:
а – начальный фрагмент интервью; *б* – финальный фрагмент интервью

Анализ матрицы сравнения кодов выходит за рамки данной статьи, исследовательскую задачу возможно будет решить на следующем этапе работы с данным материалом, отвечая на вопросы, сформулированные выше.

Литература

Аронсон Э. Теория диссонанса: прогресс и проблемы // Современная зарубежная социальная психология. М. : Изд-во МГУ, 1984. С. 111–126.

Демьянков В. З. Когнитивный диссонанс: когниция языковая и внеязыковая // Когнитивные исследования языка. Взаимодействие когнитивных языковых структур. М. ; Тамбов : Изд-во ТГУ, 2011. Вып. 9. С. 33–40.

Дроздова Т. В. Когнитивный диссонанс как лингвистическая проблема : автореф. дис. ... канд. филол. наук. Белгород, 2011. 24 с.

Ирисханова К. М. Когнитивный диссонанс в англоязычном поэтическом дискурсе // Вестник МГЛУ, 2014. Вып. 17 (703). С. 7–14.

Рикер П. Конфликт интерпретаций: очерки о герменевтике. М. : Academia-Центр, 1995. 441 с.

Ташлыкова М. Б. Языковые и дискурсивные маркеры конструирования образа прошлого // Советское как дискурсивный феномен: способы концептуализации прошлого: монография / под общ. ред. О. Л. Михалёвой. Иркутск : Изд-во ИГУ, 2013. С. 25–49.

Фестингер Л. Введение в теорию когнитивного диссонанса // Современная зарубежная социальная психология. М. : Изд-во МГУ, 1984. С. 97–110.

Cameron D. Working with spoken discourse. Sage Publ., 2001. 206 p.

Schiffrin D. Discourse markers. Cambridge : Cambridge University Press, 1987. 364 p.

УДК 81’33

У. Э. Чекмез

Иркутский государственный университет, Иркутск, Россия

Индивидуальные предпочтения говорящего как фактор, влияющий на выбор леммы для маркирования обратной связи (на примере единиц *аγа* и *мщм*)

Анализируется количественное распределение маркеров обратной связи *аγа* и *мщм* в рамках исследовательского корпуса. С использованием статистических методов выявляются значимые отличия в имеющемся распределении, формулируются отдельные гипотезы, касающиеся факторов, влияющих на разницу в частотности употребления анализируемых маркеров обратной связи. Подтверждается гипотеза о том, что разные говорящие имеют склонность к использованию в своей речи какого-либо МОС в большей степени, однако в рамках исследуемого корпуса наличие индивидуальных предпочтений зафиксировано только для МОС *аγа,* что позволяет сделать вывод о том, что индивидуальные предпочтения говорящих не являются основным предиктором, с помощью которого можно было бы объяснить преобладание в различных лингвистических корпусах МОС *мщм.*

Ключевые слова: устный дискурс, маркеры обратной связи, лингвистический корпус, статистическая гипотеза, коэффициент корреляции Пирсона.

Individual Preferences of the Speaker as a Factor Influencing the Choice of Lemma for Feedback Marking (Using the Example of Units *аγа* and *мщм*)

The article analyzes the quantitative distribution of backchannels *аγа* and *мщм* within the research body. Using statistical methods, significant differences in the existing distribution are identified, separate hypotheses are formulated concerning the factors affecting the difference in the frequency of use of the analyzed backchannels. The article confirms the hypothesis that different speakers have a tendency to use some backchannels in their speech to a greater extent. However, within the framework of the corpus under study, the presence of individual preferences was recorded only for backchannels *аγа*. This allows us to conclude that individual preferences of speakers are not the main predictor by which the predominance of backchannels *мщм* in various language corpora can be explained.

Keywords: oral discourse, backchannels, linguistic corpus, statistical hypothesis, Pearson’s r.

В рамках данной статьи рассматриваются употребления двух единиц, которые могут быть квалифицированы как маркеры обратной связи, а именно единиц *аγа* и *мщм.* Обычно под маркерами обратной связи, вслед за В. Ингве [Yngve, 1970], понимают «короткие вербальные или невербальные ответные реакции, которые используются Слушающим для выражения позитивного внимания к Говорящему без каких-либо попыток поменяться с Говорящим ролями» [Кобозева, Иванова, Захаров, 2019, с. 291].

Функционирование единиц *аγа* и *мщм* в качестве маркеров обратной связи (далее – МОС) можно наблюдать в примере, показанном в табл. 1.

Таблица 1

Line	*t* начала ЭДЕ	*t* конца ЭДЕ	В. И.	М. Б.	О. Л.
65	01:20.72	01:22.71	так а /такой-то снег он сразу замерзнет		
66	01:22.71	01:23.09		мщм	
67	01:23.41	01:24.66	это надо /кочуну		
68	01:24.74	01:25.34	вот утром /встаешь		
69	01:25.34	01:26.25	кочун \упал		
70	01:26.25				аɣа
71	01:26.56	01:26.62	его надо /наметать этот снег		
72		01:28.75			
73	01:28.27			мщм	
74	01:28.55	01:28.69			мщм
75		01:28.80			

Как видим, основным говорящим в рамках данного фрагмента дискурса выступает В. И., М. Б. и О. Л. выражают позитивное внимание к речевой партии ВИ, не пытаясь поменяться с ним ролями, поэтому можно утверждать, что в такого рода контекстах эти единицы используются в качестве МОС.

Данные единицы последовательно размечались в исследовательском корпусе, состоящем из одиннадцати фрагментов интервью (43 мин 45 с устного диалогического дискурса ≈ 7500 словоупотреблений, 2415 ЭДЕ), записанных в 2012–2014 гг. во время экспедиции в Баргузинский район Республики Бурятия. В корпусе были зафиксированы следующие маркеры обратной связи: *аɣа, мщм, но, а, м, да, конечно, так, понятно, ясно*.

Среди единиц, способных выступать в качестве маркера обратной связи, абсолютным лидером по частоте употребления в рамках исследуемого корпуса является лемма *мщм,* которая была использована говорящими 302 раза, второе место по частоте использования занимает лемма *аɣа,* которая встретилась в корпусе 98 раз, т. е. единицы *мщм* и *аɣа,* выступая в качестве отдельных ЭДЕ, составляют шестую часть (16,6 %) от всех ЭДЕ исследовательского корпуса. Несмотря на высокую частотность употребления таких единиц в рамках устного диалогического дискурса, маркеры обратной связи либо оказываются за пределами внимания лингвистов, либо описываются как нерасчленненная группа единиц с одинаковыми функциональными характеристиками. «Реплики-реакции собеседника носят в основном поддерживающий характер и дают понять говорящему, что в его сообщении заинтересованы» [Розанова, 2018, с. 16]. Маркерам обратной связи зачастую отказывают в наличии какой бы то ни было семантики: «ДМОС информационно избыточны, т. е. они не добавляют в диалог новой информации (ни пропозициональной, ни модусной)» [Кобозева, Иванова, Захаров, 2019, с. 293].

Такой подход к описанию маркеров обратной связи дает основание предполагать, что для выражения эмпатии и заинтересованности в устном диалогическом дискурсе может использоваться любая из этих единиц. Предположения такого рода легко трансформируются в статистические гипотезы, которые можно проверить, используя непараметрические методы статистического анализа, в частности, критерий хи-квадрат Пирсона. В качестве материала для тестирования специально избраны два маркера *мщм* и *ауа,* различия в употреблении которых в функции маркера обратной связи представляются наименее очевидными.

Сформулируем статистические гипотезы для тестирования.

H0. Леммы *мщм* и *ауа* равновероятно встречаются в рамках исследовательского корпуса.

H1. Распределение лемм *мщм* и *ауа* в рамках исследовательского корпуса значимо отличается от равномерного.

Опираясь на частоту употребления лемм *мщм* и *ауа* в рамках исследовательского корпуса, рассчитаем статистические значения. При значении $\chi^2 = 104{,}04$[1], df = 1 p-value стремится к нулю[2] (меньше 0,05), следовательно, можно утверждать, что верна гипотеза H1, а это значит, что распределение лемм *мщм* и *ауа* в рамках исследовательского корпуса значимо отличается от равномерного.

Можно допустить, что такое распределение характерно только для нашего исследовательского корпуса, состоящего из интервью. Чтобы убедиться в обратном, обратимся к другим устным корпусам, фиксирующим в аннотации маркеры обратной связи: демонстрационной части корпуса «Рассказы и разговоры о грушах»[3] и корпусу МУРКО в составе НКРЯ[4]. В рамках демонстрационной части корпуса «Рассказы и разговоры о грушах» обнаружено 15 употреблений леммы *ага* и 122 употребления леммы *угу*[5]. В корпусе МУРКО по запросу «угу, bdot & capital» было обнаружено 3632 употребления леммы, по запросу «ага, bdot & capital» было обнаружено 2019 употреблений леммы. Ограничения bdot & capital были введены для того, чтобы избавиться от шума[6]. Как видим, в рамках нашего исследовательского корпуса *ауа* употребляется приблизительно в три раза реже *мщм*, в МУРКО – почти в два раза, а в демонстрационном варианте корпуса «Рассказы и разговоры о грушах» – в восемь раз. Как видим, ни в одном из корпусов

[1] С поправкой Йетса $\chi^2 = 103$, df = 1, p-value = 0

[2] Для расчета p-value использовался онлайн-калькулятор. URL: https://vavilovva.shinyapps.io/dist_calc2/.

[3] Русский мультиканальный дискурс. URL: https://www.multidiscourse.ru

[4] Мультимедийный корпус. URL: https://ruscorpora.ru/new/search-murco.html

[5] Подробнее об особенностях аннотирования маркеров обратной связи в статье [Михалёва, Чекмез, 2021].

[6] Приведем примеры шума, обнаруженные в МУРКО по запросам *ага* и *угу*.

Ага / значит / щас я должен написать «банка».

Среди механизмов обратной связи – и кивки / и всякие «угу» / и направление взгляда.

Очевидно, что в этих контекстах леммы выступают в функциях, отличных от функций маркеров обратной связи.

мщм и *аγа* не встречаются равновероятно, отличие в частотности их употребления является статистически значимым[1], поэтому возникает вопрос о факторах, влияющих на выбор говорящим той или иной леммы для выражения позитивного внимания к говорящему.

В данной статье остановимся на одном из возможных предикторов, в качестве которого могут выступать индивидуальные предпочтения говорящих. Стоит отметить, что использование лексемы *предпочтения* в данном случае довольно условно, так как предполагает осуществление говорящим осознанного выбора между этими единицами, который, очевидно, отсутствует. Под индивидуальными предпочтениями в данном случае понимается, скорее, склонность того или иного говорящего употреблять в своей речи одну из этих единиц для выражения эмпатии и заинтересованности. Так, в рамках корпуса наблюдается различное соотношение *аγа* и *мщм* в речи разных говорящих: в речи одних преобладает использование единицы *мщм*, в речи других – *аγа*.

В рамках исследовательского корпуса большая часть употреблений ожидаемо содержится в речи интервьюеров (384 из 400), поэтому для проверки гипотезы о наличии корреляции между количеством употреблений лемм *мщм* и *аγа* и индивидуальными предпочтениями говорящих используем тот же критерий и проведем анализ стандартизованных остатков, применяя визуализацию.

В табл. 2 показана сопряженность для употреблений данных единиц тремя интервьюерами.

Таблица 2

	аγа	мщм	
МБ	44	219	**263**
ЕВ	21	28	**49**
ОЛ	22	50	**72**
	87	**297**	**384**

Как видим, в данном случае наблюдается довольно сильный дисбаланс классов: количество употреблений маркеров обратной связи *мщм* и *аγа* в речи М. Б. в 3,7 раз больше, чем у О. Л., и в 5,4 раз больше, чем у Е. В. Можно предположить, что дисбаланс такого рода может быть порожден представленностью речевых партий интервьюеров в разных фрагментах интервью (речь М. Б. может быть зафиксирована в большем количестве фрагментов интервью, что может обусловливать наличие большего количества употреблений МОС), однако из 11 имеющихся в исследовательском корпусе фрагментов примерно равного объема в большем количестве последних зафиксирована речевая партия О. Л. (10 из 11), в то время как речевая партия М. Б. представлена в восьми фрагментах интервью, следовательно, такое количество употреблений МОС обусловлено другой при-

[1] «Рассказы и разговоры о грушах». χ^2= 82, df = 1, p-value = 0 (с поправкой Йетса).
МУРКО. χ^2= 230.2, df = 1, p-value = 0 (с поправкой Йетса).

чиной. В частности, можно предположить, что разные говорящие в разной степени выражают заинтересованность посредством маркеров обратной связи, иными словами, одни говорящие склонны эксплицировать обратную связь через вокальный канал, в то время как другие могут маркировать ее посредством кинетической модальности (например, кивков). В рамках данного исследования невозможно подтвердить или опровергнуть такую гипотезу, так как её верификация потребовала бы анализа видеозаписи, тогда как для исследования доступны только аудиозаписи.

Даже при имеющемся дисбалансе классов мы можем проверить гипотезу о том, что разные говорящие имеют склонность использовать в своей речи один из МОС (*мщм* или *ауа*), что, в свою очередь, может влиять на количественное распределение данных единиц. Рассчитаем хи-квадрат для приведенной выше таблицы сопряженности с помощью R[1]: X-squared = 19,246, df = 2, p-value = 6,619e-05. Как видим, значение p-value меньше нуля, что позволяет говорить о наличии взаимосвязи между номинативными переменными «говорящие» и «МОС». Для того чтобы понять, в каких именно ячейках таблицы сопряженности наблюдается наиболее сильное отклонение наблюдаемых значений от ожидаемых, визуализируем данные, построив в R график с помощью функции mosaicplot (рис. 1).

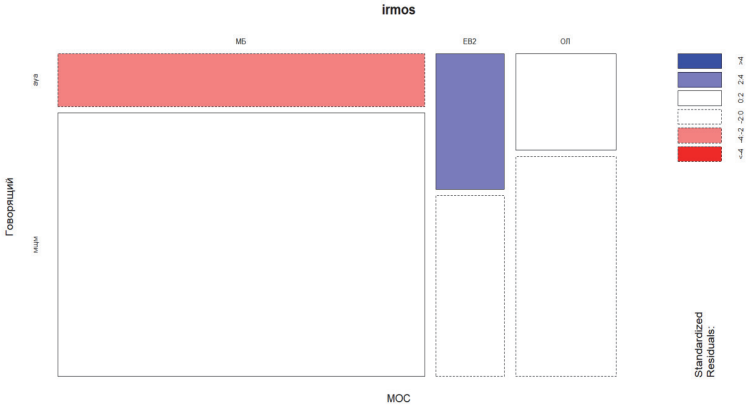


Рис. 1. Анализ стандартизованных остатков

Как видим, значимые отклонения наблюдаются только в двух ячейках таблицы сопряженности. Данные, представленные на графике, позволяют заключить, что М. Б. употребляет *ауа* значимо реже, чем предполагает теоретическое распределение с учетом имеющегося дисбаланса, в то время как Е. В. употребляет *ауа* значимо чаще, по сравнению с предсказываемыми значениями, которые наблюдались бы при отсутствии взаимосвязи между переменными.

[1] Для работы используется программа RStudio.

Стоит отметить, что проведенный тест предоставляет возможность опровергнуть ещё одно предположение, касающееся преобладания в диалогическом дискурсе МОС *мщм.* Многие исследователи объясняют преобладание *мщм* действием закона экономии речевых усилий: так как *мщм* произносится с закрытым ртом, на его произнесение тратится меньше усилий, чем на произнесение *ауа,* для которого требуется открыть рот, поэтому говорящие в большинстве случаев выбирают наименее ресурсозатратную единицу. Представляется, что, если бы данное утверждение было справедливо, невозможно было бы обнаружить носителя языка, который вопреки закону экономии речевых усилий предпочитал бы произносить наиболее энергозатратную единицу, однако данные, характеризующие речевую партию Е. В. с точки зрения употребления в ней маркеров обратной связи, наглядно демонстрируют, что в рамках речевой партии одного говорящего могут как минимум с равной вероятностью использоваться оба этих маркера.

Таким образом, можно утверждать, что на употребление маркера обратной связи *ауа* могут влиять индивидуальные предпочтения говорящих, однако данный фактор невозможно признать релевантным для употребления *мщм*, что побуждает нас искать иные предикторы для объяснения значимого преобладания употреблений *мщм* над другими маркерами обратной связи.

Очевидно, что предпочтения говорящих могут быть обусловлены как субъективными факторами (какая-то единица «нравится» больше, какая-то меньше), так и объективными факторами, например, необходимостью использовать одну из языковых единиц в рамках определенного контекста.

Приведем пример, в котором отражается наиболее частый тип контекста, в котором Е. В. использует *ауа* (табл. 3).

Таблица 3

Line	*t* начала ЭДЕ	*t* конца ЭДЕ	Г. И.	Е. В.
40	00:50.59	00:52.11		слышали такое \чаево́е?
41	00:52.11	00:53.06	да у нас-то тут больше	
42	00:53.87	00:54.44	печення	
43	00:54.56			ауа
44	00:54.74	00:54.94	печення \называли н= \печення	
45		00:56.88		
46	00:56.97			ауа

В этом фрагменте Е. В. спрашивает диалектоносителя о том, знает ли он то или иное диалектное слово. В рамках данного исследования такого рода употребления также интерпретировались как МОС, так как единица *ауа* в контекстах такого типа также выражает позитивное внимание к говорящему без попытки поменяться с ним ролями. Однако, если мы сравним этот пример с примером в табл. 1, окажутся очевидными различия, имеющиеся даже на поверхностном уровне: рассказ В. И. является целостной продолжительной речевой партией, в

то время как реплики Г. И. являются короткими ответами на вопрос интервьюера. Контексты, аналогичные данному, встречаются 6 раз в рамках корпуса, и МОС *аγа* во всех этих контекстах употребляет Е. В.

Таким образом, количественные данные и качественный анализ примеров позволяет предположить, что МОС появляются в рамках диалогического дискурса непроизвольно и неслучайно: имея значительное сходство в семантике (способность выражать позитивное внимание к говорящему), они употребляются в разном контекстуальном окружении, а следовательно, способны добавлять в диалог новую информацию, однако верификация данного предположения составляет перспективу исследования, так как требует, во-первых, увеличения количества материала (в том числе за счет увеличения числа говорящих), во-вторых, тонкого лингвистического анализа, который на следующем этапе исследования позволит определить, что может выступать в качестве предикторов, в частности, чем именно отличаются функции этих единиц в различных контекстах, в том числе аналогичных приведенным.

Литература

Кобозева И. М. Иванова О. О., Захаров Л. М. К мультимодальному моделированию верификативных дискурсивных маркеров в русском диалоге // Труды Института русского языка им. В. В. Виноградова. 2019. Вып. 21. С. 284–299.

Михалёва О. Л., Чекмез У. Э. Дискурсивная транскрипция устной речи диалогического типа // Филологические науки. Вопросы теории и практики. 2021. Т. 14, вып. 6. С. 1750–1757.

Мультимедийный русский корпус. URL: https://ruscorpora.ru/new/search-murco.html (дата обращения: 15.01.2023).

Розанова Н. Н. Введение // Русское повседневное общение: прагматика, культурология : монография / под науч. ред. И. Н. Борисовой. Екатеринбург : Гуманит. ун-т, 2018. С. 10–41.

Русский мультиканальный дискурс. URL: https://www.multidiscourse.ru (дата обращения: 15.01.2023).

Yngve V. On getting a word in edgewise // Papers from the sixth regional meeting. Chicago Linguistic Society, 1970. P. 567–578.

УДК 81’11

А. Б. Чимитова

Иркутский государственный университет, Иркутск, Россия

Использование корпусных технологий для повышения верифицируемости исследования английских фразовых глаголов

Рассматриваются возможности применения корпусных технологий в лингвистических исследованиях. Объектом изучения выступают английские фразовые глаголы с глагольными компонентами *cool, warm, hot*. Проводится анализ значения английских фразовых глаголов. Приведены примеры контекстного употребления фразовых глаголов. Рассмотрен термин «коллокация» в корпусной лингвистике. На основе проанализированных значений фразовые глаголы объединены в концептуальные области.

Ключевые слова: фразовые глаголы, верифицируемость, корпусные технологии, коллокации, послелог, концептуальная область.

The Use of Corpus Technologies To Increase the Verifiability of the Study of English Phrasal Verbs

The article discusses the possibilities of using corpus technologies in linguistic research. The object of study is English phrasal verbs with verb components cool, warm, hot. The analysis of the meaning of English phrasal verbs is carried out. Examples of the contextual use of phrasal verbs are given. The term “collocation” in corpus linguistics is considered. On the basis of the analyzed meanings, phrasal verbs are combined into conceptual domains.

Keywords: phrasal verbs, verification, corpus technologies, collocations, particle, conceptual domains.

На современном этапе развития науки лингвистические исследования трудно представить без обращения к корпусам текстов [Захаров, Богданова, 2020]. Корпус предоставляет возможность решать задачи, связанные с обследованием больших массивов текстов [Плунгян, 2009]. Применение корпусных технологий в лингвистике позволяет исследователям решать ряд задач, к числу которых относится изучение языковых единиц с опорой на контекст, в частности, изучение коллокаций.

Настоящая работа посвящена исследованию семантики английских фразовых глаголов на материале данных корпусов текстов BNC и COCA [British National Corpus … ; Corpus of Contemporary American English …]. Нами была исследована группа фразовых глаголов с глагольными компонентами cool, warm, hot. С помощью данных корпусов текстов проводится анализ прямых и переосмысленных значений фразовых глаголов.

Согласно данным словаря Oxford dictionary of phrasal verbs и анализу контекстного употребления фразовых глаголов в корпусах текстов BNC и COCA, рассматриваемые глагольные компоненты образуют фразовые глаголы путем присоединения послелогов down (cool down), off (cool off), out (cool out), through

(warm through), up (warm up, hot up). Поиск сочетаний глагольных компонентов и послелогов производился путем поиска коллокаций в корпусах текстов. Под термином «коллокации» в корпусной лингвистике понимают статистически устойчивые сочетания, которые могут быть как фразеологическими, так и свободными [Павельева, 2016].

Для поиска фразовых глаголов в окне word/phrase необходимо ввести рассматриваемый глагольный компонент, а в качестве коллоката выбрать адвербиальную частицу (рис. 1). Согласно выбранным параметрам, корпус текстов выдает список адвербиальных частиц, которые являются коллокатами для заданного глагола (рис. 2).

Рис. 1. Поиск фразовых глаголов в корпусе текстов BNC

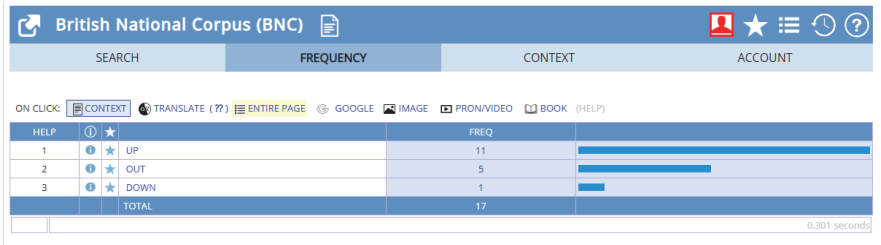


Рис. 2. Адвербиальные частицы для глагола hot

Как следует из определения термина «коллокация», рассматриваемые сочетания могут быть как фразеологическими, так и свободными. Для различения фразового глагола и свободного сочетания слов необходимо проанализировать употребление рассматриваемых единиц в контексте (рис. 3).

Например, из анализа контекстного употребления сочетания глагола hot и частицы off следует, что hot off является не фразовым глаголом, а свободным сочетанием (рис. 4).

Приведем примеры значений фразовых глаголов из словаря Oxford dictionary of phrasal verbs, каждое из значений будет подтверждено примерами контекстного употребления из корпусов текстов BNC и COCA.

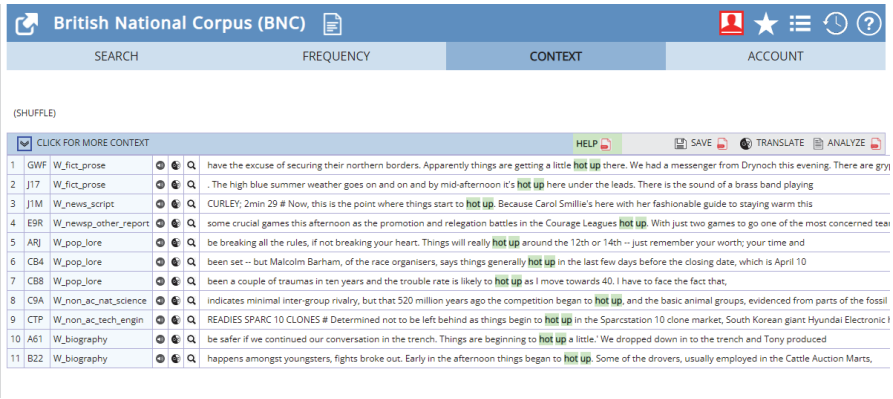


Рис. 3. Контекстное употребление глагола hot up

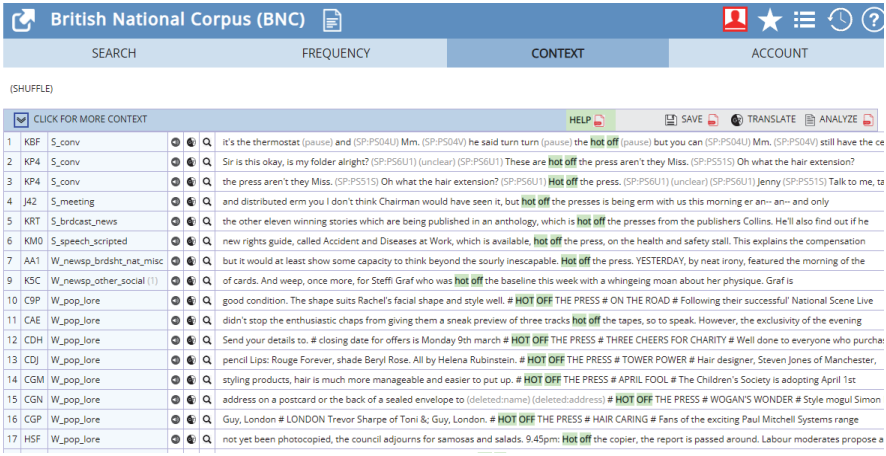


Рис. 4. Контекстное употребление сочетания глагола hot и частицы off

У фразового глагола cool down выделяется следующее значение: become or make (sb) cool or calm [Cowie, Mackin, 2011]. Согласно данным корпусов текстов, представляется возможным выделить два значения данного глагола:

1) остывать (о температуре, жидкости, газе и т. д.)

The gentle breeze managed to cool down the heat of the sun a little (BNC).

The mixture was heated to 90°C and cooled down slowly to form the hybrids (BNC).

I recently started taking walks in the evenings when the weather cools down (COCA).

2) успокаивать, успокаиваться

Besides, she probably hadn't got over being angry; she needed time to cool down (BNC).

My only choice was to make myself cool down, and try to approach the situation with patience and clarity (COCA).

You need time to cool down and relax before your session (COCA).

Фразовый глагол cool off имеет следующее значение become or make (sb) less warm, excited, ardent or interested [Cowie, Mackin, 2011]. Приведем примеры контекстного употребления рассматриваемого глагола из корпусов текстов:

1) остывать (о температуре, жидкости, газе и т. д.)

Laura placed her iron by the sink to cool off (BNC).

We cooled off on our way back to camp with a swim in a roadside lake (BNC).

A cup of coffee was in front of him, but it had long been cooled off, steam no longer rising from the mug (COCA).

2) успокаивать, успокаиваться

She had to cool off, she had to rest (BNC).

Pushing the blankets to the bottom of the bed, I tried to cool off and get a little more sleep (COCA).

There he could find a wife and spend some time until his brother's anger cooled off (COCA).

Рассмотренные фразовые глаголы являются синонимичными по каждому из выделенных значений. Глаголы cool down и cool off обозначают уменьшение свойства предмета (температуры, настроения), что находит свое выражение в послелогах down и off.

Фразовый глагол cool out также является синонимичным глаголам cool down и cool off. Однако у cool out выделяется лишь одно значение – успокаиваться, расслабляться. Важно отметить, что глагол cool out не фиксируется в словаре Oxford dictionary of phrasal verbs, примеры употребления фразового глагола cool out представлены только в корпусе COCA, что свидетельствует о его использовании преимущественно в американском варианте английского языка.

Everybody just cool out. Everybody calm down (COCA).

He worked really well like he always does and then cooled out fine (COCA).

I gave her the day off so she could just go home and cool out, you know (COCA).

Противоположные по значению отадъективные глаголы warm и hot присоединяют послелоги up и through. В словаре Oxford dictionary of phrasal verbs фразовый глагол warm up имеет следующие значения:

1) become warm or warmer

Tony measures the temperature and tells me it is warming up (BNC).

I warm up baked beans for the children's (BNC).

The weather is slowly but surely warming up and we have had wonderful sunshine and even some days of rain… (COCA).

2) reach the point, after its parts are warm and working smoothly, where it can run properly

She was warming up her car to go to work (COCA).

Vologsky made no move to close the hood, or warm up the aircraft's engines (BNC).

She tried to scoot along the pavement, hoping to reach the car she was warming up and her cellphone to call for help (COCA).

3) take exercises to loosen one's muscles etc, before a game and etc

Before you undertake any form of exercise it is very important that you warm up properly. Failure to do so is likely to cause injury (BNC).

Skipping with a rope for a few-minutes is also a good way to warm up (BNC).

Brittany, shouldn't you be stretching or warming up or something?(COCA).

4) become more lively and interested

Doctor began to sing; he was a great man at a party and he warmed up the crowd no end (BNC).

I'm warming up to the idea of a little performance for you (COCA).

She loses her parents from the start and guards herself from warming up to people in fear of losing them (COCA).

Согласно данным корпусов текстов, синонимичным глаголу warm up является глагол warm through, но только в значении нагревать, разогревать. В словаре значения данного глагола не фиксируются.

Alternatively the pie can be made well in advance and then warmed through in a moderate oven for 20 minutes (BNC).

Add in the vegetable stock and stir until warmed through (COCA).

Stir for 5 minutes, until the vegetables are warmed through and coated with the sauce (BNC).

Фразовый глагол hot up имеет только одно значение – intensify, increase [Cowie, Mackin, 2011].

The battle for territory was hotting up before the today's London peace conference (BNC).

Things are really, really hotting up around here (COCA).

The shooting war died down because the propaganda war is hotting up (COCA).

Для уточнения и конкретизации значений фразовых глаголов проведем анализ коллокаций на примере фразового глагола warm up. Сочетаемость со словами exercise, audience, crowd и подтвержденное контекстом значение позволяют отнести фразовый глагол к концептуальной области «подготовка, разминка».

Регулярное использование определенных пространственных послелогов, характеризующих разные пространственные отношения, как взаимозаменяемых или, по крайней мере, синонимичных приводит к стиранию в сознании четких границ между образами этих отношений и к их интеграции в объемную, многомерную концептуальную область, под которой понимается совокупность концептов, объединенных репрезентацией одного фрагмента действительности [Богданова, 2007, с. 10].

На основе значений фразовых глаголов в словарях, их контекстного употребления в корпусах текстов и анализа коллокатов считаем возможным объединить фразовые глаголы в концептуальные области «охлаждение», «успокоение», «нагревание», «подготовка, разминка», «усиление» (табл.).

Таблица

Концептуальные области и характеризующие их фразовые глаголы

№	Концептуальная область	Фразовые глаголы
1	Охлаждение	Cool down, cool off
2	Успокоение	Cool down, cool off, cool out
3	Нагревание	Warm up, warm through
4	Подготовка, разминка	Warm up
5	Усиление	Hot up

Корпусы текстов представляют собой инструмент, с помощью которого могут быть решены различные лингвистические задачи. Анализ контекстного употребления фразовых глаголов в корпусах текстов подтверждает значения, зафиксированные в словарях, а также позволяет выделять незафиксированные в словарях значения (на примере глаголов cool out, warm through). С помощью корпусов текстов представляется возможным уточнение прямых и переосмысленных значений фразовых глаголов. На примере исследования фразовых глаголов нами было показано, как применение корпусных технологий способствует повышению верифицируемости лингвистических исследований.

Литература

Богданова С. Ю. Концептуализация и реконцептуализация пространственных отношений (на материале английских фразовых глаголов) : дис. ... д-ра филол. наук. Волгоград, 2007. 350 с.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика : учебник. 3-е изд., перераб. и доп. СПб. : Изд-во СПб ун-та, 2020. 234 с.

Павельева Т. Ю. Изучение коллокаций на основе лингвистических корпусов текстов // Вестник Тамбовского университета. Сер. Гуманитарные науки. 2016. № 3-4. С. 56–61.

Плунгян В. А. Почему современная лингвистика должна быть лингвистикой корпусов URL: http://www.polit.ru/lectures/2009/10/23/corpus/html

British National Corpus. URL: https://www.english-corpora.org/bnc/ (дата обращения: 17.12.2022). BNC.

Corpus of Contemporary American English. URL: https://www.english-corpora.org/coca/ (дата обращения: 17.12.2022). COCA.

Cowie A. P., Mackin R. Oxford dictionary of phrasal verbs. Oxford : Oxford university press, 2011. 517 p.

УДК 81’33

А. А. Шаляпина

Новосибирский государственный университет,
Новосибирск, Россия

Определение близкого родства дикторов посредством проведения фоноскопической экспертизы

Исследуется фоноскопическая экспертиза для случая близкородственных отношений с помощью анализа формант. Рассматриваются исследования, проводимые в этом направлении ранее. Определено направление исследования и круг испытуемых. В качестве информантов выбраны родные сестры; ранее в подобных исследованиях фонограммы спонтанной и подготовленной речи людей с такими же родственными связями не рассматривались. Для решения проблемы определения имплицитных особенностей речи предлагается использование формантного анализа с помощью PLS-инструментов (для получения проекций на латентные структуры). Для описания имплицитных особенностей звукопроизношения испытуемых приводятся многомерные модели, составленные на основе полученных данных. Выявлены как общие для сестёр особенности, так и специфика для каждого информанта и каждого типа текста. Оценивается возможность применения используемого метода для продолжения исследования.

Ключевые слова: фоноскопическая экспертиза, форманты, спонтанная речь, подготовленная речь, имплицитные особенности речи, бикомпонетная модель.

Identification of Speakers’ Close Relationship Through Phonoscopic Expertise

The paper is dedicated to phonoscopic expertise for defining close-family relationships by the usage of formant analysis. We consider researches which were held in this area before. The direction of the research and the focus group were defined. Sisters were chosen as participants of our experiment; phonograms of spontaneous and prepared speech of people with the same family relationship were never checked before. We use PLS-analysis tools (Projection to Latent Structure) for formant analysis to solve the problem of identification of implicit individual speech features. The paper includes multidimensional models based on our data, which describe implicit individual speech features. The features common for the sisters, the specifics for each subject and the specifics for each type of text were revealed. The possibility of using this method to continue our research is evaluated.

Keywords: phonoscopic expertise, formants, spontaneous speech, prepared speech, implicit speech features, Two-Block PLS.

Одной из ключевых задач прикладной лингвистики и судебной экспертизы является идентификация личности по голосу и речи. Это объясняется тем, что совокупность характеристик голоса уникальна и многие из них могут быть исследованы с помощью специализированного программного обеспечения. Также в настоящее время повсеместно распространены технические устройства фиксации голосовой информации, отличающиеся высокой точностью записи, что открывает широкие возможности для сбора материала. Идентификация по голосу и речи используется в сферах защиты информации, криминалистике [Лингвистическая экспертиза … , 2017, с. 10–12].

Проблема идентификации диктора заключается в том, что получить идеальную запись голоса в обычных условиях не представляется возможным. Чаще всего на изучаемой записи присутствуют посторонние шумы, голоса других людей, также некоторые рассматриваемые характеристики речи могут быть подвержены влиянию настроения, психического и физического состояния диктора (радость, усталость, болезнь, спокойствие и т. д.). В каждом конкретном случае идентификацию затрудняют также индивидуальные акустические характеристики голоса диктора, вариативность которых зависит от ряда факторов: пола диктора, его возраста, влияния диалектных особенностей, формы и объема органов речевого тракта [Сорокин, Вьюгин, Тананыкин, 2012, с. 10].

В поле проблем идентификации личности по голосу и речи существуют также исследования, посвященные определению наличия родственных связей между дикторами. Но такие исследования, несмотря на широкие перспективы применения их результатов, единичны. А. М. Грачев в своем исследовании показал, что наиболее ярко зависимость наблюдается при сравнении голосов братьев, сравнение же голосов детей и родителей менее репрезентативно [Грачев, 2013, с. 20]. Поэтому в данном исследовании было принято решение анализировать функционально-динамические комплексы устно-речевых навыков близкородственных говорящих.

Объектом исследования являются функционально-диагностические комплексы устно-речевых навыков близкородственных говорящих, на данном этапе эксперимента это родные сёстры. Предмет исследования – форманты звуков как конкретные характеристики спонтанной и подготовленной речи испытуемых.

Цель нашей работы – при помощи специализированного ПО выяснить, существует ли связь между имплицитными особенностями подготовленной или спонтанной речи родных сестер.

Результаты данной работы могут быть использованы при разработке методики установления личности человека посредством проведения фоноскопической экспертизы.

В качестве ведущего метода данного исследования был выбран метод формантного анализа – нахождения частот F1, F2, F3, F4 для гласных фонем «а», «о», «и», «у» в сигнификативно сильной позиции и сравнения полученных характеристик.

Форманты – это резонансные частоты речевого тракта определенной формы и объема. Частоты формант (кроме ЧОТ) задаются конфигурацией речевого тракта, поэтому сведения о формантах позволяют делать определенные выводы о положении артикулирующих органов. Совокупность значений формант F0 (ЧОТ), F1, F2, F3, F4 называется формантной картиной [Князев, Пожарицкая, 2011, с. 93].

Значения формант гласных фонем напрямую связаны с традиционными классификационными характеристиками. Так, значение форманты F1 зависит от подъема языка при артикуляции. Значение форманты F2 обратно пропорционально длине ротового резонатора, т. е. зависит от ряда гласного. За формирова-

ние F3 у всех русских гласных, кроме «и», отвечает гортанный желудочек [Каганов, 2014, с. 28]. Резонансные свойства каждой конфигурации вокального тракта для F4 и более высоких формант обычно диктором не контролируются и задаются имеющимися анатомическими ограничениями на возможные изменения конфигурации артикуляторного тракта диктора, то есть нет одной конкретной конфигурации (или элемента речевого тракта), формирующей эти частоты [Кирьянов, Каганов, 2016, с. 37]. На языке формантного описания это приводит к тому, что при фиксированных значениях первых формант более высокие по частоте форманты у данного диктора могут занимать только более-менее стабильные индивидуальные положения.

Выбор метода обусловлен его точностью: так как частоты формант, как уже говорилось выше, задаются конфигурацией речевого тракта (кроме частоты основного тона), это позволяет с высокой вероятностью определить, принадлежит ли аудиозапись конкретному диктору, а в нашем случае определить характер связи имплицитных особенностей речи близкородственных говорящих.

Выбор фонем был продиктован как их частотностью в спонтанной речи, так и их классификационными характеристиками (все рассмотренные фонемы отличаются по ряду и подъему, «о» и «у» отличаются нехарактерным для данного ряда фонем свойством – лабиализованностью). Также учитывался тот факт, что формантная картина гласного «е» очень похожа на формантную картину «и», поэтому гласный «е» было решено не рассматривать.

На основе PLS-анализа полученных данных была сформирована многомерная модель для описания имплицитных особенностей звукопроизношения испытуемых.

PLS-анализ – это метод получения проекций на латентные структуры (Projection to Latent Structure), первоначальное название «метод частичных наименьших квадратов» (Partial Least Squares) [Rohlf, Corti, 2000]. Эффективным инструментом PLS-анализа является бикомпонентные модели. Они используются для изучения имплицитных когнитивных процессов путём выявления глубинных «латентных структур» (независимых механизмов или ФДК), единых для 2 блоков (матрицы B1 и B2) многомерных показателей [Assessment of the conjugation … , 2019].

При построении бикомпонентных моделей происходит центрирование рядов данных, масштабирование и повороты обоих блоков для получения максимальной ковариации между матрицами счетов (B1-score и B2-score), которые являются проекциями матриц B1 и B2 на искомые ортогональные латентные структуры [Polunin, Shtaiger, Efimov, 2019]. В один блок можно поместить переменные-признаки (состоят только из «0» и «1»), а в другой – ряды инструментальных данных.

В результате построения и анализа бикомпонентной модели мы получаем число латентных структур (новых осей координат), которое равно минимальному числу переменных из двух блоков исходных данных. Заметим, что соотношения для структур сырых данных в блоках остаются теми же самыми после любого количества (и порядка применения) таких операций, как центрирование,

масштабирование, поворот, которые применяются в PLS-моделях. Таким образом, полностью сохраняется структура сырых данных, вся информации из исходных рядов данных при построении бикомпонентной PLS-модели собирается в первых независимых латентных структурах. Модели такого типа допускают ситуацию, когда переменных больше, чем объектов, а также взаимную коррелированность исходных данных, которые могут включать в себя линейные комбинации друг друга [A PLS kernel … , 1994].

Для пилотной версии эксперимента была выбрана пара родных сестер с близкими росто-весовыми параметрами (21 год, 158 см, 45 кг; 15 лет, 153 см, 46 кг); кроме того, ни одна из девушек не имеет хронических заболеваний органов речевого аппарата и щитовидной железы, хирургических вмешательств также не было.

Эксперимент по выявлению наличия связи ФДК сестер проводится следующим образом:

1) у информантов были запрошены две аудиозаписи, на которых было зафиксировано произнесение подготовленного и спонтанного текста соответственно; нам было важно рассмотреть и спонтанную, и подготовленную речь, так как формантные характеристики в зависимости от темпа речи могут варьироваться в пределах формантного диапазона одного и того же диктора;

2) была проведена базовая шумоочистка каждой аудиозаписи в программе Audacity;

3) с помощью программы Praat в аудиозаписях были выделены слова, содержащие гласные фонемы «а», «о», «и», «у» в сигнификативно сильной позиции;

4) для каждой гласной были определены количественные значения формант F1, F2, F3, F4;

5) был проведен многомерный PLS-анализ полученных данных.

Результатом многомерного анализа стала бикомпонентная (Two-Block PLS) модель. В блоки бикомпонентной модели вошли переменные, представляющие собой форманты, частоту встречаемости фонем в спонтанном тексте (5 переменных, блок 1) и ряды признаков-вопросов (8 переменных, блок 2). Соответственно, было получено 5 латентных структур, которые перечислены в табл.

Таблица

Блоки переменных для бикомпонентной модели

Переменные	№ блока
F1..F4: 4 шкалы формантов	1
p: 1 переменная, частоты фонем в спонтанном тексте	1
fon_a, fon_o, fon_i, fon_u: 4 признака фонем	2
txt_s, txt_p: 2 признака текста (спонтанный, подготовленный)	2
subj_1, subj_2: признаки испытуемых (2 номера в первой паре сестер)	2

Представленный график осыпи латентных структур бикомпонентной модели (рис. 1) описывает наши данные. Левее первого перегиба (структура № 2) находятся общие особенности (структура № 1); между первым и вторым перегибом (структура № 4) находятся структуры частной специфики (структуры № 2 и 3). В «хвосте» находятся структуры с подавляющим влиянием «шума» (структуры № 4 и 5).

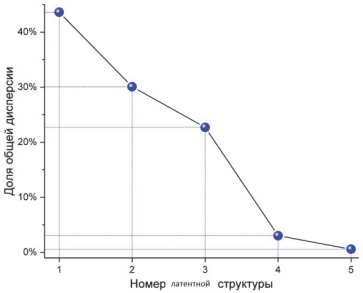


Рис. 1. График осыпи латентных структур бикомпонентной модели

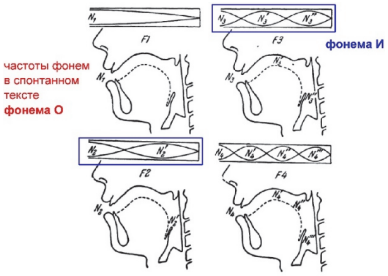


Рис. 2. Визуализация нагрузок переменных
для латентной структуры № 1 (особенности)

Таким образом, с помощью блока 2 (наши вопросы) мы задаем поворот для блока 1 (обучаем его), который обеспечивает максимально информативные ответы на вопросы из блока 2. Согласно графику осыпи латентных структур (независимых компонент, факторов), сформированных моделью (см. рис. 1), первая латентная структура должна показывать общие особенности, следующие две латентные структуры – частную специфику. Последние две латентные структуры в своей большей части выражают шумовые компоненты в данных. Таким образом, имеет смысл описывать общие особенности (латентная структура № 1 (рис. 2)) и наиболее влиятельные специфичные факторы (латентные структуры № 2 и 3).

Суммарно три первые латентные структуры, которые мы будем рассматривать далее, обусловливают 96,4 % наблюдаемой общей дисперсии.

На рис. 3, 4 приведены результаты анализа бикомпонентных моделей, устойчивые фонетические предпочтения для полученных двух условных ФДК даны соответствующим цветом (красный цвет; синий цвет).

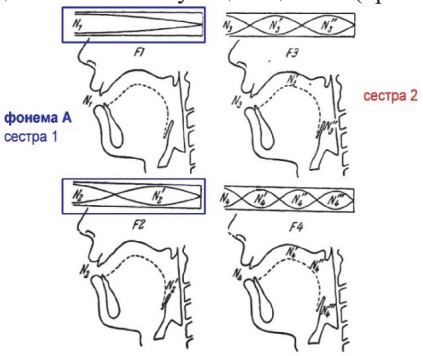


Рис. 3. Визуализация нагрузок переменных для латентной структуры № 2 (специфика)

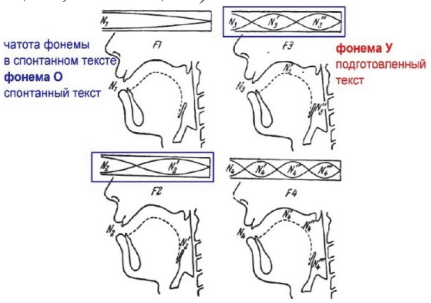


Рис. 4. Визуализация нагрузок переменных для латентной структуры № 3 (специфика)

Из рис. 2 следуют общие особенности для этой пары сестер : чем сильнее связаны F2 и F3 с фонемой «и», тем меньше будет фонем «о» в спонтанном тексте (43,6 % общей дисперсии).

Из рис. 3 следует частная специфика сестер: для сестры 1 характерна связь F1 и F3 с фонемой «а», в отличие от сестры 2, для которой это нехарактерно (30,1 % общей дисперсии).

Из рис. 4 следует частная специфика для читаемых текстов: для спонтанного текста характерна связь F2 и F3 с фонемой «о» (и более высокая частота этой фонемы в спонтанном тексте) по сравнению с фонемой «у» в подготовленном тексте (22,7 % общей дисперсии).

Таким образом, структура 1 показывает нам общие особенности для этой пары сестер. Можно предположить, что описанные особенности будут отличаться от таковых для другой пары близких родственников. Структура 2 показывает специфику ФДК озвучивания сестрами подготовленного и спонтанного текстов. Можно предположить, что это дает профилирование близких родственников для их идентификации. Структура 3 показывает специфику ФДК для этой пары сестер при озвучивании подготовленного и спонтанного текстов. Это позволит в дальнейших исследованиях готовить тексты с заранее известными критериями.

Полученные результаты позволяют предположить, что комплексное применение методов формантного анализа и получения проекций на латентные структуры перспективно в данном исследовательском поле. Для получения общих закономерностей, специфики и тенденций ФДК как пар испытуемых, так и всей совокупности информантов, планируется продолжить эксперимент с парами сестёр, а также провести аналогичный эксперимент с записями спонтанной и подготовленной речи родных братьев.

Аббревиатуры и сокращения

ФДК – функционально-динамический комплекс.
PLS – Projection to Latent Structure (Partial Least Squares).
ЧОТ – частота основного тона.
Two-Block PLS – бикомпонентная модель.

Литература

Грачев А. М. Распознавание звучащей русской речи в теоретическом и экспериментальном освещении: семейные, возрастные и гендерные аспекты лингвистической идентификации личности : автореф. дис. ... канд. филол. наук. Нижний Новгород, 2013. 25 с.

Каганов А. Ш. Об использовании спектральных характеристик речи для определения биометрических параметров речевого тракта в судебно-медицинской идентификации личности говорящего // Судебно-медицинская экспертиза. № 57(1). 2014. С. 26–29.

Кирьянов П. А., Каганов А. Ш. Применение методов спектрального анализа в задаче медико-криминалистической идентификации говорящего // Судебно-медицинская экспертиза. 2016. № 59(5). С. 36–38.

Князев С. В., Пожарицкая С. К. Современный русский литературный язык: Фонетика, орфоэпия, графика и орфография : учеб. пособие для вузов. 2-е изд., перераб. и доп. М. : Акад. проект ; Гаудеамус, 2011. 430 с.

Лингвистическая экспертиза звучащей речи : сб. учеб.-метод. материалов для направления подготовки 45.04.03. Благовещенск : Амур. гос. ун-т, 2017. 51 с.

Сорокин В. Н., Вьюгин В. В., Тананыкин А. А. Распознавание личности по голосу: аналитический обзор // Информационные процессы. 2012. Т. 12, № 1. С. 1–30.

Assessment of the conjugation of morphogenetic and molecular genetic moduli of variation in the common vole Microtus s.l. in gradient environmental conditions / V. Yu. Kovaleva, A. A. Pozdnyakov, Yu. N. Litvinov, V. M. Efimov // Ecological genetics. 2019. № 17(2). P. 21-34. DOI: 10.17816/ecogen17221-34

Polunin D., Shtaiger I., Efimov V. JACOBI4 software for multivariate analysis of biological data // bioRxiv. 2019. DOI: 10.1101/803684

A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm / S. Ränner, F. Lindgren, P. Geladi, S. Wold // Journal of Chemometrics. 1994. Vol. 8, Iss. 2. P. 111–125. DOI: 10.1002/cem.1180080204

Rohlf F. J., Corti M. The use of two-block partial least-squares to study covariation in shape // Systematic Biology. 2000. Vol. 49, N 4. P. 740–753. DOI: 10.1080/106351500750049806

Вместо послесловия

О. В. Митренина

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Нейросетевая классификация данных для диалектологов и социологов

Пять лет назад Сбербанк первым в России начал использовать технологию распознавания лиц в интернет-банке. Сегодня покупки в магазинах можно оплачивать «одним взглядом», просто посмотрев в камеру на кассе. Сбербанк обещает узнать своего пользователя в любом возрасте, с новой прической, макияжем или бородой [Биометрия в СберБанке …].

Распознавание лиц происходит с помощью заранее обученных нейронных сетей, которые сами решают, какие особенности лица им нужно учитывать. Но понятие об «особенностях лица» у компьютера свое, поскольку он любую информацию переводит в числа. Даже время компьютер считает не так, как люди. Если я спрошу у него время того момента, когда я пишу эти строки − time(), − он выдаст мне не год, месяц, день, час и минуту, а странное число: 1672223220.8778276. Это количество секунд, прошедших с 1 января 1970 г., с точностью до миллионных долей.

Точно так же − совсем не как люди − современный компьютер обрабатывает все, что ему дают для анализа.

Чтобы обучить нейросеть, необходимо передать ей корпус данных. Да и человек, чтобы по-настоящему обучиться, должен получить большой опыт общения с данными. В первоначальной версии фильма «Мимино» был эпизод, в котором два японца в московской гостинице говорят, указывая на главных героев: «До чего же все эти русские на одно лицо». Но если такие японцы проживут в Москве хотя бы год, их мозг обработает значительный корпус «московских лиц» и скорректирует свою систему распознавания образов. В IT-сфере это называется **файнтьюнинг** (от англ. *fine tuning* ‘тонкая настройка’).

Набор данных, с помощью которого обучается нейросеть, принято называть **датасетом** (от англ. *dataset* ‘набор данных’). Человеческий мозг сам находит себе материал для обучения, а компьютеру надо подавать созданный заранее датасет. Создание таких датасетов − это настолько важная задача, что для нее существует специальная IT-профессия: **дата-инженер** (от англ. *data engineer* ‘инженер данных’).

Чаще всего текстовые данные собираются в интернете с помощью заранее написанных программ. Это записи в соцсетях, отзывы на товары или фильмы и пр. Но отдельную ценность представляют данные, собранные вручную и не существующие в интернете. Данные, собранные диалектологами, как раз такими и являются: они собираются и обрабатываются вручную, а их содержание уникально.

На конференции «Цифра» (Иркутск, 14–16 ноября 2022 г.), по результатам которой составлен этот сборник, звучали доклады социологов, которые также собирают и используют интересные датасеты. Например, исследованию текстовых данных были посвящены доклады К. А. Иванова «Социальные медиа как поле для исследования практик интеграции внутренних мигрантов в сибирских региональных столицах» и Д. О. Тимошкина «“Мигрантские” цифровые медиа как узлы неформальных горизонтальных сетей». Исследование этих данных с помощью нейронных сетей может выявить такие закономерности, которые незаметны при традиционном анализе текстов.

Ученые-гуманитарии обычно считают, что исследовать текстовые данные с помощью нейросетей могут только программисты или математики. На самом деле математики и программисты нужны для того, чтобы создавать новые виды нейронных сетей. А уже существующие сети может научиться использовать ученый из других областей науки – возможностей для этого становится больше не с каждым годом, а с каждым месяцем.

Еще недавно для статистического анализа текстов требовалось выполнять сложные математические вычисления, но сейчас существует много простых и удобных инструментов, которые проведут все расчеты, сразу покажут результаты и даже нарисуют облако тегов. Например, Voyant Tools [Voyant Tools …]. Но для нейросетевых задач удобных инструментов пока не очень много. Например, нейросеть «Порфирьевич» [Нейросеть «Порфирьевич» …] помогает писать рассказы и псевдонаучные тексты. Получив предложение *Однажды я вышел из дома в 5 утра*, «Порфирьевич» продолжает: *Встретил восход солнца в окне троллейбуса и понял, что приближаюсь к жизни, которой живу с давних пор*.

Задачи, связанные с классификацией текстов, хорошо решаются с помощью нейронных сетей, которые могут выявлять особенности текстов, плохо поддающиеся формальному описанию. Например, В. И. Фирсанова [2020] обучила нейронную сеть для распознавания записей из виртуальных сообществ наркозависимых. Сейчас не существует универсальных нейросетевых классификаторов с удобным интерфейсом, подобным интерфейсу «Порфирьевича». Но существует множество готовых проектов, которые можно настроить для классификации собственных записей. Вот краткое описание основных шагов для входа в эту IT-область:

1. Для обучения нейросетей требуются мощные компьютеры, которые редко встречаются в домашнем пользовании. Здесь на помощь приходит **Google Colaboratory**, или просто Colab [Google Colaboratory …], который позволяет бесплатно подключаться к компьютерам Google и обучать на них нейросети. Достаточно выбрать меню пункты *Файл → Создать блокнот*, и вы получаете пустой блокнот, куда можно скопировать необходимый написанный на языке Python код, о котором будет сказано в п. 4.

2. Упомянутый язык программирования Python (Питон) очень простой, его код читается почти как текст на английском языке. Для Питона существует огромное количество готовых модулей (инструментов), позволяющих обрабаты-

вать тексты. Учить этот язык легко и приятно. В сети есть много уроков, инструкций и отдельных лекций для самого разного уровня программистов. Начать знакомство с Питоном можно со статьи [Митренина, 2019], она написана специально для тех гуманитариев, которые считают, что они не смогут программировать.

3. На главной странице Google Colaboratory после слов *Добро пожаловать в Colab* приводится краткое введение в современные IT-технологии. Там же есть и примеры, среди которых можно найти классификацию текста. Другие обучающие материалы находятся на портале **Hugging Face** [Учебник портала …] – одной из главных платформ для разработчиков нейросетевых систем. Не нужно стремиться понять в этих учебниках все подробности. Достаточно понять их в общих чертах. Кроме того, можно посмотреть небольшие видеокурсы или отдельные уроки по нейросетям, в открытом доступе их появляется всё больше.

4. Познакомившись с теорией, можно поискать в Сети существующие модели для классификации текстов и попытаться запустить их на Google Colaboratory самостоятельно. Для запуска нужно будет перенести в свой блокнот код на Python, который обычно прилагается к описанию модели. Некоторые такие модели представлены на Hugging Face [Классификация текстов …]. Вначале желательно испытать их на тех данных, которые приводятся в их описании. Почти все они будут на английском языке.

5. После этого в Сети можно найти похожие данные, чтобы попробовать провести их классификацию с помощью знакомой нейросети. Поскольку мы работаем с текстами, нужно искать **размеченные корпусы** текстов или создавать их самостоятельно. «Размеченные» обозначает, что у каждого фрагмента текста (например, у каждого предложения) имеется своя метка; подробнее об этом будет сказано в п. 7. Для поиска датасетов есть отдельный инструмент Google Dataset Search [Поисковая система …]. Можно посмотреть статью-путеводитель по открытым наборам данных для машинного обучения [Подборка датасетов …]. Есть и другие подборки, хотя большинство доступных датасетов не являются текстовыми. Огромная коллекция размеченных текстовых данных на 420 языках представлена на сайте **Татоэба** [Открытый многоязычный …].

6. В какой-то момент придется перейти от англоязычного датасета к русскоязычному. Как минимум для этого понадобится другая обработка исходных данных. С этой задачей можно справиться, если задавать вопросы, например, на форуме Hugging Face [Форум Hugging Face …], где в разделе для новичков сказано: «Не модерируйте сами себя, каждый должен начинать с чего-то, и каждый на этом форуме готов помогать!» Есть и другие площадки для получения советов от опытных разработчиков.

7. Для своих задач вам придется выступить в качестве дата-инженера и создать собственный размеченный датасет. Для этого каждому предложению в собранных данных надо сопоставить метку, соответствующую «классу» этой записи. Так, модели В. И. Фирсановой [2020] обучались на датасете, собранном ею в похожих по стилю сообществах «Вконтакте». Записи были разделены на два класса. Каждое предложение из сообществ наркозависимых отмечалось меткой

1, а каждое предложение из других сообществ отмечалось меткой 0. Полученный датасет подавался на обучение нейросети, которая создавала нейросетевую модель. Эта модель умела предсказывать для новых записей вероятность того, что они получены из сообщества наркозависимых.

Современные модели для обработки текстов обычно не обучаются с нуля, а используют упомянутый выше файнтьюнинг – тонкую настройку. Они используют большие существующие модели, которые обучались ранее на огромных текстовых данных, на мощных компьютерах в течение многих часов или даже дней. Именно такие модели хранятся на Hugging Face. Они уже «понимают» многие особенности естественного языка, и остается только настроить их с помощью своих данных для решения более узких задач. Как и упомянутые японцы из фильма «Мимино», опираясь на свою способность распознавать японские лица, могут настроить способность распознавания лиц в Москве.

В ближайшее время нейросетевая классификация данных может стать простым и общедоступным инструментом, ведь уже сейчас разговорная нейросеть chatGPT от OpenAI [ChatGPT на сайте …] может находить ошибки в программном коде и даже создавать новые программы. Тем интереснее оказаться среди первопроходцев в науке на стыке классических гуманитарных наук и IT-технологий нового поколения.

Литература

Биометрия в СберБанке. URL: http://www.sberbank.ru/ru/person/dist_services/bio (дата обращения: 28.12.2022).

Классификация текстов на Hugging Face. URL: https://huggingface.co/tasks/text-classification (дата обращения: 28.12.2022).

Митренина О. В. Python для тех, кто никогда не программировал // Journal of Applied Linguistics and Lexicography. 2019. № 1 (1). С. 127–135. URL: https://journall.org/index.php/main/article/view/18/19 (дата обращения: 28.12.2022).

Нейросеть «Порфирьевич». URL: https://porfirevich.ru/ (дата обращения: 28.12.2022).

Открытый многоязычный онлайн-словарь фраз Татоэба. URL: https://ru.wikipedia.org/wiki/Татоэба (на дату обращения 15.01.2023)

Подборка датасетов для машинного обучения https://habr.com/ru/post/452392/ (дата обращения: 28.12.2022).

Поисковая система для датасетов Google Dataset Search. URL: https://datasetsearch.research.google.com/ (дата обращения: 28.12.2022).

Учебник портала Hugging Face. URL: https://huggingface.co/course/ (дата обращения: 28.12.2022).

Форум Hugging Face. URL: https://discuss.huggingface.co/ (дата обращения: 28.12.2022).

ChatGPT на сайте OpenAI. URL: https://openai.com/blog/chatgpt/ (дата обращения: 15.01.2023).

Firsanova V. Automatic Recognition of Messages from Virtual Communities of Drug Addicts // Journal of applied linguistics and lexicography. 2020. Vol. 2, N 1. P. 16–27. DOI: https://doi.org/10.33910/2687-0215-2020-2-1-16-27

Google Colaboratory. URL: https://colab.research.google.com/ (дата обращения: 28.12.2022).

Voyant Tools – веб-пространство для чтения и анализа цифровых текстов. URL: https://voyant-tools.org/ (дата обращения: 28.12.2022).

Научное издание

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы конференции молодых ученых
Иркутск, 14–16 ноября 2022 г.

ISBN 978-5-9624-2180-3

Корректор *Н. А. Михайлова*
Дизайн обложки: *П. О. Ершов*

Темплан 2023. Поз. 71
Уч.-изд. 4,1

ИЗДАТЕЛЬСТВО ИГУ
664082, г. Иркутск, ул. Лермонтова, 124